Cognitive Models for Visual Commonsense

Song-Chun Zhu, Yixin Zhu

January 12, 2022

Contents

Authors

Song-Chun Zhu received his Ph.D. degree in Computer Science from Harvard University in 1996 (advised by Dr. David Mumford). He is currently a professor of Statistics and Computer Science, and director of the Center for Vision, Learning, Cognition, and Autonomy, at the University of California, Los Angeles. His research interest has been to pursue a common statistical framework for vision, and broadly intelligence. He has published over 300 papers in computer vision, statistical learning, cognition, and AI, and robot autonomy, and received a number of honors, including the David Marr Prize in 2003 for image parsing with Z. Tu *et al.*, the Marr Prize honorary nominations in 1999 for texture modeling and 2007 for object modeling with Dr. Yingnian Wu *et al.* As a junior faculty he received in 2001 the Sloan Fellow in Computer Science, NSF Career Award, and ONR Young Investigator Award. In 2008 he received the Aggarwal prize from the Intl Association of Pattern Recognition for "contributions to a unified foundation for visual pattern conceptualization, modeling, learning, and inference." In 2013 he received the Helmholtz Test-of-time prize for a paper on image segmentation. He is a fellow of IEEE Computer Society since 2011. He has been the principal investigator leading several ONR MURI and DARPA teams working on scene and event understanding and cognitive robots under a unified mathematical framework.

Yixin Zhu received his Ph.D. degree in Statistics from the University of California, Los Angeles (UCLA) in 2018 (advised by Dr. Song-Chun Zhu). His research builds interactive AI by integrating high-level common sense (functionality, affordance, physics, causality) with raw sensory inputs (pixels and haptic signals) to enable richer representation and abstract reasoning on objects, scenes, shapes, numbers, and agents. He is a co-organizer of Vision Meets Cognition (FPIC) workshops at CVPR, 3D Scene Understanding for Vision, Graphics, and Robotics workshops at CVPR, and Virtual Reality Meets Physical Reality workshops at SIGGRAPH Asia.

Preface

Introducing the Book Series

The book series consists of three parts.

The first book covers David Marr's paradigm and various underlying statistical models for vision. The mathematical foundations herein integrate three regimes of models (low-, mid-, and high-entropy regimes) and provide essential foundation for research in visual coding, recognition, cognition, and reasoning. Concepts in this book are first explained for understanding and then supported by findings in psychology and neuroscience, after which they are established by statistical models and further linked to research in other fields such as physics. A reader of this book will gain a unified, cross-disciplinary view of artificial intelligence research in vision and will accrue knowledge spanning from psychology to neuroscience to statistics.

The second book defines a stochastic grammar for parsing objects, scenes, and events, posing computer vision as a joint parsing problem. It summarizes research efforts over the past 20 years that have worked to extend King-Sun Fu's paradigm of syntactic pattern recognition. Similar to David Marr, King-Sun Fu was a pioneer and influential figure in the vision and pattern recognition community.

The third book discusses visual commonsense reasoning by connecting vision to cognition and artificial intelligence. Recent progress in deep learning is essentially based on a "big data for small tasks" paradigm, under which massive amounts of data are used to train a classifier for a single narrow task. In this work, we call for a shift that flips this paradigm upside down. Specifically, we propose a "small data for big tasks" paradigm, wherein a single AI system is challenged to develop "common sense," enabling it to solve a wide range of tasks with little training data. We illustrate the potential power of this new paradigm by reviewing models of common sense that synthesize recent breakthroughs in both machine and human vision. We identify functionality, physics, intent, causality, and utility (FPICU) as the five core domains of cognitive AI with humanlike common sense. When taken as a unified concept, FPICU is concerned with the questions of "why" and "how," beyond the dominant "what" and "where" framework for understanding vision. They are invisible in terms of pixels but nevertheless drive the creation, maintenance, and development of visual scenes. We therefore coin them the "dark matter" of vision. Just as our universe cannot be understood by merely studying observable matter, we argue that vision cannot be understood without studying FPICU. We demonstrate the power of this perspective to develop cognitive AI systems with humanlike common sense by showing how to observe and apply FPICU with little training data to solve a wide range of challenging tasks, including tool use, planning, utility inference, and social learning. In summary, we argue that the next generation of AI must embrace "dark" humanlike common sense for solving novel tasks.

The authors would like to thank many current and former Ph.D. students at UCLA for their contributions to this book: Siyuan Huang, Hangxin Liu, Mark Edmonds, Lifeng Fan, Baoxiong Jia, and Tianmin Shu. Thanks to Liangru Xiang at Tsinghua University for his editing work.

Chapter 1

Introduction

> If one hopes to achieve a full understanding of a system as complicated as a nervous system, . . . , or even a large computer program, then one must be prepared to contemplate different kinds of explanation at different levels of description that are linked, at least in principle, into a cohesive whole, even if linking the levels in complete details is impractical. — David Marr [1], pp. 20–21

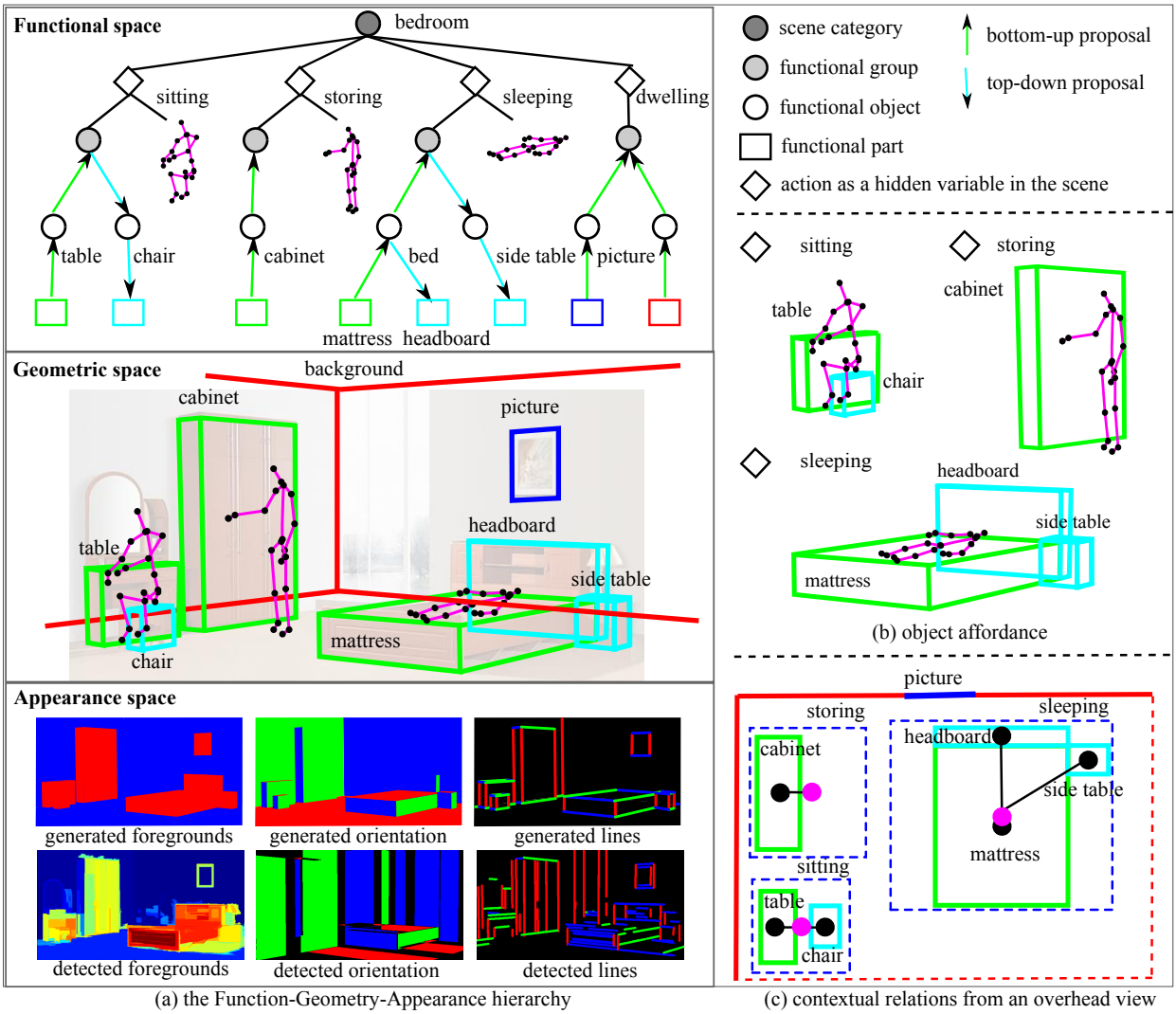
1.1 Three Types of Representation

Computer vision is the front gate to artificial intelligence (AI) and a major component of modern intelligent systems. The classic definition of computer vision proposed by the pioneer David Marr [1] is to look at "what" is "where." Here, "what" refers to object recognition (object vision), and "where" denotes three-dimensional (3D) reconstruction and object localization (spatial vision) [2]. Such a definition corresponds to two pathways in the human brain: (i) the ventral pathway for categorical recognition of objects and scenes, and (ii) the dorsal pathway for the reconstruction of depth and shapes, scene layout, visually guided actions, and so forth. This paradigm guided the geometry-based approaches to computer vision of the 1980s-1990s, and the appearance-based methods of the past 20 years.

1.1.1 Image-centered Representation

The image-based representation relies heavily on the appearance in an image and is the focus of the first book in this book series. Such a representation is primarily a bottom-up process. To provide a brief summary, David Marr [1] conjectured that the perception of a 2D image is an *explicit* multi-phase information process, involving (i) an early vision system of perceiving textures [3, 4] and textons [5, 6] to form a primal sketch as a perceptually lossless conversion from the raw image [7, 8], (ii) a mid-level vision system to construct 2.1D (multiple layers with partial occlusion) [9, 10, 11] and 2.5D [12] sketches, and (iii) a high-level vision system that recovers the full 3D [13, 14, 15].

Alternatively, since appearance in natural 2D images varies significantly due to different camera viewpoints, lighting, reflectance, occlusions, *etc*., enormous efforts have been dedicated to engineer or learn features that are robust enough to handle such large variations. Notable efforts include the engineered SIFT feature [16] and recent deep neural networks (DNNs)-based learned features [17].



(a) the Function-Geometry-Appearance hierarchy (c) contextual relations from an overhead view

Figure 1.1: An example illustrating the three types of representation. (a) On the bottom, the image- and appearance-based representation handles large variations stemmed from raw pixel input, which is the focus of the book 1 in this book series. One-level abstraction of the appearance-based representation is the scene- and geometric-based representation. Together with the appearance-based representation, they collaboratively describe contextual relations in an image. The modeling and learning of such representations is the theme of the book 2 in this book series. On the top comes the task- and functional-based presentation of the scene. Man-made scenes and objects are largely driven by these invisible tasks and functions. Together with the geometric-space, they cooperatively model a crucial concept of object affordance in the scene. In this book, we focus on modeling and learning of these "dark metters" inside the image.

1.1.2 Scene-centered Representation

Scenes in 3D world captured by camera ought to satisfy certain geometric prior. From a Bayesian perspective at a scene level, such priors, independent of any 3D scene structures, were found in the human-made scenes. A notable effect is known as the Manhattan World assumption [18].

These geometric priors are more robust than appearance features; one can view them as a one-level abstract of the appearance-based representation, as such a geometric representation removes the detailed color, reflectance, *etc.*, but only focuses on the geometric structure of the environment. This type of geometric structure can still be directly perceived and parsed from the image, although

not as straightforward as detecting a face or an object; see an illustration in Fig. 1.1(a).

The geometric-based scene-centered representation enables a stream of work called "analysis by synthesis" [19], which fuses the bottom-up proposals together with the top-down parsing that incorporates the geometric priors. The central idea is: Bottom-up process proposes by detecting primitives and grouping (similar to Gestalt laws [20, 21]) provides hypothesis of the scenes and their structures. By directly comparing with an image, in particular, the filtered geometric-focused version, these hypotheses could be accepted or rejected on the basis of how closely the proposal is compared to the filtered version.

Combining geometric- and appearance-based representations, an algorithm can learn or derive a set of contextual relations among objects in 3D within a given scene; see an example in Fig. 1.1 (c). These relations could pose additional constraints in scene understanding and synthesis, and book 2 of this book series focuses on this perspective.

1.1.3 Task-centered Representation

Man-made scenes and objects are everything but randomly generated; they have been designed to serve certain intrinsic functions for human activities and tasks. Crucially, these activities and tasks are beyond the visible pixels in a given image, and an algorithm has to infer from these missing dimensions—humanlike common sense.

These invisible dimensions drives a scene configuration of man-made environments. Take Fig. 1.1 (a) for example, a bedroom serves functions of sitting, storing, sleeping, and dwelling. Each function generates a prior of how likely a person would interact with the environment. Combining these priors with the geometric space, they form a crucial concept of object affordance (see Fig. 1.1 (b)), which provides additional constraints for parsing of the man-made environments.

This task-centered perspective of the scene is the main theme of this book. Below, we start with a more detailed introduction on task-oriented vision.

1.2 Dark Matter in Vision and AI

Over the past several years, progress has been made in object detection and localization with the rapid advancement of DNNs, fueled by hardware accelerations and the availability of massive sets of labeled data. However, we are still far from solving computer vision or real machine intelligence; the inference and reasoning abilities of current computer vision systems are narrow and highly specialized, require large sets of labeled training data designed for special tasks, and lack a general *understanding* of common facts—that is, facts that are obvious to the average human adult—that describe how our physical and social worlds work. To fill in the gap between modern computer vision and human vision, we must find a broader perspective from which to model and reason about the missing dimension, which is humanlike common sense.

This state of our understanding of vision is analogous to what has been observed in the fields of cosmology and astrophysicists. In the 1980s, physicists proposed what is now the standard cosmology model, in which the mass-energy observed by the electromagnetic spectrum accounts for less than 5% of the universe; the rest of the universe is dark matter (23%) and dark energy (72%)[1]. The properties and characteristics of dark matter and dark energy cannot be observed and must be reasoned from the visible mass-energy using a sophisticated model. Despite their invisibility, however, dark matter and energy help to explain the formation, evolution, and motion of the visible universe.

[1] `https://map.gsfc.nasa.gov/universe/`

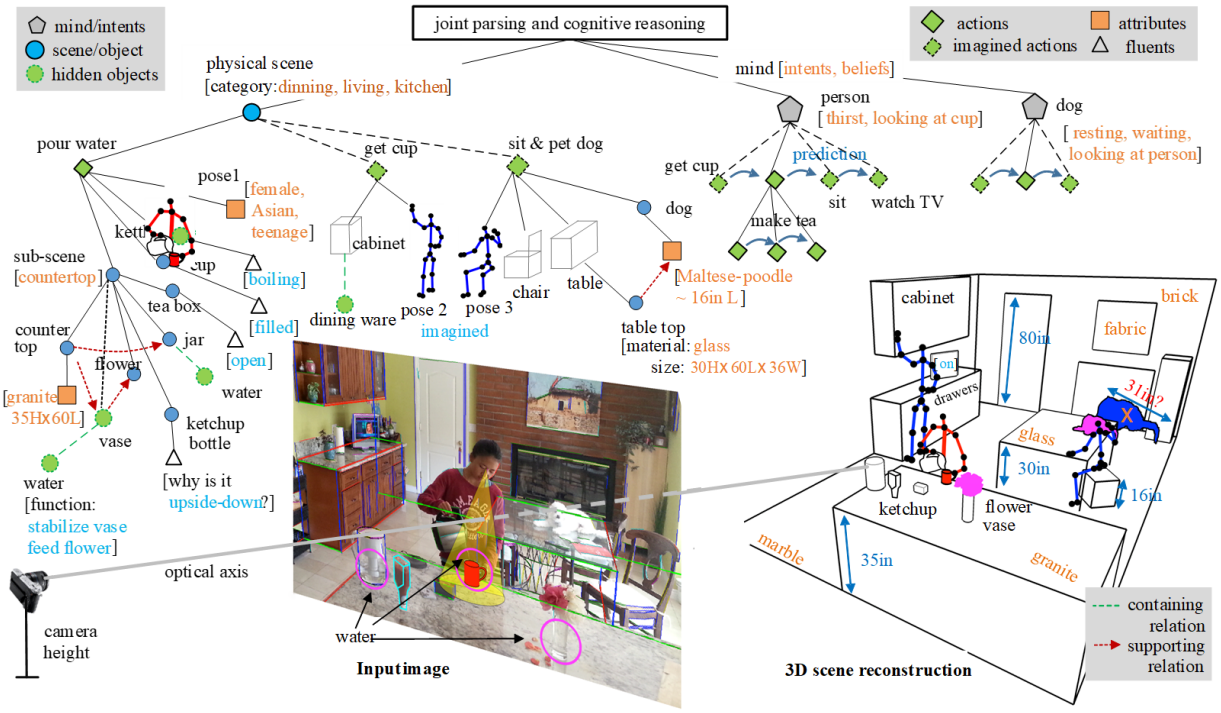


Figure 1.2: An example of in-depth understanding of a scene or event through joint parsing and cognitive reasoning. From a single image, a computer vision system should be able to jointly (i) reconstruct the 3D scene; (ii) estimate camera parameters, materials, and illumination; (iii) parse the scene hierarchically with attributes, fluents, and relationships; (iv) reason about the intentions and beliefs of agents (*e.g.*, the human and dog in this example); (v) predict their actions in time; and (vi) recover invisible elements such as water, latent object states, and so forth. We, as humans, can effortlessly (i) predict that water is about to come out of the kettle; (ii) reason that the intent behind putting the ketchup bottle upside down is to utilize gravity for easy use; and (iii) see that there is a glass table, which is difficult to detect with existing computer vision methods, under the dog; without seeing the glass table, parsing results would violate the laws of physics, as the dog would appear to be floating in midair. These perceptions can only be achieved by reasoning about unobservable factors in the scene not represented by pixels, requiring us to build an AI system with humanlike core knowledge and common sense, which are largely missing from current computer vision research. H: height; L: length; W: width. 1 in = 2.54 cm. Reproduced from Ref. [22] with permission of Elsevier, © 2020.

We intend to borrow this physics concept to raise awareness, in the vision community and beyond, of the missing dimensions and the potential benefits of joint representation and joint inference. We argue that humans can make rich inferences from sparse and high-dimensional data, and achieve deep understanding from a single picture, because we have common yet visually imperceptible knowledge that can never be understood just by asking "what" and "where." Specifically, human-made objects and scenes are designed with latent functionality, determined by the unobservable laws of physics and their down-stream causal relationships; consider how our understanding of water's flow from of a kettle, or our knowledge that a transparent substance such as glass can serve as a solid table surface, tells us what is happening in Fig. 1.2. Meanwhile, human activities, especially social activities, are governed by causality, physics, functionality, social intent, and individual preferences and utility. In images and videos, many entities (*e.g.*, functional objects, fluids, object fluents, and intent) and relationships (*e.g.*, causal effects and physical supports) are impossible to detect by most of the existing approaches considering appearance alone; these latent factors are not represented in pixels. Yet they are pervasive and govern the placement and motion of the visible

entities that are relatively easy for current methods to detect.

These invisible factors are largely missing from recent computer vision literature, in which most tasks have been converted into classification problems, empowered by large-scale annotated data and end-to-end training using neural networks. This is what we call the "big data for small tasks" paradigm of computer vision and AI.

In this book, we aim to draw attention to a promising new direction, where consideration of "dark" entities and relationships is incorporated into vision and AI research. By reasoning about the unobservable factors beyond visible pixels, we could approximate humanlike common sense, using limited data to achieve generalizations across a variety of tasks. Such tasks would include a mixture of both classic "what and where" problems (*i.e.*, classification, localization, and reconstruction), and "why, how, and what if" problems, including but not limited to causal reasoning, intuitive physics, learning functionality and affordance, intent prediction, and utility learning. We coin this new paradigm "small data for big tasks."

Of course, it is well-known that vision is an ill-posed inverse problem [1] where only pixels are seen directly, and anything else is hidden/latent. The concept of "darkness" is perpendicular to and richer than the meanings of "latent" or "hidden" used in vision and probabilistic modeling; "darkness" is a measure of the relative difficulty of classifying an entity or inferring about a relationship based on how much invisible common sense needed beyond the visible appearance or geometry. Entities can fall on a continuous spectrum of "darkness"—from objects such as a generic human face, which is relatively easy to recognize based on its appearance, and is thus considered "visible," to functional objects such as chairs, which are challenging to recognize due to their large intraclass variation, and all the way to entities or relationships that are impossible to recognize through pixels. In contrast, the functionality of the kettle is "dark;" through common sense, a human can easily infer that there is liquid inside it. The position of the ketchup bottle could also be considered "dark," as the understanding of typical human intent lets us understand that it has been placed upside down to harness gravity for easy dispensing.

Below, we start by revisiting a classic view of computer vision in terms of "what" and "where" in Section 1.2.1, in which we show that the human vision system is essentially task-driven, with its representation and computational mechanisms rooted in various tasks. In order to use "small data" to solve "big tasks," we then identify and review five crucial axes of visual common sense: **F**unctionality, **P**hysics, perceived **I**ntent, **C**ausality, and **U**tility (FPICU). Causality (Section 1.2.2) is the basis for intelligent understanding. The application of causality (*i.e.*, intuitive physics; Section 1.2.3) affords humans the ability to understand the physical world we live in. Functionality (Section 1.2.4) is a further understanding of the physical environment humans use when they interact with it, performing appropriate actions to change the world in service of activities. When considering social interactions beyond the physical world, humans need to further infer intent (Section 1.2.5) in order to understand other humans' behavior. Ultimately, with the accumulated knowledge of the physical and social world, the decisions of a rational agent are utility-driven (Section 1.2.6). In a series of studies, we demonstrate that these five critical aspects of "dark entities" and "dark relationships" indeed support various visual tasks beyond just classification. We summarize and discuss our perspectives in Section 1.2.7, arguing that it is crucial for the future of AI to master these essential unseen ingredients, rather than only increasing the performance and complexity of data-driven approaches.

1.2.1 Vision: From Data-driven to Task-driven

What should a vision system afford the agent it serves? From a biological perspective, the majority of living creatures use a *single* (with multiple components) vision system to perform *thousands*

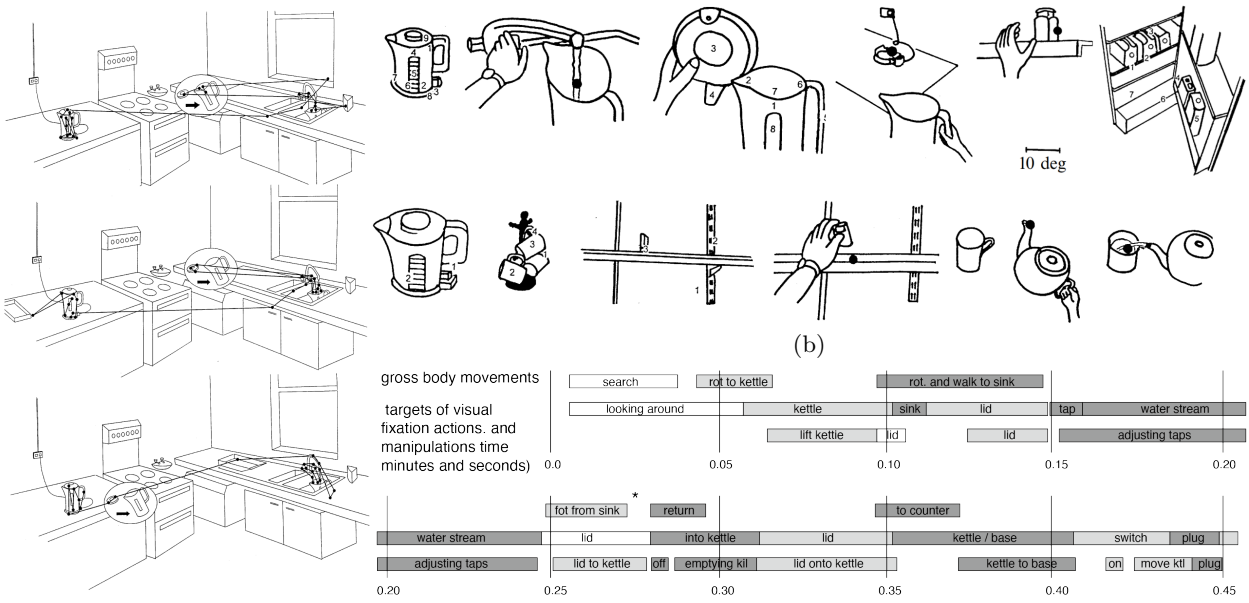
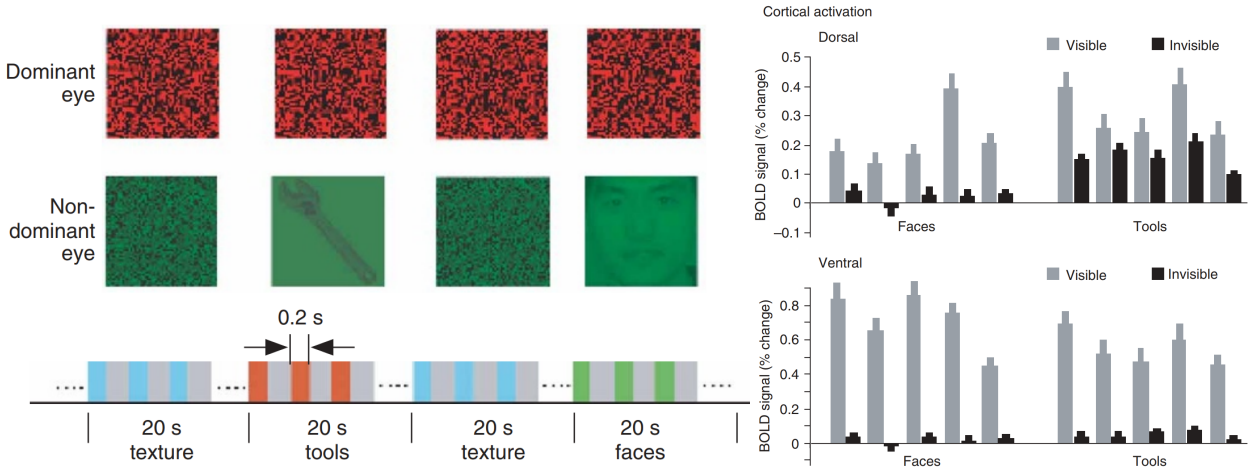


Figure 1.3: Even for as "simple" a task as making a cup of tea, a person can make use of his or her single vision system to perform a variety of subtasks in order to achieve the ultimate goal. (a) Record of the visual fixations of three different subjects performing the same task of making a cup of tea in a small rectangular kitchen; (b) examples of fixation patterns drawn from an eye-movement videotape; (c) a sequence of visual and motor events during a tea-making session. Rot: rotate; ktl: kettle. Reproduced from Ref. [24] with permission of SAGE Publication, © 1999.

of tasks. This contrasts with the dominant contemporary stream of thought in computer vision research, where a single model is designed specifically for a single task. In the literature, this organic paradigm of generalization, adaptation, and transfer among various tasks is referred to as task-centered vision [23]. In the kitchen shown in Fig. 1.3 [24], even a task as simple as making a cup of coffee consists of multiple subtasks, including finding objects (object recognition), grasping objects (object manipulation), finding milk in the refrigerator, and adding sugar (task planning). Prior research has shown that a person can finish making a cup of coffee within 1 min by utilizing a single vision system to facilitate the performance of a variety of subtasks [24].

Neuroscience studies suggest similar results, indicating that the human vision system is far more capable than any existing computer vision system, and goes beyond merely memorizing patterns of pixels. For example, Fang and He [25] showed that recognizing a face inside an image utilizes a different mechanism from recognizing an object that can be manipulated as a tool, as shown in Fig. 1.4; indeed, their results show that humans may be even more visually responsive to the appearance of tools than to faces, driving home how much reasoning about how an object can help perform tasks is ingrained in visual intelligence. Other studies [26] also support the similar conclusion that images of tools "potentiate" actions, even when overt actions are not required. Taken together, these results indicate that our biological vision system possesses a mechanism for perceiving object functionality (*i.e.*, how an object can be manipulated as a tool) that is independent of the mechanism governing face recognition (and recognition of other objects). All these findings call for a quest to discover the mechanisms of the human vision system and natural intelligence.



(a) (b)

Figure 1.4: Cortical responses to invisible objects in the human dorsal and ventral pathways. (a) Stimuli (tools and faces) and experimental procedures; (b) both the dorsal and ventral areas responded to tools and faces. When stimuli were suppressed by high-contrast dynamic textures, the dorsal response remained responsive to tools, but not to faces, while neither tools or faces evoked much activation in the ventral area. BOLD: blood oxygen level-dependent. Reproduced from Ref. [25] with permission of Nature Publishing Group, © 2005.

"What": Task-centered Visual Recognition

The human brain can grasp the "gist" of a scene in an image within 200 ms, as observed by Potter in the 1970s [27, 28], and by Schyns and Oliva [29] and Thorpe *et al.* [30] in the 1990s. This line of work often leads researchers to treat categorization as a data-driven process [31, 32, 33, 34, 35], mostly in a feed-forward network architecture [36, 37]. Such thinking has driven image classification research in computer vision and machine learning in the past decade and has achieved remarkable progress, including the recent success of DNNs [38, 39, 40].

Despite the fact that these approaches achieved good performances on scene categorization in terms of recognition accuracy in publicly available datasets, a recent large-scale neuroscience study [41] has shown that current DNNs cannot account for the image-level behavior patterns of primates (both humans and monkeys), calling attention to the need for more precise accounting for the neural mechanisms underlying primate object vision. Furthermore, data-driven approaches have led the focus of scene categorization research away from an important determinant of visual information—the categorization task itself [42, 43]. Simultaneously, these approaches have left unclear how classification interacts with scene semantics and enables cognitive reasoning. Psychological studies suggest that human vision organizes representations during the inference process even for "simple" categorical recognition tasks. Depending on a viewer's needs (and tasks), a kitchen can be categorized as an indoor scene, a place to cook, a place to socialize, or specifically as one's own kitchen (Fig. 1.5) [44]. As shown in Ref. [44], scene categorization and the information-gathering process are constrained by these categorization tasks [45, 46], suggesting a bidirectional interplay between the visual input and the viewer's needs/tasks [43]. Beyond scene categorization, similar phenomena were also observed in facial recognition [47].

In an early work, Ikeuchi and Hebert [48] proposed a task-centered representation inspired by robotic grasping literature. Specifically, without recovering the detailed 3D models, their analysis suggested that various grasp strategies require the object to afford different functional capabilities; thus, the representation of the same object can vary according to the planned task (Fig. 1.6) [48].

(a) (b) (c)

Figure 1.5: The experiment presented in Ref. [44], demonstrating the diagnostically driven, bidirectional interplay between top-down and bottom-up information for the categorization of scenes at specific hierarchical levels. (a) Given the same input image of a scene, subjects will show different gaze patterns if they are asked to categorize the scene at (b) a basic level (*e.g.*, restaurant) or (c) a subordinate level (*e.g.*, cafeteria), indicating a task-driven nature of scene categorization. Reproduced from Ref. [44] with permission of the authors, © 2014.

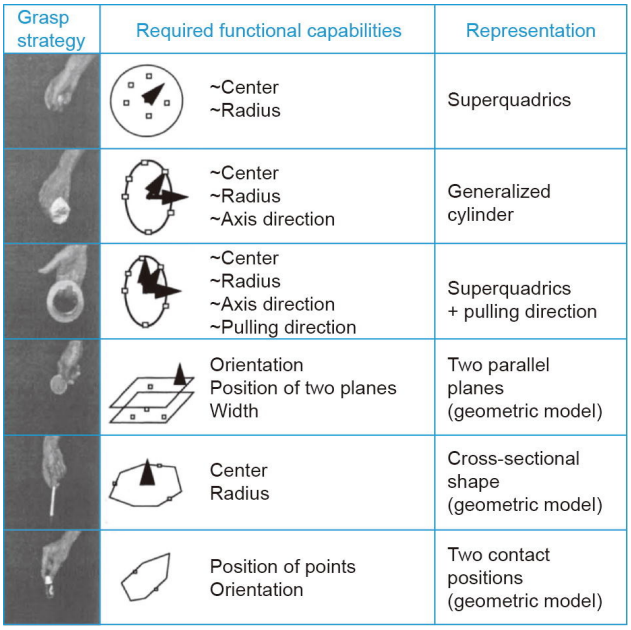


Figure 1.6: Different grasping strategies require various functional capabilities. Reproduced from Ref. [48] with permission of IEEE, © 1992.

For example, grasping a mug could result in two different grasps—the cylindrical grasp of the mug body and the hook grasp of the mug handle. Such findings also suggest that vision (in this case, identifying graspable parts) is largely driven by tasks; different tasks result in diverse visual representations.

"Where": Constructing 3D Scenes as a Series of Tasks

In the literature, approaches to 3D machine vision have assumed that the goal is to build an accurate 3D model of the scene from the camera/observer's perspective. These structure-from-motion (SfM) and simultaneous localization and mapping (SLAM) methods [49] have been the prevailing paradigms in 3D scene reconstruction. In particular, scene reconstruction from a single

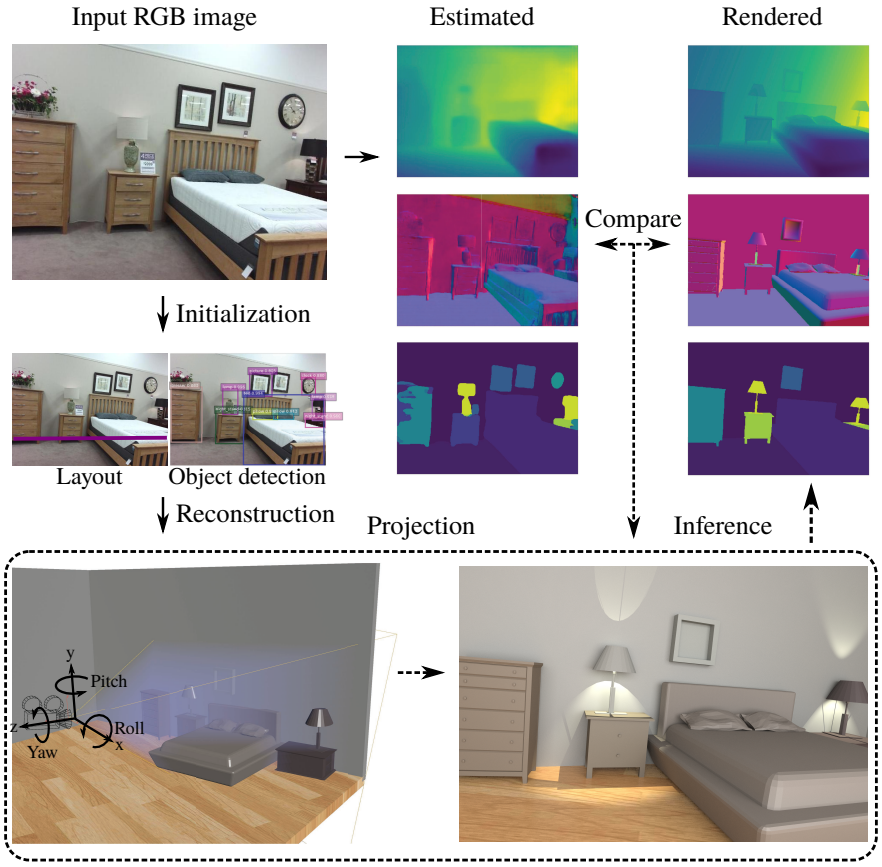


Figure 1.7: Illustration of 3D indoor scene parsing and reconstruction in an analysis-by-synthesis fashion [56]. A 3D representation is initialized by individual vision tasks (*e.g.*, object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered normal, depth, and segmentation maps and the ones estimated directly from the input RGB image, and adjusts the 3D structure iteratively. Reproduced from Ref. [56] with permission of Springer, © 2018.

two-dimensional (2D) image is a well-known ill-posed problem; there may exist an infinite number of possible 3D configurations that match the projected 2D observed images [50]. However, the goal here is not to precisely match the 3D ground-truth configuration, but to enable agents to perform tasks by generating the best possible configuration in terms of functionality, physics, and object relationships. This line of work has mostly been studied separately from recognition and semantics until recently [51, 52, 53, 54, 55, 56, 57, 58]; see Fig. 1.7 [56] for an example.

The idea of reconstruction as a "cognitive map" has a long history [59]. However, our biological vision system does not rely on such precise computations of features and transformations; there is now abundant evidence that humans represent the 3D layout of a scene in a way that fundamentally differs from any current computer vision algorithms [60, 61]. In fact, multiple experimental studies do not countenance global metric representations [62, 63, 64, 65, 66, 67]; human vision is error-prone and distorted in terms of localization [68, 69, 70, 71, 72]. In a case study, Glennerster *et al.* [73] demonstrated an astonishing lack of sensitivity on the part of observers to dramatic changes in the scale of the environment around a moving observer performing various tasks.

Among all the recent evidence, grid cells are perhaps the most well-known discovery to indicate the non-necessity of precise 3D reconstruction for vision tasks [74, 75, 76]. Grid cells encode a cognitive representation of Euclidean space, implying a different mechanism for perceiving and processing locations and directions. This discovery was later awarded the 2014 Nobel Prize in

Physiology or Medicine. Surprisingly, this mechanism not only exists in humans [77], but is also found in mice [78, 79], bats [80], and other animals. Gao *et al.* [81] and Xie *et al.* [82] proposed a representational model for grid cells, in which the 2D self-position of an agent is represented by a high-dimensional vector, and the 2D self-motion or displacement of the agent is represented by a matrix that transforms the vector. Such a vector-based model is capable of learning hexagon patterns of grid cells with error correction, path integral, and path planning. A recent study also showed that view-based methods actually perform better than 3D reconstruction-based methods in certain human navigation tasks [83].

Despite these discoveries, how we navigate complex environments while remaining able at all times to return to an original location (*i.e.*, homing) remains a mystery in biology and neuroscience. Perhaps a recent study from Vuong *et al.* [84] providing evidence for the task-dependent representation of space can shed some light. Specifically, in this experiment, participants made large, consistent pointing errors that were poorly explained by any single 3D representation. Their study suggests that the mechanism for maintaining visual directions for reaching unseen targets is neither based on a stable 3D model of a scene nor a distorted one; instead, participants seemed to form a flat and task-dependent representation.

Beyond "What" and "Where": Towards Scene Understanding with Humanlike Common Sense

Psychological studies have shown that human visual experience is much richer than "what" and "where." As early as infancy, humans quickly and efficiently perceive causal relationships (*e.g.*, perceiving that object A launches object B) [85, 86], agents and intentions (*e.g.*, understanding that one entity is chasing another) [87, 88, 89], and the consequences of physical forces (*e.g.*, predicting that a precarious stack of rocks is about to fall in a particular direction) [90, 91]. Such physical and social concepts can be perceived from both media as rich as videos [92] and much sparser visual inputs [93, 94].

To enable an artificial agent with similar capabilities, we call for joint reasoning algorithms on a joint representation that integrates (i) the "visible" traditional recognition and categorization of objects, scenes, actions, events, and so forth; and (ii) the "dark" higher level concepts of fluent, causality, physics, functionality, affordance, intentions/goals, utility, and so forth. These concepts can in turn be divided into five axes: fluent and perceived causality, intuitive physics, functionality, intentions and goals, and utility and preference, described below.

1.2.2 Fluent and Perceived Causality

A *fluent*, which is a concept coined and discussed by Isaac Newton [95] and Maclaurin [96], respectively, and adopted by AI and commonsense reasoning [97, 98], refers to a transient state of an object that is time-variant, such as a cup being empty or filled, a door being locked, a car blinking to signal a left turn, and a telephone ringing; see Fig. 1.8 for other examples of "dark" fluents in images. Fluents are linked to perceived causality [99] in the psychology literature. Even infants with limited exposure to visual experiences have the innate ability to learn causal relationships from daily observation, which leads to a sophisticated understanding of the semantics of events [100].

Fluents and perceived causality are different from the visual *attributes* [101, 102] of objects. The latter are permanent over the course of observation; for example, the gender of a person in a short video clip should be an attribute, not a fluent. Some fluents are visible, but many are "dark." Human cognition has the innate capability (observed in infants) [100] and strong inclination to perceive the *causal effects* between *actions* and *changes of fluents*; for example, realizing that flipping a switch

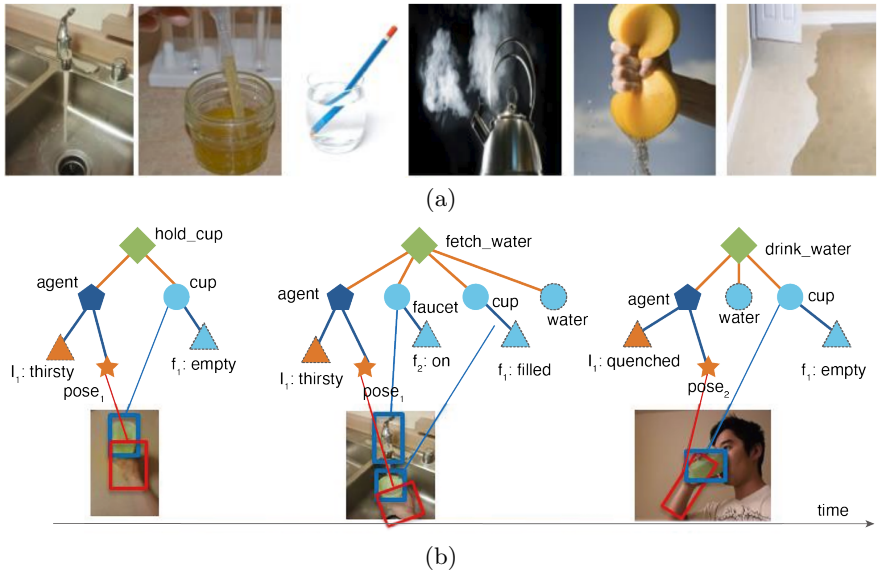


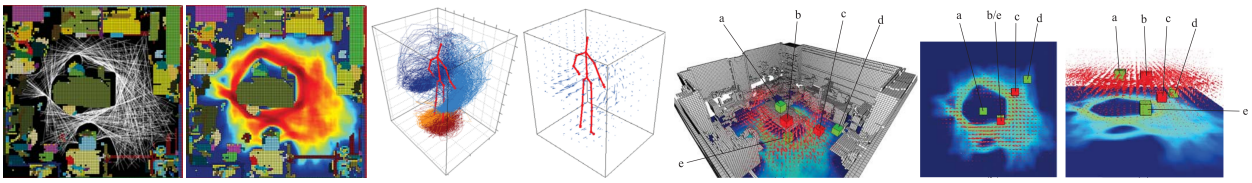
Figure 1.8: Water and other clear fluids play important roles in a human's daily life, but are barely detectable in images. (a) Water causes only minor changes in appearance; (b) the "dark" entities of water, fluents (here, a cup and faucet, represented by triangles), and the intention of a human are shown in dashed nodes. The actions (diamonds) involve agents (pentagons) and cups (objects in circles).

causes a light to turn on. To recognize the change in an object caused by an action, one must be able to perceive and evaluate the state of the object's changeable characteristics; thus, perceiving fluents, such as whether the light switch is set to the up or down position, is essential for recognizing actions and understanding events as they unfold. Most vision research on action recognition has paid a great deal of attention to the position, pose, and movement of the human body in the process of activities such as walking, jumping, and clapping, and to human-object interactions such as drinking and smoking [103, 104, 105, 106]; but most daily actions, such as opening a door, are defined by cause and effect (a door's fluent changes from "closed" to "open," regardless of how it is opened), rather than by the human's position, movement, or spatial-temporal features [107, 108]. Similarly, actions such as putting on clothes or setting up a tent cannot be defined simply by their appearance features; their complexity demands causal reasoning to be understood. Overall, the status of a scene can be viewed as a collection of fluents that *record the history of actions*. Nevertheless, fluents and causal reasoning have not yet been systematically studied in machine vision, despite their ubiquitous presence in images and videos.

1.2.3 Intuitive Physics

Psychology studies suggest that approximate Newtonian principles underlie human judgments about dynamics and stability [109, 110]. Hamrick *et al.* [91] and Battaglia *et al.* [90] showed that the knowledge of Newtonian principles and probabilistic representations is generally applied in human physical reasoning, and that an intuitive physical model is an important aspect of human-level complex scene understanding. Other studies have shown that humans are highly sensitive to whether objects in a scene violate certain understood physical relationships or appear to be physically unstable [111, 112, 113, 114, 115].

Invisible physical fields govern the layout and placement of objects in a human-made scene. By human design, objects should be physically stable and safe with respect to gravity and various other potential disturbances [116, 117, 118], such as an earthquake, a gust of wind, or the actions



(a) (b) (c) (d)

Figure 1.9: Inferring the potential for objects to fall from human actions and natural disturbances. (a) The imagined human trajectories; (b) the distribution of primary motion space; (c) the secondary motion field; (d) the integrated human action field, built by integrating primary motions with secondary motions. The five objects **a-e** are typical cases in the disturbance field: The objects **b** on the edge of a table and **c** along the pathway exhibit greater disturbance (in the form of accidental collisions) than other objects such as **a** in the center of the table, **e** below the table, and **d** in a concave corner of the room. Reproduced from Ref. [117] with permission of IEEE, © 2014.

of other humans. Therefore, any 3D scene interpretation or parsing (*e.g.*, object localization and segmentation) must be physically plausible [116, 117, 118, 119, 56, 120]; see Fig. 1.9. This observation sets useful constraints to scene understanding and is important for robotics applications [117]. For example, in a search-and-rescue mission at a disaster-relief site, a robot must be able to reason about the stability of various objects, as well as about which objects are physically supporting which other objects, and then use this information to move cautiously and avoid creating dangerous new disturbances.

1.2.4 Functionality

Most human-made scenes are designed to serve multiple human functions, such as sitting, eating, socializing, and sleeping, and to satisfy human needs with respect to those functions, such as illumination, temperature control, and ventilation. These functions and needs are invisible in images, but shape the scene's layout [121, 54], its geometric dimensions, the shape of its objects, and the selection of its materials.

Through functional magnetic resonance imaging (fMRI) and neurophysiology experiments, researchers identified mirror neurons in the pre-motor cortical area that seem to encode actions through poses and interactions with objects and scenes [122]. Concepts in the human mind are not only represented by prototypes—that is, exemplars as in current computer vision and machine learning approaches—but also by functionality [100].

1.2.5 Intentions and Goals

Cognitive studies [123] show that humans have a strong inclination to interpret events as a series of goals driven by the intentions of agents. Such a teleological stance inspired various models in the cognitive literature for intent estimation as an inverse planning problem [124, 125].

We argue that intent can be treated as the transient status of agents (humans and animals), such as being "thirsty," "hungry," or "tired." They are similar to, but more complex than, the fluents of objects, and come with the following characteristics: (i) They are hierarchically organized in a sequence of goals and are the main factors driving actions and events in a scene. (ii) They are completely "dark," that is, not represented by pixels. (iii) Unlike the instant change of fluents in response to actions, intentions are often formed across long spatiotemporal ranges. For example, in Fig. 1.10 [92], when a person is hungry and sees a food truck in the courtyard, the person decides (intends) to walk to the truck.

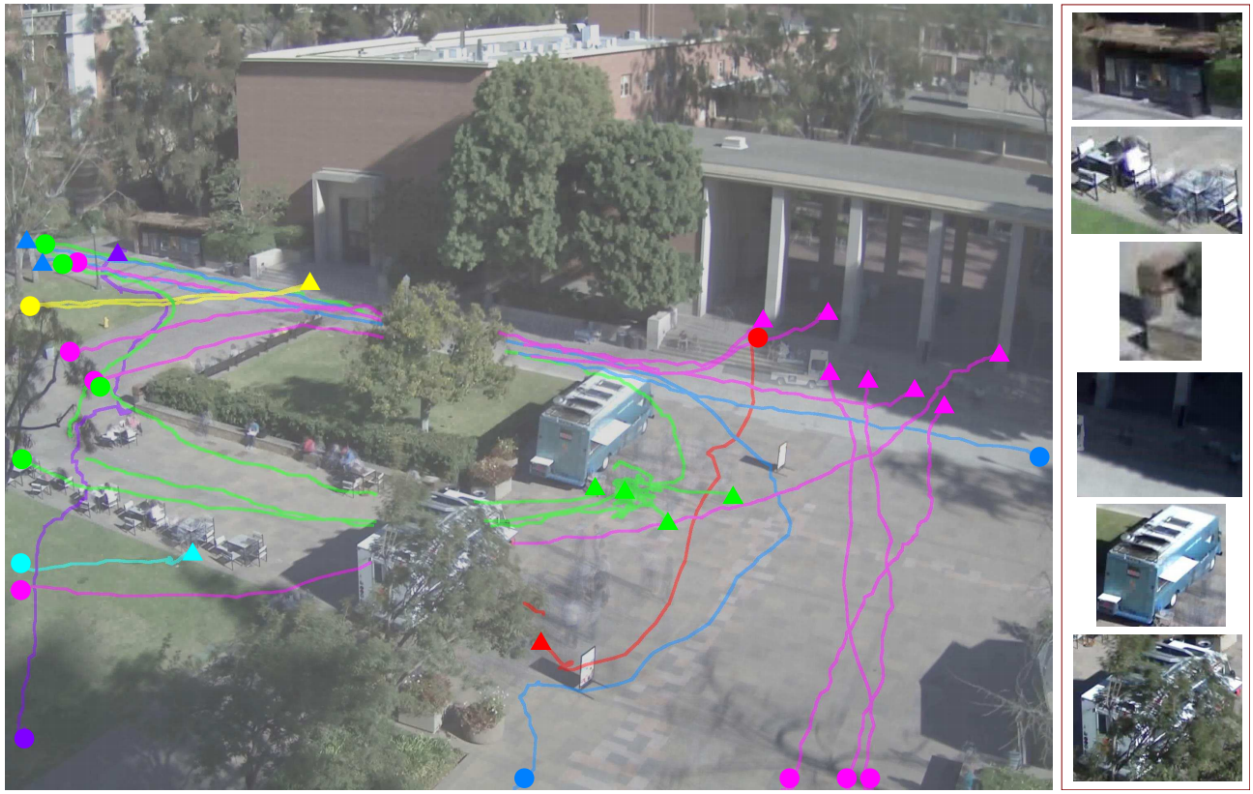


Figure 1.10: People's trajectories are color-coded to indicate their shared destination. The triangles denote destinations, and the dots denote start positions; *e.g.*, people may be heading toward the food truck to buy food (green), or to the vending machine to quench thirst (blue). Due to low resolution, poor lighting, and occlusions, objects at the destinations are very difficult to detect based only on their appearance and shape. Reproduced from Ref. [92] with permission of IEEE, © 2018.

During this process, an attraction relationship is established at a long distance. As will be illustrated later in this book, each functional object, such as a food truck, trashcan, or vending machine, emits a field of attraction over the scene, not much different from a gravity field or an electric field. Thus, a scene has many layers of attraction or repulsion fields (*e.g.*, foul odor, or grass to avoid stepping on), which are completely "dark." The trajectory of a person with a certain intention moving through these fields follows a least-action principle in Lagrange mechanics that derives all motion equations by minimizing the potential and kinematic energies integrated over time.

Reasoning about intentions and goals will be crucial for the following vision and cognition tasks: (i) early event and trajectory prediction [126]; (ii) discovery of the invisible attractive/repulsive fields of objects and recognizing their functions by analyzing human trajectories [92]; (iii) understanding of scenes by function and activity [45], where the attraction fields are longer range in a scene than the functionality maps [46, 127] and affordance maps [128, 129, 130] studied in recent literature; (iv) understanding multifaceted relationships among a group of people and their functional roles [131, 132, 133]; and (v) understanding and inferring the mental states of agents [134, 135].

1.2.6 Utility and Preference

Given an image or a video in which agents are interacting with a 3D scene, we can mostly assume that the observed agents make near-optimal choices to minimize the cost of certain tasks; that is,

we can assume there is no deception or pretense. This is known as the rational choice theory; that is, a rational person's behavior and decision-making are driven by maximizing their utility function. In the field of mechanism design in economics and game theory, this is related to the revelation principle, in which we assume that each agent *truthfully* reports its preferences; see Ref. [136] for a short introductory survey. Building computational models for human utility can be traced back to the English philosopher Jeremy Bentham, and to his works on ethics known as utilitarianism [137].

By observing a rational person's behavior and choices, it is possible to reverse-engineer their reasoning and learning process, and estimate their values. Utility, or values, are also used in the field of AI in planning schemes such as the Markov decision process (MDP), and are often associated with the states of a task. However, in the literature of the MDP, "value" is not a reflection of true human preference and, inconveniently, is tightly dependent on the agent's actions [138]. We argue that such utility-driven learning could be more invariant than traditional supervised training for computer vision and AI.

1.2.7 Summary

Despite their apparent differences at first glance, the five FPICU domains interconnect in ways that are theoretically important. These interconnections include the following characteristics: (i) The five FPICU domains usually do not easily project onto explicit visual features; (ii) most of the existing computer vision and AI algorithms are neither competent in these domains nor (in most cases) applicable at all; and (iii) human vision is nevertheless highly efficient in these domains, and human-level reasoning often builds upon prior knowledge and capability with FPICU.

We argue that the incorporation of these five key elements would advance a vision or AI system in at least three aspects:

1. Generalization. As a higher level representation, the FPICU concept tends to be globally invariant across the entire human living space. Therefore, knowledge learned in one scene can be transferred to novel situations.
2. Small sample learning. FPICU encodes essential prior knowledge for understanding the environment, events, and behavior of agents. As FPICU is more invariant than appearance or geometric features, the learning of FPICU, which is more consistent and noise-free across different domains and data sources, is possible even without "big data."
3. Bidirectional inference. Inference with FPICU requires the combination of top-down inference based on abstract knowledge and bottom-up inference based on visual pattern. This means that systems would both continue to make data-driven inferences from the observation of visible, pixel-represented scene aspects, as they do today, and make inferences based on FPICU understanding. These two processes can feed on each other, boosting overall system performance.

1.3 Cognitive Architecture for Human-Machine Communication and Teamwork

After the above brief review of key elements in humanlike common sense for computer vision and AI, a natural question arises: How could a machine master these common sense? In fact, human learning is a lifelong cognitive process of communicating with the physical and social world. Its sophistication, effectiveness, and complexity give rise to human intelligence—a phenomenon that AI is inspired to replicate. Decades of studies in cognitive psychology [139], anthropology, and communications studies [140] have revealed that human communication and learning is built on many layers of cognitive infrastructures and protocols.

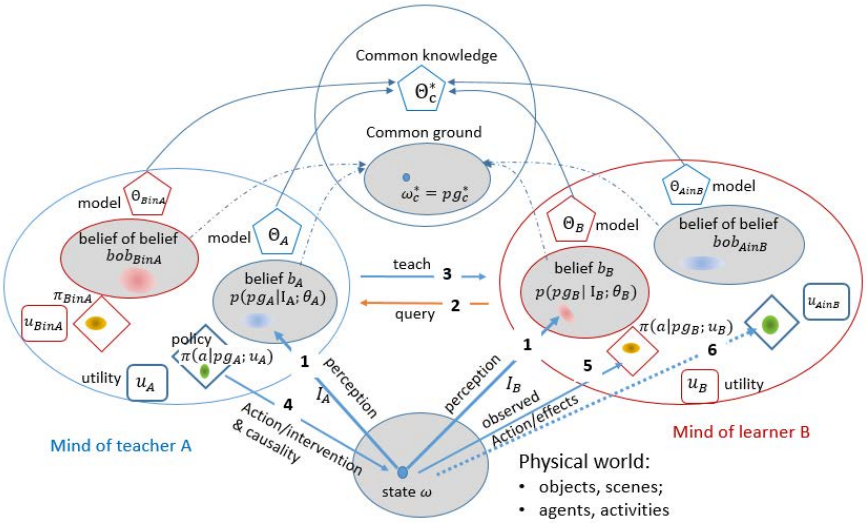


Figure 1.11: Illustration of the communicative learning protocol

To account for the complexity and sophistication in human communication and learning, we will formulate a new learning paradigm, called communicative learning (CL). Fig. 1.11 shows the key representations between two agents A for Alice and B for Bob, who can be human or machine, teacher or learner in an equal and symmetric setting, *i.e.*, they can exchange roles by turns.

Figure 2.1: A modern kitchen and an ancient kitchen with similar functions but drastically different geometry and appearances.

Chapter 2

Affordance and Functionality

Functionality refers to the property of an object or scene, especially man-made ones, which has a practical use for which it was designed. Psychologist [141] used another term, *affordance*, which refers to the property of an object that affords the opportunity for humans to perform some specific actions. From such view point, we argue that

- *objects, especially man-made ones, are defined by their functions and actions that they are involved.*
- *scenes, especially man-made ones, are defined by the activities and actions that they can provide space for.*

So, functionality is deeper than geometry and appearance and thus is a more invariant concept for scene understanding.

This represents a different philosophy that views vision tasks from the perspective of agents, that is, agents (humans, animals and robots) should perceive objects and scenes by reasoning their plausible functions.

Neuroscience studies also suggest similar ideas, indicating that the human vision system is far more capable than any existing computer vision systems and goes beyond merely memorizing the patterns based on pixels. For example, Fang and He showed that recognizing a face inside an image has a different mechanism compared to seeing an object that can be manipulated as a tool [25]; see Fig. 1.4. Other studies [26] also support the similar conclusion that the images of tool "potentiate" actions even when overt actions are not required in a task. Taking together, these results indicate

our biological vision system possesses another mechanism for perceiving object functionality (*i.e.*, how an object can be manipulated as a tool) which is independent of the mechanism in charge of face recognition (and other objects).

2.1 From data-driven scene understanding to task-driven scene understanding

Recent data-driven methods achieve remarkable performance in image classification and segmentation in computer vision during the past decade with the recent success of DNNs [38, 39, 40]. Despite the fact that these approaches achieved a good performance on scene categorization in terms of the recognition accuracy, they have led the focus of scene categorization research away from an important determinant of visual information—the categorization task itself [42, 43]. Simultaneously, these approaches have left it unclear how classification interacts with scene semantics and enables cognitive reasoning. Psychological studies suggest that human vision organizes representations during the inference process even for categorical recognition task. Depending on a viewer's needs (and tasks), a kitchen can be categorized as an indoor scene, a place to cook, a place to socialize, or specifically as my own kitchen (see Fig. 1.5). As shown in [44], scene categorization and the information gathering process are constrained by these categorization tasks [45, 46], suggesting a bidirectional interplay between the visual input and the viewer's needs/tasks [43]. In addition to the scene categorization, similar phenomenon was also found in face recognition [47].

In an early work, Ikeuchi and Herbert [23] proposed a task-centered representation inspired by robotic grasping literature. Specifically, without recovering the detailed 3D models, their analysis suggested that various grasp strategies require the object to afford different functional capabilities, and thus the representation of the same object can vary according to the tasks; see Fig. 1.6. For instance, grasping a mug could result in two different grasps—cylindrical grasp of the mug body and the hook grasp of the mug handle. Such findings also suggest that vision (identifying the parts to grasp in this case) is largely driven by tasks; different tasks result in diverse vision representations.

Therefore, in order to understand the scenes deeper with functionality, we should not only classify the visible pixels on image, but also understand the underlying actions and activities, which implies a shift from data-driven scene understanding to *task-oriented scene understanding*. The task-oriented scene understanding aims at understanding the 3D scenes by inferring the hidden functions and activities and pursuing a functional equivalence condition on the task that are involved or potentially involved. We believe the task-oriented scene understanding is a more human-like and more efficient way for understanding the scenes, and will take us to better human-like AI systems.

In this chapter we will discuss the scene functionality and object functionality. First we will show two methods that integrate functionality in 3D scene parsing and scene synthesis by inferring or considering the hidden human activities in 3D scenes. Then, we further introduce how we can make joint inference of scene parsing and human pose estimation by integrating human-object interaction and physics. To be noticed, human-object interactions are the key components for understanding the functionalities in 3D scenes since it encodes how the objects or scenes can afford the activities. Last we discuss a method that can be used to infer object functionality from videos of human-object interaction and another method that studies a specific function of objects—containment.

2.2 Hand-object Interactions: Grasping and Manipulation

In everyday life, the most common interaction that a person performs with any object is probably grasping. People grab different objects all the time, from cups to cellphones, from books to chop-

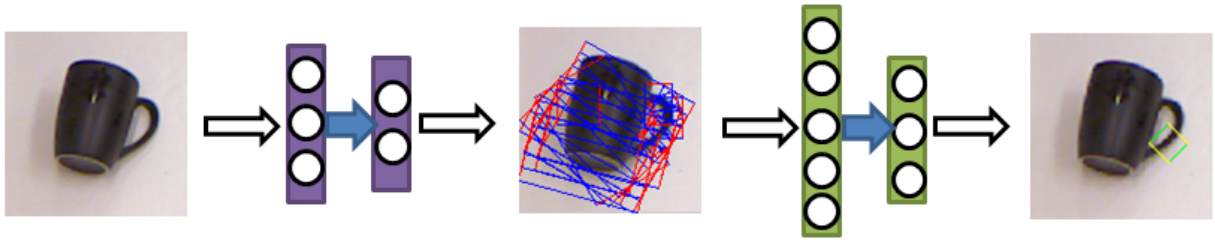


Figure 2.2: A pipeline [142] for predicting the cross-sections for parallel-jaw grippers

Figure 2.3: Different types of robotic grippers/hands.

sticks. Albeit being the most common interaction, grasping is a surprisingly complex behavior. A single-hand grasping involves controlling 16 joints of a hand, including 3 joints for each finger and a wrist joint. Such complexity makes it extremely difficulty for a robot to mimic human-level hand dexterity. To overcome the problem and empower robots for object manipulation, many simplified robotic hands and algorithms have been developed.

The most efficient robotic grasping solution might be using a parallel-jaw gripper. Parallel-jaw grippers are extremely simple and efficient at acquiring objects, and is widely used in industrial settings. For parallel-jaw grippers, the grasping strategy is simply predicting a cross-section to grip on [142] as shown in Fig. 2.2. An extension of it is the three-finger gripper as shown in Fig. 2.3

While parallel-jaw grippers are highly efficient, its simple design limits its ability of performing complex manipulations after a grasping is performed. This is especially significant for tools like cellphones and remote controls, where addition operation is desired after grasping it. To overcome this issue, four-finger and five-finger robotic hands are developed to supply more dexterous manipulation. However, the additional complexity in the state space is reflected in the difficulty of designing an algorithm for accurate grasping, not to mention further manipulations. Currently, the most commonly used method for performing grasping with a humanoid hand is by randomly sampling initial conditions, squeeze all finger joints, and then discard the ones that don't meet certain grasping criteria. This line of methods is surprisingly robust as it can always produce a grasping strategy as long as it is possible. However, these methods are extremely inefficient due to their trial-and-error nature, and they cannot guarantee to produce natural grasping.

A recent study [143] proposed to use infrared camera to capture the heat residual on objects after human demonstration of grasping, and synthesize robotic grasping by matching contact area with the collected heat residual. Such method is capable of synthesizing natural and functional grasping of objects.

It is also worth to notice that grasping is **task dependent**. Namely, we perform grasping

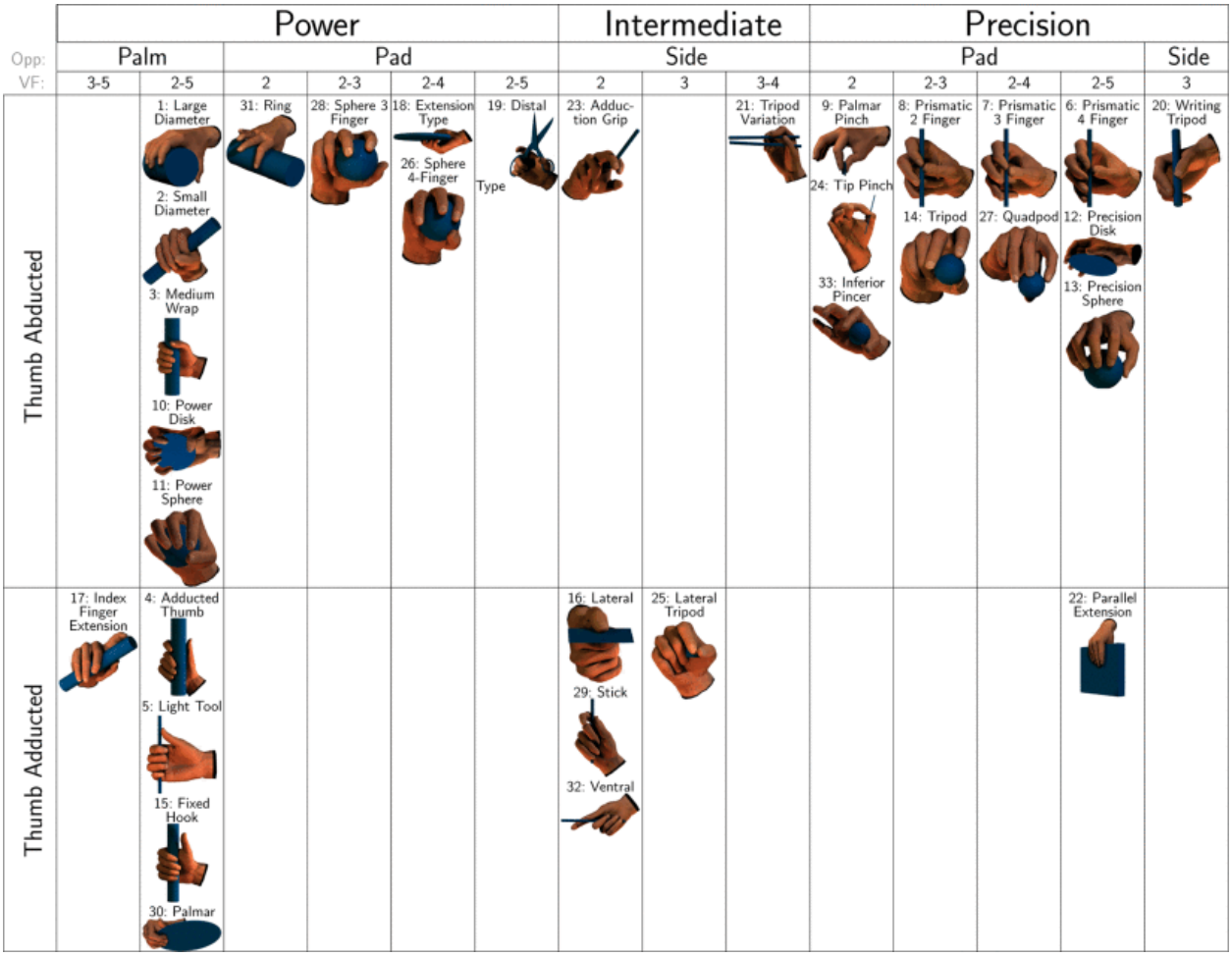


Figure 2.4: An illustration of different types of grasps [144]

differently for different tasks, even if we are grasping the same object. For example, we may use a precision grasp (see Fig. 2.4 to write with a pen, and we may naturally use a power grasp when we intend to move a pen. Both methods are valid and natural in their specific situation, and it is actually unnatural if we use the wrong grasping strategy.

To model the task-dependence nature of hand-object interaction, one must first synthesize diverse and stable grasps using a given hand. In this section, we introduce a method of grasp synthesis using differentiable force closure estimation.

2.2.1 Force Closure

A **force-closure** grasp is a grasp with contact points $\{x_i \in \mathbb{R}^3, i = 1, ..., n\}$ such that $\{x_i\}$ can resist arbitrary external wrenches with contact forces f_i, where f_i lies within the friction cones rooted from x_i. The angles of the friction cones are determined by the surface friction coefficient: The stronger the friction, the wider the cone. The force-closure metric is, therefore, irrelevant to the actual hand pose, but only relevant to the contact points and friction cones.

To test whether a set of contact points form a force-closure grasp, the first step is solving an optimization problem regarding contact forces rooted from the points [145, 146]. Although various methods have been devised, they all require iterations to jointly solve an auxiliary function, *e.g.*, a support function [147], a bilinear matrix inequality [148], or a ray shooting problem [149]. As

a result, solving force-closure grasps under the constraint of hand kinematics and force closure becomes a nested optimization problem.

Formally, given a set of n contact points $\{x_i \in \mathbb{R}^3, i = 1, ..., n\}$ and their corresponding friction cones $\{(c_i, \mu)\}$, where c_i is the friction cone axis and μ is the friction coefficient, a grasp is in *force closure* if there exists contact forces $\{f_i\}$ at $\{x_i\}$ within $\{(c_i, \mu)\}$ such that $\{x_i\}$ can resist arbitrary external wrenches. We follow the notations in Dai *et al.* [148] to define a set of contact forces to be force closure if it satisfies the following constraints:

$$GG' \geq \epsilon I_{6\times 6}, \tag{2.1a}$$

$$Gf = 0, \tag{2.1b}$$

$$f_i^T c_i > \frac{1}{\sqrt{\mu^2 + 1}} |f_i|, \tag{2.1c}$$

$$x_i \in S, \tag{2.1d}$$

where S is the object surface, and

$$G = \begin{bmatrix} I_{3\times 3} & I_{3\times 3} & ... & I_{3\times 3} \\ \lfloor x_1 \rfloor_\times & \lfloor x_2 \rfloor_\times & ... & \lfloor x_n \rfloor_\times \end{bmatrix}, \tag{2.2}$$

$$\lfloor x_i \rfloor_\times = \begin{bmatrix} 0 & -x_i^{(3)} & x_i^{(2)} \\ x_i^{(3)} & 0 & -x_i^{(1)} \\ -x_i^{(2)} & x_i^{(1)} & 0 \end{bmatrix}. \tag{2.3}$$

The form of $\lfloor x_i \rfloor_\times$ ensures the cross product $\lfloor x_i \rfloor_\times f_i = x_i \times f_i$, where $f = [f_1^T f_2^T ... f_n^T]^T \in \mathbb{R}^{3n}$ is the unknown variable of contact forces. In Eq. (2.1a), ϵ is a small constant. $A \succeq B$ means $A - B$ is positive semi-definite, *i.e.*, it is symmetric, and all its eigenvalues are non-negative. Eq. (2.1a) states that G is full rank. Eq. (2.1b) states that the contact forces cancel out each other so that the net wrench is zero. Eq. (2.1c) prevents f_i from deviating from the friction cone $\{(c_i, \mu)\}$. Eq. (2.1d) constrains contact points to be on the object surface.

2.2.2 Approximating Force Closure

Of note, Eq. (2.1b) is bilinear on x_i and f_i. Given a set of contact points $\{x_i\}$, verification of force closure requires finding a solution of $\{f_i\}$. The time complexity for computing such a solution is linear w.r.t. the number of contact points [148]. Here, we rewrite Eq. (2.1b) to

$$Gf = G(f_n + f_t) = 0, \tag{2.4a}$$

$$G\frac{f_n}{\|f_n\|_2} = -\frac{Gf_t}{\|f_n\|_2}, \tag{2.4b}$$

$$Gc = -\frac{Gf_t}{\|f_n\|_2}, \tag{2.4c}$$

where f_n and f_t are the normal and tangential components of contact force f in the force closure model, and $c = [c_1^T c_2^T ... c_n^T]^T$ is the set of friction cone axes. We obtain c_i as the surface normal of the object on x_i, which is easily accessible in many shape representations. We use Gc to approximate Gf, and therefore relax Eq. (2.1) to

$$GG' \geq \epsilon I_{6\times 6}, \tag{2.5a}$$

$$\|Gc\|_2 < \delta, \tag{2.5b}$$

$$x_i \in S, \tag{2.5c}$$

where δ is the maximum allowed error introduced from our relaxation. By adopting Eq. (2.5), we no longer need to solve the unknown variable f. The constraints of x_i becomes quadratic. Hence, the verification of force closure can now be computed extremely fast. The residual in $\|Gc\|_2$ reflects the difference between contact forces and friction cone axes.

To allow gradient-based optimization, we further cast Eq. (2.5) as a soft constraint in the form

$$FC(x, O) = \lambda_0(GG' - \epsilon I_{6\times 6}) + \|Gc\|_2 + w \sum_{x_i \in x} d(x_i, O), \tag{2.6}$$

where $\lambda_0(\cdot)$ gives the smallest eigenvalue, and $d(x, O)$ returns the distance from point x to the surface of object O. The scalar w controls the weight of the distance term. By minimizing the three terms, we are looking for $\{x_i\}$ that satisfies the three constraints in Eq. (2.5), respectively.

Using surface normal vectors to approximate contact forces implies zero friction and equal magnitude contact forces. Such an assumption may *seem* to eliminate a large pool of force-closure contact-point compositions. In practice, however, this is not the case: A residual in $\|Gc\|_2$ indicates that the existence of friction f_t and difference in force magnitude f_n on contact forces. By allowing the residual to be smaller than a reasonable threshold δ, we are allowing both the tangential and the normal components of the contact forces to deviate within reasonable range.

2.2.3 Grasp Synthesis

We formulate the grasp synthesis problem as sampling from a conditional Gibbs distribution:

$$P(H|O) = \frac{P(H, O)}{P(O)} \propto P(H, O) = \frac{1}{Z} \exp^{-E(H,O)}, \tag{2.7}$$

where Z denotes the intractable normalizing constant, H the hand, O the object, and $E(H, O)$ the energy function. We rewrite $E(H, O)$ as the minimum value of the energy function $E_{\text{grasp}}(H, x, O)$ w.r.t. contact point choices x:

$$\begin{aligned} E(H, O) &= \min_{x \subset S(H)} E_{\text{grasp}}(H, x, O) \\ &= \min_{x \subset S(H)} FC(x, O) + E_{\text{prior}}(H) + E_{\text{pen}}(H, O), \end{aligned} \tag{2.8}$$

where $S(H)$ is a set of points sampled uniformly from the surface of a hand with pose H. We denote the selected contact points from hand surface as $x \subset S(H)$. $FC(x, O)$ is the soft constraint from Eq. (2.6). $E_{\text{prior}}(H)$ is the *energy prior* of the hand pose. Its exact form depends on the hand definition. The *penetration energy* is defined as $E_{\text{pen}}(H, O) = \sum_{v \in S(H)} \sigma(v, O)$, where $\sigma(v, O)$ is a modified distance function between a point v and an object O:

$$\sigma(v, O) = \begin{cases} 0 & \text{if } v \text{ outside } O \\ |d| & \text{otherwise} \end{cases}, \tag{2.9}$$

where d is the distance from v to surface of O.

Due to the complexity of human hand kinematics, our grasp energy suffers from a complex energy landscape. A naïve gradient-based optimization algorithm is likely to stop at sub-optimal local minima. We use a modified Metropolis-adjusted Langevin algorithm (MALA) to overcome this issue; see the algorithm details in Algorithm 1. The random walk aspect of Langevin dynamics provides the chance of escaping bad local minima. Our algorithm starts with random initialization of hand pose H and contact points $x \subset S(H)$. Next, we run our algorithm L iterations to update H, x and maximize $P(H, O)$. In each iteration, our algorithm randomly decides to update either the

Algorithm 1: Modified MALA Algorithm

Input: Energy function E_{grasp}, object shape O, step size η, Langevin steps L, switch probability ρ
Output: grasp parameters H, x

1 Initialize H, x
2 **for** step $= 1 : L$ **do**
3 **if** rand() $< \rho$ **then**
4 Propose H^* according to Langevin dynamics

$$H^* = H - \frac{\eta^2}{2}\frac{\partial}{\partial H}E_{\text{grasp}}(H, x, O) + \eta\epsilon,$$

  where $\epsilon \sim N(0, 1)$ is a Gaussian noise
5 **else**
6 Propose x^* by sampling from $S(H)$
7 **end**
8 Accept $H \leftarrow H^*, x \leftarrow x^*$ by Metropolis-Hastings algorithm using energy function E_{grasp}
9 **end**

hand pose by Langevin dynamics or one of the contact points to a point uniformly sampled from the hand surface. The updates are accepted or rejected according to the Metropolis-Hastings algorithm, in which a lower-energy update is more likely to be randomly accepted than a higher-energy one.

Of note, different compositions of contact points in fact correspond to different grasp types as they contribute to some of the classification basis of the grasp taxonomy, including virtual finger assignment and opposition type. Hence, sampling contact points on Line 6 is crucial for exploring different types of grasps. In practice, we also empirically find that this step is essential for escaping bad local minima.

While our modified MALA algorithm can produce realistic results, we still observe physical inconsistencies in the synthesized examples such as penetrations and gaps between contact points and object surface. To resolve these issues, we further refine the synthesized results by minimizing E_{grasp} using gradient descent on H. We do not update the contact point selection x in this step, since we hope to focus on optimizing the physical consistency in this step rather than exploring the grasp landscape for diverse grasp types.

2.2.4 Results

Fig. 2.5 shows a collection of synthesis results with and without the refinement step: Higher values of our force closure estimation corresponds to non-grasps, whereas force closure estimations close to zero are as good as the ones with force closure estimations equal to zero. The last column shows two cases when the synthesis is trapped in bad local minima; these examples exhibit large values in our force closure estimation. Such errors happened because of the non-convexity of the optimization problem; one cannot avoid every bad minimum with gradient-based methods. Fortunately, we can identify these examples by their high force closure scores.

The diversity of the synthesized grasp is qualitatively evaluated by inspecting the energy landscape plotted by the ADELM algorithm [150]. The landscape is projected to a disconnectivity graph in Fig. 2.6. In the disconnectivity graph, each circle at the bottom represents a local minima group. The size of the circle indicates how many synthesized examples fall into this group. The height of the horizontal bar between two groups represent the maximum energy (or energy barrier) along the minimum energy pathways (MEPs) between two groups. The MEPs with lowest barriers connect smaller groups into larger groups, and this process is repeated until all examples are connected.

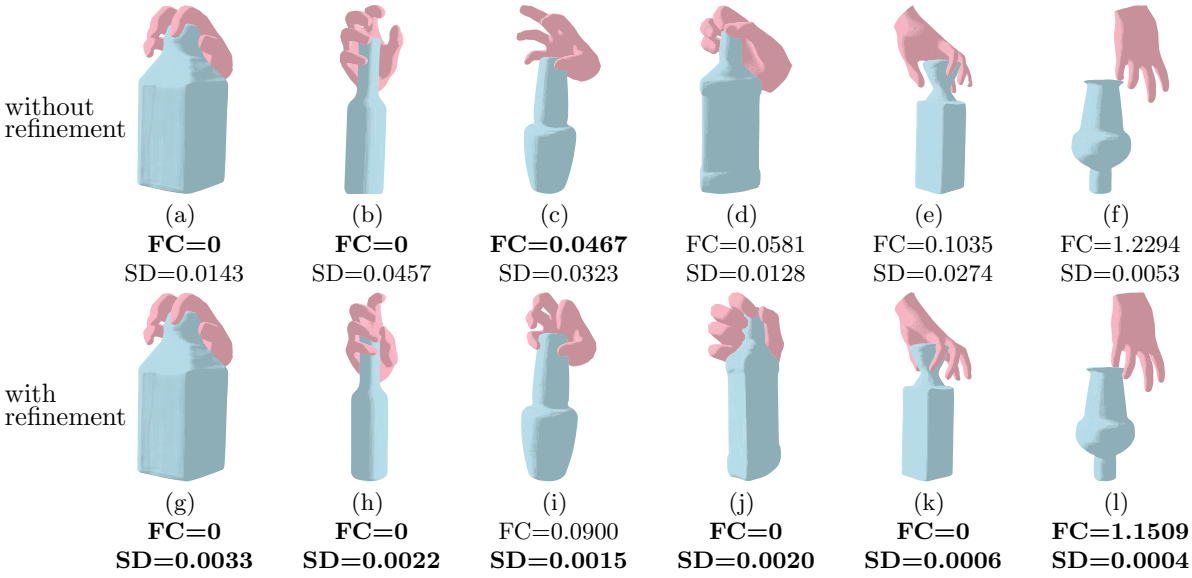


Figure 2.5: **Examples of synthesized grasps.** Top: synthesized grasps before refinement. Bottom: the same set of synthesized grasps after refinement. FC: estimated force closure error. SD: mean distance from each contact point to the object surface. Left to right: examples with zero FC error, small FC error, and high FC error qualitatively illustrate how our estimation of force closure correlates to grasp quality.

The produced disconnectivity graph is an estimation of the true landscape of the energy function. Energy landscape mapping in Fig. 2.6 shows that the local minima with low energy barriers between them have similar grasps, and those with high energy barriers between them tend to have different grasps. We also observe that the energy landscape contains all three categories in the power/precision dimension as described in Feix *et al.* [144].

The algorithm can also synthesize rare grasps that are not presented in existing grasping taxonomies, as shown in Fig. 2.7. These grasps are rarely collected in any of the modern 3D grasp datasets (*e.g.*, [151, 152]), since they do not belong to any type as defined in the grasp taxonomy. However, these grasps are valid grasps and could well exist in physical interactions. For example, the left example in Fig. 2.7 is commonly used to twist-open a bottle when some of the fingers are occupied or injured. The second example would occur if one is already holding something in the palm while picking up another bottle.

These grasps occur because the human hand is excellent in doing multiple tasks simultaneously, which have not been recognized in grasp literature as we always assumed otherwise. Such limitation would hinder a robotic hand's capacity from developing to its full potential.

2.2.5 Grasp Synthesis for Arbitrary Hand Structure

Although the shown results are all in the form of a humanoid form, this algorithm in fact makes no assumption on the hand kinematics except for having a differentiable mapping between pose and shape. As a result, we can synthesize grasps for arbitrary hand so long as there exists such a mapping. In Fig. 2.8, we show the synthesis results of applying this algorithm to synthesize grasps of a humanoid hand with its thumb removed and a Robotiq 3-finger gripper. These examples demonstrate that this method can explore a wide range of grasps for arbitrary hand structure, which could provide valuable insights for understanding the task affordance of prosthetic or robotic hands, and hands with injuries or disabilities. This method is also applicable to animations, wherein grasps of non-standard hands or claws are common.

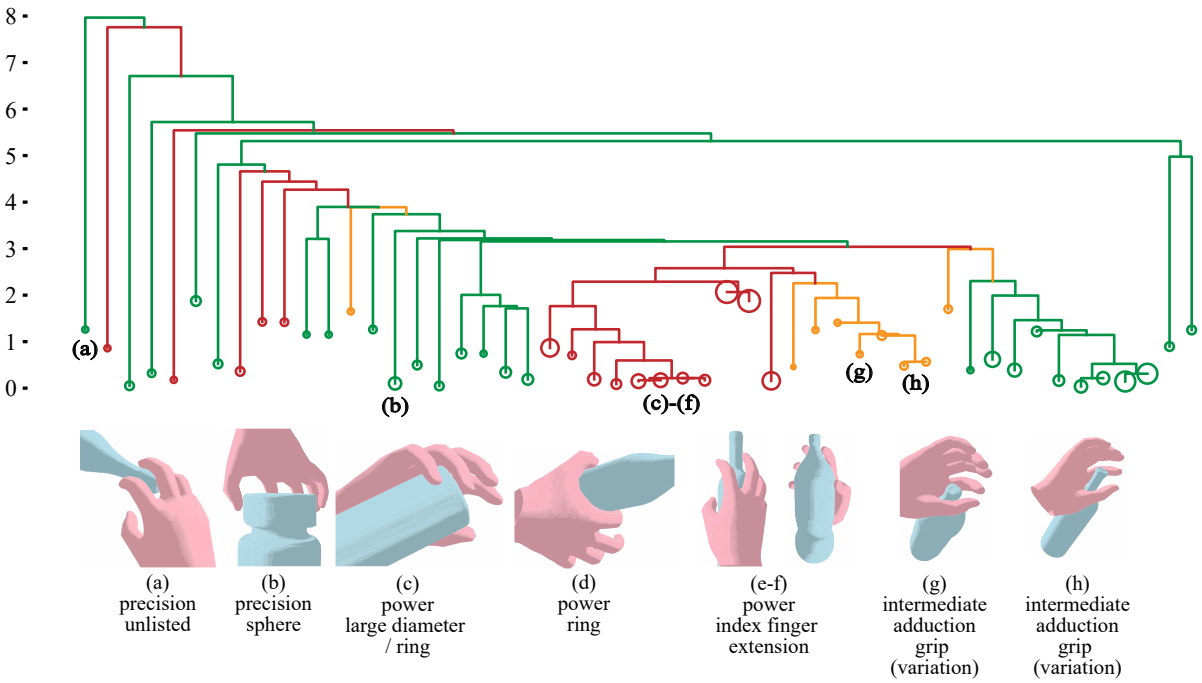


Figure 2.6: **Energy landscape mapping generated by the ADELM algorithm [150] (best viewed in color).** Top: disconnectivity diagram of the energy landscape of our energy function $E(H, O)$. Green minima denote precision grasps, red power grasps, and yellow intermediate grasps. Bottom: examples from selected local minima; minima with lower energy barriers in between have similar grasps. We also label the grasp taxonomy of each example according to [144]. Examples marked as *unlisted* do not belong to any manually classified type.

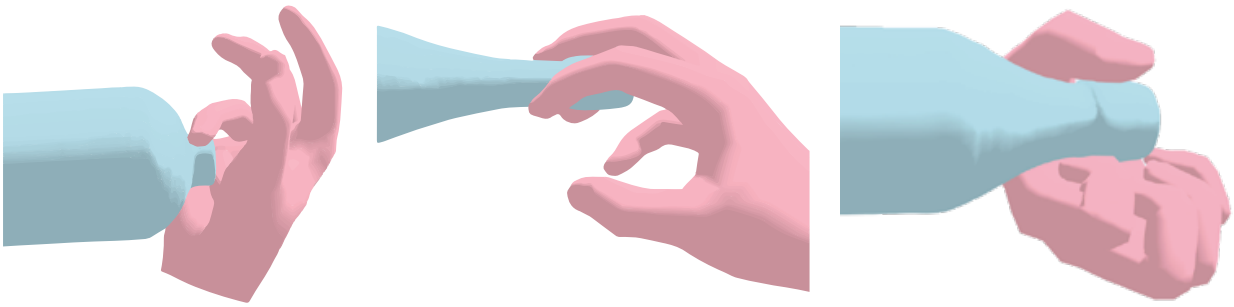


Figure 2.7: **Examples of novel grasp poses.** To the best of our knowledge, these newly discovered grasp poses do not correspond to any grasp types in existing human-designed grasp taxonomy.

2.2.6 Limitations

We show two representative failure cases in Fig. 2.9, wherein an unstable or unrealistic grasp receives a low force closure score. Most failure cases of this method are caused by concavities in object shapes. For concave shapes, the force closure requirement is sometimes satisfied with a single finger in the concavity, providing contact forces in opposing directions. The issue may be eliminated with manually defined heuristics, such as enforcing contact points on different fingers or encouraging contact points to have larger distances between each other. Another common failure comes from model intersections. The original implementation tests penetration by computing the signed distance between hand surface vertices and the object shape. When the vertices are sparse,

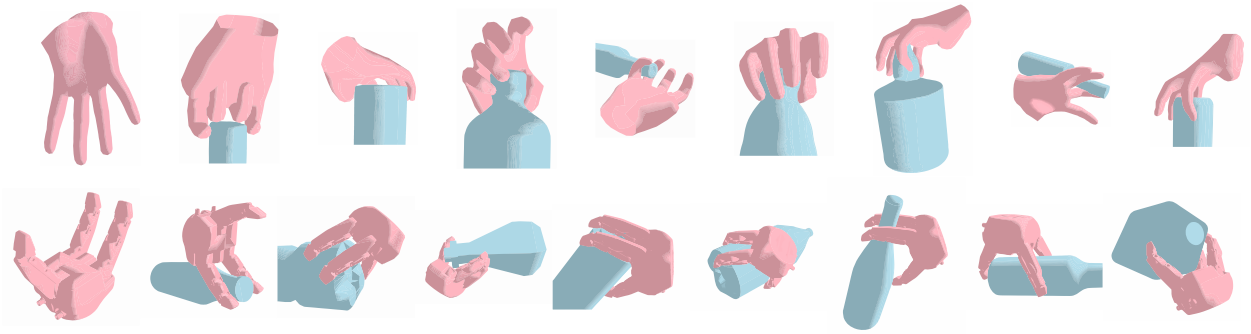
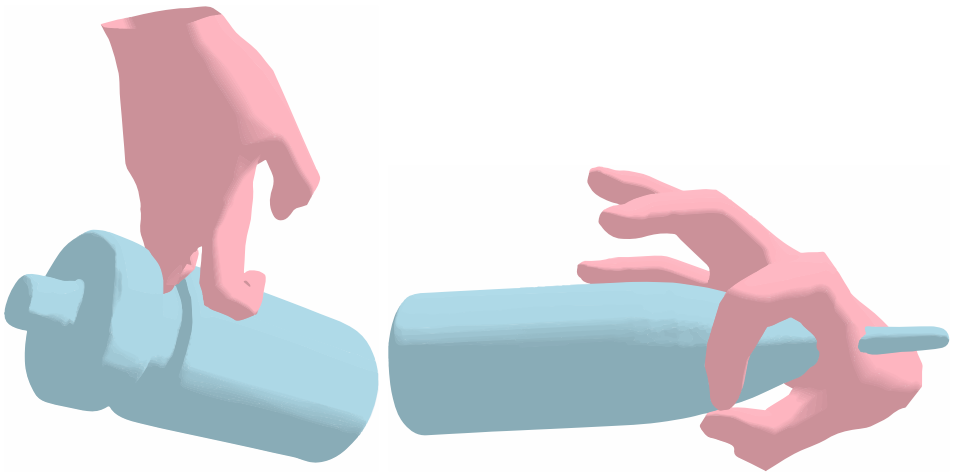


Figure 2.8: **Synthesized grasps of different hands using our formulation.** Top: A MANO hand with its thumb removed. Bottom: A Robotiq 3-finger gripper. The left-most figure shows the hand used in each row.



(a) (b)

Figure 2.9: **Failure case examples.** (a) Concave shape results in force closure configuration that is not a grasp. (b) Sparse penetration detection leads to intersection.

or the object has a pointy part, it is possible for the object to penetrate the hand without being detected. This issue can be addressed by using a dense sample of hand surface vertices or adopting a differentiable mesh intersection algorithm.

2.3 Human-object Interactions

2.3.1 Functional Object Parts

To investigate the functionality and affordances for different parts of objects, efforts have been devoted to the compositionality and hierarchy of objects. ShapeNet is a large, information-rich repository of 3D models. It contains models spanning a multitude of semantic categories. Many objects, especially manmade artifacts such as furniture and appliances, can be used by humans. See Fig. 2.10 for example.

Functional annotations describe these usage patterns. Such annotations are often highly correlated with specific regions of an object. In addition, it is often related with the specific type of human action. ShapeNet aims to store functional annotations at the global shape level and at the object part level.

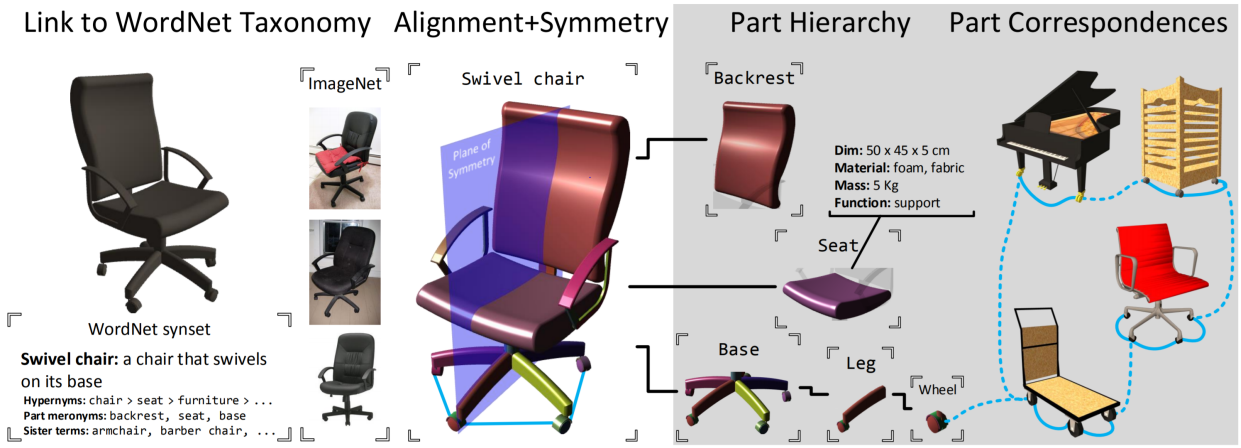


Figure 2.10: Examples of part-based annotation and hierarchy in ShapeNet.

1. Functional Parts: Parts are critical for understanding object structure, human activities involving a 3D shape, and ergonomic product design. We plan to annotate parts according to their function — in fact the very definition of parts has to be based on both geometric and functional criteria.
2. Affordances: We are interested in affordance annotations that are function and activity specific. Examples of such annotations include supporting plane annotations, and graspable region annotations for various object manipulations.

Physical Annotations: Real objects exist in the physical world and typically have fixed physical properties such as dimensions and densities. Thus, it is important to store physical attribute annotations for 3D shapes.

1. Surface Material: We are especially interested in the optical properties and semantic names of surface materials. They are important for applications such as rendering and structural strength estimation.
2. Weight: A basic property of objects which is very useful for physical simulations, and reasoning about stability and static support.

2.3.2 Synthetic Human Activities with Dynamic Environment Dataset

We collect SHADE (Synthetic Human Activities with Dynamic Environment), a self-annotated dataset that consists of dynamic 3D human skeletons and objects, to learn the prior model for each HOI. It is collected from a video game Grand Theft Auto V with various daily activities and HOIs. Currently, there are over 29 million frames of 3D human poses, where 772,229 frames are annotated. On average, each annotated frame is associated with 2.03 action labels and 0.89 HOIs. The SHADE dataset contains 19 fine-grained HOIs for both indoor and outdoor activities. By selecting most frequent HOIs and merging similar HOIs, we choose 6 final HOIs: *read [phone, notebook, tablet], sit-at [human-table relation], sit [human-chair relation], make-phone-call, hold, use-laptop*. Fig. 2.11 shows some typical examples and relations in the dataset.

2.3.3 4DHOI

In this section we introduce a new formulation of 4-dimensional human object interaction (4DHOI). The 4DHOI model describes the 3-dimensional geometrical relationship between human and object during an interaction over time. The name **4D** comes from the combination of regular 3D space

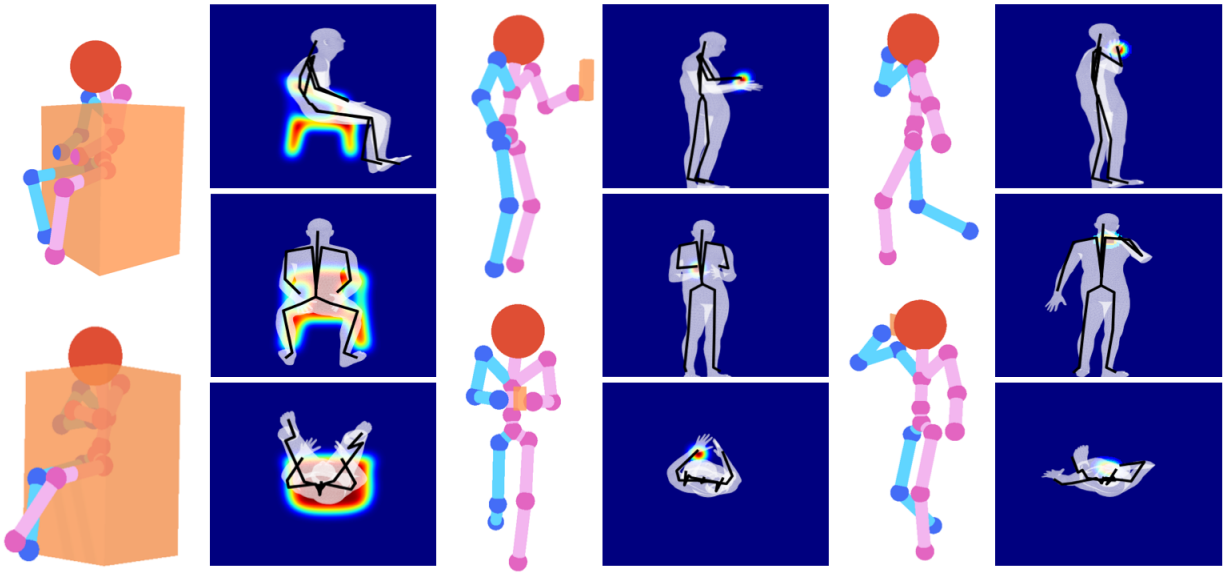


Figure 2.11: Examples of typical HOIs and examples from the SHADE dataset. The heatmap indicates the probable locations of HOI.

and the additional dimension of time. A 4DHOI unit (shown in Fig. 2.12) is a node in an attributed spatial-temporal parse graph (X-ST-pg) of a video that bridges the spatial and the temporal.

- **Spatial relationship.** Each 4DHOI unit describes a set of geometric relationships between an object and a human part.
- **Temporal relationship.** Each 4DHOI unit is a basic component of human activities.

Fig. 2.13 shows an example of 4DHOI units in real-life images.

Before jumping into the definition of a 4DHOI unit, we need to first define the attributed spatial-temporal parse graph that the 4DHOI units live in. Formally, an X-ST-pg $G = \langle V_N^S, V_T^S, E^S, X^S, V_N^T, V_T^T, E^T \rangle$ is the parsing result of an attributed spatial-temporal AOG.

- V_N^O is the set of non-terminal nodes in the spatial parse graph of objects.
- V_T^O is the set of terminal nodes in the spatial parse graph of objects.
- E^O is the set of edges in the spatial parse graph of objects.
- V_N^H is the set of non-terminal nodes in the spatial parse graph of human.
- V_T^H is the set of terminal nodes in the spatial parse graph of human.
- E^H is the set of edges in the spatial parse graph of human.
- V_N^T, V_T^T, E^T are the respective nodes and edges in the temporal parse graph.
- X^S is a function $X^S : V^S \longrightarrow \mathbb{X}$ whose input is a node in S-pg and whose output is its geometrical attributes. $V^S = V_N^O \cup V_T^O \cup V_N^H \cup V_T^H$ is the set of all nodes in the S-pg and $\mathbb{X}$ denotes the space of geometrical attributes.

Based on the definition of X-ST-pg, a 4DHOI unit is defined as $U = \langle v^T, \mathcal{R} \rangle$, where $\mathcal{R} = \{R : R = \langle f, p_O, p_H \rangle\}$ is a set of geometrical relationship R's. Each geometrical relationship R contains a pointer p_O to an object node in S-pg, a pointer p_H to a human node in S-pg, and an energy function $f : \mathbb{X}^2 \longrightarrow \mathbb{R}$ to describe the geometrical relationship between the two nodes. The 4DHOI units replace the temporal terminal nodes V_T^T.

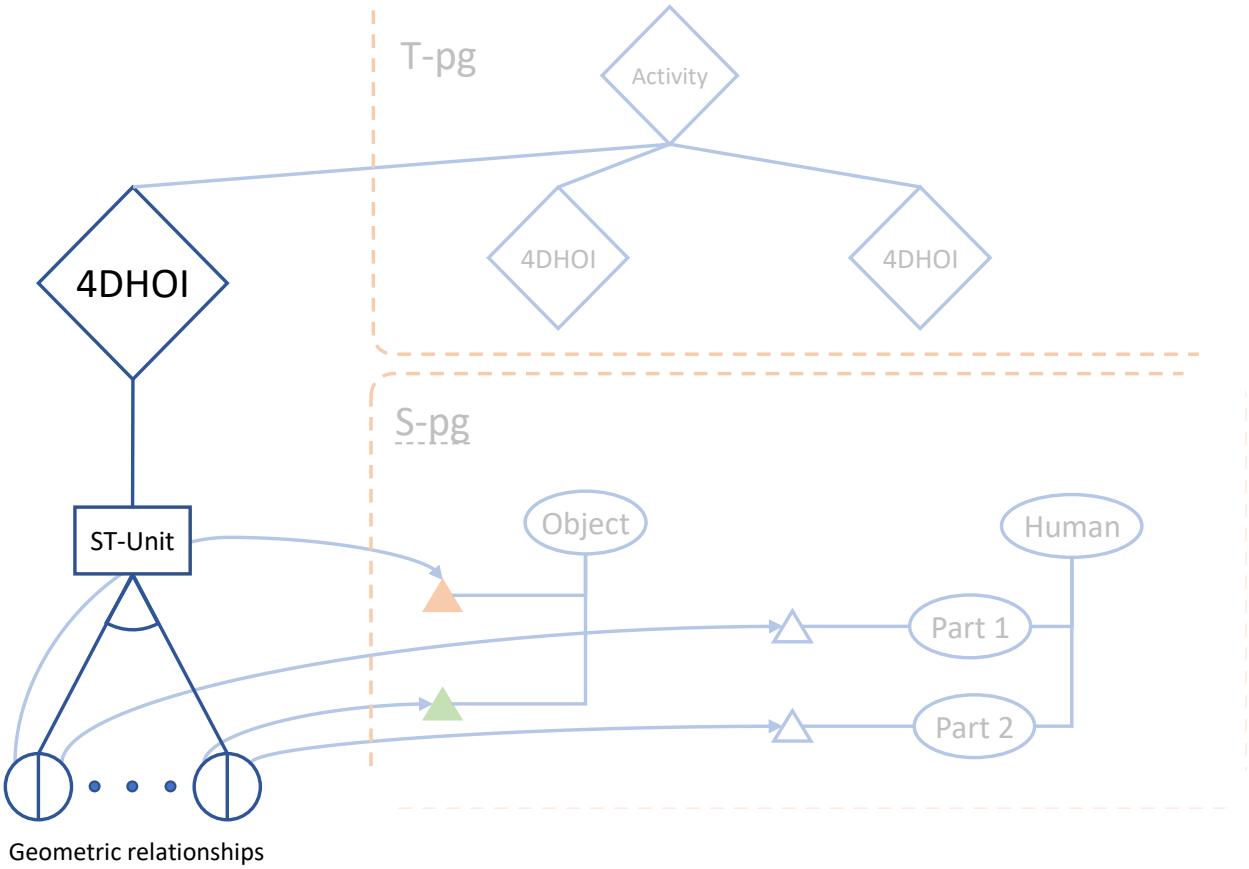


Figure 2.12: Structure of a 4DHOI unit

2.3.4 Learning 4DHOI from Video

In [153], the authors proposed a method of learning 4DHOI from video. In this work, the authors formulate 4DHOI with a stochastic hierarchical graph similar to And-Or Graph [154]. Let $V = (f_1, ..., f_\tau)$ be a video sequence in the time interval $[1, \tau]$, where $f_t = (I_t, h_t)$ is the frame at time t. I_t is the RGB-D data. h_t is the human pose feature extracted from the 3D skeletons estimated by motion capture technology [155].

The sequence V is interpreted by the hierarchical graph $G =< E, L >$ as follows.

i) $E \in \Delta = \{e_i | i = 1, ..., |\Delta|\}$ is the event category such as *fetch water from dispenser*. Δ is the set of event categories.

ii) $L = (l_1, ..., l_\tau)$ is a sequence of frame labels. $l_t = (a_t, o_t)$ is the interpretation of the frame f_t. $a_t \in \Omega_E = \{\omega_i | i = 1, ..., K_E\}$ is the atomic event label such as *fetch water*.

Ω_E is the atomic event set of E. Each event category e_i has its distinct atomic event set Ω_{e_i}, i.e. the relations between an event and its atomic events are hard constraints.

$o_t = (o_t^1, ..., o_t^{n_t})$ are the objects interacting with the human at time t, where n_t is the number of objects. Each object has a class label and a 3D location.

Similar to the graphical formulation in [154], the energy that the video V is interpreted by the graph G is defined as

$$\mathrm{En}(G|V) = \sum_{t=1}^{\tau} \Phi(f_t, l_t) + \sum_{t=2}^{\tau} \Psi(l_{1:t-1}, l_t). \tag{2.10}$$

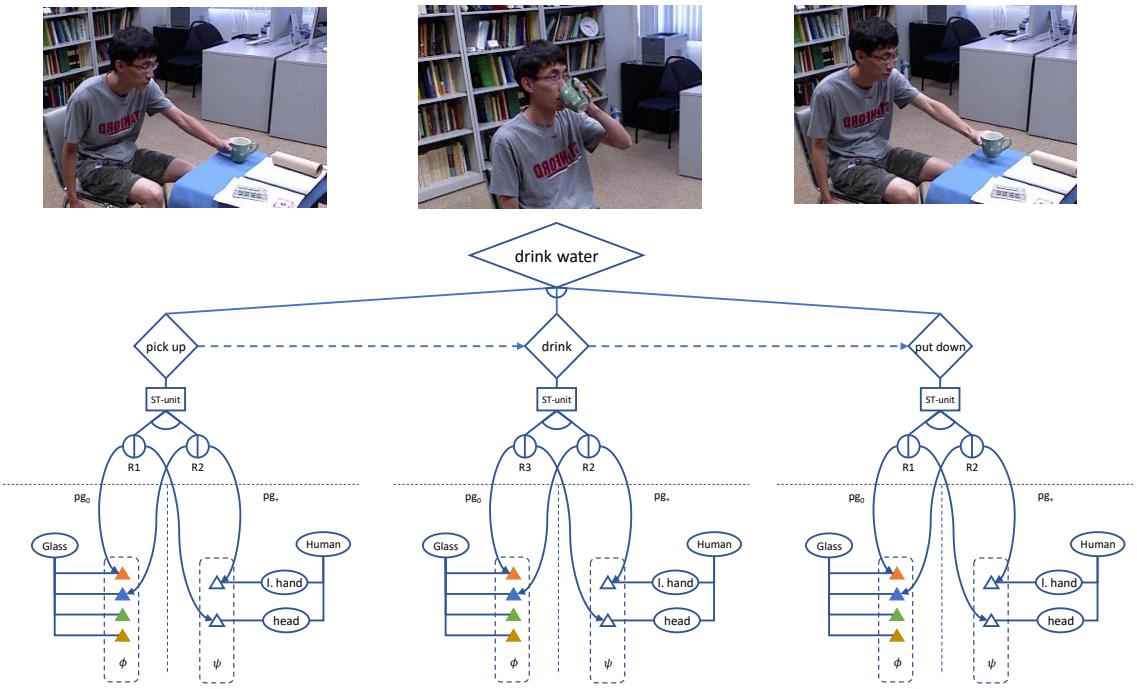


Figure 2.13: Structure of a 4DHOI unit

$\Phi(\cdot)$ is the spatial energy term of a single frame, encoding the human-object interactions in 3D space.

$\Psi(\cdot)$ is the temporal energy term encoding the temporal relations between the current frame l_t and all previous frames $l_{1:t-1}$. This is different from conventional HMM [156]. The variable E is omitted in the right side of Eq. (2.10)) since each event has its own distinct atomic event set.

$\Phi(f_t, l_t)$ in Eq. (2.10) describes the human-object interactions in 3D space. The interactions include:

- semantic co-occurrence between a specific type of human pose and the object classes; and
- geometric compatibility describing the 3D spatial constraints between the human pose and the objects.

Thus, $\Phi(f_t, l_t)$ is further decomposed into three terms which will be defined in the remainder of the subsection,

$$\Phi(f_t, l_t) = \phi_1(a_t, h_t) + \phi_2(a_t, o_t, I_t) + \phi_3(a_t, h_t, o_t), \tag{2.11}$$

where $\phi_1(a_t, h_t)$ is the pose model, $\phi_2(a_t, o_t, I_t)$ is the contextual object model, and $\phi_3(a_t, h_t, o_t)$ is the 3D geometrical relationship between human and object.

Applications

Object Functionality The functionality of an object, under the definition of 4DHOI, can be described as the set of 4DHOI units that can connect with the object and produce low energy expectation with regard to interacting human.

3D Scene Reconstruction Each 4DHOI unit describes a combination of geometrical relationships between nodes in S-pg. These relationships serve as constraints in 3D reconstruction. Given a noisy 3D reconstruction of a scene and a noisy 3D pose estimation of a person in that scene, we

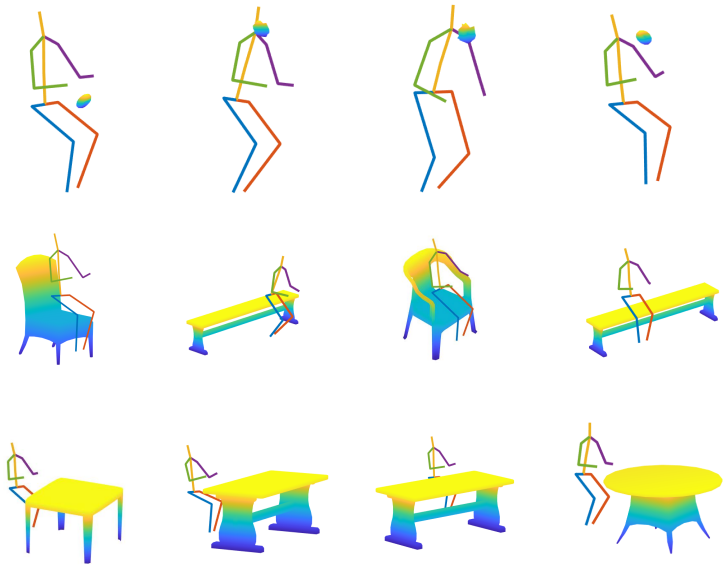


Figure 2.14: Imagined objects. Row 1: Eating. Row 2: Sitting. Row 3: Sitting-in-front-of.

can use such constraints to jointly optimize for 3D scene and 3D human. An example of utilizing 4DHOI in 3D scene reconstruction is illustrated in the next section.

Contextual Object Localization Objects, especially those that are interacting with human, are often occluded by the interacting agent and therefore are very difficult, if not impossible, to detect. With 4DHOI, we can infer the location of an interacting object if we know the type of the interaction and the 3D pose of the interacting agent. Fig. 2.14 shows an example of imagined objects given 3D poses and three different interaction types for each pose.

2.4 Functionality Grammar for 3D Scene Synthesis and Analysis

2.4.1 Scene Parsing with Functionality Grammar

In this example [56], we propose a computational framework to jointly parse a single RGB image and reconstruct a holistic 3D configuration composed by a set of CAD models using a stochastic grammar model. Specifically, we introduce a Holistic Scene Grammar to represent the 3D scene structure, which characterizes a joint distribution over the functional and geometric space of indoor scenes. The proposed Holistic Scene Grammar captures three essential and often latent dimensions of the indoor scenes: i) latent human context, describing the affordance and the functionality of a room arrangement, ii) geometric constraints over the scene configurations, and iii) physical constraints that guarantee physically plausible parsing and reconstruction. We solve this joint parsing and reconstruction problem in an analysis-by-synthesis fashion, seeking to minimize the differences between the input image and the rendered images generated by our 3D representation, over the space of depth, surface normal, and object segmentation map. The optimal configuration, represented by a parse graph, is inferred using Markov chain Monte Carlo, which efficiently traverses through the non-differentiable solution space, jointly optimizing object localization, 3D layout, and hidden human context.

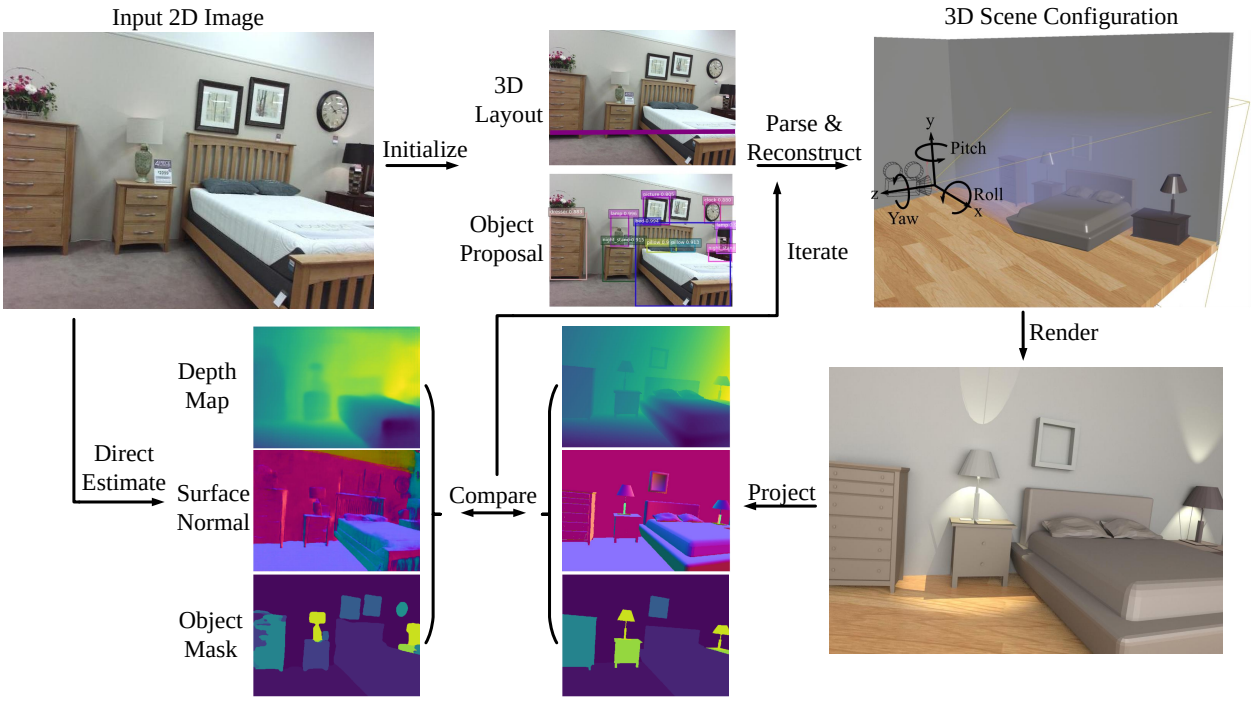


Figure 2.15: Illustration of the proposed holistic 3D indoor scene parsing and reconstruction in an analysis-by synthesis fashion. A 3D representation is initialized by individual vision modules (*e.g.*, object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered normal, depth, and segmentation map with the ones estimated directly from the input RGB image, and adjust the 3D structure iteratively.

Analysis-by-Synthesis

We embrace the concept of vision as inverse graphics, and propose a holistic 3D indoor scene parsing and reconstruction algorithm that simultaneously reconstructs the functional hierarchy and the 3D geometric structure of an indoor scene from a single RGB image. Fig. 2.15 schematically illustrates the analysis-by-synthesis inference process. The joint inference algorithm takes proposals from various vision modules and infers the 3D structure by comparing various projections (*i.e.*, depth, normal, and segmentation) rendered from the recovered 3D structure with the ones directly estimated from an input image.

Holistic Scene Grammar

We represent the hierarchical structure of indoor scenes by a Holistic Scene Grammar (HSG). An Holistic Scene Grammar consists of a latent hierarchical structure in the functional space $\mathbb{F}$ and terminal object entities in the geometric space $\mathbb{G}$. The intuition is that for human environments, the object arrangement in the geometric space can be viewed as a projection from the functional space (*i.e.*, human activities). The functional space as a probabilistic context free grammar (PCFG) captures the hierarchy of the functional groups, and the geometric space captures the spatial contexts among objects by defining an Markov random field (MRF) on the terminal nodes. The two spaces together form a stochastic context-sensitive grammar (SCSG). The HSG starts from a root scene node and ends with a set of terminal nodes. An indoor scene is represented by a parse graph **pg** as illustrated in Fig. 2.16.

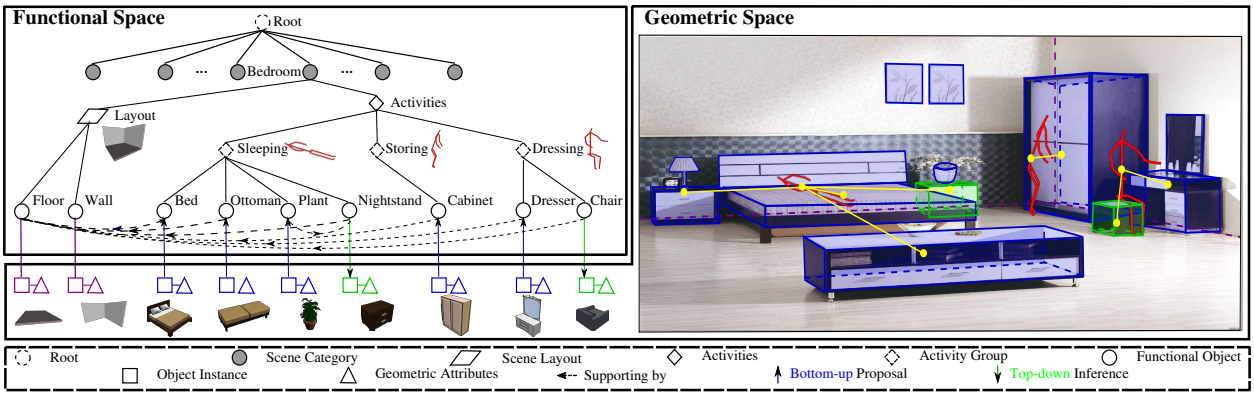


Figure 2.16: An indoor scene represented by a parse graph (***pg***) of the HSG that spans across the functional space and the geometric space. The functional space characterizes the hierarchical structure and the geometric space encodes the spatial entities with contextual relations.

Definition: The stochastic context-sensitive grammar HSG is defined as a 5-tuple $\langle S, V, R, E, P \rangle$. S denotes the root node of the indoor scene. V is the vertex set that includes both non-terminal nodes $V_f \in \mathbb{F}$ and terminal nodes $V_g \in \mathbb{G}$. R denotes the production rule, and E the contextual relations among the terminal nodes, which are represented by the horizontal links in the **pg**. P is the probability model defined on the **pg**.

Functional Space: The non-terminal nodes $V_f = \{V_f^c, V_f^a, V_f^o, V_f^l\} \in \mathbb{F}$ consist of the scene category nodes V_f^c, activity group nodes V_f^a, objects nodes V_f^o, and layout nodes V_f^l.

Geometric Space: The terminal nodes $V_g = \{V_g^o, V_g^l\} \in \mathbb{G}$ are the CAD models of object entities and room layouts. Each object $v \in V_g^o$ is represented as a CAD model, and the object appearance is parameterized by its 3D size, location, and orientation. The room layout $v \in V_g^l$ is represented as a cuboid which is further decomposed into five planar surfaces of the room (left wall, right wall, middle wall, floor, and ceiling with respect to the camera coordinate).

The following production rules R are defined for HSG:

Production Rule	Semantic Meaning	Instances
$r1: \ S \rightarrow V_f^c$	scene → category 1 | category 2 | ...	scene → office| kitchen
$r2: \ V_f^c \rightarrow V_f^a \cdot V_f^l$	category → activity groups · layout	office → (walking, reading) · layout
$r3: \ V_f^a \rightarrow V_f^o$	activity group → functional objects	sitting → (desk, chair)

where $\cdot$ denotes the deterministic decomposition, | alternative explanations, and () combination. Contextual relations E capture relations among objects, including their relative positions, relative orientations, grouping relations, and supporting relations. The objects could be supported by either other objects or the room layout; *e.g.*, a lamp could be supported by a night stand or the floor.

Finally, a scene configuration is represented by a **pg**, whose terminals are room layouts and objects with their attributes and relations. As shown in Fig. 2.16, a **pg** can be decomposed as **pg** $= (pg_f, pg_g)$, where pg_f and pg_g denote the functional part and geometric part of the **pg**, respectively. $E \in pg_g$ denotes the contextual relations in the terminal layer.

Probabilistic Formulation

The objective of the holistic scene parsing is to find an optimal **pg** that represents all the contents and relations observed in the scene. Given an input RGB image I, the optimal **pg** could be derived by an maximum a posteriori probability (MAP) estimator,

$$p(\mathbf{pg}|I) \propto p(\mathbf{pg}) \cdot p(I|\mathbf{pg}) \tag{2.12}$$

$$\propto p(pg_f) \cdot p(pg_g|pg_f) \cdot p(I|pg_g) \tag{2.13}$$

$$= \frac{1}{Z} \exp\left\{-\mathcal{E}(pg_f) - \mathcal{E}(pg_g|pg_f) - \mathcal{E}(I|pg_g)\right\}, \tag{2.14}$$

where the prior probability $p(\mathbf{pg})$ is decomposed into $p(pg_f)p(pg_g|pg_f)$, and $p(I|\mathbf{pg}) = p(I|pg_g)$ since the image space is independent of the functional space given the geometric space. We model the joint distribution with a Gibbs distribution; $\mathcal{E}(pg_f)$, $\mathcal{E}(pg_g|pg_f)$ and $\mathcal{E}(I|pg_g)$ are the corresponding energy terms.

Functional Prior $\mathcal{E}(pg_f)$ characterizes the prior of the functional aspect in a **pg**, which models the hierarchical structure and production rules in the functional space. For production rules of alternative explanations | and combination (), each rule selects child nodes and the probability of the selections is modeled with a multinomial distribution. The production rule $\cdot$ is deterministically expanded with probability 1. Given a set of production rules R, the energy could be written as:

$$\mathcal{E}(pg_f) = \sum\nolimits_{r_i \in R} -\log p(r_i). \tag{2.15}$$

Note we do not model the production rule $\cdot$ since it is deterministically expanded.

Geometric Prior $\mathcal{E}(pg_g|pg_f)$ characterizes the prior of the geometric aspect in a **pg**. Besides modeling the size, position and orientation distribution of each object, we also consider two types of contextual relations $E = \{E_s, E_a\}$ among the objects: i) relations E_s between supported objects and their supporting objects (*e.g.*, monitor and desk); ii) relations E_a between imagined human and objects in an activity group (*e.g.*, relation between imagined human and the chair in an activity group of reading).

We define different potential functions for each type of contextual relations, constructing an MRF in the geometric space including four terms:

$$\mathcal{E}(pg_g|pg_f) = \mathcal{E}_{sc}(pg_g|pg_f) + \mathcal{E}_{spt}(pg_g|pg_f) + \mathcal{E}_{grp}(pg_g|pg_f) + \mathcal{E}_{phy}(pg_g). \tag{2.16}$$

Specifically, • *Size Consistency* $\mathcal{E}_{sc}$ constrains the size of an object. We model the distribution of object scale using a non-parametric way, *i.e.*, kernel density estimation (KDE),

$$\mathcal{E}_{sc}(pg_g|pg_f) = \sum\nolimits_{v_i \in V_g^o} -\log p\left(s_i|V_f^o\right), \tag{2.17}$$

where s_i denotes the size of object v_i. Empirically, we find that KDE fits better than a parametric estimation (*e.g.*, multivariate normal), and it is easier to sample from.

• *Supporting Constraint* $\mathcal{E}_{spt}$ characterizes the contextual relations between supported objects and supporting objects (including floors, walls and ceilings). We model the distribution with their relative heights and overlapping areas:

$$\mathcal{E}_{spt}(pg_g|pg_f) = \sum\nolimits_{(v_i,v_j)\in E_s} \mathcal{K}_o(v_i,v_j) + \mathcal{K}_h(v_i,v_j) - \lambda_s \log p\left(v_i, v_j \mid V_f^l, V_f^o\right), \tag{2.18}$$

Figure 2.17: Illustration of imagined human in scene parsing. We learn the distribution of the human-object relation and utilize it to sample human poses.

where $\mathcal{K}_o(v_i, v_j) = 1 - area(v_i \cup v_j)/area(v_i)$ defines the overlapping ratio in xy-plane, and $\mathcal{K}_h(v_i, v_j)$ defines the relative height between the lower surface of v_i and the upper surface of v_j. $\mathcal{K}_o(\cdot)$ and $\mathcal{K}_h(\cdot)$ is 0 if supporting object is floor and wall, respectively. $p(v_i, v_j|V_f^l, V_f^o)$ is the prior frequency of the supporting relation modeled by multinoulli distributions. λ_s is a balancing constant.

• *Human-Centric Grouping Constraint* $\mathcal{E}_{grp}$. For each activity group, we imagine the invisible and latent human poses to help parse and understand the scene. The intuition is that the indoor scenes are designed to serve human daily activities, thus the indoor images should be jointly interpreted by the observed entities and the unobservable human activities. This is known as the *Dark Matter* [157] in computer vision that drives the visible components in the scene. Prior methods on scene parsing often merely model the object-object relations. In this work, we go beyond passive observations to model the latent human-object relations, thereby proposing a human-centric grouping relationship and a joint inference algorithm over the visible scene and invisible latent human context. Specifically, for each activity group $v \in V_f^a$, we define correspondent imagined human with a six tuple $< y, \mu, t, r, s, \tilde{\mu} >$, where y is the activity type, $\mu \in \mathbb{R}^{25\times 3}$ is the mean pose of activity type y, t denotes the translation, r denotes the rotation, s denotes the scale, and $\tilde{\mu}$ is the imagined position to place a person: $\tilde{\mu} = \mu \cdot r \cdot s + t$. The energy among the imagined human and objects is defined as:

$$\begin{aligned}\mathcal{E}_{grp}(pg_g|pg_f) &= \sum\nolimits_{v_i \in V_f^a} \mathcal{E}_{grp}(\tilde{\mu}_i|v_i) \\ &= \sum\nolimits_{v_i \in V_f^a} \sum\nolimits_{v_j \in ch(v_i)} \mathcal{D}_d(\tilde{\mu}_i, \nu_j; \bar{d}) + \mathcal{D}_h(\tilde{\mu}_i, \nu_j; \bar{h}) + \mathcal{D}_o(\tilde{\mu}_i, \nu_j; \bar{o}),\end{aligned} \tag{2.19}$$

where $ch(v_i)$ denotes the set of child nodes of v_i, ν_j denotes the 3D position of v_j. $\mathcal{D}_d(\cdot)$, $\mathcal{D}_h(\cdot)$ and $\mathcal{D}_o(\cdot)$ denote geometric distances, heights and orientation differences, respectively, calculated by the center of the imagined human pose to the object center subtracted by their mean (*i.e.*, $\bar{d}$, $\bar{h}$ and $\bar{o}$). Fig. 2.17 shows some examples of the imagined human.

This reflects the human-object interaction in 3D space and it could be formulated as the relative geometric relation between the hallucinated human and the objects as following:

$$\phi(e|F_a, F_o) \propto \exp -\{l_d(h, g_o) + l_o(h, g_o) + l_h(h, g_o)\} \tag{2.20}$$

where e connects the hallucinated human node g and object node g_o. The geometric distance between the center of the object and hallucinated human is defined as $l_d = \|d(x_i, x_j) - \bar{d}(x_i, x_j)\|^2$, and $d(x_i, x_j)$ is the mean distance learned from the data. Similarly, $l_o = \|\theta(x_i, x_j) - \bar{\theta}(x_i, x_j)\|^2$ defines the angle and $l_h = \|h(x_i, x_j) - \bar{h}(x_i, x_j)\|^2$ defines the height difference.

• *Physical Constraints:* Additionally, in order to avoid violating physical laws during parsing, we define the physical constraints $\mathcal{E}_{phy}(pg_g)$ to penalize physical violations. Exceeding the room cuboid or overlapping among the objects are defined as violations. This term is formulated as:

$$\mathcal{E}_{phy}(pg_g) = \sum\nolimits_{v_i \in V_g^o} (\sum\nolimits_{v_j \in V_g^o \setminus v_i} \mathcal{O}_o(v_i, v_j) + \sum\nolimits_{v_j \in V_g^l} \mathcal{O}_l(v_i, v_j)), \tag{2.21}$$

where $\mathcal{O}_o(\cdot)$ denotes the overlapping area between objects, and $\mathcal{O}_l(\cdot)$ denotes the area of objects exceeding the layout.

Likelihood $\mathcal{E}(I|pg_g)$ characterizes the similarity between the observed image and the rendered image generated by the parsing results. Since there is still a difference between the two images due to various lighting conditions, textures, and material properties, we solve the problem in an *analysis-by-synthesis* fashion. By combining generative models and discriminative models, this approach tries to reverse-engineer the hidden factors that generate the observed image.

Specifically, we first use discriminative methods to project the original image I to various feature spaces. In this work, we directly estimate three intermediate images including the depth map $\Phi_d(I)$, surface normal map $\Phi_n(I)$ and object segmentation map $\Phi_m(I)$, as the feature representation of the observed image I.

Meanwhile, a **pg** inferred by our method represents the 3D structure of the observed image. Thus, we can use the inferred **pg** to reconstruct image I', and recover the corresponding depth map $\Phi_d(I')$, surface normal map $\Phi_n(I')$, and object segmentation map $\Phi_m(I')$ through a forward graphics rendering.

Finally, we compute the likelihood term by comparing these rendered results from the generative model with the directly estimated results calculated by the discriminative models. Specifically, the likelihood is computed by pixel-wise differences between the two sets of maps,

$$\mathcal{E}(I|pg_g) = \mathcal{D}_p(\Phi_d(I), \Phi_d(I')) + \mathcal{D}_p(\Phi_n(I), \Phi_n(I')) + \mathcal{D}_p(\Phi_m(I), \Phi_m(I')), \tag{2.22}$$

where function $\mathcal{D}_p(\cdot)$ indicates the summation of pixel-wise Euclidean distances between the two maps.

Inference

Given a single RGB image as the input, the goal in the inference phrase is to find the optimal **pg** that best explains the hidden factors that generate the observed image while recovering the 3D scene structure.

The inference process includes three major steps.

- *Room geometry estimation:* estimate the room geometry by predicting the 2D room layout and the camera parameter, and projecting the estimated 2D layout to 3D. Details are provided in [56].
- *Objects initialization:* detect objects and retrieve CAD models correspondingly with the most similar appearance, then roughly estimate their 3D poses, positions, sizes, and initialize the support relations. See [56] for details.
- *Joint inference:* optimize the objects, layout and hidden human context in the 3D scene in an analysis-by-synthesis fashion by maximizing the posterior probability of the **pg**. Details are provided in next section.

Joint Inference

Given an image I, we first estimate the room geometry, object attributes and relations as described in the above two subsections. The goal of joint inference is to (1) optimize the objects and layout; (2) group objects, assign activity label and imagine human pose in each activity group; and (3) optimize the objects, layout and human pose iteratively.

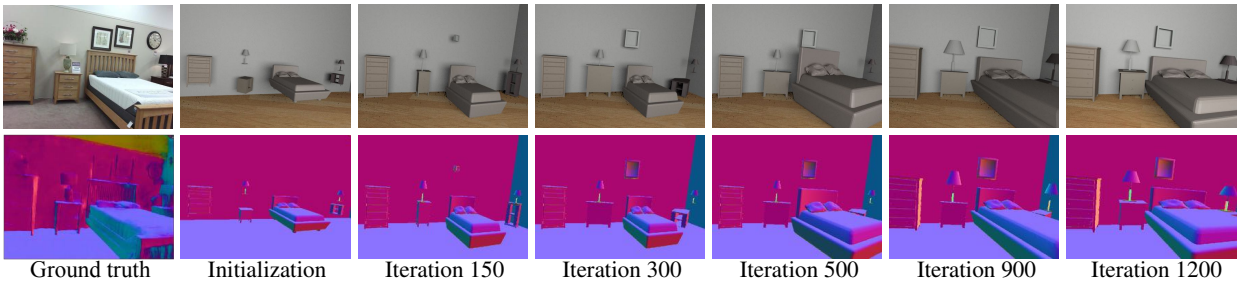


Figure 2.18: The process of joint inference of objects and layout by MCMC with simulated annealing. The first row contains rendered RGB images (for visualization), the second row contains rendered surface normal maps. During the joint inference, objects and layout are optimized iteratively.

In each step, we use distinct MCMC processes. Specifically, to travel through non-differentiable solution spaces, we design Markov Chain dynamics $\{q_1^o, q_2^o, q_3^o\}$ for objects, $\{q_1^l, q_2^l\}$ for layout, and $\{q_1^h, q_2^h, q_3^h\}$ for human pose. Specifically,

• *Object Dynamics:* Dynamics q_1^o adjusts the position of a random object, which translates the object center in one of the three coordinate directions. Instead of translating the object center and changing the object size directly, Dynamics q_2^o translates one of the six faces of the cuboid to generate a smoother diffusion. Dynamics q_3^o proposes rotation of the object with a specified angle. Each dynamic can diffuse in two directions, *e.g.*, each object can translate in direction of '$+x$' and '$-x$', or rotate in direction of clockwise and counterclockwise. By computing the local gradient of $P(\mathbf{pg}|I)$, the dynamics propose to move following the direction of the gradient with a proposal probability of 0.8, or the inverse direction of the gradient with proposal probability of 0.2.

• *Layout Dynamics:* Dynamics q_1^l translates the faces of the layout, which also optimizes the predefined camera height while translating the floor. Dynamics q_2^l proposes to rotate the layout.

• *Human pose Dynamics* q_1^h, q_2^h and q_3^h are designed to translate, rotate and scale the human pose, respectively.

Given a current $\mathbf{pg}$, each dynamic will propose a new $\mathbf{pg}'$ according to a proposal probability $p(\mathbf{pg}'|\mathbf{pg}, I)$. The proposal is accepted according to an acceptance probability $\alpha(\mathbf{pg} \rightarrow \mathbf{pg}')$ defined by the Metropolis-Hasting algorithm [158]:

$$\alpha(\mathbf{pg} \rightarrow \mathbf{pg}') = \min(1, \frac{p(\mathbf{pg}|\mathbf{pg}', I)p(\mathbf{pg}'|I)}{p(\mathbf{pg}'|\mathbf{pg}, I)p(\mathbf{pg}|I)}). \tag{2.23}$$

Fig. 2.18 shows the process of step (1).

In step (2), we design an algorithm to group objects and assign activity labels. For each type of activity, there is a major object category which has the highest occurrence frequency (*i.e.*, chair in activity 'reading'). Intuitively, the correspondence between objects and activities should be n-to-n but not n-to-one, which means each object can belong to several activity groups. In order to find out all possible activity groups, for each type of activity, we define an activity group around each major object and incorporate nearby objects (within a distance threshold) with prior larger than 0. For each activity group $v_i \in V_f^a$, the pose of the imagined human is estimated by maximizing the likelihood $p(v_i|\tilde{\mu}_i)$, which is equivalent to minimize the grouping energy $\mathcal{E}_{grp}(\tilde{\mu}_i|v_i)$ defined in Eq. (2.19),

$$y_i^*, m_i^*, t_i^*, r_i^*, s_i^* = \underset{y_i, m_i, t_i, r_i, s_i}{\arg\min} \mathcal{E}_{grp}(\tilde{\mu}_i|v_i), \tag{2.24}$$

Fig. 2.19 shows the results of sampled human poses in various indoor scenes. Fig. 2.20 shows more qualitative parsing results (top 20%).

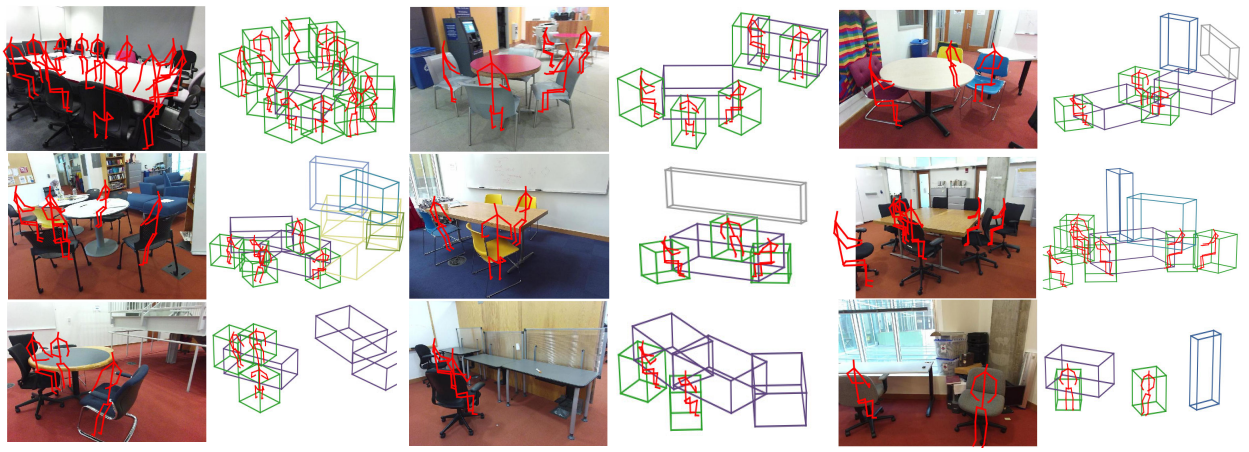


Figure 2.19: Sampled human poses in various indoor scenes. Objects in multiple activity groups have multiple poses. We visualize the pose with the highest likelihood.

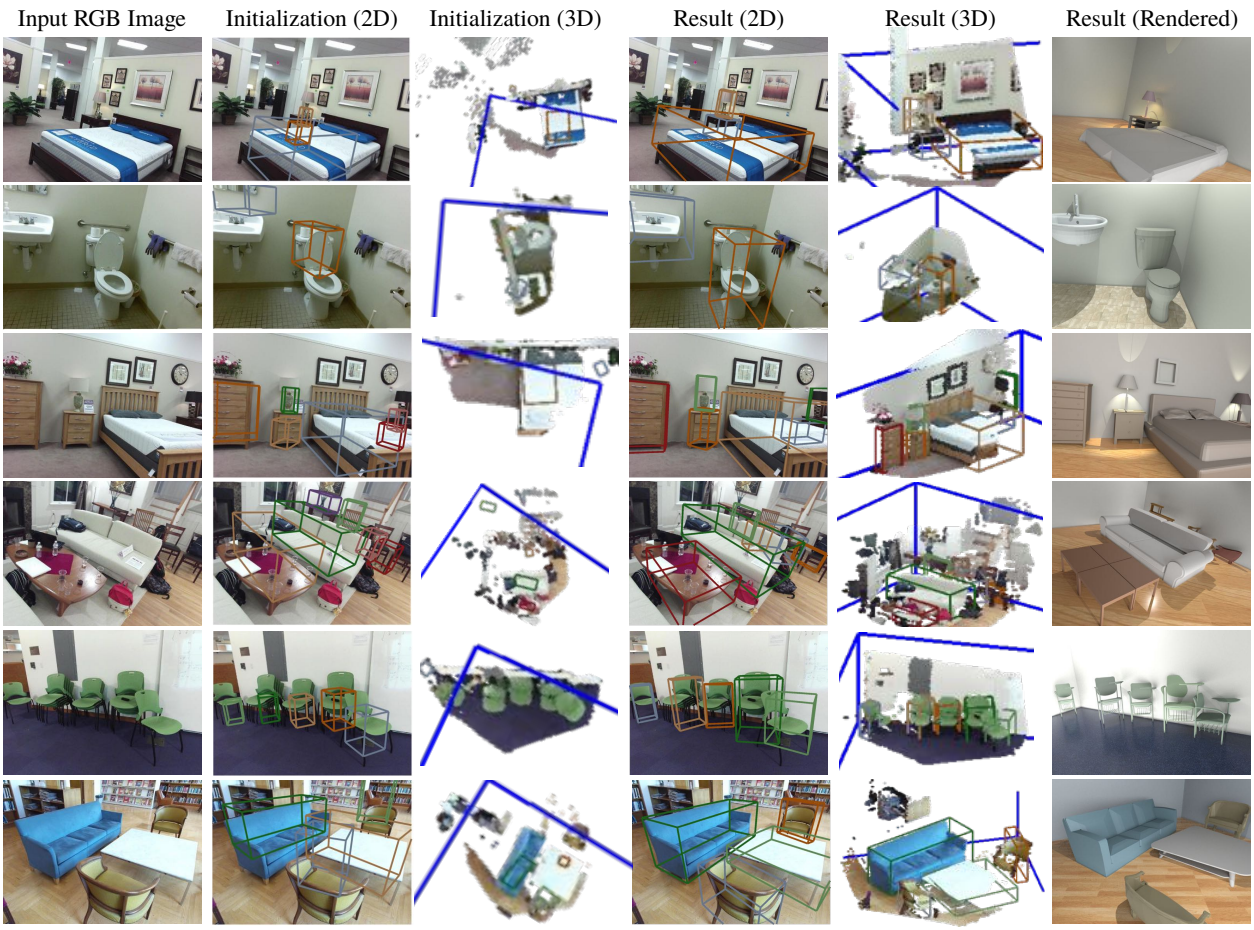


Figure 2.20: Qualitative results of our method on SUN RGB-D dataset. The joint inference significantly improves the performance over individual modules.

2.4.2 Scene Synthesis with Functionality Grammar

In this example [119], we present a human-centric method to sample and synthesize 3D room layouts and 2D images thereof, to obtain large-scale 2D/3D image data with the perfect per-pixel ground

truth. An attributed spatial And-Or graph (S-AOG) is proposed to represent indoor scenes. The S-AOG is a probabilistic grammar model, in which the terminal nodes are object entities including room, furniture, and supported objects. Human contexts as contextual relations are encoded by Markov Random Fields (MRF) on the terminal nodes. We learn the distributions from an indoor scene dataset and sample new layouts using Monte Carlo Markov Chain.

Synthesizing indoor scenes is a non-trivial task. It is often difficult to properly model either the relations between furniture of a functional group, or the relations between the supported objects and the supporting furniture. Specifically, we argue there are four major difficulties. (i) In a functional group such as a dining set, the number of pieces may vary. (ii) Even if we only consider pair-wise relations, there is already a quadratic number of object-object relations. (iii) What makes it worse is that most object-object relations are not obviously meaningful. For example, it is unnecessary to model the relation between a pen and a monitor, even though they are both placed on a desk. (iv) Due to the previous difficulties, an excessive number of constraints are generated. Many of the constraints contain loops, making the final layout hard to sample and optimize.

To address these challenges, we propose a human-centric approach to model indoor scene layout. It integrates human activities and functional grouping/supporting relations. This method not only captures the human context but also simplifies the scene structure. Specifically, we use a probabilistic grammar model for images and scenes [154] – an attributed spatial And-Or graph (S-AOG), including vertical hierarchy and horizontal contextual relations. The contextual relations encode functional grouping relations and supporting relations modeled by object affordances [141]. For each object, we learn the affordance distribution, *i.e.*, an object-human relation, so that a human can be sampled based on that object. Besides static object affordance, we also consider dynamic human activities in a scene, constraining the layout by planning trajectories from one piece of furniture to another.

Representation

We use an attributed S-AOG [154] to represent an indoor scene. An attributed S-AOG is a probabilistic grammar model with attributes on the terminal nodes. It combines i) a probabilistic context free grammar (PCFG), and ii) contextual relations defined on an Markov Random Field (MRF), *i.e.*, the horizontal links among the nodes. The PCFG represents the hierarchical decomposition from scenes (top level) to objects (bottom level) by a set of terminal and non-terminal nodes, whereas contextual relations encode the spatial and functional relations through horizontal links. The structure of S-AOG is shown in Fig. 2.16.

Formally, the S-AOG is defined as a 5-tuple: $\mathcal{G} = \langle S, V, R, P, E \rangle$, where we use notations S the root node of the scene grammar, V the vertex set, R the production rules, P the probability model defined on the attributed S-AOG, and E the contextual relations represented as horizontal links between nodes in the same layer. **Vertex Set** V can be decomposed into a finite set of non-terminal and terminal nodes: $V = V_{NT} \cup V_T$.

• $V_{NT} = V^{And} \cup V^{Or} \cup V^{Set}$. The non-terminal nodes consists of three subsets. i) A set of **And-nodes** V^{And}, in which each node represents a decomposition of a larger entity (*e.g.*, a bedroom) into smaller components (*e.g.*, walls, furniture and supported objects). ii) A set of **Or-nodes** V^{Or}, in which each node branches to alternative decompositions (*e.g.*, an indoor scene can be a bedroom or a living room), enabling the algorithm to reconfigure a scene. iii) A set of **Set nodes** V^{Set}, in which each node represents a nested And-Or relation: a set of Or-nodes serving as child branches are grouped by an And-node, and each child branch may include different numbers of objects.

• $V_T = V_T^r \cup V_T^a$. The terminal nodes consists of two subsets of nodes: regular nodes and address nodes. i) A **regular terminal node** $v \in V_T^r$ represents a spatial entity in a scene (*e.g.*, an office

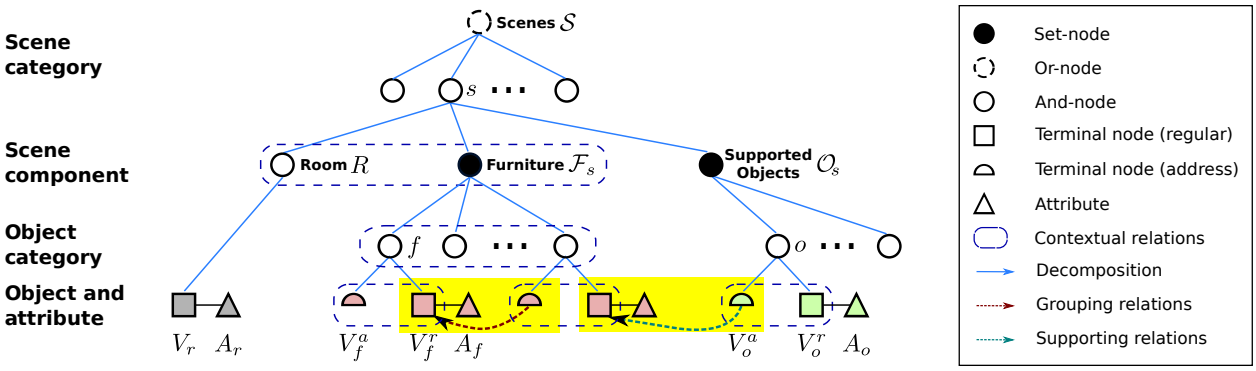


Figure 2.21: Scene grammar as an attributed S-AOG. A scene of different types is decomposed into a room, furniture, and supported objects. Attributes of terminal nodes are internal attributes (sizes), external attributes (positions and orientations), and a human position that interacts with this entity. Furniture and object nodes are combined by an address terminal node and a regular terminal node. A furniture node (*e.g.*, a chair) is grouped with another furniture node (*e.g.*, a desk) pointed by its address terminal node. An object (*e.g.*, a monitor) is supported by the furniture (*e.g.*, a desk) it is pointing to. If the value of the address node is null, the furniture is not grouped with any furniture, or the object is put on the floor. Contextual relations are defined between the room and furniture, between a supported object and supporting furniture, among different pieces of furniture, and among functional groups.

chair in a bedroom) with attributes. In this work, the attributes include internal attributes A_{int} of object sizes (w, l, h), external attributes A_{ext} of object position (x, y, z) and orientation $(x - y$ plane) θ, and sampled human positions A_h. ii) To avoid excessively dense graphs, an **address terminal node** $v \in V_T^a$ is introduced to encode interactions that only occur in a certain context but are absent in all others [159]. It is a pointer to regular terminal nodes, taking values in the set $V_T^r \cup \{\text{nil}\}$, representing supporting or grouping relations as shown in Fig. 2.21.

Contextual Relations E among nodes are represented by the horizontal links in S-AOG forming MRFs on the terminal nodes. To encode the contextual relations, we define different types of potential functions for different cliques. The contextual relations $E = E_f \cup E_o \cup E_g \cup E_r$ are divided into four subsets: i) relations among furniture E_f; ii) relations between supported objects and their supporting objects E_o (*e.g.*, a monitor on a desk); iii) relations between objects of a functional pair E_g (*e.g.*, a chair and a desk); and iv) relations between furniture and the room E_r. Accordingly, the cliques formed in the terminal layer could also be divided into four subsets: $C = C_f \cup C_o \cup C_g \cup C_r$. Instead of directly capturing the object-object relations, we compute the potentials using affordances as a bridge to characterize the object-human-object relations.

A hierarchical parse tree pt is an instantiation of the S-AOG by selecting a child node for the Or-nodes as well as determining the state of each child node for the Set-nodes. A parse graph pg consists of a parse tree pt and a number of contextual relations E on the parse tree: $pg = (pt, E_{pt})$. Fig. 2.22 illustrates a simple example of a parse graph and four types of cliques formed in the terminal layer.

Probabilistic Formulation

A scene configuration is represented by a parse graph pg, including objects in the scene and associated attributes. The prior probability of pg generated by an S-AOG parameterized by Θ is

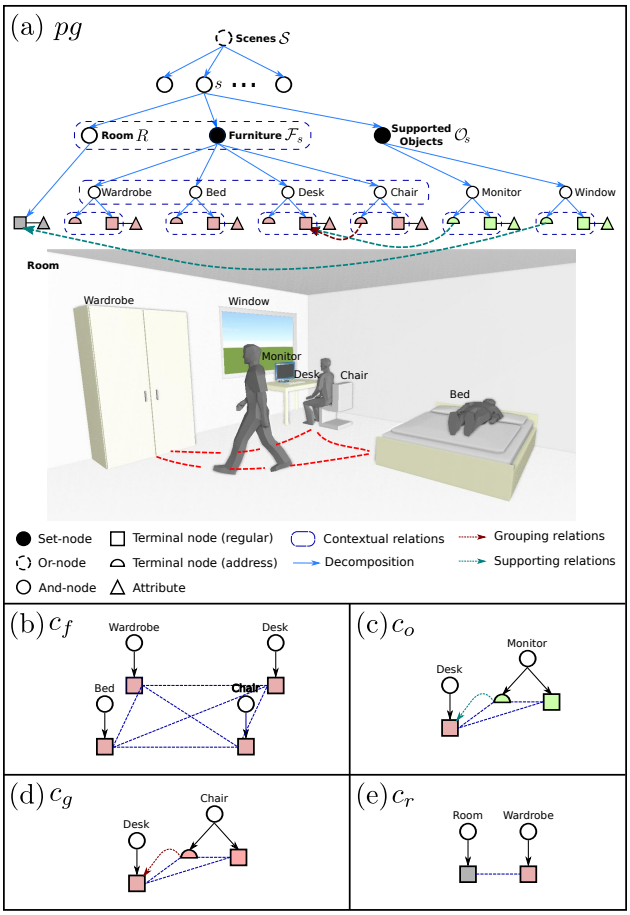


Figure 2.22: (a) A simplified example of a parse graph of a bedroom. The terminal nodes of the parse graph form an MRF in the terminal layer. Cliques are formed by the contextual relations projected to the terminal layer. Examples of the four types of cliques are shown in (b)-(e), representing four different types of contextual relations.

formulated as a Gibbs distribution:

$$p(pg|\Theta) = \frac{1}{Z}\exp\{-\mathcal{E}(pg|\Theta)\} \tag{2.25}$$

$$= \frac{1}{Z}\exp\{-\mathcal{E}(pt|\Theta) - \mathcal{E}(e_{pt}|\Theta)\}, \tag{2.26}$$

where $\mathcal{E}(pg|\Theta)$ is the energy function of a parse graph, $\mathcal{E}(pt|\Theta)$ is the energy function of a parse tree, and $\mathcal{E}(e_{pt}|\Theta)$ is the energy term of the contextual relations. Here, the energy function of a parse tree is defined as combinations of probability distributions with closed-form expressions and non-parametric distributions, and the energy of the contextual relations E is defined as potential functions regarding to the attributes of the terminal nodes.

$\mathcal{E}(pt|\Theta)$ can be further decomposed into the energy functions of different types of non-terminal nodes, and the energy functions of internal attributes of both regular and address terminal nodes:

$$\mathcal{E}(pt|\Theta) = \underbrace{\sum_{v\in V^{Or}} \mathcal{E}_{\Theta}^{Or}(v) + \sum_{v\in V^{Set}} \mathcal{E}_{\Theta}^{Set}(v)}_{\text{non-terminal nodes}} + \underbrace{\sum_{v\in V_T^r} \mathcal{E}_{\Theta}^{A_{in}}(v)}_{\text{terminal nodes}}, \tag{2.27}$$

where the choice of the child node of an Or-node $v \in V^{Or}$ and the child branch of a Set-node $v \in V^{Set}$ follow different multinomial distributions. Since the And-nodes are deterministically expanded, we do not have an energy term for the And-nodes here. The internal attributes A_{in} (size) of terminal nodes follows a non-parametric probability distribution learned by kernel density estimation.

$\mathcal{E}(e_{pt}|\Theta)$ combines the potentials of the four types of cliques formed in the terminal layer, integrating human attributes and external attributes of regular terminal nodes:

$$p(E_{pt}|\Theta) = \frac{1}{Z}\exp\{-\mathcal{E}(e_{pt}|\Theta)\} \tag{2.28}$$

$$= \prod_{c\in C_f}\phi_f(c)\prod_{c\in C_o}\phi_o(c)\prod_{c\in C_g}\phi_g(c)\prod_{c\in C_r}\phi_r(c). \tag{2.29}$$

Human-Centric Potential Functions:

- Potential function $\phi_f(c)$ is defined on relations between furniture (Fig. 2.22 (b)). The clique $c = \{f_i\} \in C_f$ includes all the terminal nodes representing furniture:

$$\phi_f(c) = \frac{1}{Z}\exp\{-\lambda_f \cdot \langle \sum_{f_i \neq f_j} l_{\text{col}}(f_i, f_j), l_{\text{ent}}(c)\rangle\}, \tag{2.30}$$

  where λ_f is a weight vector, $< \cdot, \cdot >$ denotes a vector, and the cost function $l_{\text{col}}(f_i, f_j)$ is the overlapping volume of the two pieces of furniture, serving as the penalty of collision. The cost function $l_{\text{ent}}(c) = -H(\Gamma) = \Sigma_i p(\gamma_i)\log p(\gamma_i)$ yields better utility of the room space by sampling human trajectories, where Γ is the set of planned trajectories in the room, and $H(\Gamma)$ is the entropy. The trajectory probability map is first obtained by planning a trajectory γ_i from the center of every piece of furniture to another one using bi-directional rapidly-exploring random tree (RRT) [160], which forms a heatmap. The entropy is computed from the heatmap as shown in Fig. 2.23.
- Potential function $\phi_o(c)$ is defined on relations between a supported object and the supporting furniture (Fig. 2.22 (c)). A clique $c = \{f, a, o\} \in C_o$ includes a supported object terminal node o, the address node a connected to the object, and the furniture terminal node f pointed by a:

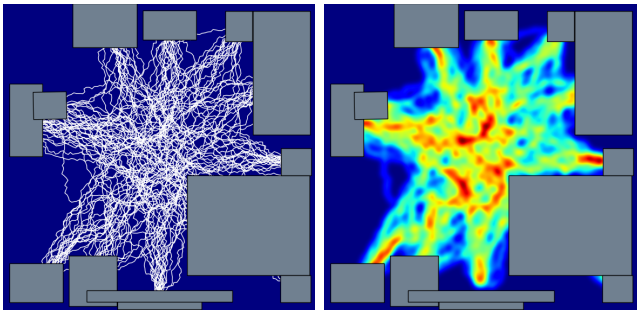
$$\phi_o(c) = \frac{1}{Z}\exp\{-\lambda_o \cdot \langle l_{\text{hum}}(f, o), l_{\text{add}}(a)\rangle\}, \tag{2.31}$$

  where the cost function $l_{\text{hum}}(f, o)$ defines the human usability cost—a favorable human position should enable an agent to access or use both the furniture and the object. To compute the usability cost, human positions h_i^o are first sampled based on position, orientation, and the affordance map of the supported object. Given a piece of furniture, the probability of the human positions is then computed by:

$$l_{\text{hum}}(f, o) = \max_i p(h_i^o|f). \tag{2.32}$$

  The cost function $l_{\text{add}}(a)$ is the negative log probability of an address node $v \in V_T^a$, treated as a certain regular terminal node, following a multinomial distribution.
- Potential function $\phi_g(c)$ is defined on functional grouping relations between furniture (Fig. 2.22 (d)). A clique $c = \{f_i, a, f_j\} \in C_g$ consists of terminal nodes of a core functional furniture f_i, pointed by the address node a of an associated furniture f_j. The grouping relation potential is defined similarly to the supporting relation potential

$$\phi_g(c) = \frac{1}{Z}\exp\{-\lambda_c \cdot \langle l_{\text{hum}}(f_i, f_j), l_{\text{add}}(a)\rangle\}. \tag{2.33}$$



(a) Planned trajectories (b) Probability map

Figure 2.23: Given a scene configuration, we use bi-directional RRT to plan from every piece of furniture to another, generating a human activity probability map.

Other Potential Functions:

- Potential function $\phi_r(c)$ is defined on relations between the room and furniture (Fig. 2.22 (e)). A clique $c = \{f, r\} \in C_r$ includes a terminal node f and r representing a piece of furniture and a room, respectively. The potential is defined as

$$\phi_r(c) = \frac{1}{Z} \exp\{-\lambda_r \cdot \langle l_{\text{dis}}(f, r), l_{\text{ori}}(f, r) \rangle\}, \tag{2.34}$$

  where the distance cost function is defined as $l_{\text{dis}}(f, r) = -\log p(d|\Theta)$, in which $d \sim \ln \mathcal{N}(\mu, \sigma^2)$ is the distance between the furniture and the nearest wall modeled by a log normal distribution. The orientation cost function is defined as $l_{\text{ori}}(f, r) = -\log p(\theta|\Theta)$, where $\theta \sim p(\mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$ is the relative orientation between the model and the nearest wall modeled by a von Mises distribution.

Learning S-AOG

We use the SUNCG dataset [161] as training data. It contains over 45K different scenes with manually created realistic room and furniture layouts. We collect the statistics of room types, room sizes, furniture occurrences, furniture sizes, relative distances, orientations between furniture and walls, furniture affordance, grouping occurrences, and supporting relations. The parameters Θ of the probability model P can be learned in a supervised way by maximum likelihood estimation (MLE).

Weights of Loss Functions: Recall that the probability distribution of cliques formed in the terminal layer is

$$p(E_{pt}|\Theta) = \frac{1}{Z} \exp\{-\mathcal{E}(e_{pt}|\Theta)\} = \frac{1}{Z} \exp\{-\langle \lambda, l(E_{pt}) \rangle\}, \tag{2.35}$$

where λ is the weight vector and $l(E_{pt})$ is the loss vector given by four different types of potential functions.

To learn the weight vector, the standard MLE maximizes the average log-likelihood:

$$\mathcal{L}(e_{pt}|\Theta) = -\frac{1}{N} \sum_{n=1}^{N} \langle \lambda, l(E_{pt_n}) \rangle - \log Z. \tag{2.36}$$

This is usually maximized by following the gradient:

$$\frac{\partial \mathcal{L}(e_{pt}|\Theta)}{\partial \lambda} = -\frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) - \frac{\partial \log Z}{\partial \lambda} \tag{2.37}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) - \frac{\partial \log \sum_{pt} \exp\{-\langle \lambda, l(E_{pt})\rangle\}}{\partial \lambda} \tag{2.38}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) + \sum_{pt} \frac{1}{Z} \exp\{-\langle \lambda, l(E_{pt})\rangle\} l(E_{pt}) \tag{2.39}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) + \frac{1}{\widetilde{N}}\sum_{\tilde{n}=1}^{\widetilde{N}} l(E_{pt_{\tilde{n}}}), \tag{2.40}$$

where $\{E_{pt_{\tilde{n}}}\}_{\tilde{n}=1,\cdots,\widetilde{N}}$ is the set of synthesized examples from the current model.

It is usually computationally infeasible to sample a Markov chain that burns into an *equilibrium distribution* at every iteration of gradient ascent. Hence, instead of waiting for the Markov chain to converge, we adopt the contrastive divergence (CD) learning that follows the gradient of difference of two divergences [162]

$$\mathrm{CD}_{\widetilde{N}} = \mathrm{KL}(p_0||p_\infty) - \mathrm{KL}(p_{\tilde{n}}||p_\infty), \tag{2.41}$$

where $\mathrm{KL}(p_0||p_\infty)$ is the Kullback-Leibler divergence between the data distribution p_0 and the model distribution p_∞, and $p_{\tilde{n}}$ is the distribution obtained by a Markov chain started at the data distribution and run for a small number $\tilde{n}$ of steps. In this work, we set $\tilde{n} = 1$.

Contrastive divergence learning has been applied effectively to addressing various problems; one of the most notable work is in the context of Restricted Boltzmann Machines [163]. Both theoretical and empirical evidences shows its efficiency while keeping bias typically very small [164]. The gradient of the contrastive divergence is given by

$$\frac{\partial \mathrm{CD}_{\widetilde{N}}}{\partial \lambda} = \frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) - \frac{1}{\widetilde{N}}\sum_{\tilde{n}=1}^{\widetilde{N}} l(E_{pt_{\tilde{n}}}) - \frac{\partial p_{\tilde{n}}}{\partial \lambda}\frac{\partial \mathrm{KL}(p_{\tilde{n}}||p_\infty)}{\partial p_{\tilde{n}}}. \tag{2.42}$$

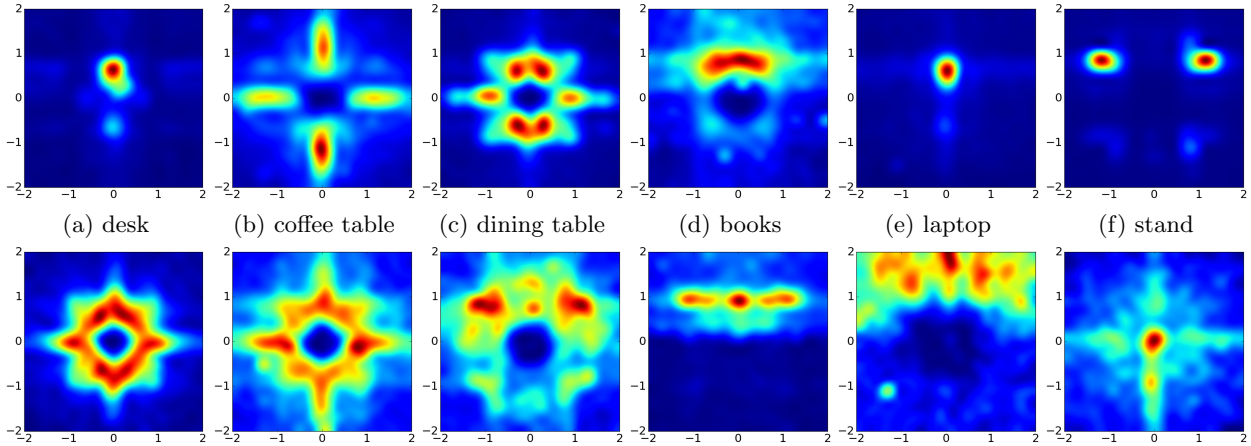
Extensive simulations [162] showed that the third term can be safely ignored since it is small and seldom opposes the resultant of the other two terms.

Finally, the weight vector is learned by gradient descent computed by generating a small number $\widetilde{N}$ of examples from the Markov chain

$$\lambda_{t+1} = \lambda_t - \eta_t \frac{\partial \mathrm{CD}_{\widetilde{N}}}{\partial \lambda} = \lambda_t + \eta_t \left(\frac{1}{\widetilde{N}}\sum_{\tilde{n}=1}^{\widetilde{N}} l(E_{pt_{\tilde{n}}}) - \frac{1}{N}\sum_{n=1}^{N} l(E_{pt_n}) \right). \tag{2.43}$$

Branching Probabilities: The MLE of the branch probabilities ρ_i of Or-nodes, Set-nodes and address terminal nodes is simply the frequency of each alternative choice [154]:

$$\rho_i = \frac{\#(v \to u_i)}{\sum_{j=1}^{n(v)} \#(v \to u_j)} \tag{2.44}$$



(a) desk (b) coffee table (c) dining table (d) books (e) laptop (f) stand

(g) fruit bowl (h) vase (i) floor lamp (j) wall lamp (k) fireplace (l) ceiling fan

Figure 2.24: Examples of the learned affordance maps. Given the object positioned in the center facing upwards, *i.e.*, coordinate of $(0, 0)$ facing direction $(0, 1)$, the maps show the distributions of human positions. The affordance maps accurately capture the subtle differences among desks, coffee tables, and dining tables. Some objects are orientation sensitive, *e.g.*, books, laptops, and night stands, while some are orientation invariant, *e.g.*, fruit bowls and vases.

Grouping Relations: The grouping relations are hand-defined (*i.e.*, nightstands are associated with beds, chairs are associated with desks and tables). The probability of occurrence is learned as a multinomial distribution, and the supporting relations are automatically extracted from SUNCG.

Room Size and Object Sizes: The distribution of the room size and object size among all the furniture and supported objects is learned as a non-parametric distribution. We first extract the size information from the 3D models inside SUNCG dataset, and then fit a non-parametric distribution using kernel density estimation. The distances and relative orientations of the furniture and objects to the nearest wall are computed and fitted into a log normal and a mixture of von Mises distributions, respectively.

Affordances: We learn the affordance maps of all the furniture and supported objects by computing the heatmap of possible human positions. These position include annotated humans, and we assume that the center of chairs, sofas, and beds are positions that humans often visit. By accumulating the relative positions, we get reasonable affordance maps as non-parametric distributions as shown in Fig. 2.24.

Synthesizing Scene Configurations

Synthesizing scene configurations is accomplished by sampling a parse graph pg from the prior probability $p(pg|\Theta)$ defined by the S-AOG. The structure of a parse tree pt (*i.e.*, the selection of Or-nodes and child branches of Set-nodes) and the internal attributes (sizes) of objects can be easily sampled from the closed-form distributions or non-parametric distributions. However, the external attributes (positions and orientations) of objects are constrained by multiple potential functions, hence they are too complicated to be directly sampled from. Here, we utilize a Markov chain Monte Carlo (MCMC) sampler to draw a typical state in the distribution. The process of each sampling can be divided into two major steps:

- Directly sample the structure of pt and internal attributes A_{in}: (i) sample the child node for the Or-nodes; (ii) determine the state of each child branch of the Set-nodes; and (iii) for each regular terminal node, sample the sizes and human positions from learned distributions.

- Use an MCMC scheme to sample the values of address nodes V^a and external attributes A_{ex} by making proposal moves. A sample will be chosen after the Markov chain converges.

We design two simple types of Markov chain dynamics which are used at random with probabilities $q_i, i = 1, 2$ to make proposal moves:

- Dynamics q_1: translation of objects. This dynamic chooses a regular terminal node, and samples a new position based on the current position x: $x \rightarrow x + \delta x$, where δx follows a bivariate normal distribution.
- Dynamics q_2: rotation of objects. This dynamic chooses a regular terminal node, and samples a new orientation based on the current orientation of the object: $\theta \rightarrow \theta + \delta\theta$, where $\delta\theta$ follows a normal distribution.

Adopting the Metropolis-Hastings algorithm, the proposed new parse graph pg' is accepted according to the following acceptance probability:

$$\alpha(pg'|pg, \Theta) = \min(1, \frac{p(pg'|\Theta)p(pg|pg')}{p(pg|\Theta)p(pg'|pg)}) \tag{2.45}$$

$$= \min(1, \exp(\mathcal{E}(pg|\Theta) - \mathcal{E}(pg'|\Theta))), \tag{2.46}$$

where the proposal probability rate is canceled since the proposal moves are symmetric in probability. A simulated annealing scheme is adopted to obtain samples with high probability.

Some qualitative results are shown in Fig. 2.25.

2.4.3 Joint Inference of Scene and Human

In this section we introduce how to jointly tackle two tasks: (i) holistic scene parsing and reconstruction—3D estimations of object bounding boxes, camera pose, and room layout, and (ii) 3D human pose estimation, which is a challenging 3D scene understanding problem from a single RGB image. The intuition behind is to leverage the coupled nature of the two tasks by exploiting two critical and essential connections between these two tasks: (i) HOI to model the fine-grained relations between human agents and objects in the scene, and (ii) physical commonsense to model the physical plausibility. The optimal configuration of the 3D scene, represented by a parse graph, is inferred using MCMC, which efficiently traverses through the non-differentiable joint solution space.

Representation

We represent the configuration of an indoor scene by a parse graph $pg = (pt, E)$ as shown in Fig. 2.26. It combines a parse tree pt and contextual relations E among the leaf nodes. Here $pt = (V, R)$ and we denote $V = V_r \cup V_m \cup V_t$ the vertex set and R the decomposing rules. The tree has three levels. The first level is the root node V_r that represents the scene, and the second level V_m has three nodes (objects, human, and room layout). The third level (terminal nodes V_t) contains child nodes of the second level nodes, representing the detected instances of the parent node in this scene. $E \subset V_t \times V_t$ is the set of contextual relations among the terminal nodes, represented by horizontal links.

Terminal Nodes $\mathbf{V_t}$ in pg can be further decomposed as $V_t = V_{\text{layout}} \cup V_{\text{object}} \cup V_{\text{human}}$:

- The room layout $v \in V_{\text{layout}}$ is represented by a 3D bounding box $X^L \in \mathbb{R}^{3\times 8}$ in the world coordinate. The 3D bounding box is parametrized by the node's attributes, including its 3D size $S^L \in \mathbb{R}^3$, center $C^L \in \mathbb{R}^3$, and orientation $Rot(\theta^L) \in \mathbb{R}^{3\times 3}$. See the supplementary for the parametrization of the 3D bounding box.
- Each 3D object $v \in V_{\text{object}}$ is represented by a 3D bounding box with its semantic label. We keep the same parameterization of the 3D bounding box as the one for room layout.

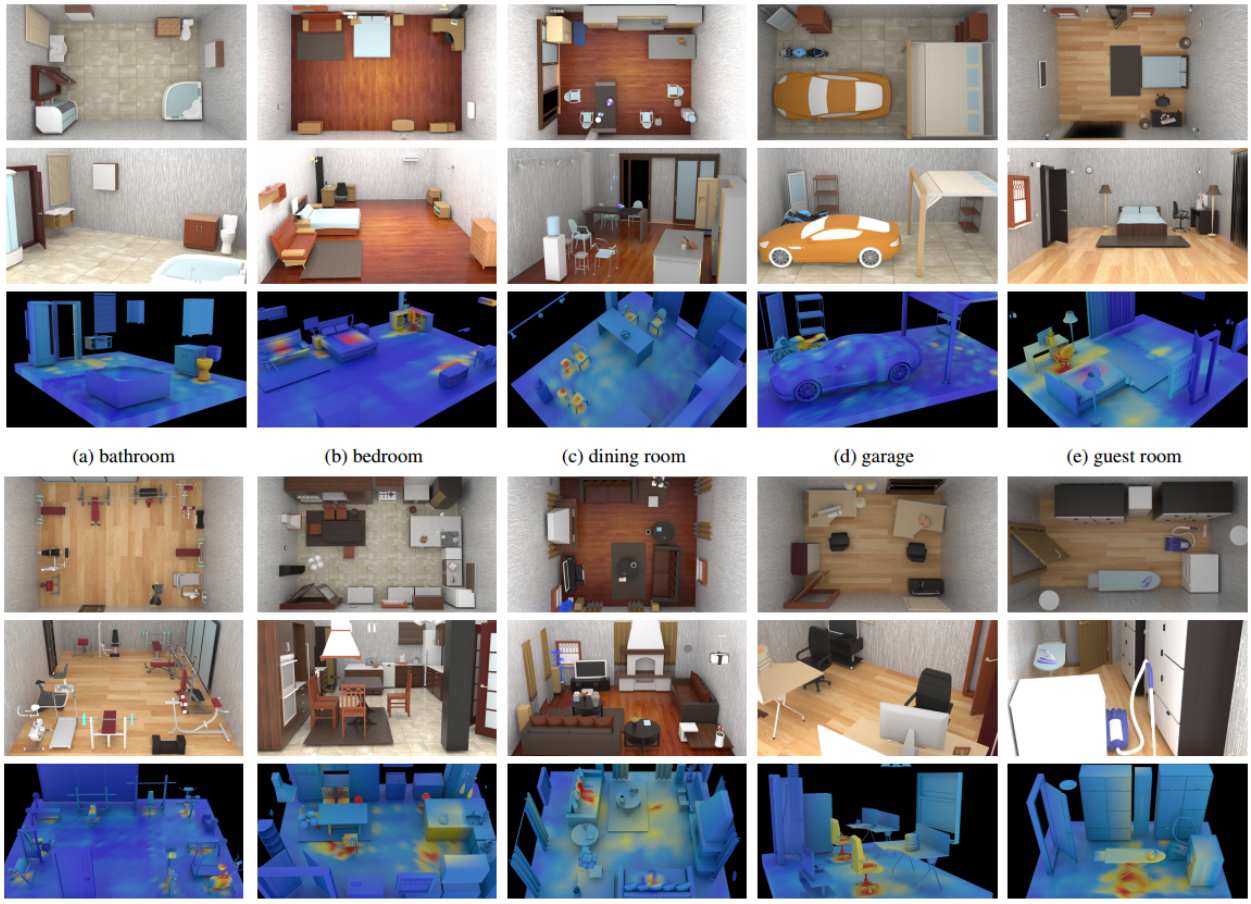


Figure 2.25: Examples of scenes in ten different categories. Top: top-view. Middle: a side-view. Bottom: affordance heatmap.

- Each human $v \in V_{\text{human}}$ is represented by 17 3D joints $X^H \in \mathbb{R}^{3\times 17}$ with their action labels. These 3D joints are parametrized by the pose scale $S^H \in \mathbb{R}$, pose center (*i.e.*, hip) $C^H \in \mathbb{R}^3$, local joint position $Rel^H \in \mathbb{R}^{3\times 17}$, and pose orientation $Rot(\theta^H) \in \mathbb{R}^{3\times 3}$. Each person is also attributed by a concurrent action label a, which is a multi-hot vector representing the current actions of this person: one can "sit" and "drink," or "walk" and "make phone call" at the same time.

Contextual Relations E contains three types of relations in the scene $E = \{E_s, E_c, E_{hoi}\}$. Specifically:

- E_s and E_c denote support relation and physical collision, respectively. These two relations penalize the physical violations among objects, between objects and layout, and between human and layout, resulting in a physically plausible and stable prediction.
- E_{hoi} models HOI and gives us more constraints to reconstruct 3D from 2D. For instance, if a person is detected as sitting on the chair, we can constrain the relative 3D positions between this person and chair using a pre-learned spatial relation of "sitting."

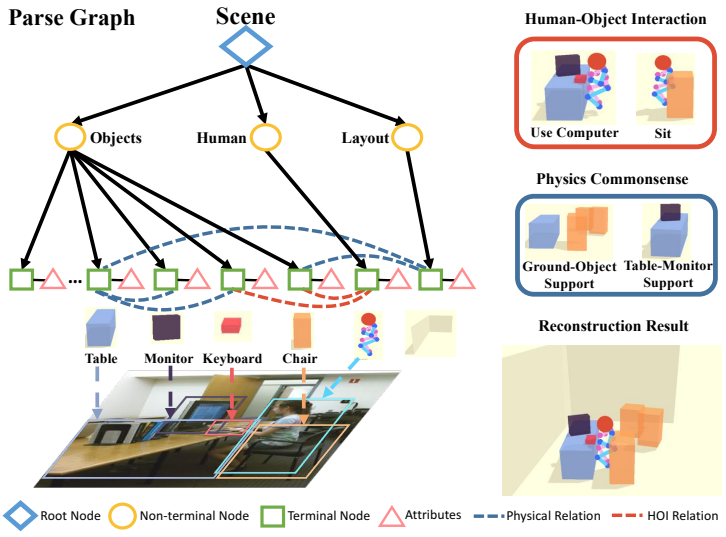


Figure 2.26: Jointly recovering a parse graph that represents the scene, including human poses, objects, camera pose, and room layout, all in 3D. Reasoning HOI helps reconstruct the detailed spatial relations between humans and objects. Physical commonsense (*e.g.*, physical property, plausibility, and stability) further refines relations and improves predictions.

Probabilistic Formulation

The parse graph pg is a comprehensive interpretation of the observed image I. The goal of the holistic^{++} scene understanding is to infer the optimal parse graph pg^* given I by a MAP estimation:

$$\begin{aligned} pg^* &= \arg\max_{pg} p(pg|I) = \arg\max_{pg} p(pg) \cdot p(I|pg) \\ &= \arg\max_{pg} \frac{1}{Z} \exp\{-\mathcal{E}_{phy}(pg) - \mathcal{E}_{hoi}(pg) - \mathcal{E}(I|pg)\}, \end{aligned} \tag{2.47}$$

We model the joint distribution by a Gibbs distribution, where the prior probability of parse graph can be decomposed into physical prior and HOI prior.

Physical Prior $\mathcal{E}_{phy}(pg)$ represents physical commonsense in a 3D scene. We consider two types of physical relations among the terminal nodes: support relation E_s and collision relation E_c. Therefore, the energy of physical prior is defined as $\mathcal{E}_{phy}(pg) = \lambda_s \mathcal{E}_s(pg) + \lambda_c \mathcal{E}_c(pg)$, where λ_s and λ_c are balancing factors. Specifically:

- *Support Relation* $\mathcal{E}_s(pg)$ defines the energy between the supported object/human and the supporting object/layout:

$$\mathcal{E}_s(pg) = \sum_{(v_i, v_j) \in E_s} \mathcal{E}_o(v_i, v_j) + \mathcal{E}_{height}(v_i, v_j), \tag{2.48}$$

where $\mathcal{E}_o(v_i, v_j) = 1 - \text{area}(v_i \cap v_j)/\text{area}(v_i)$ is the overlapping ratio in the xy-plane, and $\mathcal{E}_{height}(v_i, v_j)$ is the absolute height difference between the lower surface of the supported object v_i and the upper surface of the supporting object v_j. We define $\mathcal{E}_o(v_i, v_j) = \mathcal{E}_{height}(v_i, v_j) = 0$ if the supporting object is floor or wall.

- *Physical Collision* $\mathcal{E}_c(pg)$ denotes the physical violations. We penalize the intersection among human, objects, and room layout except the objects in HOI and objects that could be a container.

The potential function is defined as:

$$\mathcal{E}_c(pg) = \sum_{v\in(V_{object}\cup V_{human})} \mathcal{C}(v, V_{layout}) + \sum_{\substack{v_i\in V_{object} \\ v_j\in V_{human} \\ (v_i,v_j)\notin E_{hoi}}} \mathcal{C}(v_i, v_j) + \sum_{\substack{v_i,v_j\in V_{object} \\ v_i,v_j\notin V_{container}}} \mathcal{C}(v_i, v_j) \ , \tag{2.49}$$

where $\mathcal{C}()$ denotes the volume of intersection between entities. $V_{container}$ denotes the objects that can be a container, such as a cabinet, desk, and drawer.

Human-object Interaction Prior $\mathcal{E}_{hoi}(pg)$ is defined on the interactions between human and objects:

$$\mathcal{E}_{hoi}(pg) = \sum_{(v_i,v_j)\in E_{hoi}} \mathcal{K}(v_i, v_j, a_{v_j}), \tag{2.50}$$

where $v_i \in V_{object}, v_j \in V_{human}$, and $\mathcal{K}$ is an HOI function that evaluates the interaction between an object and a human given the action label a:

$$\mathcal{K}(v_i, v_j, a_{v_j}) = -\log l(v_i, v_j|a_{v_j}), \tag{2.51}$$

where $l(v_i, v_j|a_{v_j})$ is the likelihood of the relative position between node v_i and v_j given an action label a, and λ_a the balancing factor. We formulate the action detection as a *multi-label classification*. The likelihood $l(\cdot)$ models the distance between key joints and the center of the object; *e.g.*, for "sitting", it models the relative spatial relation between the hip and the center of a chair. The likelihood can be learned from 3D HOI datasets with a multivariate Gaussian distribution $(\Delta x, \Delta y, \Delta z) \sim \mathcal{N}_3(\mu, \Sigma)$, where $\Delta x, \Delta y$, and Δz are the relative distances in the directions of three axes.

Likelihood $\mathcal{E}(I|pg)$ characterizes the consistency between the observed 2D image and the inferred 3D result. The projected 2D object bounding boxes and human poses can be computed by projecting the inferred 3D objects and human poses onto a 2D image plane. The likelihood is obtained by comparing the directly detected 2D bounding boxes and human poses with projected ones from inferred 3D results:

$$\mathcal{E}(I|pg) = \sum_{v\in V_{object}} \lambda_o \cdot \mathcal{D}_o(b(v), B'(v)) + \sum_{v\in V_{human}} \lambda_h \cdot \mathcal{D}_h(Po(v), Po'(v)), \tag{2.52}$$

where $B()$ and $B'()$ are the bounding boxes of detected and projected 2D objects, $Po()$ and $Po'()$ the poses of detected and projected 2D humans, $\mathcal{D}_o(\cdot)$ the IOU between the detected 2D bounding box and the convex hull of the projected 3D bounding box, and $\mathcal{D}_h(\cdot)$ the average pixel-wise Euclidean distance between two 2D poses.

Joint Inference

Given a single RGB image as the input, the goal of joint inference is to find the optimal parse graph that maximizes the posterior probability $p(pg|I)$. The joint parsing is a four-step process: (i) 3D scene initialization of the camera pose, room layout, and 3D object bounding boxes, (ii) 3D human pose initialization that estimates rough 3D human poses in a 3D scene, (iii) concurrent action detection, and (iv) joint inference to optimize the objects, layout, and human poses in 3D scenes by maximizing the posterior probability.

3D Scene Initialization

Following [120], we initialize the 3D objects, room layout, and camera pose cooperatively, where the room layout and objects are parametrized by 3D bounding boxes. For each object $v_i \in V_{object}$, we find its supporting object/layout by minimizing the supporting energy:

$$v_j^* = \arg\min_{v_j} \mathcal{E}_o(v_i, v_j) + \mathcal{E}_{height}(v_i, v_j) - \lambda_s \log p_{spt}(v_i, v_j), \tag{2.53}$$

where $v_j \in (V_{object}, V_{layout})$ and $p_{spt}(v_i, v_j)$ are the prior probabilities of the supporting relation modeled by multinoulli distributions, and λ_s a balancing constant.

3D Human Pose Initialization

We take 2D poses as the input and predict 3D poses in a local 3D coordinate following [165], where the 2D poses are detected and estimated by [166]. The local 3D coordinate is centered at the human hip joint, and the z-axis is aligned with the up direction of the world coordinate. To transform this local 3D pose into the world coordinate, we find the 3D world coordinate $\mathbf{v_{3D}} \in \mathbb{R}^3$ of one visible 2D joint $\mathbf{v_{2D}} \in \mathbb{R}^2$ (*e.g.*, head) by solving a linear equation with the camera intrinsic parameter K and estimated camera pose R. Per the pinhole camera projection model, we have

$$\alpha \begin{bmatrix} \mathbf{v_{2D}} \\ 1 \end{bmatrix} = K \cdot R \cdot \mathbf{v_{3D}}, \tag{2.54}$$

where α is a scaling factor in the homogeneous coordinate. To make the function solvable, we assume a pre-defined height h_0 for the joint position $\mathbf{v_{3D}}$ in the world coordinate. Lastly, the 3D pose initialization is obtained by aligning the local 3D pose and the corresponding joint position with $\mathbf{v_{3D}}$.

Concurrent Action Detection

We formulate the concurrent action detection as a multi-label classification problem to ease the ambiguity in describing the action. We define a portion of the action labels (*e.g.*, "eating", "making phone call") as the HOI labels, and the remaining action labels (*e.g.*, "standing", "bending") as general human poses without HOI. The mixture of HOI actions and non-HOI actions covers most of the daily human actions in indoor scenes. We manually map each of the HOI action labels to a 3D HOI relation learned from the SHADE dataset, and use the HOI actions as cues to improve the accuracy of 3D reconstruction by integrating it as prior knowledge in our model. The concurrent action detector takes 2D skeletons as the input and predicts multiple action labels with a three-layer MLP.

Inference

Given an initialized parse graph, we use MCMC with simulated annealing to jointly optimize the room layout, 3D objects, and 3D human poses through the non-differentiable energy space; see Algorithm 2 as a summary. To improve the efficiency of the optimization process, we adopt a scheduling strategy that divides the optimization process into following four phases with different focuses: (i) Optimize objects, room layout, and human poses without HOIs. (ii) Assign HOI labels to each human in the scene, and search the interacting objects of each human. (iii) Optimize objects, room layout, and human poses jointly with HOIs. (iv) Generate possible miss-detected objects by top-down sampling.

Algorithm 2: Joint Inference Algorithm

Given: Image I, initialized parse graph pg_{init}
Phase 1: **for** *Different temperatures* **do**
  Inference with physical commonsense $\mathcal{E}_{phy}$ but without HOI $\mathcal{E}_{hoi}$: randomly select from room layout, objects, and human poses to optimize pg
end
Phase 2: Match each agent with their interacting objects
Phase 3: **for** *Different temperatures* **do**
  Inference with total energy $\mathcal{E}$, including physical commonsense and HOI: randomly select from layout, objects, and human poses to optimize pg
end
Phase 4: Top-down sampling by HOIs

Dynamics. In Phase (i) and (iii), we use distinct MCMC processes. To traverse non-differentiable energy spaces, we design Markov chain dynamics q_1^o, q_2^o, q_3^o for objects, q_1^l, q_2^l for room layout, and q_1^h, q_2^h, q_3^h for human poses.

• Object Dynamics: Dynamics q_1^o adjusts the position of an object, which translates the object center in one of the three Cartesian coordinate axes or along the depth direction. The depth direction starts from the camera position and points to the object center. Translation along depth is effective with proper camera pose initialization. Dynamics q_2^o proposes rotation of the object with a specified angle. Dynamics q_3^o changes the scale of the object by expanding or shrinking corner positions of the cuboid with respect to object center. Each dynamic can diffuse in two directions: each object can translate in the direction of '$+x$' and '$-x$,' or rotate in the direction of clockwise and counterclockwise. To better traverse in energy space, the dynamics may propose to move along the gradient descent direction with a probability of 0.95 or the gradient ascent direction with a probability of 0.05.

• Human Dynamics: Dynamics q_1^h proposes to translate 3D human joints along x, y, z, or depth direction. Dynamics q_2^h is designed to rotate the human pose with a certain angle. Dynamics q_3^h adjusts the scale of human poses by a scaling factor on the 3D joints with respect to the pose center.

• Layout Dynamics: Dynamics q_1^l translates the wall towards or away from the layout center. Dynamics q_2^l adjusts the floor height, equivalent to change the camera height.

In each sampling iteration, the algorithm proposes a new pg' from current pg under the proposal probability of $q(pg \to pg'|I)$ by applying one of the above dynamics. The generated proposal is accepted with respect to an acceptance rate $\alpha(\cdot)$ as in the Metropolis-Hastings algorithm [158]:

$$\alpha(pg \to pg') = \min(1, \frac{q(pg' \to pg) \cdot p(pg'|I)}{q(pg \to pg') \cdot p(pg|I)}), \tag{2.55}$$

A simulated annealing scheme is adopted to obtain pg with high probability.

Top-down sampling. By top-down sampling objects from HOIs, the proposed method can recover the interacting 3D objects that are too small or novel to be detected by the state-of-the-art 2D object detector. In Phase (iv), we propose to sample an interacting object from the person if the confidence of HOI is higher than a threshold. Specifically, we minimize the HOI energy in Eq. (2.50) to determine the category and location of the object; see examples in Fig. 2.27.

Implementation Details. In Phase (ii), we search the interacting objects for each agent involved in HOI by minimizing the energy in Eq. (2.50). In Phase (iii), after matching each agent

Figure 2.27: Illustration of the top-down sampling process. The object detection module misses the detection of the bottle held by the person, but our model can still recover the bottle by reasoning HOI.

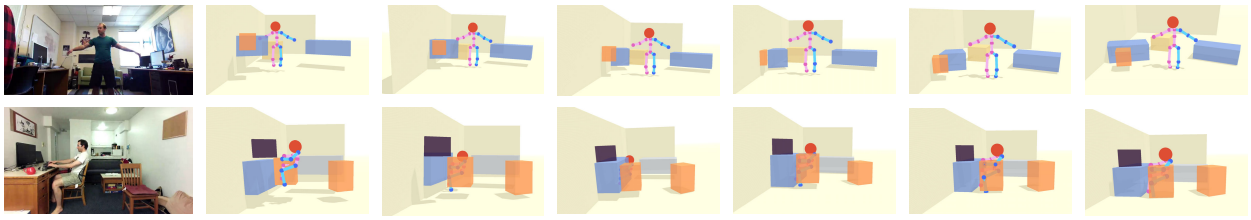


Figure 2.28: The optimization process of the scene configuration by simulated annealing MCMC. Each step is the number of accepted proposal.

with their interacting objects, we can jointly optimize objects, room layout, and human poses with the constraint imposed by HOI. Fig. 2.28 shows examples of the simulated annealing optimization process.

Some qualitative results are shown in Fig. 2.29.

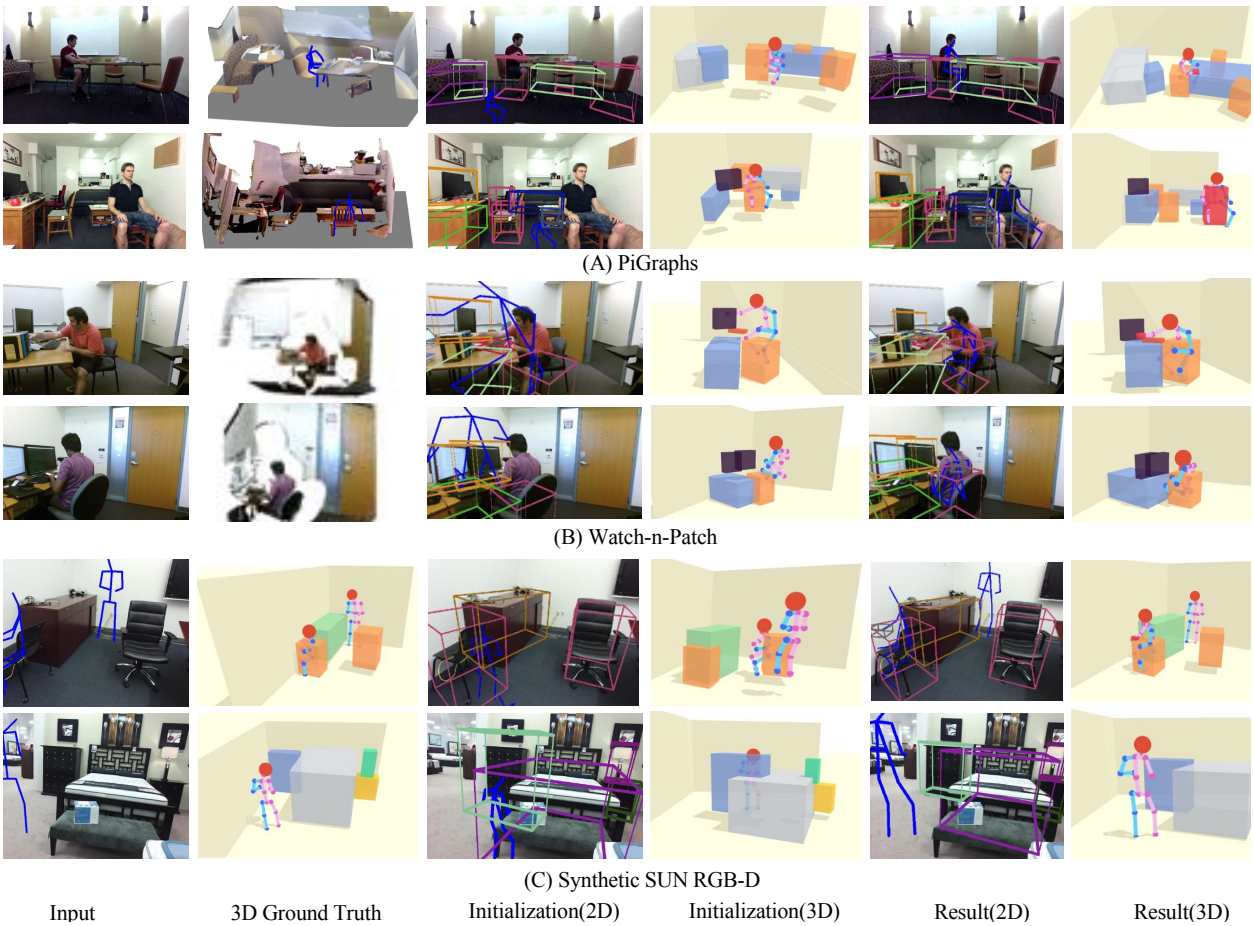


Figure 2.29: Qualitative results of the proposed method on three datasets. The proposed model improves the initialization with accurate spatial relations and physical plausibility and demonstrates an outstanding generalization across various datasets.

Chapter 3

Physical Commonsense Reasoning

3.1 Commonsense of Newtonian Physics

Perceiving causality, and using this perception to interact with an environment, requires a commonsense understanding of how the world operates at a physical level. Physical understanding does not necessarily require us to precisely or explicitly invoke Newton's laws of mechanics; instead, we rely on intuition, built up through interactions with the surrounding environment. Humans excel at understanding their physical environment and interacting with objects undergoing dynamic state changes, making approximate predictions from observed events. The knowledge underlying such activities is termed *intuitive physics* [167]. The field of intuitive physics has been explored for several decades in cognitive science and was recently reinvigorated by new techniques linked to AI.

Surprisingly, humans develop physical intuition at an early age [100], well before most other types of high-level reasoning, suggesting the importance of intuitive physics in comprehending and interacting with the physical world. The fact that physical understanding is rooted in visual processing makes visual task completion an important goal for future machine vision and AI systems. We begin this section with a short review of intuitive physics in human cognition, followed by a review of recent developments in computer vision and AI that use physics-based simulation and physical constraints for image and scene understanding.

3.1.1 Intuitive Physics in Human Cognition

Early research in intuitive physics provides several examples of situations in which humans demonstrate common misconceptions about how objects in the environment behave. For example, several studies found that humans exhibit striking deviations from Newtonian physical principles when asked to explicitly reason about the expected continuation of a dynamic event based on a static image representing the situation at a single point in time [168, 167, 169]. However, humans' intuitive understanding of physics was shown later to be much more accurate, rich, and sophisticated than previously expected once *dynamics* and proper *context* were provided [170, 171, 172, 173, 174].

These later findings are fundamentally different from prior work that systematically investigated the development of infants' physical knowledge [175, 176] in the 1950s. The reason for such a difference in findings is that the earlier research included not only tasks of merely reasoning about physical knowledge, but also other tasks [177, 178]. To address such difficulties, researchers have developed alternative experimental approaches [179, 112, 180, 181] to study the development of infants' physical knowledge. The most widely used approach is the violation-of-expectation method, in which infants see two test events: an expected event, consistent with the expectation shown, and an unexpected event, violating the expectation. A series of these kinds of studies have provided

strong evidence that humans—even young infants—possess expectations about a variety of physical events [182, 183].

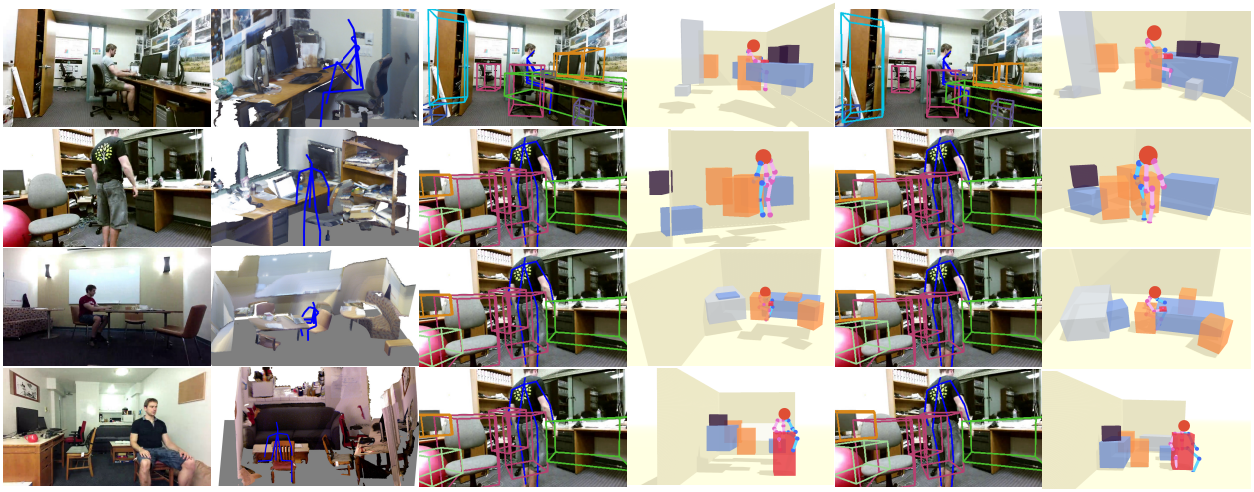
In a single glance, humans can perceive whether a stack of dishes will topple, whether a branch will support a child's weight, whether a tool can be lifted, and whether an object can be caught or dodged. In these complex and dynamic events, the ability to perceive, predict, and therefore appropriately interact with objects in the physical world relies on rapid physical inference about the environment. Hence, intuitive physics is a core component of human commonsense knowledge and enables a wide range of object and scene understanding.

In an early work, Achinstein [184] argued that the brain builds mental models to support inference through mental simulations, analogous to how engineers use simulations for the prediction and manipulation of complex physical systems (*e.g.*, analyzing the stability and failure modes of a bridge design before construction). This argument is supported by a recent brain imaging study [185] suggesting that systematic parietal and frontal regions are engaged when humans perform physical inferences even when simply viewing physically rich scenes. These findings suggest that these brain regions use a generalized mental engine for intuitive physical inference—that is, the brain's "physics engine." These brain regions are much more active when making physical inferences relative to when making inferences about *nonphysical* but otherwise highly similar scenes and tasks. Importantly, these regions are not exclusively engaged in physical inference, but are also overlapped with the brain regions involved in action planning and tool use. This indicates a very intimate relationship between the cognitive and neural mechanisms for understanding intuitive physics, and the mechanisms for preparing appropriate actions. This, in turn, is a critical component linking perception to action.

To construct humanlike commonsense knowledge, a computational model for intuitive physics that can support the performance of *any* task that involves physics, not just one narrow task, must be explicitly represented in an agent's environmental understanding. This requirement stands against the recent "end-to-end" paradigm in AI, in which neural networks directly map an input image to an output action for a specific task, leaving an implicit internal task representation "baked" into the network's weights.

Recent breakthroughs in cognitive science provide solid evidence supporting the existence of an intuitive physics model in human scene understanding. This evidence suggests that humans perform physical inferences by running probabilistic simulations in a mental physics engine akin to the 3D physics engines used in video games [186]. Human intuitive physics can be modeled as an approximated physical engine with a Bayesian probabilistic model [90], possessing the following distinguishing properties: (i) Physical judgment is achieved by running a coarse and rough forward physical simulation; and (ii) the simulation is stochastic, which is different from the deterministic and precise physics engine developed in computer graphics. For example, in the tower stability task presented in Ref. [90], there is uncertainty about the exact physical attributes of the blocks; they fall into a probabilistic distribution. For every simulation, the model first samples the blocks' attributes, then generates predicted states by recursively applying elementary physical rules over short-time intervals. This process creates a distribution of simulated results. The stability of a tower is then represented in the results as the probability of the tower not falling. Due to its stochastic nature, this model will judge a tower as stable only when it can tolerate small jitters or other disturbances to its components. This single model fits data from five distinct psychophysical tasks, captures several illusions and biases, and explains core aspects of mental models and commonsense reasoning that are instrumental to how humans understand their everyday world.

More recent studies have demonstrated that intuitive physical cognition is not limited to the understanding of rigid bodies, but also expands to the perception and simulation of the physical properties of liquids [187, 188] and sand [189]. In these studies, the experiments demonstrate that



(a) Input (b) Ground-truth (c) No physics (d) No physics (e) With physics (f) With physics

Figure 3.1: Scene parsing and reconstruction by integrating physics and human-object interactions. (a) Input image; (b) ground truth; (c and d) without incorporating physics, the objects might appear to float in the air, resulting in an incorrect parsing; (e and f) after incorporating physics, the parsed 3D scene appears physically stable. The system has been able to perceive the "dark" physical stability in which objects must rest on one another to be stable. Reproduced from Ref. [57] with permission of IEEE, © 2019.

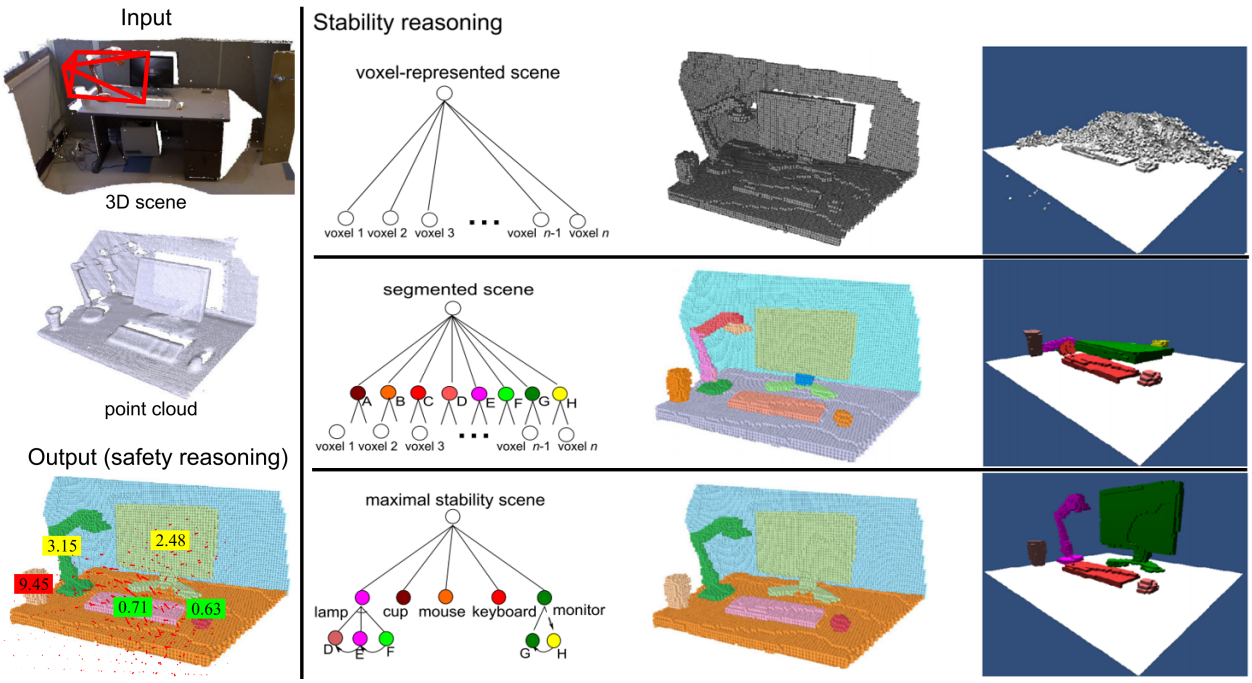
humans do not rely on simple qualitative heuristics to reason about fluid or granular dynamics; instead, they rely on perceived physical variables to make quantitative judgments. Such results provide converging evidence supporting the idea of mental simulation in physical reasoning. For a more in-depth review of intuitive physics in psychology, see Ref. [190].

3.1.2 Physics-based Reasoning in Computer Vision

Classic computer vision studies focus on reasoning about appearance and geometry—the highly visible, pixel-represented aspects of images. Statistical modeling [191] aims to capture the "patterns generated by the world in any modality, with all their naturally occurring complexity and ambiguity, with the goal of reconstructing the processes, objects and events that produced them [192]." Marr conjectured that the perception of a 2D image is an *explicit* multiphase information process [1], involving (i) an early vision system for perceiving [3, 4] and textons [5, 6] to form a primal sketch [7, 8]; (ii) a mid-level vision system to form 2.1D [9, 10, 11] and 2.5D [12] sketches; and (iii) a high-level vision system in charge of full 3D scene formation [13, 14, 15]. In particular, Marr highlighted the importance of different levels of organization and the internal representation [193].

Alternatively, perceptual organization [194, 195] and Gestalt laws [196, 20, 21, 197, 198, 199, 200, 201] aim to resolve the 3D reconstruction problem from a single RGB image without considering depth. Instead, they use priors—groupings and structural cues [202, 203] that are likely to be invariant over wide ranges of viewpoints [204]—resulting in feature-based approaches [16, 107].

However, both appearance [205] and geometric [49] approaches have well-known difficulties resolving ambiguities. In addressing this challenge, modern computer vision systems have started to account for "dark" aspects of images by incorporating physics; as a result, they have demonstrated dramatic improvements over prior works. In certain cases, ambiguities have been shown to be extremely difficult to resolve through current state-of-the-art data-driven classification methods, indicating the significance of "dark" physical cues and signals in our ability to correctly perceive and operate within our daily environments; see examples in Fig. 3.1 [57], where systems perceive



(a) (b)

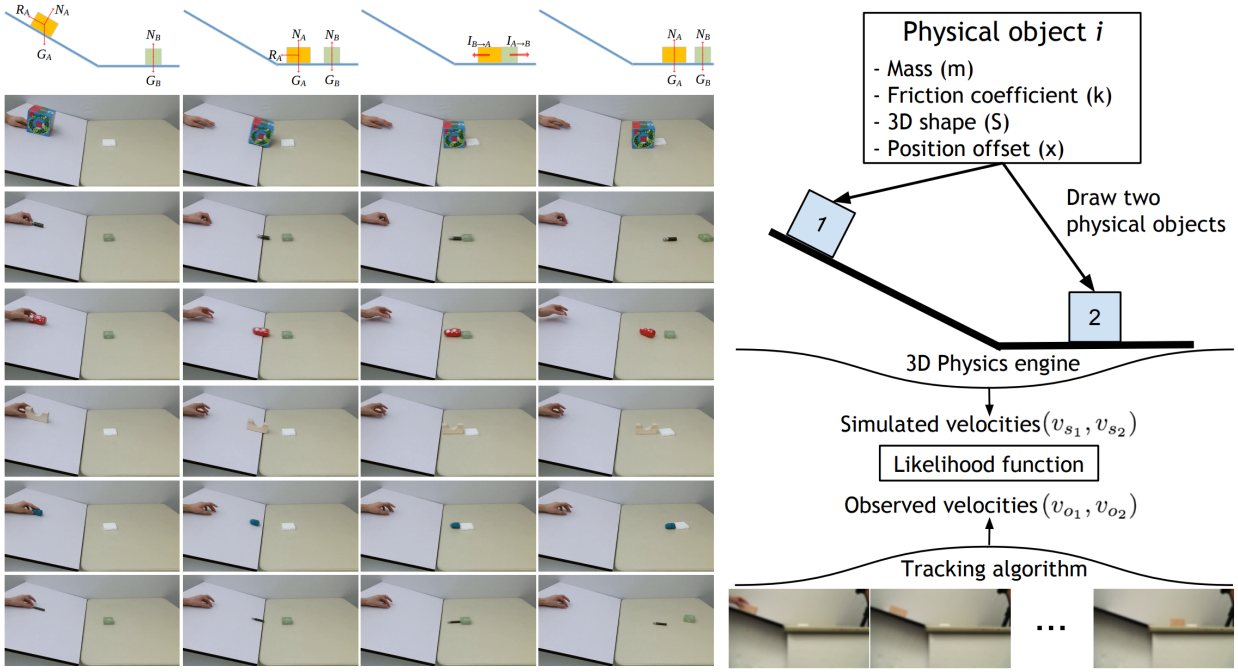
Figure 3.2: An example explicitly exploiting safety and stability in a 3D scene-understanding task. Good performance in this task means that the system can understand the "dark" aspects of the image, which include how likely each object is to fall, and where the likely cause of falling will come from. (a) Input: reconstructed 3D scene. Output: parsed and segmented 3D scene comprised of stable objects. The numbers are "unsafety" scores for each object with respect to the disturbance field (represented by red arrows). (b) Scene-parsing graphs corresponding to three bottom-up processes: voxel-based representation (top), geometric pre-process, including segmentation and volumetric completion (middle), and stability optimization (bottom). Reproduced from Ref. [118] with permission of Springer Science+Business Media New York, © 2015.

which objects must rest on each other in order to be stable in a typical office space.

Through modeling and adopting physics into computer vision algorithms, the following two problems have been broadly studied:

1. Stability and safety in scene understanding. As demonstrated in Ref. [118], this line of work is mainly based on a simple but crucial observation in human-made environments: by human design, objects in static scenes should be stable in the gravity field and be safe with respect to various physical disturbances. Such an assumption poses key constraints for physically plausible interpretation in scene understanding.
2. Physical relationships in 3D scenes. Humans excel in reasoning about the physical relationships in a 3D scene, such as which objects support, attach, or hang from one another. As shown in Ref. [56], those relationships represent a deeper understanding of 3D scenes beyond observable pixels that could benefit a wide range of applications in robotics, virtual reality (VR), and augmented reality (AR).

The idea of incorporating physics to address vision problems can be traced back to Helmholtz and his argument for the "unconscious inference" of probable causes of sensory input as part of the formation of visual impressions [206]. The very first such formal solution in computer vision dates back to Roberts' solutions for the parsing and reconstruction of a 3D block world in 1963 [207]. This work inspired later researchers to realize the importance of both the violation of physical laws for scene understanding [208] and stability in generic robot manipulation tasks [209, 210].



(a) Snapshots of datasets (b) Galileo model

Figure 3.3: Inferring the dynamics of the scenes. (a) Snapshots of the dataset; (b) overview of the Galileo model that estimates the physical properties of objects from visual inputs by incorporating the feedback of a physics engine in the loop. Reproduced from Ref. [220] with permission of Neural Information Processing Systems Foundation, Inc., © 2015

Integrating physics into scene parsing and reconstruction was revisited in the 2010s, bringing it into modern computer vision systems and methods. From a single RGB image, Gupta *et al.* proposed a qualitative physical representation for indoor [51, 121] and outdoor [211] scenes, where an algorithm infers the volumetric shapes of objects and relationships (such as occlusion and support) in describing 3D structure and mechanical configurations. In the next few years, other work [212, 213, 214, 215, 216, 129, 52, 217, 218, 54] also integrated the inference of physical relationships for various scene understanding tasks. In the past two years, Liu *et al.* [55] inferred physical relationships in joint semantic segmentation and 3D reconstruction of outdoor scenes. Huang *et al.* [56] modeled support relationships as edges in a human-centric scene graphical model, inferred the relationships by minimizing supporting energies among objects and the room layout, and enforced physical stability and plausibility by penalizing the intersections among reconstructed 3D objects and room layout [120, 57].

The aforementioned recent work mostly adopts simple physics cues; that is, very limited (if any) physics-based simulation is applied. The first recent work that utilized an actual physics simulator in modern computer vision methods was proposed by Zheng *et al.* in 2013 [116, 117, 118]. As shown in Fig. 3.2 [118], the proposed method first groups potentially unstable objects with stable ones by optimizing for stability in the scene prior. Then, it assigns an "unsafety" prediction score to each potentially unstable object by inferring hidden potential triggers of instability (the disturbance field). The result is a physically plausible scene interpretation (voxel segmentation). This line of work has been further explored by Du *et al.* [219] by integrating an end-to-end trainable network and synthetic data.

Going beyond stability and support relationships, Wu *et al.* [220] integrated physics engines with deep learning to predict the future dynamic evolution of static scenes. Specifically, a generative

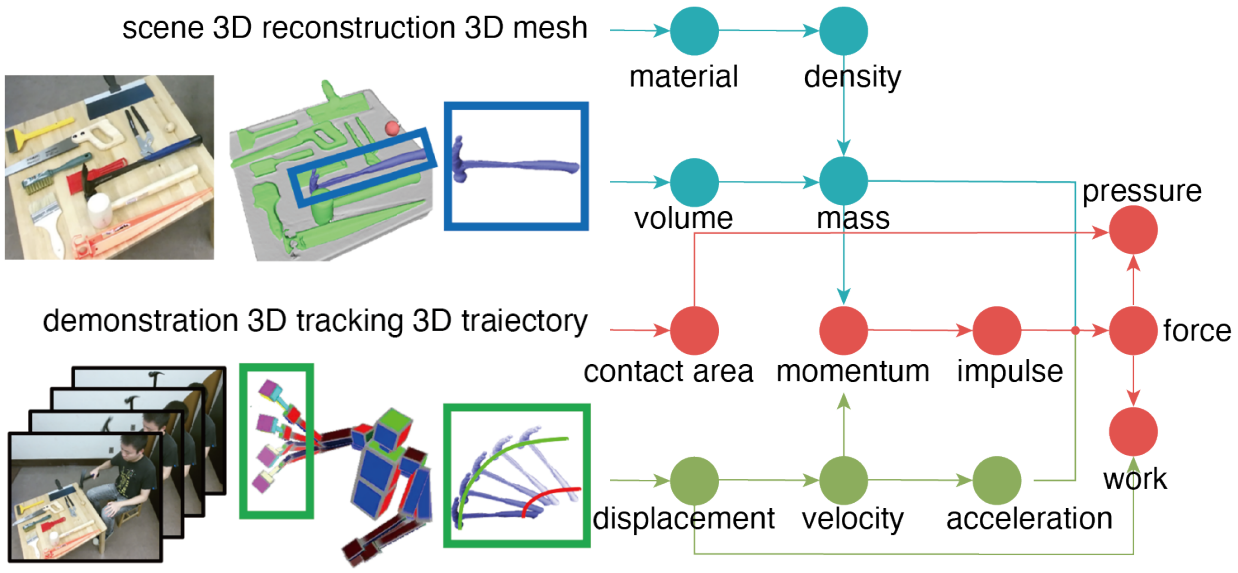


Figure 3.4: Thirteen physical concepts involved in tool use and their compositional relationships. By parsing a human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool attributes (blue), trajectories of tool use (green), or both together (red). Higher level physical concepts can be further derived recursively. Reproduced from Ref. [222] with permission of the authors, © 2015.

model named Galileo was proposed for physical scene understanding using real-world videos and images. As shown in Fig. 3.3, the core of the generative model is a 3D physics engine, operating on an object-based representation of physical properties including mass, position, 3D shape, and friction. The model can infer these latent properties using relatively brief runs of markov chain monte carlo (MCMC), which drive simulations in the physics engine to fit key features of visual observations. Wu *et al*. [221] further explored directly mapping visual inputs to physical properties, inverting a part of the generative process using deep learning. Object-centered physical properties such as mass, density, and the coefficient of restitution from unlabeled videos could be directly derived across various scenarios. With a new dataset named *Physics 101* containing 17 408 video clips and 101 objects of various materials and appearances (*i.e*., shapes, colors, and sizes), the proposed unsupervised representation learning model, which explicitly encodes basic physical laws into the structure, can learn the physical properties of objects from videos.

Integrating physics and predicting future dynamics opens up quite a few interesting doors in computer vision. For example, given a human motion or task demonstration presented as a RGB-D image sequence, *et al*. [222] built a system that calculated various physical concepts from just a single example of tool use (Fig. 3.4), enabling it to reason about the essential physical concepts of the task (*e.g*., the force required to crack nuts). As the fidelity and complexity of the simulation increased, Zhu *et al*. [223] were able to infer the forces impacting a seated human body, using a finite element method (FEM) to generate a mesh estimating the force on various body parts.

Physics-based reasoning can not only be applied to scene understanding tasks, as above, but have also been applied to pose and hand recognition and analysis tasks. For example, Brubaker *et al*. [224, 225, 226] estimated the force of contacts and the torques of internal joints of human actions using a mass-spring system. Pham *et al*. [227] further attempted to infer the forces of hand movements during human-object manipulation. In computer graphics, soft-body simulations based on video observation have been used to jointly track human hands and calculate the force of

contacts [228, 229]. Altogether, the laws of physics and how they relate to and among objects in a scene are critical "dark" matter for an intelligent agent to perceive and understand; some of the most promising computer vision methods outlined above have understood and incorporated this insight.

3.2 Case Study: Commonsense of Particle and Fluid Stuff

3.2.1 Introduction

Consider *KerPlunk*, a children's game in which marbles are suspended in the air by a lattice of straws within a cylindrical tube. The goal of the game is for each player to take turns removing straws while minimizing the number of marbles that fall through the lattice. The task requires players to reason about the interaction between rigid bodies and obstacles in 3D space. But what if the marbles were replaced by balls of liquid or sand? Could humans predict how those substances would move? Would those predictions agree with a generative model based on ground-truth, Newtonian physics? Recent computational evidence has demonstrated that human predictions *do* agree with Newtonian physics, given noisy perception and prior beliefs about spatially represented variables: *i.e.*, the *noisy Newton* hypothesis [187, 90, 230, 231, 188, 232, 233, 171]. The hypothesis suggests that humans rationally infer the values of physical variables and utilize normative conservation principles (approximately) to make predictions about future scene states. Computationally, this is achieved by sampling the initial locations, motions from noisy sensory input, and sampling physical attributes in a physical scene, propagating these variables forward in time according to approximated physical principles, and aggregating queries on the final scene states to form predicted response distributions.

[187] extended the noisy Newton framework from block tower judgments [90] to liquid dynamics using an intuitive fluid engine (IFE). In their IFE, ground-truth physics was approximated using smoothed particle hydrodynamics (SPH [234], a particle-based computational method for simulating non-solid dynamics. Their model predictions matched human judgments about future fluid states and outperformed alternative models that did not employ probabilistic simulation or account for physical uncertainty. Furthermore, the authors found that their participants' predictions were sensitive to latent fluid attributes (stickiness and viscosity), suggesting that humans have rich knowledge about the intrinsic properties of liquid.

The present study argues for the same general class of model as Bates *et al.*'s (2015) IFE and extends their work by examining (1) whether human predictions about future states of multiple substances (*i.e.*, rigid balls, liquid, and sand) differ, and (2) whether those differences can be consistently modeled using approximate, probabilistic simulation based on a hybrid particle/grid simulator adapted from previous work ([188]). Although granular materials (*e.g.*, sand) are encountered in everyday life, they are far less common than liquid; can humans accurately predict how sand will interact with obstacles and support surfaces? We present two experiments exploring the human capacity to predict the dynamics of substances varying in familiarity and physical properties, examining how human judgments and model predictions vary for different substances. Experiment 1 examines human predictions about the resting composition of sand after pouring from a funnel. In Experiment 2, participants make predictions about the flow of liquid, sand, and rigid balls past obstacles using a design similar to Bates *et al.*'s [187] study.

3.2.2 Computational Models

MPM Physical Simulator

The Material Point Method (MPM) [235] is commonly used in computer graphics to simulate the behavior of solids and fluids. The MPM has produced physically accurate and visually realistic simulations of the dynamics of liquid [236] and sand [237], in addition to general continuum materials such as stiff elastic objects [238].

The Appendix presents a mathematical overview of our MPM simulator, which provides a unified, particle-based simulation framework that handles rigid balls, liquid, and sand with essentially the same numerical algorithm, albeit with appropriately differing material parameters. The MPM method is physically accurate, numerically stable, and computationally efficient, enabling us to synthesize a large set of stimuli in a short amount of time by simply varying material parameters and the locations of the initial objects and colliding geometries. Running all the simulations in the same framework for the purposes of the present study also enables fair comparisons among the three types of substances, since we avoid potential inconsistencies in the numerical accuracies of multiple simulators specialized to particular materials.

Intuitive Substance Engine

Although the MPM simulator provides accurate and stable kinematics and dynamics for liquid, sand, and rigid balls using a unified framework, this high-precision, deterministic process does not account for the variability of human judgments in various intuitive physics tasks. Inspired by previous implementations of the noisy Newton framework (*e.g.*, [187, 90]), we combined our MPM simulator with noisy inputs, yielding an Intuitive Substance Engine (ISE) that accounts for uncertainty in human perception and reasoning in physical situations involving the three substances examined in this study. Details on how noisy perceptual inputs are defined and sampled are provided in the *Model Results* section of each experiment.

It is important to note that our ISE (employing MPM simulation) is roughly equivalent to Bates *et al.*'s (2015) IFE (employing SPH simulation) in that both models apply the noisy Newton framework to substance dynamics. Indeed, SPH is a viable method for simulating the dynamics of both granular materials and liquids, although MPM provides a more efficient and accurate means of doing so. We do not envision that the predictions of the two methods would differ substantially from one another when applied to a given set of stimuli.

Data-Driven Models

Two data-driven models based on statistical learning methods were constructed as competing models—the generalized linear model (GLM) [239] and Extreme Gradient Boosting (XGBoost) [240]. GLM is a classic machine learning method, commonly expressed by $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, where $\mathbf{X}$ is the feature input matrix, $\mathbf{B}$ is the parameter matrix (learned using a training dataset), and $\mathbf{U}$ is the error between the ground truth matrix $\mathbf{Y}$ and prediction $\mathbf{XB}$.

XGBoost is a recently-published machine learning method which has been utilized by multiple research teams to achieve outstanding performance in several Kaggle competitions. Essentially, it is a type of tree ensemble model: *i.e.*, a set of classification and regression trees (CART). Formally, $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$, where K is the number of trees, f_k is a function in the functional space $\mathbf{F}$ comprising the set of all possible CARTS. The objective function is defined as $R(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$, where θ includes the model parameters to be learned during training, l is the loss function, which

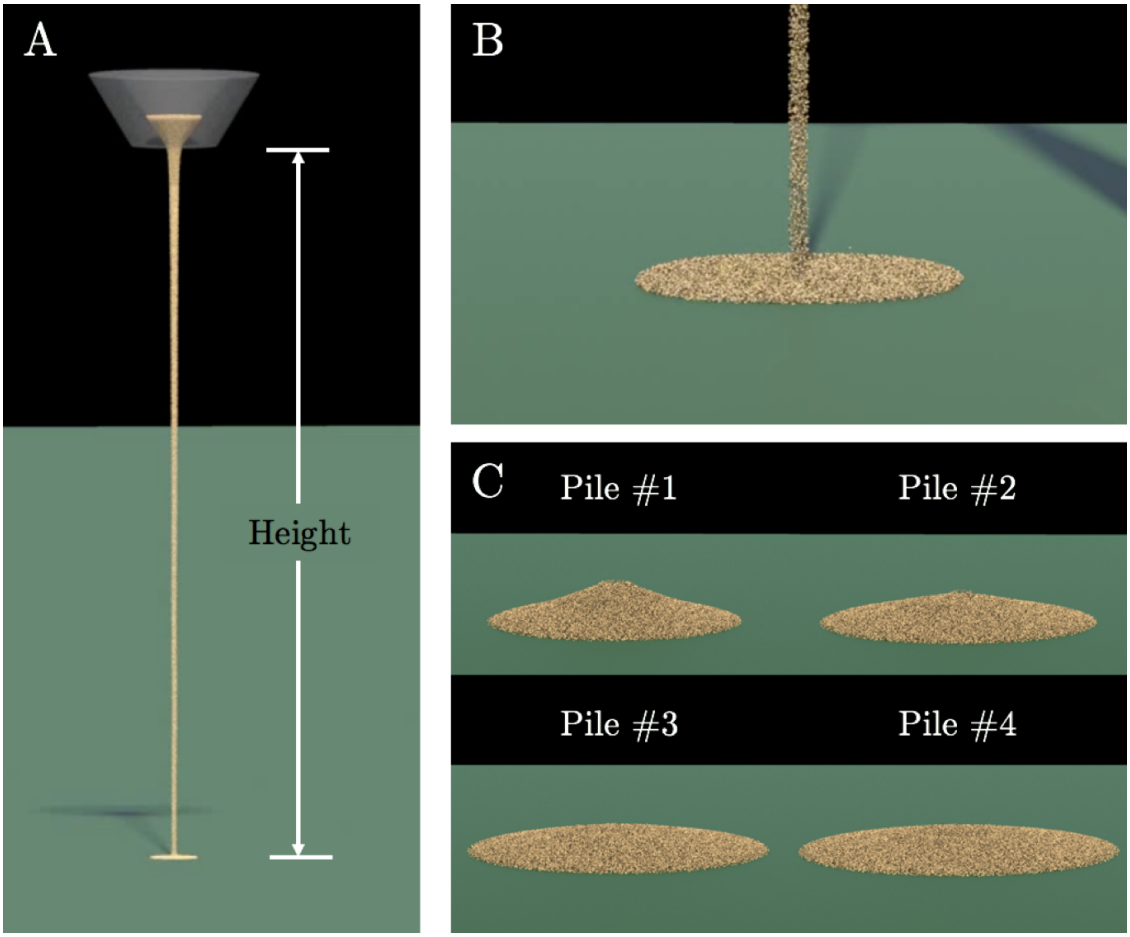


Figure 3.5: Intermediate frames from the demonstration video in Experiment 1 from the (A) zoomed-out and (B) zoomed-in perspective. (C) Sand pile choices in Experiment 1's judgment task.

measures the cost between ground truth y_i and prediction $\hat{y}_i$, and $\sum_{k=1}^{K} \Omega(f_k)$ is a regularization term that prevents the model from over-fitting the training data.

3.2.3 Experiment 1

The first experiment was designed to determine whether humans are able to predict the resting geometry of sand after it is poured from a funnel onto a surface, and whether dynamic visualizations of the pouring behavior facilitate mental simulation of sand-surface interactions.

Participants

A total of 108 undergraduate students (81 females), of mean age = 20.2 years, were recruited from the University of California, Los Angeles (UCLA), Department of Psychology subject pool and were compensated with course credit.

Materials and Procedure

Participants first viewed a demonstration video of sand falling from a funnel suspended 10 cm above a level surface. The pouring event was viewed three times from a zoomed-out perspective (Fig. 3.5A) and then a zoomed-in perspective (Fig. 3.5B). The duration of the video was 29 sec. After viewing

the demonstration video, participants were presented with a sand-filled funnel suspended 1/2, 1, 2, and 4 cm above the surface in a randomized order.

Forty-three participants were assigned to the Static Condition and viewed a static image (zoomed-out) in which the funnel was positioned at a particular height. Sixty-five were assigned to the Dynamic Condition and viewed a video (zoomed in and out; looped three times; 35 sec duration) of sand pouring from a funnel that was positioned at different heights above the surface. In the Dynamic Condition, the region of the surface where the sand fell was occluded by a gray rectangle.

After viewing each situation, participants were asked to indicate which of four sand piles would result from the sand pouring from the funnel at the indicated height (Fig. 3.5C). For each trial, the stimulus images (for the Static Condition) and final video frames (Dynamic Condition) remained on the screen until a response was made. The pile choices were shown from the zoomed-in perspective and represented the ground-truth resting geometries resulting from each situation: *i.e.*, Piles 1, 2, 3, and 4 correspond with the pile resulting from funnels suspended 1/2, 1, 2, and 4 cm above the surface, respectively. The experiment consisted of 4 trials. The stimulus videos can be viewed at *https://vimeo.com/216585992*.

Human Results

At each funnel height, the proportion of participants choosing each sand pile did not differ between the Dynamic and Static Conditions: $\chi^2(3) = 2.21$, 2.34, 2.41, and 1.13 for funnel heights of 1/2, 1, 2, and 4 cm, respectively. These results suggest that dynamic visualizations of sand pouring from the funnel in each situation did not alter participants' judgments about the sand's resting geometry. However, the participants' pile choices did vary across different heights ($\chi^2(9) = 176.54$), indicating that funnel height influenced their predictions on the resting geometry of falling sand.

As shown in Fig. 3.6, participants' pile choices shifted toward higher-numbered, flatter piles as funnel height increased. These results indicate that participants' predictions were sensitive to funnel height, but inconsistent with ground-truth resting states. In the next section, predictions from the three computational models (ISE, GLM, and XGBoost) are compared to human performance to determine whether the noisy Newton framework can account for participants' deviations from ground-truth judgments.

Model Results

ISE Predictions: The input variables for our ISE in Experiment 1 were funnel height (*i.e.*, initial sand height) with perceptual uncertainty and sand friction angle with mental simulation uncertainty. Given the ground-truth values of initial funnel height and friction angle (H_{iT}, θ_{iT}), $N = 10,000$ noisy samples $\{(H_i, \theta_i), \quad i = 1, ..., N\}$ were generated and passed to our MPM simulator, which returned the final height of the sand pile for each sample. Instead of choosing from 4 piles (*i.e.*, the task presented to the participants), the MPM simulator compares the estimated height of the final sand pile, formally $D(H_i, \theta_i) = H_p \in \mathbb{R} > 0$, with the heights of the 4 pile options given to human participants. The pile option with the minimum height difference was chosen as the predicted judgment for each sample. Finally, by aggregating predictions across the $10,000$ samples, our ISE outputs a predicted response distribution for each trial.

To model physical uncertainty in participants' mental simulations, our ISE sampled funnel heights and friction angles from noisy distributions. Gaussian noise (0 mean, σ_H^2 variance) was added to the ground-truth funnel height in each situation. Gaussian noise was also added to the ground-truth friction angle θ_{iT}, but in logarithmic space (see [233]): $\theta_i = f^{-1}(f(\theta_{iT}) + \epsilon)$, where

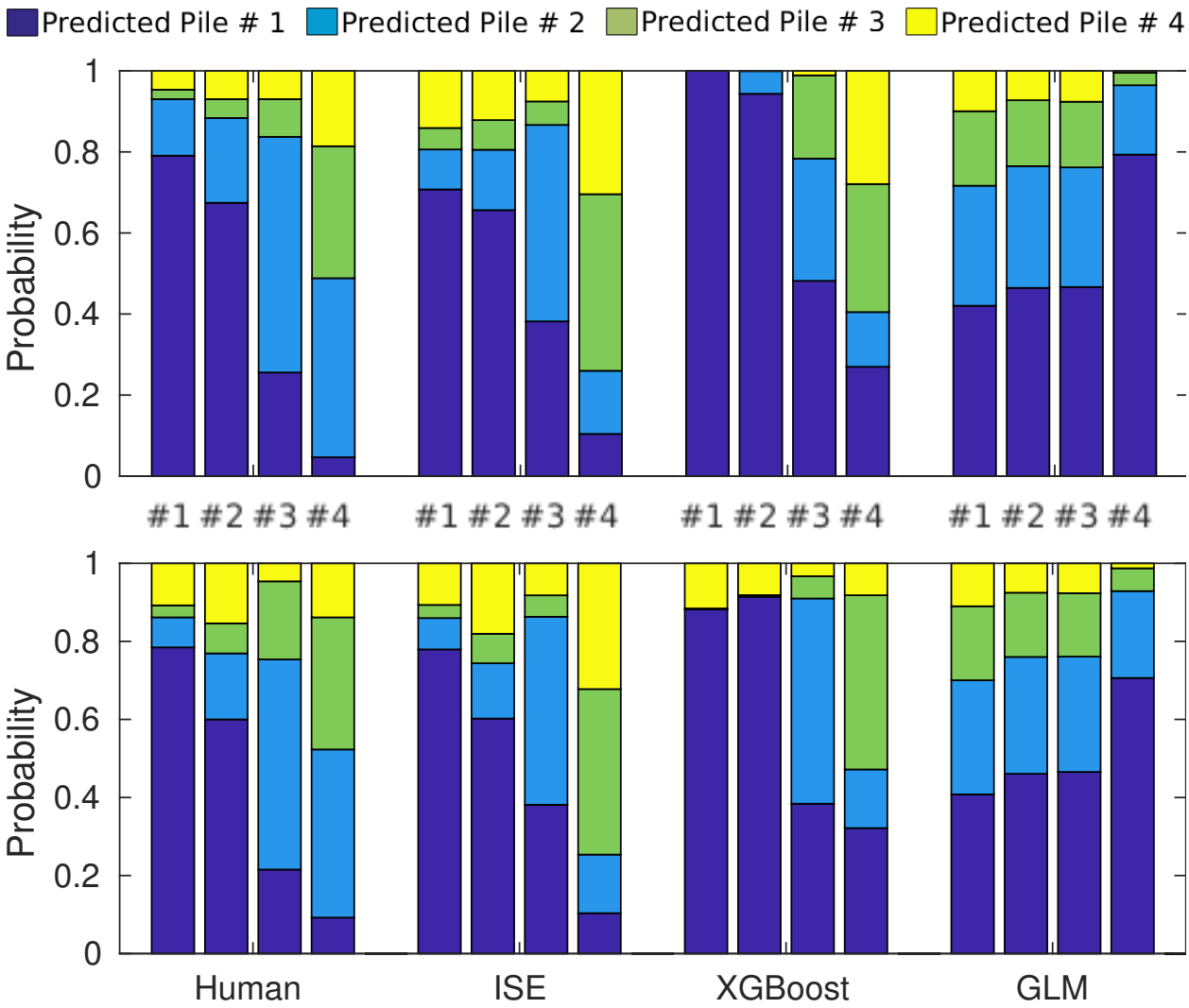


Figure 3.6: Model prediction results compared to human judgments. (Upper) Static Condition. (Lower) Dynamic Condition. Each bar, 1, 2, 3, and 4, corresponds to testing trials with funnel height 1/2, 1, 2, and 4 cm, respectively.

θ_{iT} is the ground truth value of the initial sand height, $f(\theta_{iT}) = \log(\omega \cdot \theta_{iT} + k)$, and ϵ represents Gaussian noise with 0 mean and σ_ϵ^2 variance. The results reported herein used the following model parameters: $\sigma_H = 0.12H_{iT}$, $\sigma_\epsilon = 0.6$, $\omega = 0.8$ and $k = 1.5$.

Data-Driven Predictions: To predict human judgments, both GLM and XGBoost were tested on the ith pile ($i = 1, 2, 3, 4$) and trained on the remaining three piles. During training, $10,000$ samples were drawn for each remaining pile ($30,000$ samples) and passed to our MPM simulator. Samples were generated using the sampling method described in the previous section. After training on the $30,000$ samples, both data-driven models were tested on another $10,000$ samples generated from noisy input based on the configuration of pile i. The final distribution was formed by aggregating the predictions across the $10,000$ samples.

Model Comparisons: Fig. 3.6 depicts the predictions of the ISE, XGBoost, and GLM models compared to human judgments. All four models achieved high correlations with human performance

Table 3.1: Root-mean-square deviation (RMSD) values for the ground-truth (GT), ISE, GLM, and XGBoost models for Experiments 1 and 2. Lower values of RMSD indicate better model fits.

	GT	ISE	XGBoost	GLM
Experiment 1 (Static)	0.458	0.101	0.267	0.171
Experiment 1 (Dynamic)	0.445	0.104	0.237	0.148
Experiment 2 (Liquid)	0.145	0.081	1.382	0.077
Experiment 2 (Sand)	0.170	0.080	1.422	0.120
Experiment 2 (Balls)	0.186	0.102	2.067	0.191

(Static: $r(12) = 0.91$, 0.84, and 0.27; Dynamic: $r(12) = 0.88$, 0.88, and 0.30 for ISE, XGBoost, and GLM, respectively). Human performance was much less correlated with ground-truth predictions (Static: $r(12) = 0.17$; Dynamic: $r(12) = 0.19$). The ISE model predictions were more correlated with the human data than the competing data-driven model predictions in the Static condition but were only slightly more correlated than XGBoost predictions in the Dynamic condition. Hence, this paper uses The root-mean-square deviation (RMSD) between human responses and model results to compare the model fits. We found that RMSD between human responses and ISE predictions for the 4 judgment trials was less than that between ground-truth predictions in both Static and Dynamic Conditions (see Table 3.1). We also examined modeling performance using the Bayesian information criterion (BIC) to account for the different number of free parameters in each model. We found that the ISE provides a better fit to the human data than the ground-truth and data-driven models in both conditions. For ground-truth, ISE, XGBoost, and GLM models, Static BIC $= -25.0$, -62.3, -31.2, -45.4, and Dynamic BIC $= -25.9, -61.3, -35.0, -50.0$, respectively. The model with the lowest BIC value is preferred.

Although XGBoost captures most of the trends in the human judgments, it appears to overfit the data in some cases. In the Static Condition, XGBoost's predicted response proportion for Pile 1 in the Trial 1 (1/2 cm funnel height) is greater than the proportion in Trial 2 (1 cm funnel height), which is consistent with human judgments. In the Dynamic Condition, however, XGBoost's predicted response proportion for Pile 1 is greater in Trial 1 than in Trial 2, which is inconsistent with trends in human performance. Alternatively, GLM showed very poor performance, predicting an increasing probability of Pile 1 choices for larger funnel heights. This trend is in the opposite direction of that observed in the human data, most likely due to the small number of training trials used to make each prediction.

3.2.4 Experiment 2

Our results from the first experiment indicate that humans are able to predict the resting geometry of sand piles, even though they may not have very rich experience interacting with sand in daily-life. The second experiment was conducted to determine 1) whether humans can reason about complex interactions between sand and rigid obstacles and 2) whether their predictions about the resting state of sand in novel situations differ from predictions about other substances, such as liquid and rigid balls.

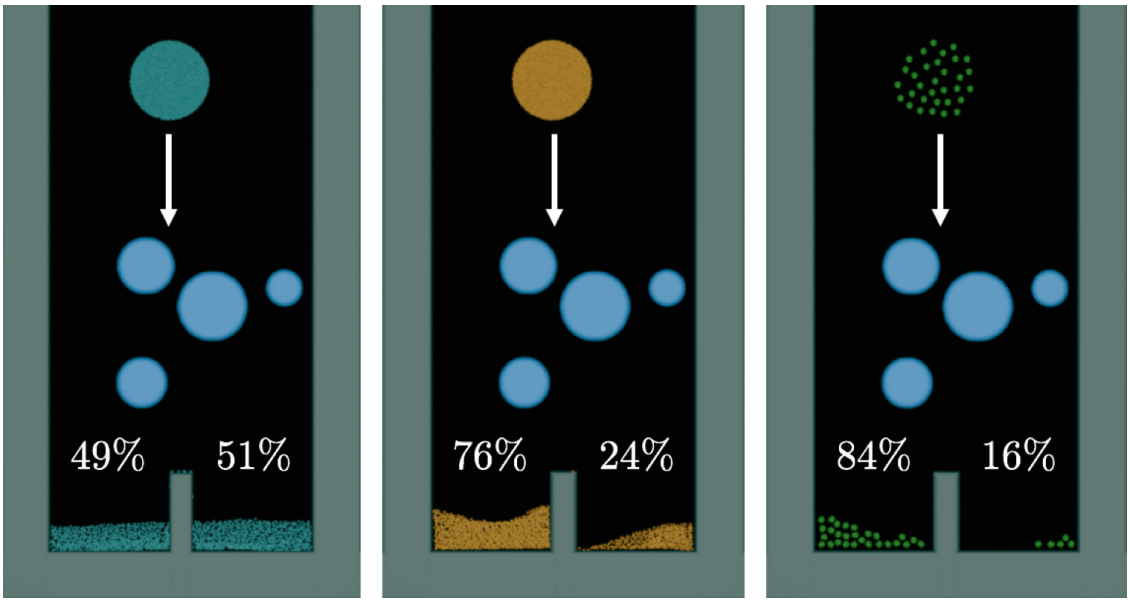


Figure 3.7: Initial (top) and final (bottom) state of liquid (left), sand (middle), and a set of rigid balls (right) for a testing trial in Experiment 2 with 5 obstacles. The percentages indicate the amount of each substance that fell into the left and right basins. Only the initial state of each substance was shown in the testing trials.

Participants

A total of 90 undergraduate students (66 females), mean age 20.9, were recruited from the UCLA Department of Psychology subject pool, and were compensated with course credit.

Materials and Procedure

The procedure in Experiment 2 was similar to the design in Bates *et al.*'s (2015) experiment: *i.e.*, participants viewed a volume of a substance suspended in the air above obstacles and were asked to predict the proportion that would fall into two basins separated by a vertical divider below (Fig. 3.7). The present experiment differed from previous work in that participants reasoned about the resting state of one of three different substances: liquid, sand, or sets of rigid balls. Also, whereas the previous study used polygonal obstacles, those in the present study were circles varying in size. Depth information was also not present in the rendered situations. The stimulus videos can be viewed at *https://vimeo.com/216585992*.

Situations were generated by sampling between 2 and 5 obstacle locations from a uniform distribution bounded by the width and height of the chamber. The diameter, d, of each obstacle was sampled from a uniform distribution bounded by $[0.15, 0.85]$ relative to the randomly-generated center points. The center points were generated by uniformly sampling the entire space. If the generated obstacles were placed outside the boundary, the configuration was rejected and re-sampled. Our MPM simulator was used to determine the ground-truth proportion of each substance in the left and right basins for each of the generated situations. For each substance, forty testing trials (10 trials with 2, 3, 4, and 5 obstacles) were chosen from the generated set such that the ground-truth proportion of substance in the left basin was approximately uniform across trials. The testing trials were the same for each substance.

Participants were randomly assigned to either the liquid, sand, or rigid balls condition. Thirty participants were assigned to each condition in a between-subjects experimental design. Prior to the testing trials, participants completed five practice trials with two obstacles in each situation

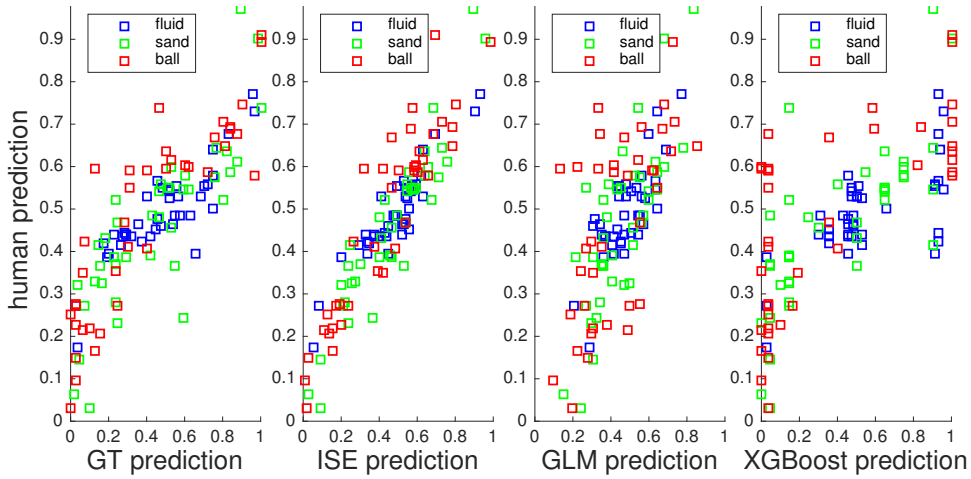


Figure 3.8: Model prediction results compared to human predictions. From left to right: Ground-truth (GT), ISE, GLM, and XGBoost.

in a randomized order. After answering 1) which basin the majority of the substance would fall into and 2) the expected proportion that would fall into the indicated basin, participants viewed a video (13 second duration) of the situation unfolding and were told the resulting proportion in the ground-truth simulation. After completing the practice trials, participants completed 40 testing trials in a randomized order by answering the same two questions in each trial. No feedback was given following the completion of each testing trial.

Human Results

Participants' predicted proportions in the testing trials were strongly correlated with ground-truth predictions in the liquid, sand, and rigid balls conditions ($r(38)$=0.86, 0.82, and 0.88; RMSD = 0.145, 0.170, 0.186, respectively). The deviation for each trial was calculated by subtracting the ground-truth proportion from each participant's proportion response. The deviation differed significantly between the three substance conditions ($F(2) = 3.64$, $p = 0.03$), indicating that the difference between human predictions and the ground-truth status varied according to the substance type. To determine whether participants' response proportions differed between substances, a random factor ANOVA was conducted for a chosen set of trials. The chosen set excluded those trials where the majority of each substance fell into the same basin (left or right) according to the ground-truth simulation. We found that the response proportions showed significant differences depending on substance type ($F(2) = 8.43$, $p < 0.01$). The next section examines whether an ISE and two data-driven models can capture differences in human performance between the three substances.

Model Results

ISE Predictions: In Experiment 2, the observable input variables for our ISE for each substance were 1) the initial, horizontal position of the substance, and 2) the positions of the circular obstacles in each situation. The latent substance attributes accepted by the engine were viscosity, friction angle, and restitution coefficient for liquid, sand, and the rigid balls, respectively. Gaussian noise was added to the substance's (ground-truth) horizontal position (0 mean, 0.35 variance) and the obstacles' (ground-truth) positions in 2D space (0 mean, 0.4 variance). Logarithmic Gaussian noise was added to each substance's ground-truth attribute value via the logarithmic transformation

specified in Experiment 1. The results reported here utilized the following model parameters for all three substances: $\sigma_\epsilon = 0.5$, $\omega = 0.8$, $k = 1.2$. Two thousand samples (40 situations $\times$ 50 noisy samples) were used for each substance.

Data-Driven Predictions: Similar to Experiment 1, both GLM and XGBoost were tested. The training data were randomly generated situations with basin proportions calculated using resting state output from our MPM simulator. Input features were the collection of both the observable input variables and latent substance attributes used in the ISE prediction. In total, 6000 samples were used for training.

Model Comparisons: Fig. 3.8 depicts the comparison between human and model basin predictions from the ground-truth (GT), ISE, GLM, and XGBoost models, and Table 1 depicts the root-mean-square deviation (RMSD) of each model's predictions from human ones. The human data were highly consistent with ISE predictions ($r(38) = 0.93, 0.93, 0.93$; RMSD = 0.081, 0.080, 0.102 for liquid, sand, and rigid balls, respectively). The ISE model predictions deviated from the human data to a lesser degree than the GT model predictions ($r(38) = 0.87, 0.85, 0.88$; RMSD = 0.145, 0.170, 0.186 for liquid, sand, and rigid balls, respectively), indicating a superior account of human predictions across a range of substances. In comparison, GLM and XGBoost predictions were less consistent with human predictions (GLM: $r(38) = 0.77, 0.78, 0.65$, RMSD = 0.077, 0.120, 0.191; XGBoost: $r(38) = 0.67, 0.74, 0.71$, RMSD = 1.382, 1.422, 2.067 for liquid, sand and rigid balls, respectively). As in the previous experiment, we compared each model's BIC measure in each condition to account for the number of free parameters in each model. We found that the BIC values for the ground-truth, GLM, and XGBoost models (GT: BIC = -154.5, -141.8, -134.6; GLM: BIC = -194.0, -158.6, -121.4; XGBoost: BIC = 36.9, 39.2, 69.2 for liquid, sand, and rigid balls, respectively) were consistently greater than the values for the ISE model (BIC = -190.0, -191.0, -171.6 for liquid, sand, and rigid balls, respectively), further reinforcing the superior performance of our simulation-based model.

It is worth noting that our ISE achieved consistent performance across all three substances, whereas GLM and XGBoost were less capable of predicting human judgments about rigid balls and liquid. In addition, our ISE used only one third of the training samples that XGBoost and GLM needed, demonstrating that a generative physical model with noisy perceptual inputs is capable of learning with a smaller number of samples than data-driven methods.

3.2.5 Discussion

Results from Experiments 1 and 2 provide converging evidence that humans can predict outcomes of novel physical situations by propagating approximate spatial representations forward in time using mental simulation. This stands in contrast to early research in rigid-body collisions suggesting that human physical predictions do not obey ground-truth physics, instead relying on heuristics (*e.g.*, [241, 242]). ISE predictions entailing the noisy Newton framework outperformed both ground-truth and data-driven models in both experiments, further confirming the role of perceptual noise and physical dynamics in human intuitive physical predictions.

Previous work has demonstrated that humans spontaneously employ mental simulation strategies when reasoning about novel physical situations [243, 244, 245]. Recent fMRI results suggest that intuitive physical inferences are made using an internal physics engine encoded in the brain's "multiple demand" network [185]. Although our ISE employed herein accounted for perceptual uncertainty in each situation, the simulations themselves closely approximated normative physical principles. Adding "stochastic noise" to physical dynamics, however, has been shown to increase

model performance when predicting human responses in simple physical situations [246]. While dynamic uncertainty can easily be built into rigid-body collisions, employing this strategy in the present physical simulations would preclude stable numerical evaluation. Thus, future computational work should explore methods for adding dynamic uncertainty into complex physical simulations while preserving their accuracy and stability.

Results from the present study demonstrate that human predictions about substance dynamics can be accurately predicted by a unified simulation method with uncertainty implemented into underlying physical variables. It is unlikely, however, that the human brain numerically evaluates partial differential equations to discern whether physical quantities (*e.g.*, mass and momentum) are conserved, nor is it likely that the brain stores the locations of vast numbers of particles to form physical predictions and judgments. Instead, our results provide evidence that humans approximate the dynamics of substances in a manner consistent with ground-truth physics but succumb to biases invoked by perceptual noise when inferring future environmental states. It remains unclear, however, whether the dynamics of rigid objects, liquids, and granular materials are approximated using separable mechanisms or a single cognitive architecture with different assumptions and constraints. The success of our unified simulation model across different substance-types supports the latter perspective.

Acknowledgments Support for the present study was provided by a NSF Graduate Research Fellowship, NSF grant BCS-1353391, DARPA XAI grant N66001-17-2-4029, DARPA SIMPLEX grant N66001-15-C-4035, ONR MURI grant N00014-16-1-2007, and DoD CASIT grant W81XWH-15-1-0147.

3.2.6 Appendix: Details of Our MPM Simulator

The governing partial differential equations utilize the principles of conservation of mass and momentum:

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{v} = 0, \quad \frac{D\mathbf{v}}{Dt} = \nabla\cdot\sigma + \rho\mathbf{g}, \tag{3.1}$$

where σ is the stress imparted on a particle, $\mathbf{g}$ is the gravitational acceleration, and $\frac{D}{Dt}$ is the material derivative with respect to time. The equations are discretized spatially and temporally with a collection of Lagrangian particles (or material points) and a background Eulerian grid. The material type of the simulated substances is naturally specified from the constitutive model, which defines how a material exerts internal stress (or forces) as a result of deformation.

Rigid balls are simulated as highly stiff elastic objects with the neo-Hookean hyperelasticity model, described through the elastic energy density function

$$\Psi(\mathbf{F}) = \frac{\mu}{2}(tr(\mathbf{F}^T\mathbf{F}) - d) - \mu\log(J) + \frac{\lambda}{2}\log^2(J), \tag{3.2}$$

where d is the dimension (2 or 3), $\mathbf{F}$ is the deformation gradient (*i.e.*, the gradient of the deformation from undeformed space to deformed space), J is the determinant of $\mathbf{F}$, and μ and λ are Lamé parameters that describe the material's stiffness.

Liquid is modeled as a nearly incompressible fluid, with its state governed by the Tait equation [247]:

$$p = k\left[\left(\frac{\rho_0}{\rho}\right)^{\gamma} - 1\right], \tag{3.3}$$

where p is the pressure, ρ and ρ_0 are the current and original densities of the particles, $\gamma = 7$ for water, and k is the bulk modulus (*i.e.*, how incompressible the fluid is). Through this Equation-of-State (EOS), the stress inside a non-viscous fluid is given by $\sigma = -p\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We further adopt the Affine Particle-In-Cell method (APIC) [236] to greatly reduce numerical error and artificial damping. This enables us to simulate fluids with better accuracy compared to alternative computer graphics methods.

The motion of dry sand is largely determined by the frictional contact between grains. In the theory of elastoplasticity, the modeling of large deformation (*e.g.*, frictional contact) can be based on a constitutive

law that follows the Mohr-Coulomb friction theory. Following [237], we simulate dry sand based on the Saint Venant Kirchhoff (StVK) elasticity model combined with a Drucker-Prager non-associated flow rule. Plasticity models the material response as a constraint projection problem, where the feasible region (or yield surface) of the final material stress is restricted to be inside

$$tr(\sigma)c_F + \left\| \sigma - \frac{tr(\sigma)}{d} \right\|_F \leqslant 0, \tag{3.4}$$

where d is the dimension and c_F is the coefficient of internal friction between sand grains. The stress (and thus deformation gradient) of each sand particle is projected onto the yield surface so as to satisfy the second law of thermodynamics.

3.2.7 Introduction

Imagine that you are preparing to pour pancake batter onto a griddle. To pour the correct amount, you must decide where to hold the container, at what angle, and for how long. We encounter similar situations frequently in our daily lives when interacting with viscous fluids ranging from water to honey, and with different volumes, contained in receptacles of various shapes and sizes.

In the majority of these situations, we are able to reason about fluid-related physical processes so as to implicitly predict how far a filled container can be tilted before the fluid inside begins to spill over its rim. However, people perform significantly worse when asked to make explicit reasoning judgments in similar situations [248, 249]. In the well-known Piagetian water level task (WLT; [250]), participants receive instructions to draw the water level at indicated locations on the inside of tilted containers. Surprisingly, about 40% of adults predict water levels that deviate from the horizontal by 5 degrees or more (*e.g.*, [248]). [249] modified the WLT to include two containers, one wider than the other. The investigators asked participants to judge which container would need to be tilted farther before the water inside begins to pour out. Only 34% of the participants correctly reported that the thinner container would need to be tilted farther than the wider one. However, when instructed to complete the task by closing their eyes and imagining the same situation, nearly all (95% of) the participants rotated a thinner, imaginary container (or a real, empty one) farther. These findings suggest that people are able to reason successfully about relative pour angles by mentally simulating the tilting event. An apparent contrast in human performance between an explicit reasoning task and a simulated-doing task has also been found in people's inferences about the trajectories of falling objects [251, 171]. Thus, empirical findings in the literature of physical reasoning suggest that people employ both explicit knowledge about physical rules *and* and mental simulation when making inferences [244].

The *noisy Newton* framework for physical reasoning hypothesizes that inferences about dynamical systems can be generated by combining noisy perceptual inputs with the principles of classical (*i.e.*, Newtonian) mechanics, given prior beliefs about represented variables [187, 90, 230, 232, 233, 246]. In this framework, the locations, motions and physical attributes of objects are sampled from noisy distributions and propagated forward in time using an *intuitive physics engine*. The resulting predictions are queried and averaged across simulations to determine the probability of the associated human judgment. [187] extended the framework from physical scene understanding (*e.g.*, [90]) to fluid dynamics using an *intuitive fluid engine* (IFE), where future fluid states are approximated by probabilistic simulation via a Smoothed Particle Hydrodynamics (SPH) method [234]. The particle-based IFE model matched human judgments about final fluid states and provided a better quantitative fit than alternative models that did not employ simulation or account for physical uncertainty.

The present study aims to determine whether a particle-based IFE model coupled with noisy input variables can account for human judgments about the relative pour angle of two containers

filled with fluids differing in their volume and viscosity. The experiment reported here was inspired by previous empirical findings in water-pouring tasks; *e.g.*, participants tilt containers filled with imagined molasses farther than those with an equal volume of water, suggesting that people are able to take physical attributes such as viscosity into account when making fluid-related judgments [249]. [187] also found that their participants' judgments were sensitive to latent attributes of the fluid (*e.g.*, stickiness and viscosity).

To quantify the extent that humans employ their perceived viscosity of fluids in subsequent reasoning tasks, we utilized a recent development in graphical fluid simulation [252, 236] to simulate the dynamic behavior of fluids in vivid animations. Previous work has shown that realistic animations can facilitate representation of *dynamic* physical situations [253]. Furthermore, recent research on human visual recognition indicates that latent attributes of fluids (*e.g.*, viscosity) are primarily perceived from visual motion cues [254], so displaying realistic fluid movement is needed to provide the input of key physical properties that enable mental fluid simulations. The present study, which uses a modification of [249]'s [249] water-pouring problem coupled with advanced techniques in computer graphics, aims to test the hypothesis that animated demonstrations of flow behavior facilitate inference of latent fluid attributes, which inform mental simulations and enhance performance in subsequent reasoning tasks.

3.2.8 Experiment

Participants A total of 152 participants were recruited from the Department of Psychology subject pool at the University of California, Los Angeles, and were compensated with course credit.

Materials and Procedure Prior to the reasoning task, participants viewed animated demonstrations of the movement of a moderately viscous fluid in two situations. The fluid used in the demonstrations was colored orange and was not observed in the judgment task. In the first demonstration, the fluid pours over two torus-shaped obstructions in a video looped three times and lasting for 11.5 seconds. The flow demonstration videos were presented to provide visual motion cues to inform participants' perceived viscosity. Following the flow demonstration, participants viewed a video of a cylindrical container filled with the same orange fluid tilting at a constant angular rate ($\omega = 22$ deg $\cdot s^{-1}$; see Fig. 3.9) from the upright orientation of the container and moving towards the horizontal. The video was looped three times for a duration of 14.7 seconds.

Following the demonstration videos, two new fluids were introduced, one with low viscosity (LV; similar to water) and one with high viscosity (HV; similar to molasses). The LV and HV fluids were colored either red or green (counterbalanced across subjects). As shown in the top panel of Fig. 3.9, participants viewed a flow video of both the HV and LV fluids (looped three times) for a duration of 11.5 seconds before each judgment trial. The two flow videos were presented side by side for comparison, and the relative position of each fluid was counterbalanced across subjects. The LV and HV fluids were selected to readily distinguish the two fluids based on their perceived viscosities, which were inferred from visual motion cues in the flow videos [254].

In the subsequent reasoning task, participants viewed a static image of two containers side by side filled with the LV and HV fluids (see bottom panel of Fig. 3.9). Participants were instructed to assume that each container was tilted simultaneously in the same way as observed earlier for the orange fluid in the tilting demonstration. They were informed that both containers were tilted at the same rate, and were provided with the quantity of fluid in each container. Participants were then asked to report "which container will need to be tilted with a larger angle before the fluid inside begins to pour out" and received no feedback following completion of each trial. The experiment manipulated the volume of the LV and HV fluids (V_{LV} and V_{HV}, respectively) in each

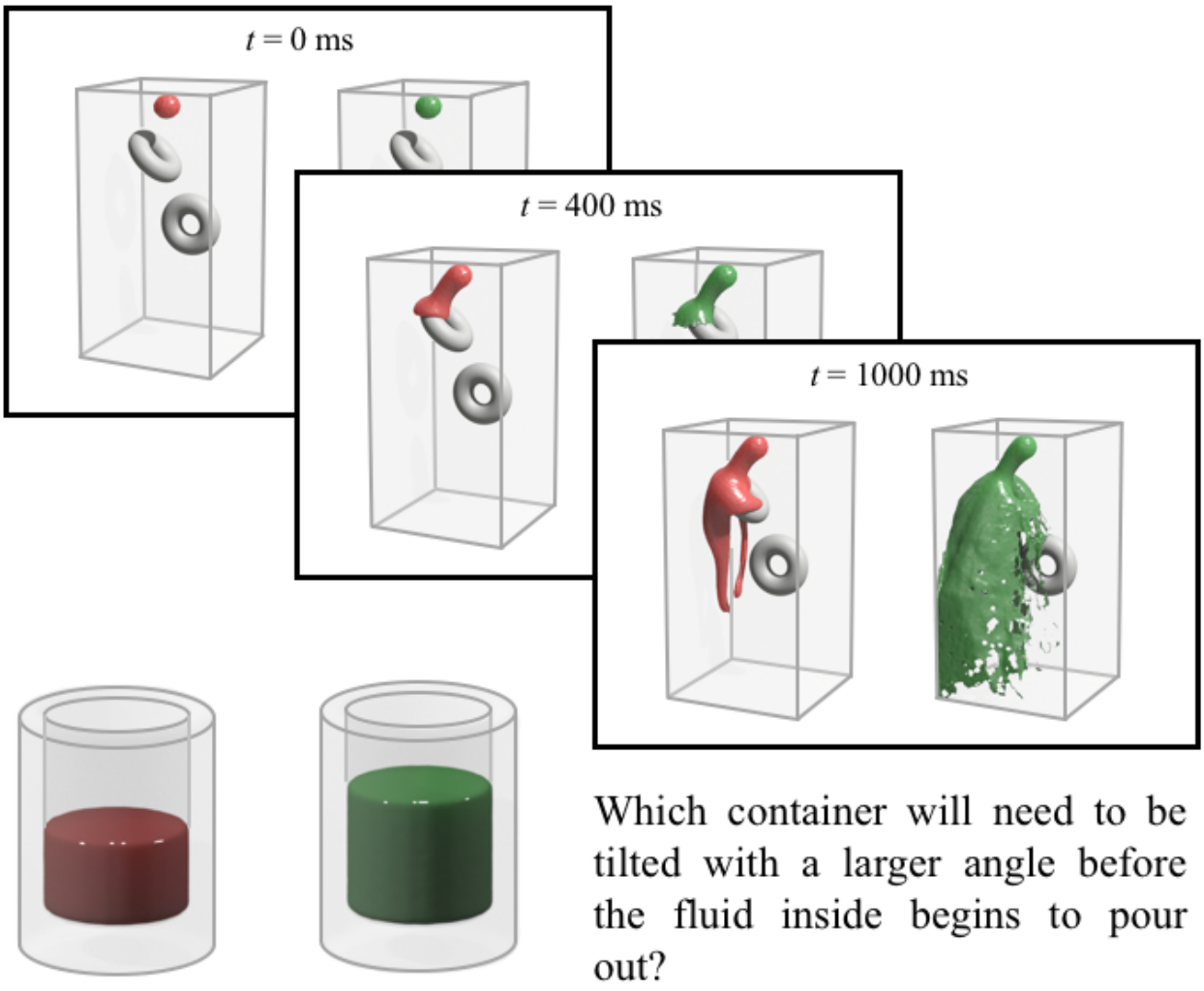


Figure 3.9: Illustration of flow demonstration video and judgment trial. (Top) Sample frames from the HV (red) and LV (green) flow video. (Bottom) Tilt judgment trial, where $V_{HV} = 40\%$ and $V_{LV} = 60\%$.

container across the values 20%, 40%, 60%, and 80%, representing the proportion of the container filled. Hence, the experiment consisted of 16 trials presented in a randomized order, including all possible volume pairs between the LV and HV fluids. The experiment lasted approximately 10 minutes.

Human Results

Fig. 3.10 depicts human performance for all 16 fluid volume combinations. To assess the relationship between HV fluid volume and human judgments, we performed a logistic regression analysis and compared estimated coefficients for each participant across LV fluid volume conditions. We found that coefficients for each participant varied significantly between V_{LV} conditions ($F(3, 149) = 113.89, p < .001$). In the highest LV fluid volume condition ($V_{LV} = 80\%$), V_{HV} had a minor impact on participants' responses ($\bar{\beta}_{80} = .04, \sigma_{\beta_{80}} = .25$) relative to the lowest LV fluid volume condition ($V_{HV} = 20\%$; $\bar{\beta}_{20} = .57, \sigma_{\beta_{20}} = .43$). These results demonstrate that participants' responses were increasingly sensitive to HV fluid volume for the greater V_{LV} conditions.

Next, we examined whether humans rely on heuristic-based reasoning to make their judgments.

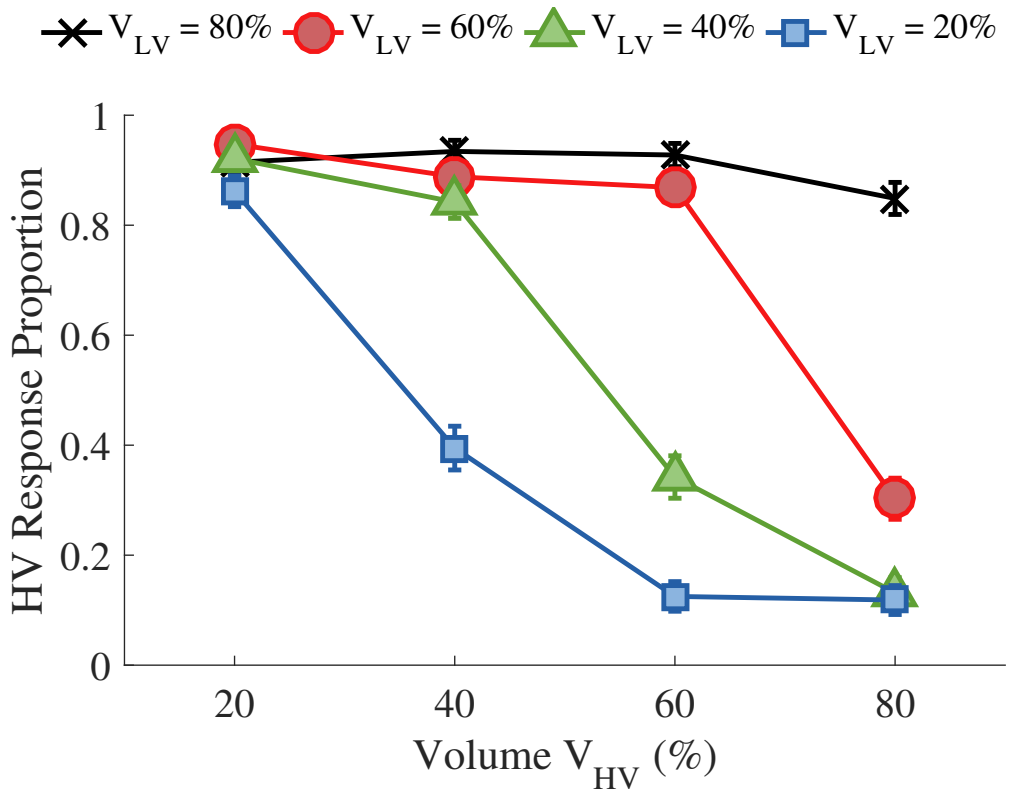


Figure 3.10: Human *HV* response proportions for all experimental conditions. The volume of the *HV* fluid (V_{HV}) is plotted on the horizontal axis, and separate lines indicate the four possible volumes for the *LV* fluid (V_{LV}).

One potential heuristic is that given two containers filled with different volumes of each fluid, the container with lesser fluid volume requires a greater rotation before beginning to pour. While participants consistently adhered to this rule for trials where $V_{HV} < V_{LV}$, their judgments for each of the $V_{LV} < V_{HV}$ trials did not accord to the same heuristic. For example, in trials where $V_{HV} =$ 40%, 60%, and 80% and $V_{LV} = 20\%$, 40%, and 60%, respectively, the lesser-volume heuristic predicts *LV* fluid responses. However, *HV* response proportions for those trials were significantly greater thane zero ($t(151) = 9.92, 8.86, 8.10$, $p < .001$). A second potential heuristic is to always choose the *HV* fluid as requiring a greater rotation since it moves slower than the *LV* fluid. The above three cases also disagreed with this heuristic since the rule would predict unity response proportions. In summary, response proportions in the specified trials reveals that participants attended to latent fluid attributes (*e.g.*, viscosity) *and* volume difference when making their tilt angle judgments (see Fig. 3.11).

3.2.9 Models

Fluid Simulation with Physical Dynamics

The simulation of incompressible flows through numerical evaluation of physical equations has become one of the most significant topics in computer graphics and mechanical engineering. The velocity field of simulated fluids is determined according to the constraints specified in the Navier-

Stokes equations:

$$\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} + \frac{1}{\rho} \nabla p = \boldsymbol{g} + \mu \nabla \cdot \nabla \boldsymbol{v}, \tag{3.5}$$

$$\nabla \cdot \boldsymbol{v} = 0, \tag{3.6}$$

where $\boldsymbol{v}$ is the velocity, ρ is the density, p is the pressure, $\boldsymbol{g}$ is the gravitational acceleration (approximately 9.8 m s^{-2}), and μ is the viscosity. Eq. (3.5) is called the momentum equation—it is simply Newton's second law (*i.e.*, $F = ma$) applied to fluid dynamics. Eq. (3.6) is the incompressibility constraint on fluid velocity, where the null divergence of the velocity field corresponds to constant density within volumetric regions. Most liquids need to satisfy this constraint in order to maintain constant volume while moving.

To numerically solve these partial differential equations, we adopt the Fluid Implicit Particle/Affine Particle in Cell (FLIP/APIC) method [255, 236, 256], which has become standard in physics-based simulation calculations due to its accuracy, stability and efficiency. Unlike Smoothed Particle Hydrodynamics (SPH), which purely relies on particles to discretize the computational domain, FLIP/APIC uses both particles and a background Eulerian grid. The Navier-Stokes equations are solved on the grid, allowing for accurate derivative calculations, well-defined free surface and solid boundary conditions, and accurate first-order approximation of physical reality. The FLIP/APIC method also circumvents common artifacts of SPH; *e.g.*, underestimated density near free surfaces and weakly compressible artifacts. In fact, the requirement for incompressibility is crucial in the fluid-pouring problem studied in this paper. We choose not to use SPH because it does not guarantee a divergence-free velocity field unless additional computational components are included. FLIP/APIC, however, maintains the benefits of particle-based methods due to its hybrid particle/grid nature. The presence of particles in the current model serves to facilitate visualization and the tracking of material properties. Besides modeling fluid, the state-of-the-art physics-based simulation methods have provided realistic cues for modeling complex tool and tool-uses [222], generic containers [257] and soft human body [223].

Realistic visualization of simulated fluids is particularly important for the flow demonstration animations displayed before each trial in our viscous fluid-pouring task. To provide vivid impressions of the motion of the *LV* and *HV* fluids, we adopt OpenVDB [258] and utilize the latest particle fluid surfacing methods developed in the field of computer graphics to reconstruct a smooth level set surface from the simulated fluid particles based on their locations.

The favorable efficiency and precision of the FLIP/APIC method allows for effortless generation of ground truth responses for our task, given that fluid viscosity and volume are known. The particle locations and vessel tilt angles are explicitly recorded at each time step as simulation outputs. Since the FLIP/APIC method does not involve any stochastic processes, the output of each simulation is deterministic. In each simulation, $40,000$ particles (with around 8 particles per grid cell) were used to ensure both stability and convergence.

Intuitive Fluid Engine

Fluid simulation with physical dynamics provides deterministic fluid movements if the ground-truth values of viscosity and volume are known. Hence, the decisions directly derived from the FLIP/APIC fluid simulator are binary judgments (Fig. 3.11), which implies that the physical simulation with high precision cannot explain humans' probabilistic judgment in the fluid-pouring task. Inspired by the approach of [187] and the noisy Newton model (*e.g.*, [233]), we combine the physical simulator of FLIP/APIC with noisy input variables (*i.e.*, viscosity and volume) to form the Intuitive Fluid

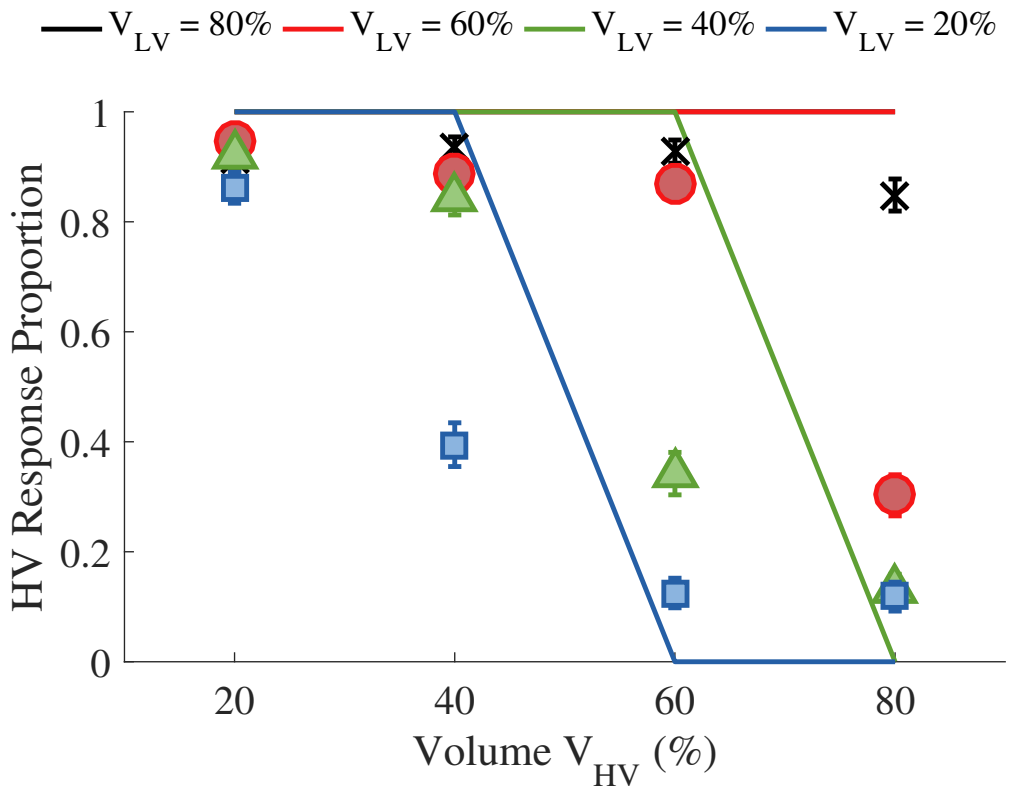


Figure 3.11: Simulation results from the FLIP/APIC model ($RMSD = 0.6747$). Separate lines indicate model predictions, given ground-truth volume/viscosity values for each fluid. Symbols indicate human response proportions. The deterministic simulation method returns binary predictions and does not provide probabilistic response proportion estimates.

Engine (IFE) model, thereby accounting for physical uncertainty and the influence of viscosity and volume on our reasoning task.

The input variables for our IFE are the ground-truth values of volume and viscosity (V_T, μ_T) for the two fluids in each experimental trial. The sampling process of the IFE then samples $N = 10,000$ noisy viscosity and volume pair values $\{(V_i, \mu_i), i = 1, 2, \cdots, N\}$ for each fluid. Each sample (V_i, μ_i) of the 10000 generated noisy input variables is later passed to the FLIP/APIC simulator to produce a binary decision $B_i(V_i, \mu_i) \in \{L, R\}$. The decision is that either the left L or the right R container needs to be tilted with a larger angle before the fluid inside begins to pour out. By aggregating the prediction from all $10,000$ samples and dividing the sum by the number of samples, the IFE outputs the distribution $\boldsymbol{P}(V_T, \mu_T)$ for the given ground-truth values of viscosity and volume:

$$\boldsymbol{P}(V_T, \mu_T) = \begin{cases} P(V_T, \mu_T)_L = \dfrac{\sum_i H(B_i(V_i, \mu_i), L)}{N} \\ \\ P(V_T, \mu_T)_R = \dfrac{\sum_i H(B_i(V_i, \mu_i), R)}{N}, \end{cases} \tag{3.7}$$

where $H(\Psi, \Theta) = 1$ when $\Psi = \Theta$, and it is 0 otherwise.

To model physical uncertainty, the sampling process of the IFE model is implemented by adding perceptual noises to the ground-truth values of the physical input variables (*i.e.*, viscosity and volume). Noisy volume is generated by adding an offset to its ground-truth value from a Gaussian distribution with mean 0 and variance σ_μ, whereas the noisy viscosity is generated by adding a

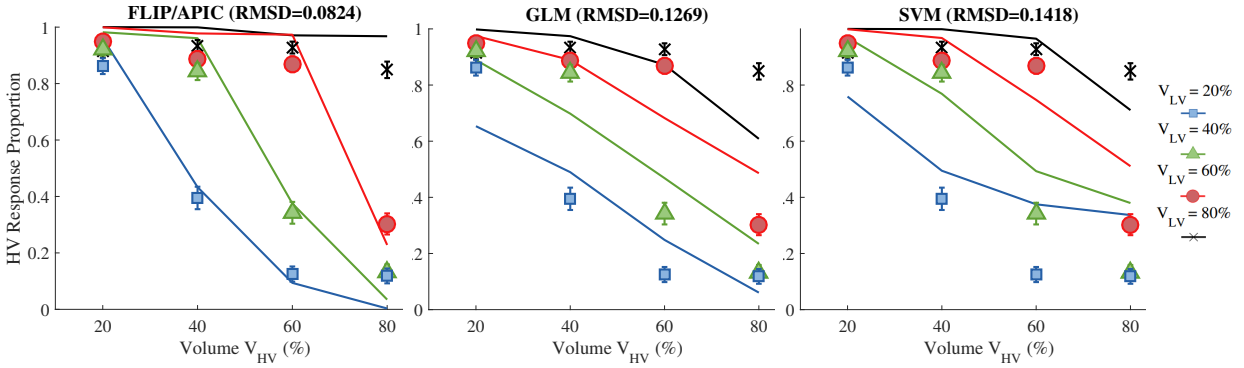


Figure 3.12: Comparison of results between our three prediction models: (Left) IFE, (Middle) Regression, and (Right) SVM with perceptual noise. Horizontal axes denote HV fluid volume; vertical axes denote the predicted proportion of HV fluid responses associated with a greater rotation angle. The IFE simulation model outperforms competing data-driven models.

fixed amount of Gaussian noise on a logarithmic scale [233]: $V_i = f^{-1}(f(V_T) + \epsilon)$, where V_T is the ground-truth value, $f(V_T) = \log(\omega \cdot V_T + k)$, f^{-1} is the inverse of f, and $\epsilon \sim \text{Gaussian}(0, \sigma_V)$. The results reported herein chooses $\sigma_\mu = 0.1$, $\sigma_V = 0.1$, $k = 1.5$, $\omega = 1$.

Each simulation required approximately 10 minutes to run on a modern single-core CPU. In order to run a large number of simulations during the sampling process, we discretized the viscosity and volume spaces into finite sets. Specifically, the viscosities are discretized into 8 different cases $(0.1, 1, 10, 100, 200, 500, 1000, 2000)$ and the volumes into 21 different cases (0% to 100% with a step-size 5%). We pre-computed the simulation results for each discretized case, and stored the results in a database. During the sampling process, the sampled inputs are numerically rounded to the pre-computed discretized cases, where the results can be immediately retried from the database without re-computing.

Non-Simulation Models

To examine whether fluid simulation is necessary to account for how humans reason about fluid behavior, we compare the simulation model with two statistical learning methods—the generalized linear model (GLM) [239] and the support vector machine (SVM) [259]. These models are purely data-driven and do not involve any explicit knowledge of physical laws or physical simulation. The selected features for these models include (i) the volumes of fluids in both containers, and (ii) the viscosity ratio between the LV and HV fluids.

To predict human judgment for the $i^\star$th trial $J_{i^\star}$, both non-simulation models were tested with the $i^\star$th trial, and trained with the remaining 15 trials $\{J_i, i = 1, 2, \cdots, 16, i \neq i^\star\}$. The trained GLM model is directly applied to the test case to predict which container will need to be tilted to a larger angle before the fluid inside begins to pour out. Since the SVM is a discriminative classification method which can only predict discretized labels (*i.e.*, $+1$ indicating selection of the left container and -1 indicating selection of the right container), we introduced perceptual noise (the same method for the IFE) to each test trial to model physical uncertainty. For each test trial, a set with $10,000$ samples was generated. The trained SVM model is then applied to predict the labels ($+1$ or -1) in each sample, which are then aggregated to form the probability distribution for each test trial.

Model Results

We first compared how well different computational models account for human performance for the 16 trials. Fig. 3.12 depicts results from the IFE, GLM, and SVM models with perceptual noise. Human judgments and model predictions were highly correlated ($r(14) = 0.9954, 0.9488$, and 0.9251, respectively). RMSD (root-mean-squared deviation) between human judgments and the models' predictions are $0.0824, 0.1269$, and 0.1418, respectively. Compared to the purely data-driven models (*i.e.*, the GLM and SVM models), the simulation-based IFE model encodes material properties (*e.g.*, viscosity) and perceptual features (*e.g.*, volume) and provides better approximations to human judgments in the viscous fluid-pouring task. These results again support the role of simulation as a potential mental model that supports human inference in physical reasoning tasks.

Next, we examined whether the IFE model captures human performance on the three trials where the heuristic rule outlined earlier provides incorrect predictions; *i.e.*, those trials where V_{HV} = 40%, 60%, and 80% and V_{LV} = 20%, 40%, and 60%, respectively. Here, the ground-truth model predicts that the HV fluid will require a greater angle of rotation before beginning to pour, while the heuristic rule discussed earlier suggests the opposite. HV response proportions for these trials were .39, .34, and .30, respectively, and the IFE model returned consistent predictions of .43, .37, and .23. Alternatively, the GLM model predicted response proportions of .49, .47, and .49, and the SVM model predicted response proportions of .49, .49, and .51. Thus, our IFE model captured human performance on the specified trials while competing data-driven methods returned predictions biased toward the ground-truth model and away from the lesser-volume heuristic.

3.2.10 Discussion

Our results from the viscous fluid-pouring task agree with the findings of [187] in that our probabilistic, simulation-based IFE model outperformed two non-simulation models (SVM and GLM). Our behavioral experiment also indicates that people naturally attend to latent attributes (*e.g.*, viscosity) when reasoning about fluid states following observation of realistic flow demonstration animations. By extending our probabilistic IFE method to a reasoning task, we demonstrated that the noisy Newton framework can account for human performance in a fluid-related judgment task that traditionally precludes mental simulation strategies.

While simulation has been demonstrated as the default strategy in other mechanical reasoning tasks [244, 243], the participants in [249] experiments failed to spontaneously represent and simulate physical properties relevant to the water-pouring problem when making their judgments. It is important to note, however, that the present task differs from the traditional water-pouring task in several ways: (i) fluid viscosity and volume (rather than container diameter) varied across trials, (ii) a cup-tilting demonstration was displayed to visualize the rate of simulated rotation, and (iii) motion cues from flow demonstrations informed the perception of latent fluid attributes (*e.g.*, viscosity). Comparison of our study to previous water-pouring studies suggests that the dissociation between explicit physical prediction and implicit judgment reported in the intuitive physics literature could be resolved in some situations by modifying task characteristics and instructions in ways that motivate simulated representation. While our viscous water-pouring problem indicates a set of simulation-inducing task characteristics, further research should aim to determine specific experimental factors that trigger simulation strategies. Specifically, can the conditions employed in the present task extend to classical rigid-body and fluid mechanics problems to resolve the discrepancy between people's explicit predictions and tacit judgments, and if so, what additional task characteristics serve to facilitate mental simulation?

Classical research in artificial intelligence has traditionally dismissed robust mental simulation

as a strategy for physical reasoning due to its inherent complexity, often proposing simplified qualitative models instead [260]. While the computational fluid simulations employed in the present study require extensive numerical evaluation to make predictions about future fluid states, humans appear to do so with precision and accuracy in comparatively small amounts of time. Furthermore, their performance in our reasoning task suggests representation of physical quantities that extends beyond qualitative process theory. While human results are generally consistent with physics-based simulation models coupled with noisy input variables, there remain discrepancies between model predictions and human judgments. Hence, future research should aim to address whether humans *simulate* fluid movements using mental models that accord to physical laws or *emulate* fluid dynamics by drawing on their everyday interactions with liquids across diverse physical situations [261].

3.2.11 Acknowledgments

Support for the present study was provided by an NSF Graduate Research Fellowship, DoD CASIT grant W81XWH-15-1-0147, DARPA SIMPLEX grant N66001-15-C-4035 and ONR MURI grant N00014-16-1-2007.

3.3 Case Study: Physical Stability as Grouping Principle

This work presents a new perspective for 3D scene understanding by reasoning object stability and safety using intuitive mechanics. Our approach utilizes a simple observation that, by human design, objects in static scenes should be stable in the gravity field and be safe with respect to various physical disturbances such as human activities. This assumption is applicable to all scene categories and poses useful constraints for the plausible interpretations (parses) in scene understanding. Given a 3D point cloud captured for a static scene by depth cameras, our method consists of three steps: i) recovering solid 3D volumetric primitives from voxels; ii) reasoning stability by grouping the unstable primitives to physically stable objects by optimizing the stability and the scene prior; and iii) reasoning safety by evaluating the physical risks for objects under physical disturbances, such as human activity, wind or earthquakes.

Zhao and Zhu adopt a novel intuitive physics model and represent the energy landscape of each primitive and object in the scene by a disconnectivity graph (DG). Zhao and Zhu construct a contact graph with nodes being 3D volumetric primitives and edges representing the supporting relations. Then Zhao and Zhu adopt a Swendson-Wang Cuts Algorithm to group/partition the contact graph into groups. Each group is a stable object. In order to detect unsafe objects in a static scene, our method infers hidden and situated causes (disturbances) of the scene, and then introduces intuitive physical mechanics to predict possible effects (*e.g.*, falls) as consequences of the disturbances.

In experiments, Zhao and Zhu demonstrate that the algorithm achieves substantially better performance for i) object segmentation, ii) 3D volumetric recovery, and iii) scene understanding in comparison to state-of-the-art methods. Zhao and Zhu also compare the safety prediction from the intuitive mechanics model with human ratings.

3.3.1 Introduction

Intuitive Physics

Interacting with the world requires a commonsense understanding of how it operates at a physical level, which does not necessarily require us to precisely or explicitly invoke Newton's laws of

mechanics; instead, we rely on intuition, built up through active interactions with the surrounding environment. Humans excel at understanding their physical environment and interacting with objects undergoing dynamic state changes, making approximate predictions from observed events. The knowledge underlying such activities is termed *intuitive physics* [167]. The field of intuitive physics has been explored for several decades in cognitive science and recently reinvigorated by new techniques linked to AI.

Early research in intuitive physics provides several examples of situations where humans demonstrate common misconceptions about how objects in the environment behave. For example, several studies found that humans exhibit striking deviations from Newtonian physical principles when asked to explicitly reason about the expected continuation of a dynamic event based on a static image representing the situation at a single time point [168, 167, 169]. However, humans' intuitive understanding of physics is much more accurate, rich, and sophisticated than previously expected if *dynamics* and proper *context* were provided [170, 171, 172, 173, 174].

Surprisingly, humans develop physical intuitions at an early age [100, 262], well before most other types of high-level reasoning. At the age of two months, human infants expect inanimate objects to follow principles of solidity, cohension, continuity, and persistence, while at round six month old, infants will have developed different expectations for diffirent bodies, *e.g.*, rigid body and liquids [263, 264]. While a widely accepted computational account of these observations is lacking, Lake *et al.* [262] suggests the recent approach of physics software engine, or Intuitive Physics Engine (IPE), could be a promising approach towards this problem [90]. In their hypothesis, people also construct a scene using internal properties of objects, such as mass, gravity, stiffness, and friction, and simulate its dynamics after external perturbations.

To verify the hypothesis, Battaglia *et al.* built a probabilistic model of IPE based on a block world and ask people to judge whether a shown block tower configuration is stable enough. Though the proposed model makes probabilistic simulations and predictions, the results are strongly correlated with those made by human subjects regarding both stability prediction and stability rating, regardless of whether the guesses reflect Newtonian physics or not.

The intuitive physics engine approach is approximate and probabilistic, sometimes oversimplified and incomplete. However, it also exhibits important features, such as flexibility and generality, compared to current pattern recognition approach, *e.g.*, deep learning. The model could be generalized to a wide range of daily lives without huge amounts of training data, in contrast to a deep learning model [265] which fails when a few more layers of blocks are added.

However, human perception might not always cohere with the physics law. A typical example is the refraction of water that easily tricks people into a mis-measurement of its depth. Another factor that affects one's perception is his physiological state. In Witt and Proffitt's experiments [266], it's noticed that people tend to overestimate the incline of mountains by using merely visual clues, especially when they are in poor physical conditions. Height measurement is also deeply affected depending on whether one is afraid of height [267]. Unfortunately, a computational model that explains why human makes these wrong perceptions is lacking as well.

Motivation and Objectives

Traditional approaches, *e.g.*, [268, 269], for scene understanding have been mostly focused on segmentation and object recognition from 2D/3D images. Such representations lack important physical information, such as the stability of the objects, potential physical safety, and supporting relations which are critical for scene understanding, situation awareness and especially robot vision. The following scenarios illustrate the importance of this information.

- *Stability and safety understanding.* Our approach utilizes a simple observation that, by human

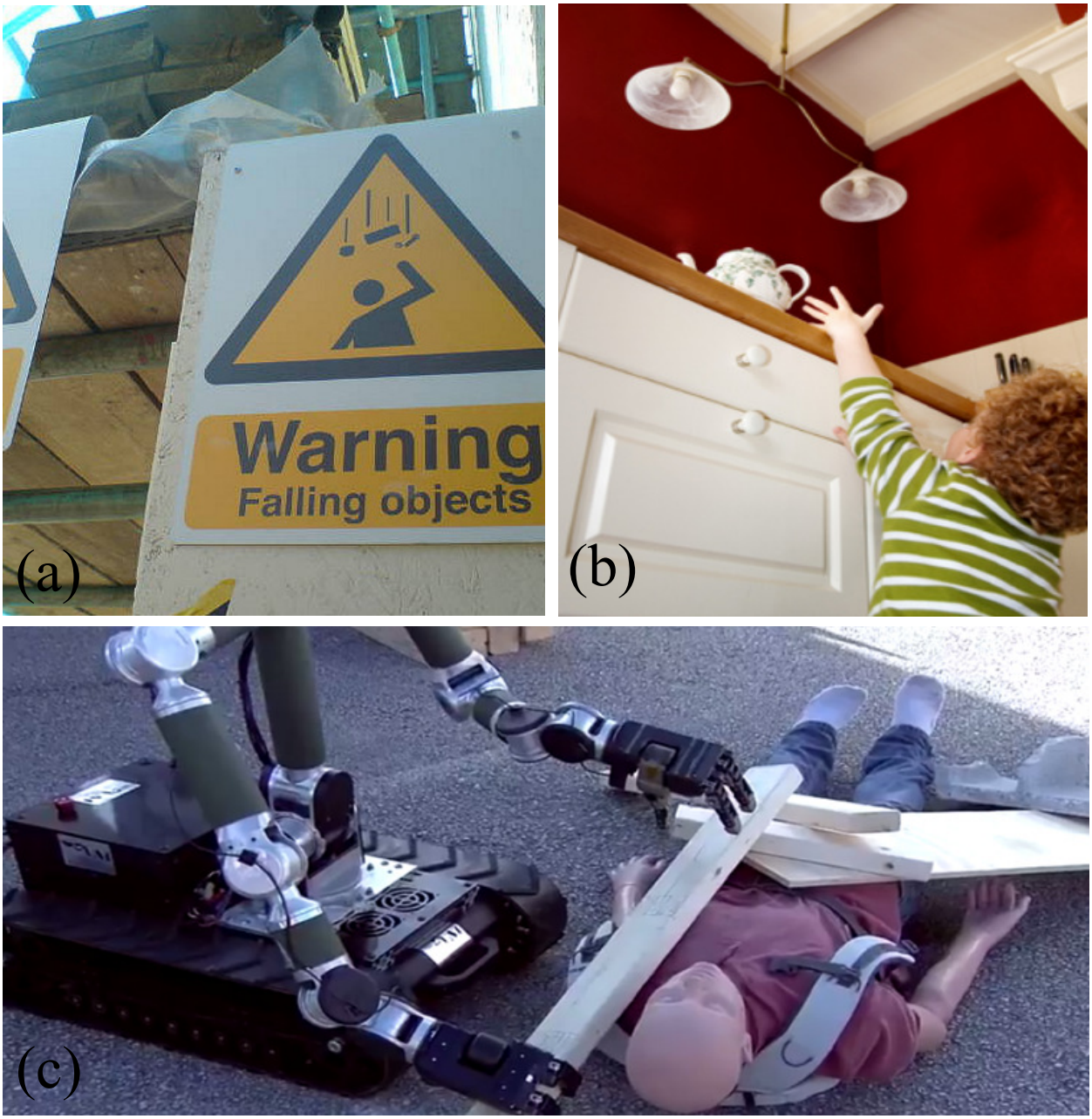


Figure 3.13: A safety-aware robot can be used to detect potentially physically unstable objects in a variety of situations: (a) falling objects at a constructions site, (b) the human assistant for baby proofing, and (c) the disaster rescue (from the recent DARPA Robotics Challenge), where the Multi-Arm robot needs to understand the physical relationships between obstacles.

design, objects in static scenes should be stable in the gravity field and be safe respect to various physical disturbances such as human activities. This assumption poses useful constraints for the plausible interpretations (parses) in scene understanding.

- *Human assistant robots.* Objects have the potential to fall onto or hit people at workplaces, as the warning sign shows in Fig. 3.13 (a). To prevent objects from falling freely from one level to another, safety surveillance ensures that objects be stored in safe places, especially for children, elders and people with disabilities. As the example shows in Fig. 3.13 (b), Zhao and Zhu can

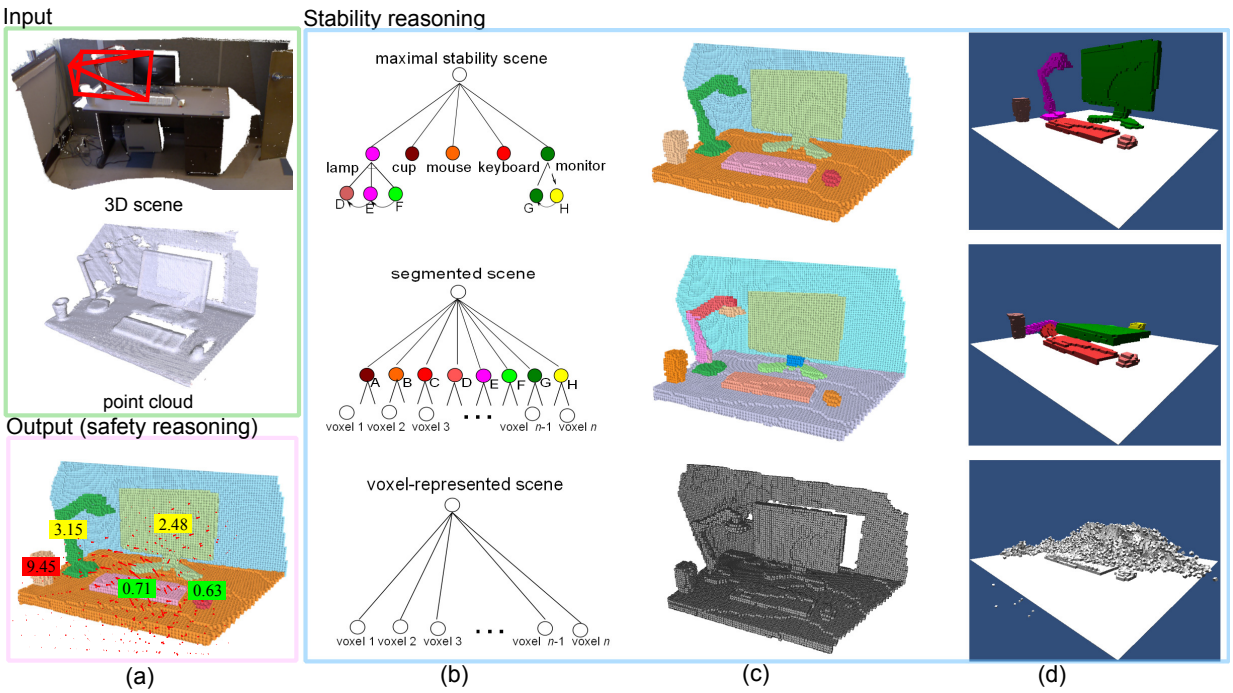


Figure 3.14: Overview of our method. (a) Input: 3D scene reconstructed by SLAM technique and Output: parsed 3D scene as stable objects with supporting relations. The number are unsafety scores for each object under the disturbance field (in red arrows), (b) scene parsing graphs corresponding to 3 bottom-up processes: voxel based representation (bottom), geometric preprocess including segmentation and volumetric completion (middle), and stability optimization (top). (c) result at each step. (d) physical simulation result of each step.

predict a possible action of the child—he is reaching for something - and then infer possible consequences of his action—he might be struck by the falling teapot.

- *Disaster rescue robots.* Fig. 3.13 (c) shows a demonstration of a HDR-IAI Multi-Arm robot rescuing people during a mock disaster in the DARPA robot challenge [270]. Before planning how to rescue people, the robot needs to understand the physical information, such as which wood block is unsafe or unstable, and the support relations between them.

In this work, Zhao and Zhu present an approach for reasoning physical stability and safety of 3D volumetric objects reconstructed from either a depth image captured by a range camera, or a large scale point cloud scene reconstructed by the SLAM technique [271]. Zhao and Zhu utilize a simple observation that, by human design, objects in static scenes should be "stable" but might not be "safe" with respect to gravity and various physical disturbances caused by wind, a mild earthquake or human activities. For example, a parse graph is said to be valid if the objects, according to its interpretation, do not fall under gravity. If an object is not stable on its own, it must be grouped with neighbors or fixed to its supporting base. In addition, while objects are stable physically, they might be potentially unsafe if the places where they stay are prone to collisions with human bodies during common activities. These assumptions are applicable to all scene categories and thus pose powerful constraints for the plausible interpretations (parses) in scene understanding.

Overview

As Fig. 3.14 shows, given the input point cloud, our method consists of two main steps: stability reasoning and safety reasoning.

- *Stability reasoning*: hierarchically pursuing a physically stable scene understanding in two substeps:
  - *Geometric preprocessing*: recovering solid 3D volumetric primitives from a defective point cloud. Firstly Zhao and Zhu segment and fit the input $2\frac{1}{2}$D depth map or point cloud to small simple (*e.g.*, planar) surfaces; secondly, Zhao and Zhu merge convexly connected segments into shape primitives; and thirdly, Zhao and Zhu construct 3D volumetric shape primitives by filling the missing (occluded) voxels, so that each shape primitive has physical properties: volume, mass and supporting areas to allow the computation of the potential energies in the scene.
  - *Reasoning maximum stability*: grouping the primitives to physically stable objects by optimizing the stability and the scene prior. Zhao and Zhu build a contact graph for the neighborhood relations of the primitives. For example, as shown in Fig. 3.14 (c) in the second row, the lamp on the desk originally was divided into 3 primitives and would fall under gravity (see result simulated using a physical simulation engine in Fig. 3.14 (d)), but becomes stable when they are group into one object—the lamp. So is the computer screen grouped with its base.
- *Safety reasoning*—Given a static scene consisting of stable objects, our method first infers hidden and situated causes (disturbance field, red arrows in Fig. 3.14 (a)) of the scene, and then introduces intuitive physical mechanics to predict the unsafety scores (*e.g.*, falls) as the consequences of the causes. As shown in Fig. 3.14 (a) Output), since the cup is unsafe (falls off the table) under the act of the disturbance field, it gets a high unsafety score and a red label.

Our method adopts a novel intuitive physics model based on an energy landscape representation using disconnectivity graph (DG). Based on the energy landscape, it defines the physical stability function explicitly by studying the minimum energy (physical work) needed to change the pose and position of an object from one equilibrium to another, and thus release potential energy. For optimizing the scene stabilities, Zhao and Zhu propose to construct a contact graph and adopt the cluster sampling method, Swendsen-Wang Cut, introduced in image segmentation [272]. The algorithm groups/partitions the contact graph into groups, each being a stable object.

In order to detect unsafe objects in a static scene, our method first infers the "cause"—disturbance field, such as human activities or natural effects. To model the field of human disturbance, Zhao and Zhu collect the motion capture data of human actions, and apply it to the 3D scene (walkable areas) to estimate the statistical distribution of human disturbance. In order to generate a meaningful human action field, Zhao and Zhu first predict primary motions on the 2D ground plane which recodes the visiting frequency and walking direction for each walkable position, and add detailed secondary body part motions in 3D space. In addition, Zhao and Zhu explore two natural disturbances: wind and earthquakes. Zhao and Zhu then reason the "effects" (*e.g.*, falling) of each possible disturbance by our intuitive physics model. In this case, Zhao and Zhu calculate the minimum kinetic energy to move an entity from one stable point to a local maximum, *i.e.*, knocking it off equilibrium, and then Zhao and Zhu further evaluate the risk by calculating the energy released in reaching a deeper minimum. That is, the greater the energy it releases, the higher the risk is.

In experiments, Zhao and Zhu demonstrate that the algorithms achieve a substantially better performance for i) object segmentation, ii) 3D volumetric recovery of the scene, and iii) scene understanding in comparison to state-of-the-art methods in both public datasets [215]. Zhao and Zhu evaluate the accuracy of potentially unsafe object detection by ranking the falling risk w.r.t. human judgments.

Related Work

Our work is related to 6 research streams in the vision and robotics literature.

- *Geometric segmentation and grouping.* Our approach for geometric pre-processing is related to a set of segmentation methods, *e.g.*, [273, 274, 275]. Most of the existing methods are focused on classifying point clouds for object category recognition, not for 3D volumetric completion. For work in 3D geometric reasoning, [274] extracts 3D geometric primitives (planes or cylinders) from a 3D mesh. In comparison, our method is more faithful to the original geometric shape of object in the point cloud data. There has also been interesting work in constructing 3D scene layouts from 2D images for indoor scenes, such as [276, 213, 51, 277]. [278] also performed volumetric reasoning with the Manhattan-world assumption on the problem of multi-view stereo. In comparison, our volumetric reasoning is based on complex point cloud data and provides more accurate 3D physical properties, *e.g.*, masses, gravity potentials, contact area, *etc.*
- *Physical reasoning.* The vision communities have studied the physical properties based on a single image for the "block world" in the past three decades [208, 211, 121, 276, 213, 51]). *e.g.*, Biederman *et al.* [208] studied human sensitivity of objects that violate certain physical relations. Our goal of inferring physical relations is most closely related to [211] who infer volumetric shapes, occlusion, and support relations in outdoor scenes inspired by physical reasoning from a 2D image, and Silberman *et al.* [215, 130, 217] who infers the support relations between objects from a single depth image using supervised learning with many prior features. In contrast, our work is the first that defines explicitly the mathematical model for object stability. Without a supervised learning process, our method is able to infer the 3D objects with maximum stability.
- *Intuitive physics model.* The intuitive physics model is an important perspective for human-level complex scene understanding. However, to our best knowledge, there is little work that mathematically defines intuitive physics models for real scene understanding. [129] adopts an intuitive physics model in [167], however this model lacks deep consideration on complex physical relations. In our recent work [116, 117], Zhao and Zhu propose a novel intuitive physics model based on gravity potential energy transfer. In this work, Zhao and Zhu extend this intuitive physics model by combining specific physical disturbance fields. While Physics engines in graphics can accurately simulate the motion of objects under the influence of gravity, it is computationally too expensive for the purpose of measuring object stability.
- *Safe Motion Planning.* As motion planning is a classic problem in robotics, [279] tackled the problem of safe motion planning in the presence of moving obstacles. They consider the moving obstacles as a real-time constraint inherent to the dynamic environment. Zhao and Zhu first argue that a robot needs to be aware of potential dangers even in a static environment due to possible incoming disturbances.
- *Human in the loop.* This stream of research emphasizes a human-centric representation, differing from the classic feature-classifier paradigm of object recognition. Some recent work utilized the notion of "affordance." [128] recognized chairs by hallucinating a "sitting" actor interacting with the scene. [121] predicted the "workspace" of a human given an estimated 3D scene geometry. [280] and [281] demonstrated that observing people performing different actions can significantly improve estimates of scene geometry and scene semantics. [130] and [282] proposed scene labeling algorithms by considering humans as the hidden context.
- *Cognitive studies.* Recent psychology studies suggested that approximate Newtonian principles underlie human judgments about dynamics and stability [109, 91]. Hamrick *et al.* [91] showed that knowledge of Newtonian principles and probabilistic representations are generally applied for human physical reasoning. These intuitive models are studied for understanding human behaviors, not for vision robotics.

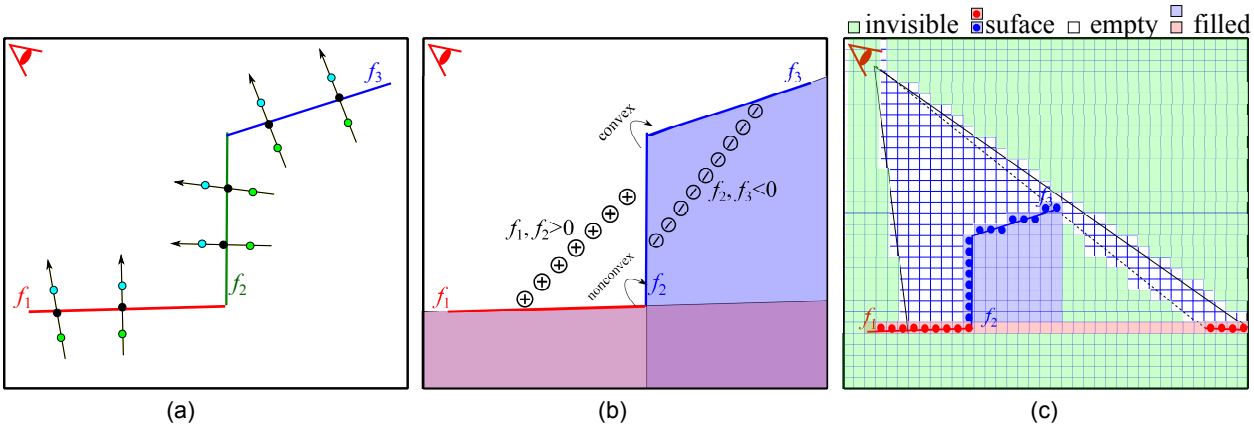


Figure 3.15: (a) Splitting. Two 1-degree IAMs f_1, f_2 and f_3 (in red, green and blue lines respectively) are fitted to the 3-Layer point cloud. Points in green and blue are the extra layer points generated from original points in black. (b) Merging. the segments fitted by f_2 and f_3 are merged together, because they are convexly connected. The convexity can be detected by drawing a line (in circular points) between any two connected segments and checking if their function values are negative. (c) Volumetric completion. Four types of voxels are estimated in volumetric space: invisible voxels (light green), empty voxels (white), surface voxels (red and blue dots), and the voxels filled in the invisible space (colored square in light red or blue).

Contributions

This work makes the following contributions.

- It defines the physical stability function explicitly by studying minimum forces and thus physical work needed to change the pose and position of an primitive (or object) from one equilibrium to another, and thus to release potential energy.
- It introduces a novel disconnectivity graph (DG) from physics [283] to represent the energy landscapes of objects.
- It solves the complex optimization problem by applying the cluster sampling method Swendsen-Wang cut used in image segmentation [272] to physical reasoning.
- It collects a new dataset for large scenes using depth sensors for scene understanding and the data and annotations will be released to the public.

Over the well-defined intuitive physics model in our previous work [116], Zhao and Zhu extend it to a safety model by introducing various disturbance fields.

The rest of this work is organized as: Section 2 presents our geometric preprocessing method that first forms solid object primitives from raw point clouds; then the method for reasoning the maximal stability for a static scene is described in Section 3; and reasoning the safety for each object in the scene is presented in Section 4 followed by experimental results and discussions in Sections 5 and 6 respectively.

3.3.2 Preprocessing: Computing Solid Volumes from Point Clouds

In order to infer the physical properties (*e.g.*, mass, gravity potential energy, supporting area) of objects from point clouds, Zhao and Zhu first compute a 3D volumetric representation for each object part. Zhao and Zhu proceed in two steps: 1) point cloud segmentation, and 2) volumetric completion.

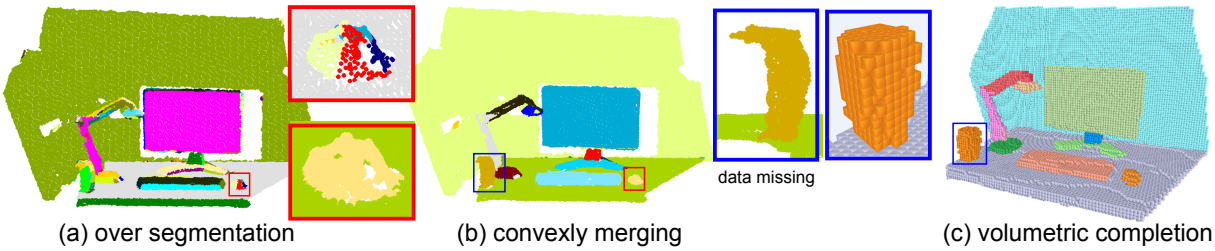


Figure 3.16: (a) Over-segmentation result obtained by splitting with IAMs. (b) Result after merging the convexly connected faces. (see the difference on "mouse" object). (c) Result after volumetric completion. (see the difference on "cup" object and hole on the back wall).

Segmentation with Implicit Algebraic Models

Zhao and Zhu adopt a segmentation method using implicit algebraic models (IAMs) [284] which fits IAMs to point clouds with simple geometry.

$$f_i(\mathbf{p}) \approx 0, \tag{3.8}$$

where $\mathbf{p} = \{x, y, z\}$ is a 3D point and f_i is defined by an n-degree polynomial:

$$f_i(\mathbf{p}) = \sum_{0 \leqslant i,j,k; i+j+k \leqslant n} a_{ijk} x^i y^j z^k, \tag{3.9}$$

where a_{ijk} are the unknown coefficients of the polynomial. The main advantage of IAM is that it is convenient for accessing the "inside" ($f_i < 0$) or "outside" ($f_i > 0$) of a surface fitted by an IAM.

Our method is in 2 steps as Fig. 3.15 (a) and (b) illustrated: 1) splitting step: over-segmenting the point cloud into simple regions approximated by IAMs, and then 2) merging step: merging them together with respect to their convexly connected relations.

Splitting Step The objective in this step can be considered to be finding the maximal 3D regions, each of them well fitted by an IAM. The IAM fitting for each region is formulated in least squares optimization using the 3-Layer method proposed by [284].

As shown in Fig. 3.15 (a), it first generates two extra point layers along the surface normals. Then, the IAM can be fitted to the point set constrained by 3 layers with linear least squared fitting.

Zhao and Zhu adopt a region growing scheme [275] in our segmentation. Thus our method can be described as: starting from several given seeds, the regions grow until there is no point that can be merged into the region fitted by an IAM. Zhao and Zhu adopt the IAM of 1 or 2 degree, *i.e.*, planes or second order algebraic surfaces and use the IAM fitting algorithm proposed by Zheng *et al.* [285] to select the models in a degree-increasing manner.

Merging Step The above segmentation method over-segments the objects into pieces. This is still a poor representation for objects, since only the segments viewed as faces of objects are obtained. According to a common observation that an object should be composed of several convex hulls (primitives) whose faces are convexly connected, Zhao and Zhu propose a merging step that merges the convexly connected segments together to approach the representation of object primitives.

To detect the convex connection, as shown in Fig. 3.15 (b), Zhao and Zhu first sample the points on a line which connects two adjacent regions (the circle lines in Fig. 3.15 (b)) as: $\{\mathbf{p}_l | \mathbf{p}_l \in L\}$, where

L denotes a line segment whose ends are on the two connected regions respectively. To detect the convexly connected relationship, Zhao and Zhu take a condition as the judgment:

$$\frac{\#\{\mathbf{p}|\mathbf{p}_l \in L \wedge f_i(\mathbf{p}_l) < 0 \wedge f_j(\mathbf{p}_l) < 0\}}{\#\{\mathbf{p}|\mathbf{p}_l \in L\}} > \delta_2, \tag{3.10}$$

where the ratio threshold δ_2 is set as 0.6. As illustrated in Fig. 3.15 (b), since the circular points drawn between f_2 and f_3 are negative, the segments should be merged. Fig. 3.16 (a) and (b) shows the difference before and after merging the convexly connected regions.

Volumetric Space Completion

The primitives output from the above method are still insufficient to reason the physical properties, *e.g.*, in Fig. 3.16 (b), the wall and table have hollow surfaces with holes and the cup has missing volume. To overcome this, Zhao and Zhu first generate a voxel-based representation for the point cloud such that each voxel can be viewed as a small mass unit with its own volume, gravity and contact region (contact faces of the cube). Secondly, Zhao and Zhu fill out the hidden voxels for each incomplete volumetric primitive obtained by the segmentation result above.

Voxel Generation and Gravity Direction Our voxel based representation is generated by constructing the octree of the point cloud as proposed by Sagawa *et al*. [286], after which the point cloud is regularized into the coordinate system under the Manhattan world assumption [278], supposing many visible surfaces orient along one of three orthogonal directions. To detect gravity direction, 1) Zhao and Zhu first calculate the distributions of the principal orientations of the 3D scene by clustering the surface normals into K ($K > 3$) clusters; 2) Then Zhao and Zhu extract three biggest clusters and take their corresponding normals as three main orientations; 3) After the orthogonalization of these three orientations, Zhao and Zhu choose the one with smallest angle to the Y-axis of camera plane as the gravity direction.

Invisible Space Estimation As light travels in straight lines, the space behind the point clouds and beyond the view angles is not visible from the camera's perspective. However this invisible space is very helpful for completing the missing voxels from occlusions. Inspired by Furukawa's method in [278], the Manhattan space is carved by the point cloud into three parts, as shown in Fig. 3.15 (c): Object surface $\mathbb{S}$ (colored-dots voxels), Invisible space $\mathbb{U}$ (light green voxels) and Visible space $\mathbb{E}$ (white voxels).

Voxels Filling After obtaining labels by the above point cloud segmentation, first each voxel on surface $\mathbb{S}$ inherits the labels from the points that it enclosed. Then the completion of the missing parts for the volumetric primitives can be considered as guessing the label for each voxel which are invisible but should be belong to the object. As Fig. 3.15 (b) illustrates, the algorithm can be described as:
Loop: for each invisible voxel $v_i \in \mathbb{U}$, $i = 1, 2, \ldots$

- Starting from v_i to search the voxels, along 6 directions, until reach a voxel $v_j, j = 1 \ldots, 6$ that $v_j \in \mathbb{S}$. or v_j belongs to boundary of the whole space.
- Checking the labels of v_js, if there are more than two same labels exist, then assign this label to current voxel.

Fig. 3.16 (c) shows an example of volumetrically completing the primitives from (b). With the voxel representation, the primitives' mass, center of gravity (CoG) can be efficiently calculated.

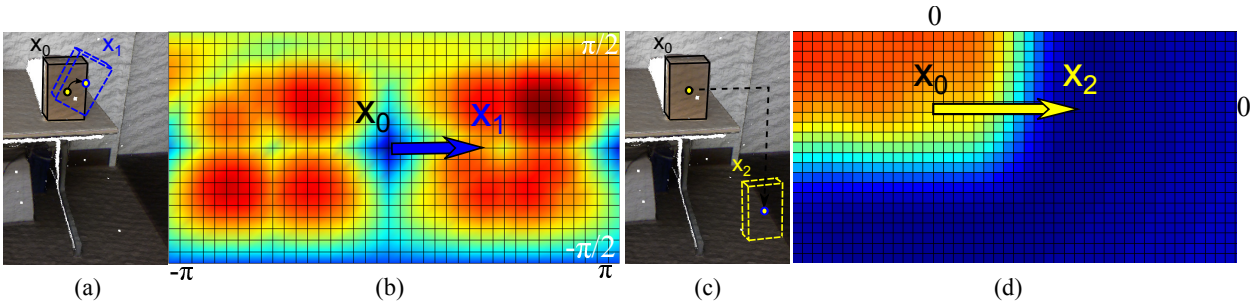


Figure 3.17: An example of potential energy map determined by pose and position changes: (a) the box on desk changes pose from state x_0 to x_1. Mass center trajectory is shown as black arrow. (b) the energy map of changing box poses in arbitrary directions. State x_0 is at local minimum on the map. (c) the box on desk changes position from state x_0 to x_2; (d) the energy map of changing box position. Due to friction is considered, State x_0 is at local minimum on the map.

3.3.3 Modeling Physical Stability and Safety

Energy Landscapes

Since any object (or primitive) has potential energy determined by its mass and height to the ground, Zhao and Zhu can generate its potential energy landscape according to the environment where it stays.

The object is said to be *in equilibrium* when its current state is a local minimum (stable) or non-local minimum (unstable) of this potential function (See Fig. 3.17 for illustration). This equilibrium can be broken after the object has absorbed external energy, and then the object moves to a new equilibrium and releases energy. Note that if too much uncontrolled energy is released, the object is perceived to be "unsafe," which Zhao and Zhu will discuss later. Without loss of generality, Zhao and Zhu divide the change into two cases.

- *Case I: pose change.* In Fig. 3.17 (a), the box on a desk is in a stable equilibrium and its pose is changed with external work to raise its center of mass. Zhao and Zhu define the energy change needed for the state change $\mathbf{x}_0 \rightarrow \mathbf{x}_1$ by

$$\mathcal{E}_r(\mathbf{x}_0 \rightarrow \mathbf{x}_1) = (R\mathbf{c} - \mathbf{t}_1) \cdot m\mathbf{g}, \tag{3.11}$$

  where $\cdot$ denotes inner product, R is rotation matrix; $\mathbf{c}$ is the center of mass, $\mathbf{g} = (0, 0, 1)^T$ is the gravity direction, $\mathbf{t}_1$ is the lowest contact point on the support region (its corners). Suppose the support region is flat, only the rotations of roll and pitch change the object CoM. Thus Zhao and Zhu can visualize the energy landscape in a spherical coordinate system (ϕ, θ): $S^2 \rightarrow \mathbb{R}$ with two pose angles $\{\phi \in [-\pi\ \pi], \theta \in [-\pi/2, \pi/2]\}$. In Fig. 3.17 (b), the blue color means lower energy and red means high energy. Such energy can be computed for any rigid objects by bounding the object with a convex hull. Zhao and Zhu refer to the early work of Kriegman [287] for further details.
- *Case II: position change.* Zhao and Zhu consider the position change when object is viewed as a mass point and can move to different position in its environment. For example, as shown in Fig. 3.17 (c), the box on desk at stable equilibrium state $\mathbf{x}_0$, one can push it to the edge of the desk. Then it falls to the ground and releases energy to reach a deeper minimum state $\mathbf{x}_2$. The total energy change need to consider the gravity potentials and the frictions which is overcome by a work absorbed.

$$\mathcal{E}_t(\mathbf{x}_0 \rightarrow \mathbf{x}_2) = -(\mathbf{c} - \mathbf{t}) \cdot m\mathbf{g} + W_f, \tag{3.12}$$

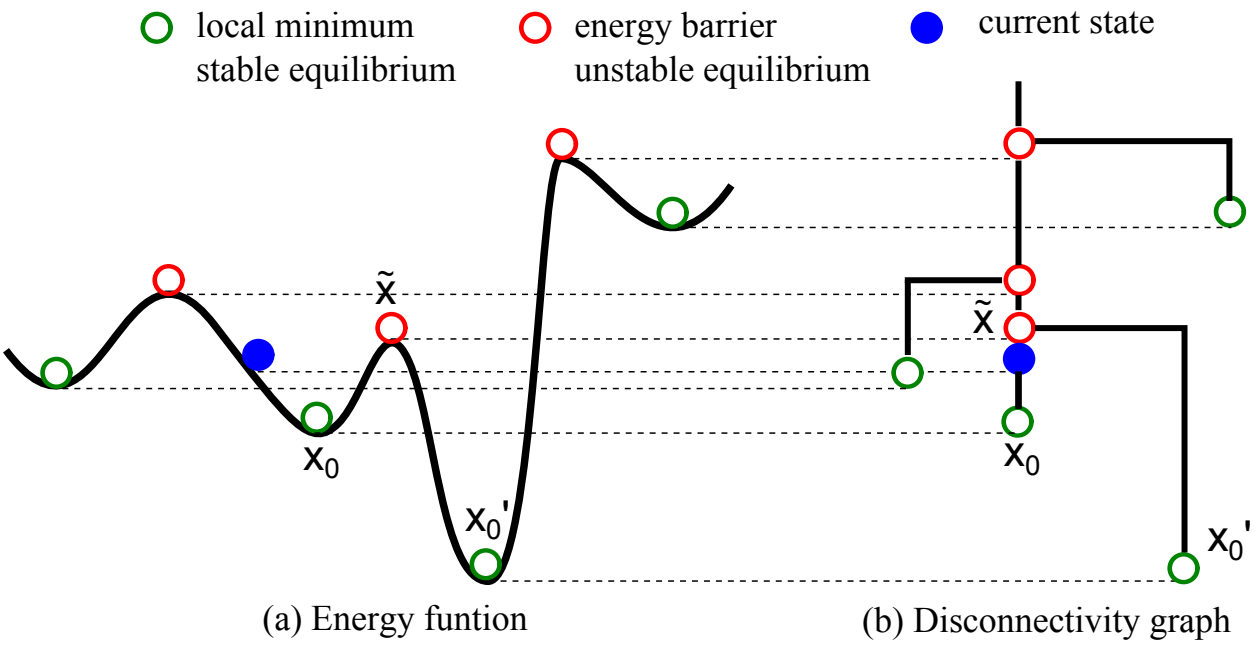


Figure 3.18: (a) Energy landscapes and its corresponding disconnectivity graph (b).

where $\mathbf{t} \in \mathbb{R}^3$ is the translation parameter (shortest path to the final position $\mathbf{x}_2$), and W_f is the absorbed energy for overcoming the frictions: $W_f = f_c \cdot mg\sqrt{(t_1 - c_1)^2 + (t_2 - c_2)^2}$ given the friction coefficient f_c. Note for common indoor scenes, Zhao and Zhu choose f_c as 0.3 as common material such as wood. Therefore the energy landscape can be viewed as a map from 3D space $\mathbb{R}^3 \rightarrow \mathbb{R}$.

Disconnectivity Graph Representation

The energy map is continuously defined over the object position and pose. For our purpose, Zhao and Zhu are only interested in how deep its energy basin is at the current state (according to the current interpretation of the scene). As the interpretation changes during optimization process, the energy landscape for each object will be updated. Therefore, Zhao and Zhu represent the energy landscape by a so-called disconnectivity graph (DG) which has been used in studying spin-glass models in physics [283]. As Fig. 3.18 illustrates that, in the DG, the vertical lines represent the depth of the energy basins and the horizontal lines connect adjacent basins. The DG can be constructed by an algorithm scanning energy levels from low to high and checking the connectivity of components at each level [283].

From the disconnectivity graph, Zhao and Zhu can conveniently calculate two quantities: *Energy absorption* and *Energy release* during the state changes.

Definition 1 The energy absorption $\Delta\mathcal{E}(\mathbf{x}_o \rightarrow \tilde{\mathbf{x}})$ *is the energy absorbed from the perturbations, which moves the object from the current state* $\mathbf{x}_0$ *to an unstable equilibrium* $\tilde{\mathbf{x}}$ *(say a local maximum or an energy barrier).*

For the box on the desk in Fig. 3.17, its energy absorption is the work needed to push it in one direction to an unstable state $\mathbf{x}_1$. For the state $\mathbf{x}_2$, its energy barrier is the work needed (to overcome friction) to push it to the edge. In both cases, the energy depends on the direction and path of movement.

Definition 2 Energy release $\Delta\mathcal{E}(\tilde{\mathbf{x}} \rightarrow \mathbf{x}'_o)$ *is the potential energy released when an object moves from its unstable equilibrium* $\tilde{\mathbf{x}}$ *to a minimum* $\mathbf{x}'_0$ *which is lower but connected by the energy barrier.*

For example, when the box falls off from the edge of the table to the ground, energy is released. The higher the table, the larger the released energy.

Definition of Stability

With DG, Zhao and Zhu define object stability in 3D space.
Definition 3 The instability $S(a, \mathbf{x}_0, W)$ *of an object* a *at state* $\mathbf{x}_0$ *in the presence of a disturbance work* W *is the maximum energy that it can release when it moves out of the energy barrier by the external work* W*.*

$$\begin{aligned} &\mathrm{S}(a, \mathbf{x}_0, W) \\ &= \max_{\mathbf{x}'_0} \triangle\mathcal{E}(\widetilde{\mathbf{x}} \rightarrow \mathbf{x}'_{\mathrm{o}})\delta([\min_{\widetilde{\mathbf{x}}} \triangle\mathcal{E}(\mathbf{x}_{\mathrm{o}} \rightarrow \widetilde{\mathbf{x}})] \leqslant W), \end{aligned} \tag{3.13}$$

where $\delta()$ is an indicator function and $\delta(z) = 1$ if condition z is satisfied, otherwise $\delta(z) = 0$. $\triangle\mathcal{E}(\mathbf{x}_{\mathrm{o}} \rightarrow \widetilde{\mathbf{x}})$ is the energy absorbed, if it is overcome by W, then $\delta() = 1$, and thus the energy $\triangle\mathcal{E}(\widetilde{\mathbf{x}} \rightarrow \mathbf{x}'_{\mathrm{o}})$ is released. Zhao and Zhu find the easiest direction $\widetilde{\mathbf{x}}$ to minimize the energy barrier and the worst direction $\mathbf{x}'_0$ to maximize the energy release. Intuitively, if $S(a, \mathbf{x}_0, W) > 0$, then the object is said to be unstable at level W disturbance.

Definition of Safety

Zhao and Zhu measure the safety by supposing a specific disturbance field as potentially existing in the scene, such human activities, winds or earthquakes. This specific disturbance field should have nonuniform and directional energy distribution.

Definition 4 The risk $R(a, \mathbf{x}_0)$ *of an entity* a *at position* $\mathbf{x}_0$ *in the presence of a disturbance field* $p(W, \mathbf{x})$ *is the expected risk with respect to the disturbance distribution.*

$$R(a, \mathbf{x}_0) = \int p(W, \mathbf{x}_0) S(a, \mathbf{x}_0, W) dW, \tag{3.14}$$

For example, it is more unsafe if there exist a disturbance that makes the box in Fig. 3.17 fall off from the desk than just fall down on the desk.

With the definition of the instability and risk, Zhao and Zhu first present the algorithm for static scene understanding by reasoning the stability, and then Zhao and Zhu introduce the inference of the disturbance field in Section 3.3.5 and the calculation of potential energy and initial kinetic energy given a disturbance in Section 3.3.5

3.3.4 Reasoning Stability

Stability Optimization

Given a list of 3D volumetric primitives obtained by our geometric reasoning step, Zhao and Zhu first construct the contact graph, and then the task of physical reasoning can be posed as a well-known graph labeling or partition problem, through which the unstable primitives can be grouped together and assigned the same label to achieve global stability of the whole scene at a certain disturbance level W.

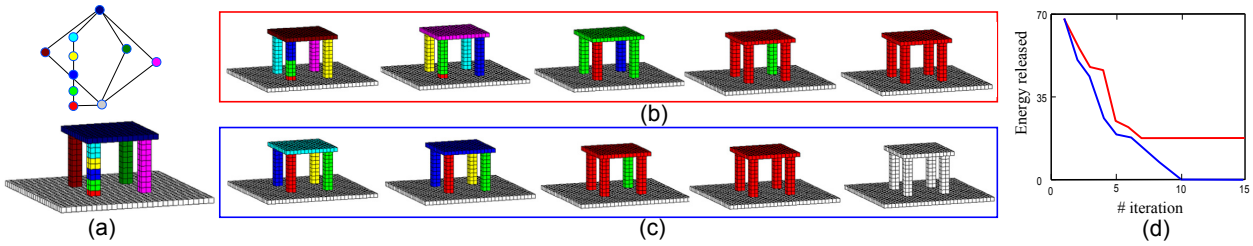


Figure 3.19: Example of illustrating the Swendsen-Wang cut sampling process. (a) Initial state with corresponding contact graph. (b) shows the grouping proposals accepted by SWC at different iterations. (c) convergence under increasingly (from left to right) larger disturbance W and consequently the table is fixed to the ground. (d) shows two curves of Energy released v.s. number of iteration in SWC sampling corresponding to (b) and (c).

Contact Graph and Group Labeling

The contact graph is an adjacency graph $G =< V, E >$, where $V = \{v_1, v_2, ..., v_k\}$ is a set of nodes representing the 3D primitives, and E is a set of edges denoting the contact relation between the primitives. An example is shown in Fig. 3.19 (a) top where each node corresponds to a primitive in Fig. 3.19 (a) bottom. If a set of nodes $\{v_j\}$ share a same label, that means these primitives are fixed to a single rigid object, denoted by O_i, and their instability is re-calculated according to O_i.

The optimal labeling L^* can be determined by minimizing a global energy function, for a disturbance level W

$$E(L|G; W) = \sum_{O_i \in L} (\mathrm{S}(O_i, \mathbf{x}(O_i), W) - \mathcal{F}(O_i)), \tag{3.15}$$

where $\mathbf{x}(O_i)$ is the current state of grouped object O_i. The new term $\mathcal{F}$ represents a penalty function expressing the scene prior and can be decomposed into three terms.

$$\mathcal{F}(O_i) = \lambda_1 f_1(O_i) + \lambda_2 f_2(O_i) + \lambda_3 f_3(O_i), \tag{3.16}$$

where f_1 is the total number of voxels in object O_i; f_2 is the geometric complexity of O_i, which can be simply computed as the summation of the difference of normals for any two connected voxels on its surface; and f_3 is the freedom of object movement on its support area. f_3 can be calculated as the ratio between the support plane and the contact area $\frac{\#S}{\#CA}$ of each pair of primitives $\{v_j, v_k \in \mathcal{O}_i\}$, where one of them is supported by the other. After they are regularized to the scale of objects, the parameters λ_1, λ_2 and λ_3 are set as 0.1, 0.1, and 0.7 in our experiment. Note, the third penalty is designed from the observation that, *e.g.*, a cup should have freedom of movement supported by a desk, and therefore the penalty arises if the cup is assigned the same label as the desk, as shown in Fig. 3.14. Therefore under the stable conditions, objects should have maximal freedom of movement.

Inference of Maximum Stability

As the labels of primitives are coupled with each other, Zhao and Zhu adopt the graph partition algorithm Swendsen-Wang Cut (SWC) [272] for efficient MCMC inference. To obtain the globally optimal L^* by the SWC, the next 3 main steps work iteratively until convergence.

- *Edge turn-on probability.* Each edge $e \in E$ is associated with a Bernoulli random variable $\mu_e \in \{\mathrm{on}, \mathrm{off}\}$ indicating whether the edge is turned on or off, and a weight reflecting the possibility of doing so. In this work, for each edge $e =< v_i, v_j >$, Zhao and Zhu define its turn-on probability as:

$$q_e = p(\mu_e = on|v_i, v_j) = \exp\left(-(F(v_i, v_j)/T\right), \tag{3.17}$$

where T is temperature factor and $F(\cdot,\cdot)$ denotes the feature between two connected primitives. Here Zhao and Zhu adopt a feature using the ratio between contact area (plane) and object planes as: $F = \frac{\#CA}{max(\#A_i,\#A_j)}$, where CA is the contact area, A_i and A_j are the areas of v_i and v_j on the same plane of CA.

- *Graph Clustering.* Given the current label map, it removes all edges between nodes of different categories. Then all the remaining edges are turned on independently with probability q_e. Thus, Zhao and Zhu have a set of connected components (CCPs) Π's, in which all nodes have the same category label.
- *Graph Flipping.* It randomly selects a CCP Π_i from the set formed in step (ii) with a uniform probability, and then flips the labels of all nodes in Π_i to a category $c \in \{1, 2, ..., C\}$. The flip is accepted with probability [272]:

$$\alpha(L \to L') = \min\left(1, \frac{\prod_{e\in\mathcal{C}(V_o, V_{L'} \setminus V_o)}(1-q_e)}{\prod_{e\in\mathcal{C}(V_o, V_L - V_o)}(1-q_e)} \cdot \frac{p(L'|G;W)}{p(L|G;W)}\right), \tag{3.18}$$

where $p = \frac{1}{z}\exp(-E)$. Fig. 3.19 illustrates the process of labeling a number of primitives of a table into a single object. SWC starts with an initial graph in (a), and some of the sampling proposals are accepted by the probability (Section 3.3.4) shown in (b) and (c), resulting in the energy v.s. iterations in (d). It is worth noticing that i) in case of Fig. 3.19 (b), the little chair is not grouped to floor, since the penalty term A_3 penalizes the legs grouping with the floor; and ii) with increased disturbance W, the chair is fixed to the floor.

3.3.5 Reasoning Safety

While the objects are stable in the gravity field of a static scene after reasoning the stability, they might be unsafe under a potential specific physical disturbance, such as human activities. For example, all the objects shown in Fig. 3.20 (a) can be parsed correctly to be stable in the scene, but if the physical disturbance generated from human common activities is applied, the objects show different safety levels.

Our method infers the disturbance field caused by an earthquake or wind, as well as the human action disturbance field. Given the scene geometry and walkable area, Zhao and Zhu detect the potential falling objects by calculating its expected falling risk given a disturbance field in Fig. 3.20 (b).

Safety Under Different Disturbances

Natural Disturbance Field Aside from the gravity applying a constant downward force to all the voxels, other natural disturbances such as earthquakes and winds are also present in a natural scene.

- *Earthquake* transmits energy by forces of interactions between contacting surfaces, typically by the frictions in our scenes. Here, Zhao and Zhu estimate the disturbance field by generating random horizontal forces to the voxels along the contacting surfaces. Zhao and Zhu use a certain constant to simulate the strength of the earthquake and the work W it generates.
- *Wind* applies fluid forces to exposed voxels in the space. A precise simulation needs to simulate the fluid flow in the space. Here, Zhao and Zhu simplify it as a uniformly distributed field over the space.

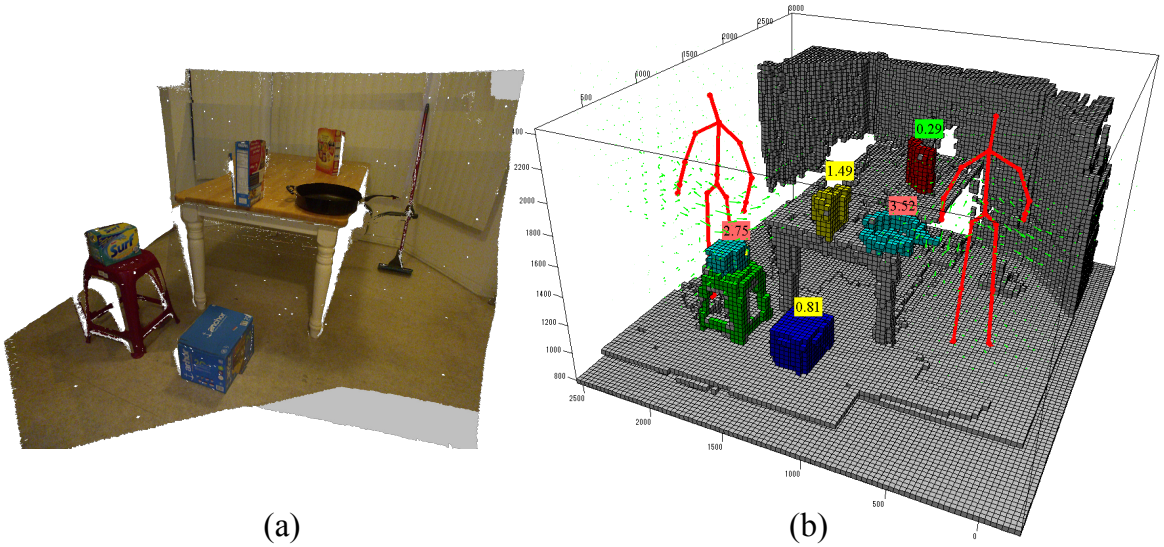


Figure 3.20: (a) The input point cloud; (b) Hallucinated human action field and detected potential falling objects with red tags.

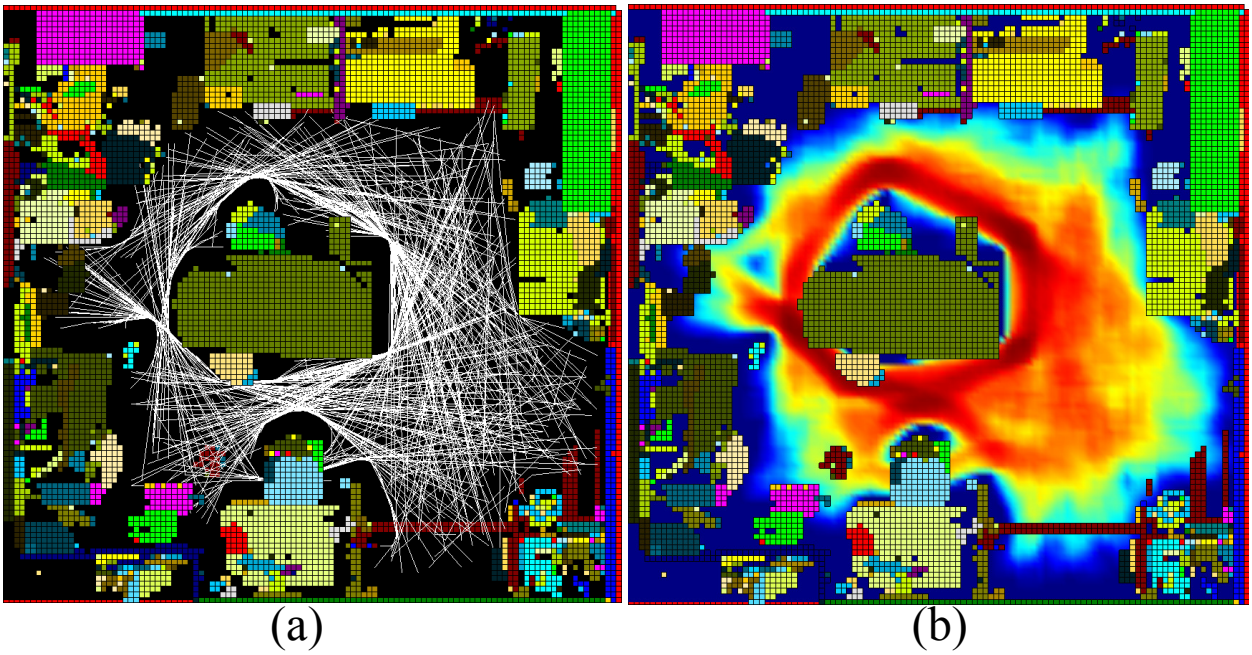


Figure 3.21: Primary motion field: (a) The hallucinated human trajectories (white lines); (b) The distribution of the primary motion space. The red represents high probability to be visited.

Human Action Disturbance Field In order to generate a meaningful disturbance field of human actions, Zhao and Zhu decompose the human actions into the primary motions *i.e.*, the center of mass movements in Fig. 3.21 and the secondary motions *i.e.*, the body parts' movements in Fig. 3.22 Zhao and Zhu first predict a human primary motion field on the 2D ground plan, and add detailed secondary motions in 3D space on top. The disturbance field is characterized by the moving frequency and moving velocity for each quantized voxel.

The primary motion field captures the movement of human body as a particle. Zhao and Zhu estimate the distribution of primary human motion space by synthesizing human motion trajectories

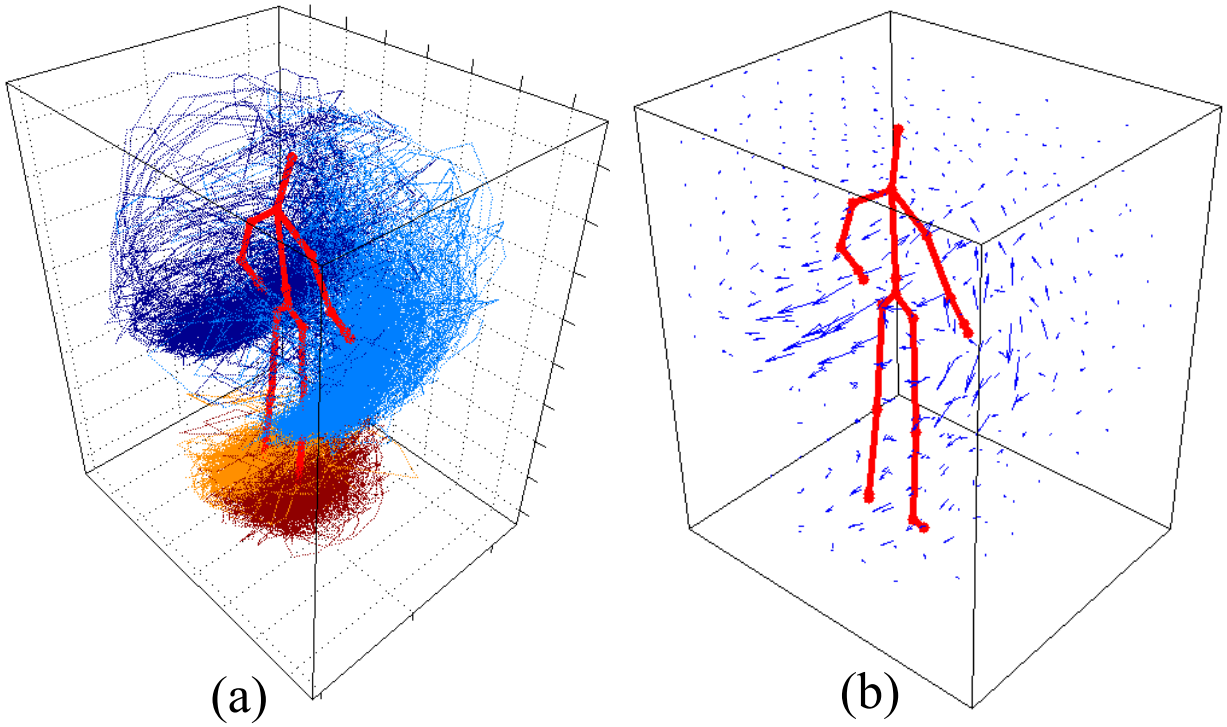


Figure 3.22: Secondary motion field: (a) Secondary motion trajectories from motion capture data; (b) Distribution of the secondary motion field. Long vectors represent large velocity of body movement.

following two simple observations:

- A rational agent mostly walks along a shortest path with minimal effort.
- An agent has a basic need to travel between any two walkable positions in the scene.

Therefore, Zhao and Zhu randomly pick 500 pairs of positions in the walkable space, Zhao and Zhu calculate the shortest path connecting these two positions as shown in Fig. 3.21 (a), and Zhao and Zhu calculate the walking frequency as well as walking directions based on the synthesized trajectories. Fig. 3.21 (b) demonstrates a distribution of walkable space; the red color means the position has high probability to be visited, and the length of the small arrows shows the probability of moving directions.

In Fig. 3.21 (b), Zhao and Zhu can see that convex corners, *e.g.*, table corners, are more likely to be visited, and objects in these busy area may have higher risk than the ones in concave corners. A hallway connecting two walkable area is also frequently visited, and objects in the hallway are less safe too. Note the distribution of moving directions is also very distinctive. It helps to locate human body movement in the right direction for generating the human disturbance field.

The secondary motion field is the movement that is not part of the main action, for example, arms swinging while walking. The secondary motion is important to capture the random disturbance; for example, people may push objects off the edge of the table by hand or kick objects on the ground by foot. Zhao and Zhu also the Kinect camera to collect human motion capture data Fig. 3.22 (a), and then calculate the distribution of moving velocities as shown in Fig. 3.22 (b).

The primary motion field further convolves with secondary motion field, thus generating a dense disturbance field that captures the distribution of motion velocity for each voxel in the space. The disturbance field is then represented by a probability distribution over the entire space for the velocities along different directions and frequencies that they occur. For example, a box in the middle of a large table will not be reachable by a walking person and thus the distribution of velocity

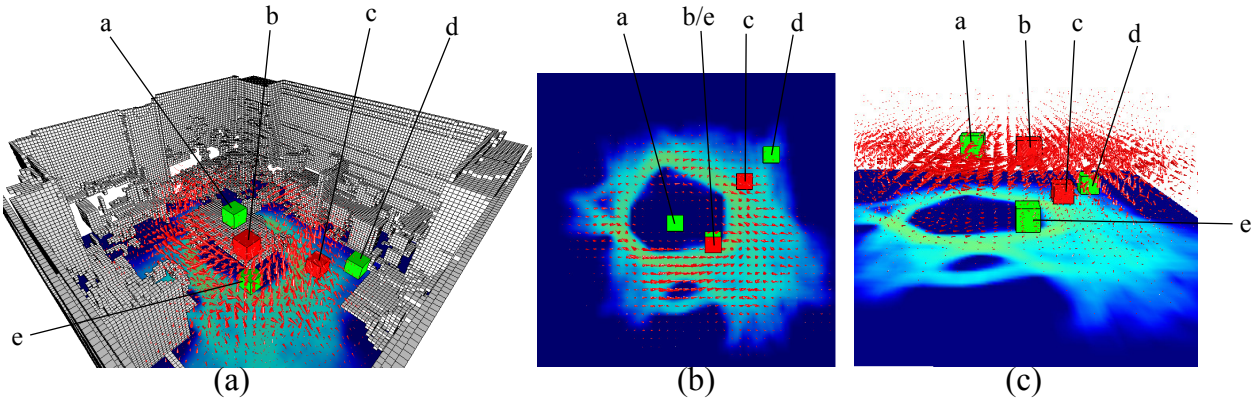


Figure 3.23: The integrated human action field by convolving primary motions with secondary motions. The objects *a*-*e* are five typical cases in the disturbance field: the object *b* on edge of table and the object *c* along the passway exhibit more disturbances (accidental collisions) than other objects such as *a* in the center of the table, *e* below the table and *d* in a concave corner of space.

above the table center, or any unreachable points, is zero. Five typical cases in the integrated field is demonstrated in Fig. 3.23

Calculating the Physical Energy

Given the disturbance field, in this section Zhao and Zhu present a feasible way for calculating input work (energy) that might lead to an object falling. However, building sophisticated physical engineering models is not feasible, as it becomes intractable if Zhao and Zhu consider complex object shapes and material properties, *e.g.*, to detect a box falling off from a table, a huge amount of actions need to be simulated until meeting the case of the human body acting on the box.

The relation between intuitive physical models and human psychology was discussed by a recent cognitive study [91].

In this work, for simplicity, Zhao and Zhu make following assumptions: 1) All the objects in the scene are rigid; 2) All the objects are made from same material, such as wood (friction coefficient: 0.3, uniform density: $700kg/m^3$); and 3) A scene is a dissipative mechanical system such that total mechanical energy along any trajectory is always decreasing due to friction, while kinetic and potential energy may be traded off at different states due to elastic collision.

Given the human motion distribution with velocity of each body part, Zhao and Zhu intuitively calculate the kinetic energy of human motion, as the input work. Here, Zhao and Zhu simplify the parts of body as mass points and at each location in 3D space its kinetic energy can be calculated given the mass of parts. For example, supposing the mass of right hand with upper arm is about 700g, Zhao and Zhu can simply calculate out the kinetic energy distribution by multiplying half of the velocity squares.

3.3.6 Experimental Result

Zhao and Zhu quantitatively evaluate our method in four criteria: i) single depth image segmentation, ii) volumetric completion evaluation, iii) physical inference accuracy evaluation, and iv) safety ratings for objects in scene.

All these evaluations are based on three datasets:

- the NYU depth dataset V2 [215] including 1449 RGBD images with manually labeled ground truth.

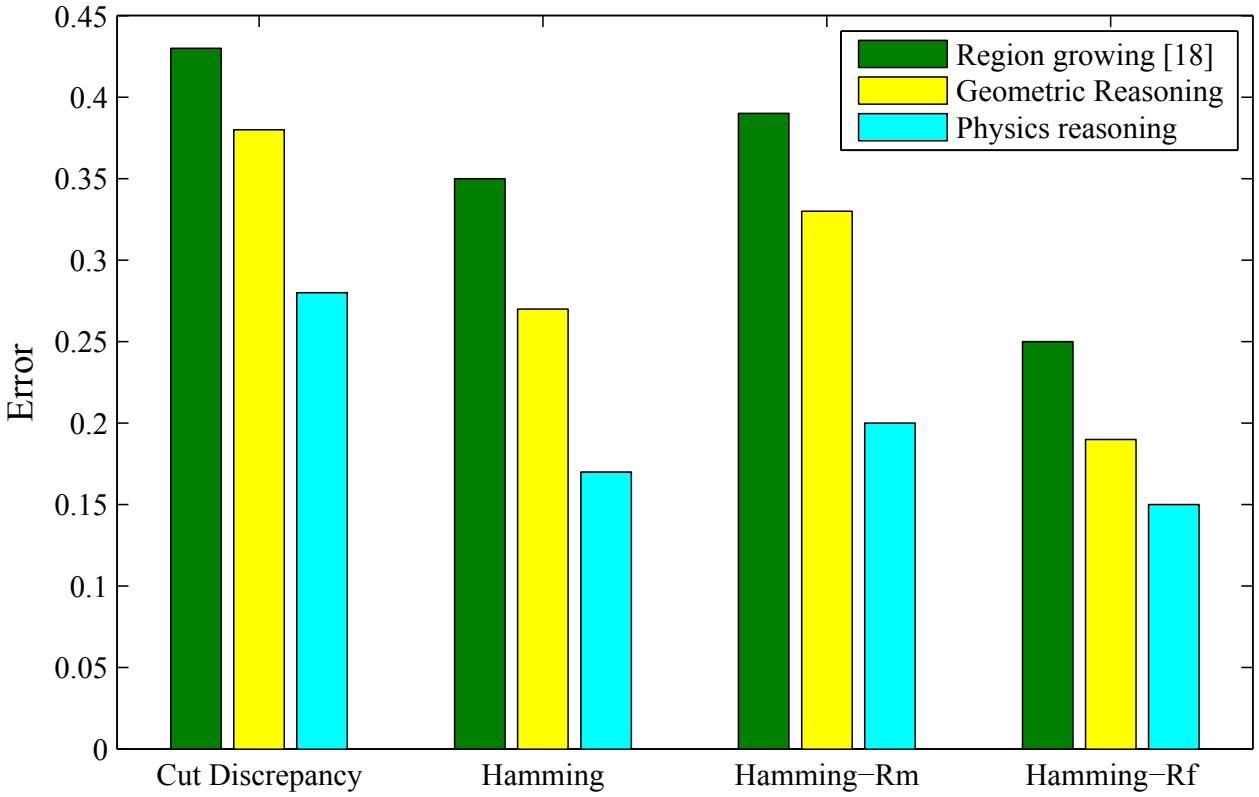


Figure 3.24: Segmentation accuracy comparison of three methods: Region growing method [275], result of our geometric reasoning and physical reasoning by one "Cut Discrepancy" and three "Hamming Distance."

- synthesized depth map and volumetric images simulated from CAD scene data.
- 13 reconstructed 3D scene data captured by Kinect Fusion [271] gathered from office and residential rooms with ground truth labeled by a dense mesh coloring.

Evaluating Single Depth Image Segmentation

Two evaluation criteria: "Cut Discrepancy" and "Hamming Distance" mentioned in [288] are adopted. The former measures errors of segment boundaries to ground truth, and the latter measures the consistency of segment interiors to ground truth. As shown in Fig. 3.24, our segmentation by physical reasoning has a lower error rate than the other two: region growing segmentation [275], and our geometric reasoning.

Fig. 3.25 shows some examples of comparing another point cloud segmentation result [275] and our result. However, it is worth noticing that, beyond the segmentation task, our method can provide richer information such as volumetric information, physical relations, stability, *etc*.

Evaluating Volumetric Completion

For evaluating the accuracy of volumetric completion, Zhao and Zhu densely sample point clouds from a set of CAD data including 3 indoor scenes. Zhao and Zhu simulate the volumetric data (as ground truth) and depth images from a certain view (as test images). Zhao and Zhu calculate the precision and recall which evaluates voxel overlapping between ground truth and the volumetric completion of testing data. Table 3.2 shows the result that our method has much better accuracy than traditional Octree methods such as [286].

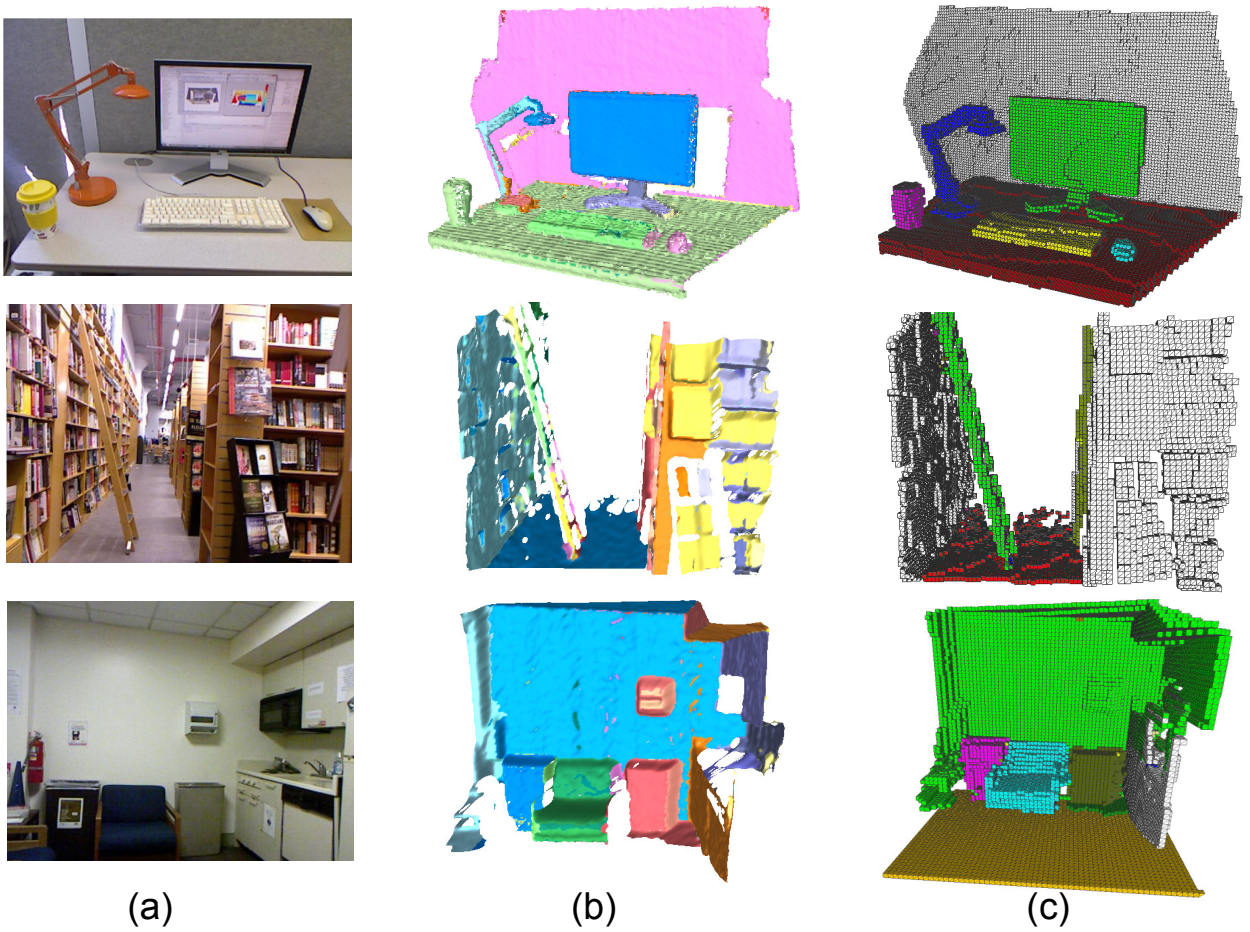


Figure 3.25: Segmentation result for single depth images. (a) RBG images for reference. (b) segmentation result by region growing [275]. (c) stable volumetric objects by physical reasoning.

	Octree	Invisible space	Vol. com.
Precision	98.5%	47.7%	*94.1%*
Recall	7.8%	95.1%	*87.4%*

Table 3.2: Precision and recall of Volumetric completion. Comparison of three method: 1) voxel-based representation generated by Octree algorithm [286], 2) voxels in surface and invisible space, and 3) our volumetric completion.

Evaluating Physical Inference Accuracy

Because the physical relations are defined in terms of our contact graph, Zhao and Zhu map the ground-truth labels to the nodes of contact graphs obtained by geometric reasoning. Than Zhao and Zhu evaluate our physical reasoning against two baselines: discriminative methods using 3D feature priors similar to the method in [215], and greedy inference methods such as the marching pursuit algorithm for physical inference. The result shown in Table 3.3 is evaluated by the average over 13 scene data captured by Kinect Fusion.

Fig. 3.26 (a)–(d) and (e)–(j) show two examples from the results. Here Zhao and Zhu discuss some irregular cases illustrated by close-ups of the figures.

- *Case I: Fig. 3.26 (c)* the ball is fixed onto the handle of sofa. The reason can be considered as: stability of the "ball" is very low measured by (Section 3.3.3). The unstable state is calculated

relations	Discriminative	Greedy	SWC
fixed joint	20.5%	66%	*81.8%*
support	42.2%	60.3%	*78.1%*

Table 3.3: Results of inferring the fixed joints and support relations between primitives. Accuracy is measured by nodes of the contact graph whose label is correctly inferred divided by the total number of labeled nodes.

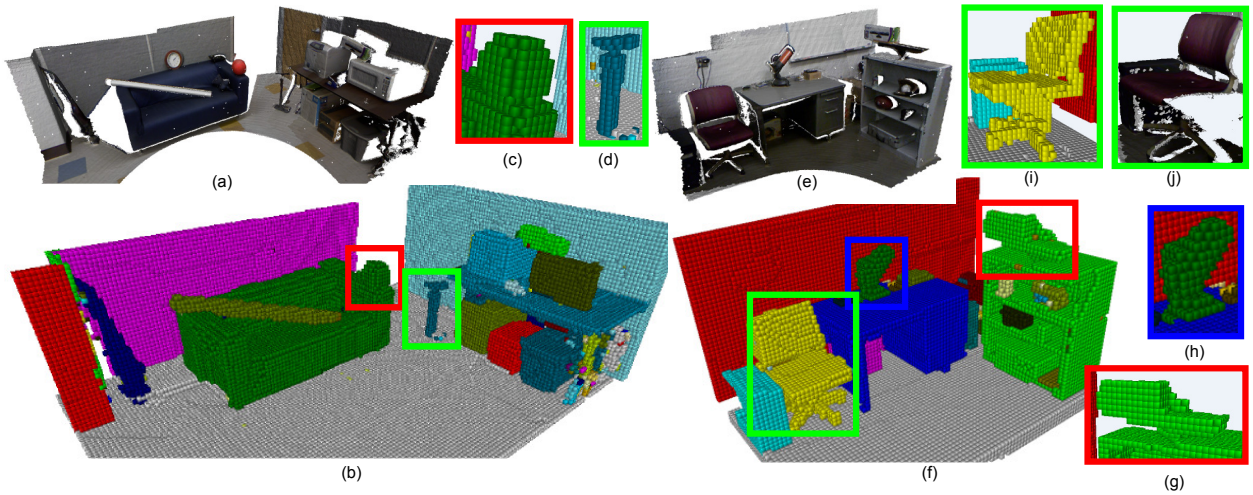


Figure 3.26: Example result. (a) and (e): data input. (b) and (f): volumetric representation of stable objects. (c): the ball is fixed onto the handle of sofa. (d): the "pump" is unstable (see text). (i): a irregular case of (g). (j): hidden voxels under chair compared to (h).

out as that it trends to release much potential energy (draw from the sofa) by absorbing little possible energy (*e.g.*, the disturbance by human activity).

- *Case II: Fig. 3.26 (d)* the "air pump" unstably stands on floor but is an independent object, because although its stability is very low, the penalty penalized it to be fixed onto the floor. The lamp is not affixed for the same reason, as shown in Fig. 3.26 (h).
- *Case III: Fig. 3.26 (g)* the "empty Kinect box" with its base is fixed together with the shelf, because of the mis-segmentation of the base, *i.e.*, the lower part of base is mis-merged to the top of the shelf.
- *Case IV: Fig. 3.26 (i)* voxels under the "chair" are completed with respect to stability.

The reasons are: 1) our algorithm reasons the hidden part occluded in invisible space. 2) the inference of the hidden part is not accurate geometrically, but it helps to form a stable object physically. In contrast, the original point cloud shown in Fig. 3.26 (j) misses more data.

Evaluating Safety Ratings

First Zhao and Zhu provide a selected qualitative result shown in Fig. 3.27. Zhao and Zhu compare the potential falling objects under three different disturbance fields: 1) The human action field in Fig. 3.27 (b,e); 2) The wind field (a uniform directional field) in Fig. 3.27 (c,f) and 3) earthquake (random forces on contacting object surface) in Fig. 3.27 (d,g). As Zhao and Zhu can see the cups with red tags are detected as potential falling objects, which is very close to human judgments: (i) objects around the table corner are not safe w.r.t. human walking action; (ii) objects along the edge of wind direction are not safe w.r.t. wind disturbance; and (iii) object along all the edges are not safe w.r.t. earthquake disturbance.

Next Zhao and Zhu report selected results in different 3D scenes, as shown in Fig. 3.28 top row:

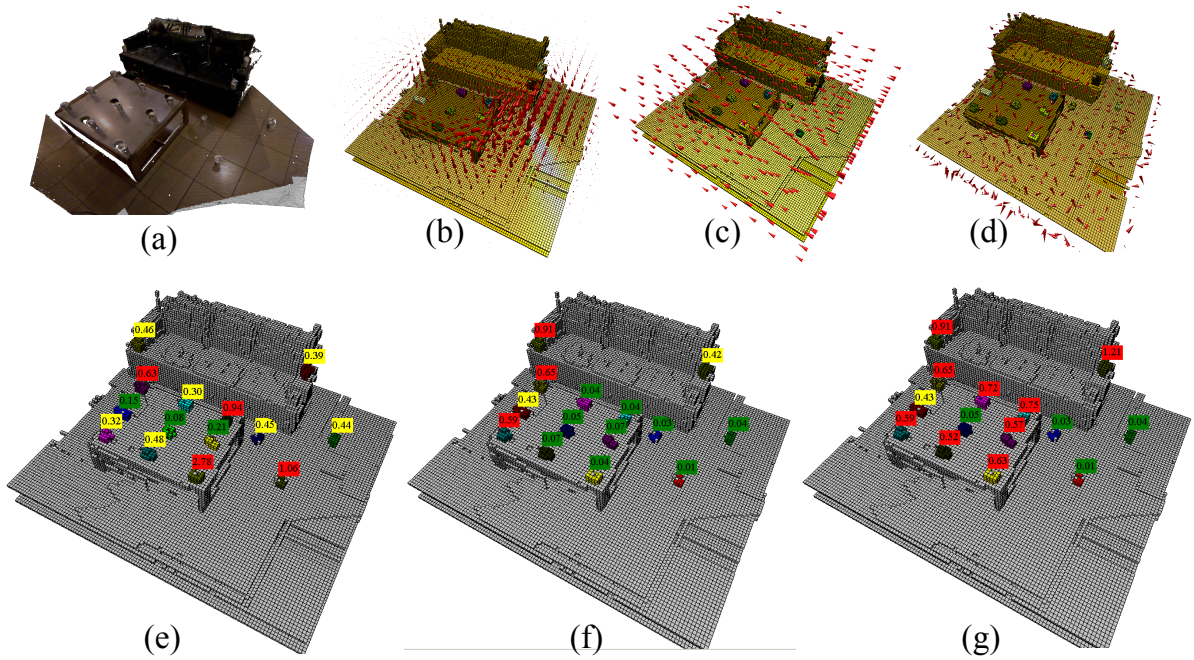


Figure 3.27: The potential falling objects (with red tags) under the human action field (b,e), the wind field (c,f) and the earthquake field (d,g) respectively. The results match with human perception: (i) objects around table corner are not safe w.r.t. human walking action; (ii) object along the edge of wind direction are not safe w.r.t. wind disturbance; and (iii) object along all the edges are not safe w.r.t. earthquake disturbance.

vending machine room and bottom row: copy machine room. Zhao and Zhu can see that, according to human motions, the cans on vending machine room at risky of being kicked off, while the can near the window is considered stable, since people can rarely reach there. In the copy room, the objects put on the edges of table are at more risk than others.

Discussion

For evaluating safety ratings, Zhao and Zhu rank object unsafeness in a scene in comparison with human subjects. Fig. 3.29 (a) shows a 3D scene (constructed in CAD design), from which Zhao and Zhu pick 8 objects and ask 14 participants to rank the unsafeness of these objects considering gravity, common life activity and the risk of falling. Zhao and Zhu compare the human ranking with our unsafeness function $R(a, \mathbf{x})$ in Fig. 3.29 (b). Zhao and Zhu found that 1) humans got big variations while considering the safeness, due to deeper consideration of information such as material; 2) however, the model got similar ranking scores with the average of human rankings. As shown in Fig. 3.29 (b), the average of human vs. model scores for each object lies near to the diagonal line.

3.3.7 Conclusion

Zhao and Zhu present a novel approach for scene understanding by reasoning their instability and risk using intuitive mechanics with the novel representations of the disconnectivity graph and disturbance fields. Our work is based on a seemingly simple but powerful observation that objects, by human design, are created to be stable and have maximum utility (such as freedom of movement). Zhao and Zhu demonstrated its feasibility in experiments and show that this provides a new method

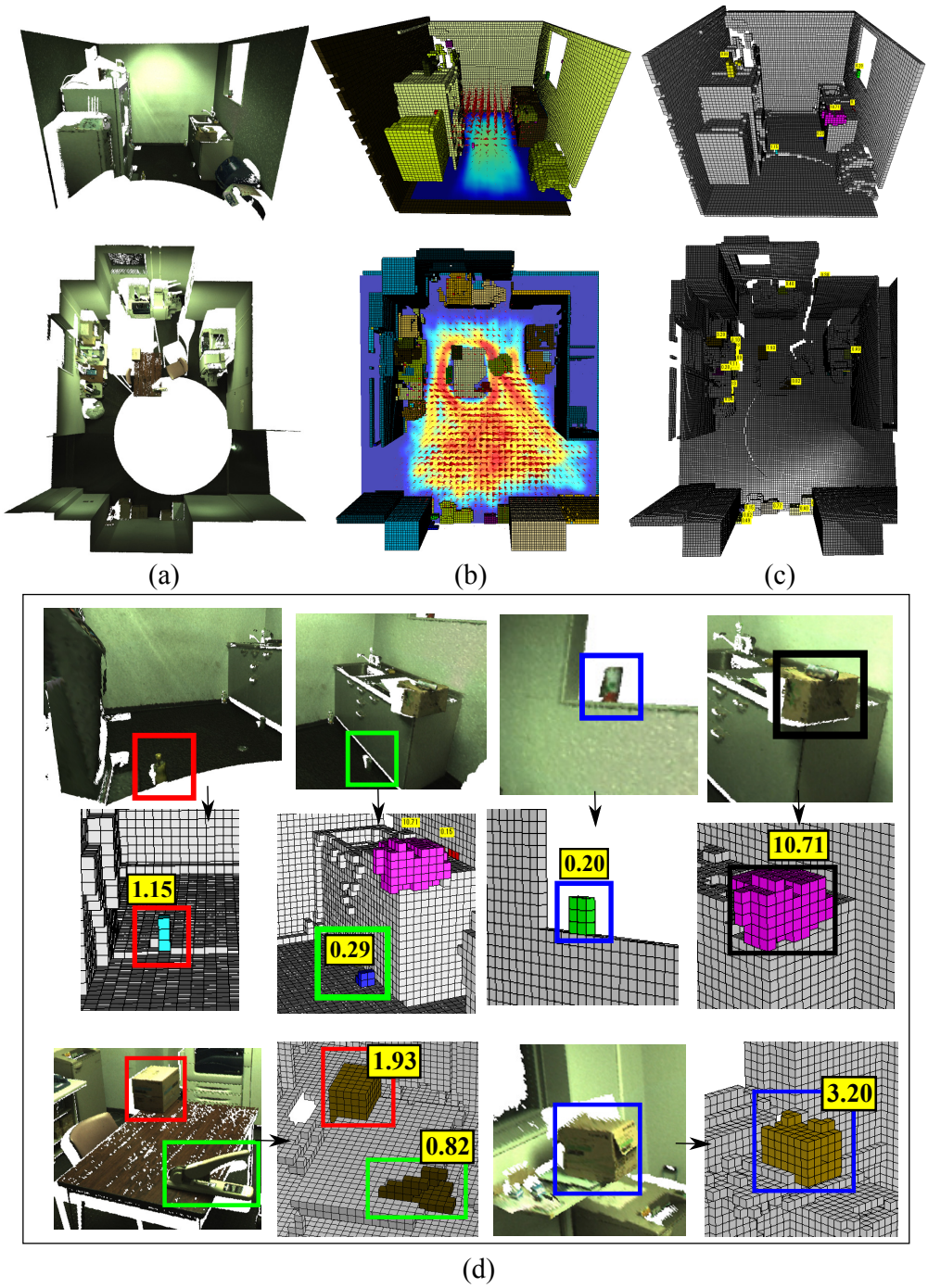


Figure 3.28: (a) Input 3D scene point clouds; (b) Segmented volumetric objects in different colors and inferred disturbance fields of human activity; (c) objects with risk scores. (d) Zoom-in details of detected potential risky objects.

for object grouping when it is hard to pre-define all possible object shapes and appearance in an object category.

This work also presents a novel approach for detecting potential unsafe objects. Zhao and Zhu demonstrated that, by applying various disturbance fields, our model achieves a human level recognition rate of potential falling objects on a dataset of challenging and realistic indoor scenes.

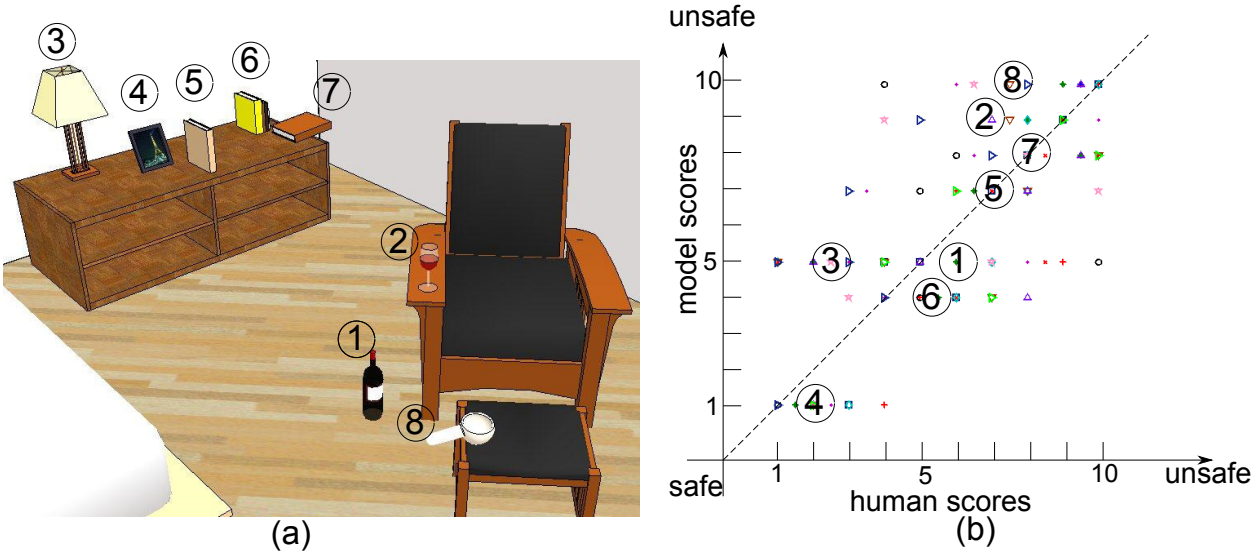


Figure 3.29: Scoring object unsafeness in a scene (a) with 8 objects. Zhao and Zhu show the correlation graph (b) with human score against our measurement $R(a, \mathbf{x})$ which is normalized from 1 to 10. Color/shape points show human vs. model scores corresponding to different persons. Circle points with numbers inside show the average of human vs. model scores for each object corresponding to (a).

Differing from the traditional object classification paradigm, our approach goes beyond the estimation of 3D scene geometry. The approach is implemented by making use of "causal physics." It first infers hidden and situated "causes" (disturbance) of the scene, and introduces intuitive mechanics to predict possible "effects" (falls) as consequences of the causes. Our approach revisits classic physics-based representation, and uses the state-of-the-art algorithms. Further studies along this way, including friction, material properties, causal reasoning, can be very interesting dimensions of vision research.

In future research, Zhao and Zhu plan to explore several directions: i) Connecting our work to human psychology models like the one in [91], and to compare our results with human experiments; ii) Studying material properties in typical indoor scenes, and thus to reason about the materials jointly with stability, especially if Zhao and Zhu can see object movements in video; iii) Combing the physical cues with other appearance and geometric informations for scene parsing; and iv) Studying other specific action distributions to reason about whether a room is safe to children and infants.

Chapter 4

Causality in Daily Activities

4.1 Introduction

4.1.1 Why is Causality important?

The core of knowledge is rooted in causation - Aristotle believed "we do not have knowledge of a thing until we have grasped its why, that is to say, its cause" [289] and Mackie stated "causality is the cement of the universe" [290]. Causal learning is the basis through which humans have learned to master and control our observable universe. Such knowledge allows humans to ask counterfactual questions like *what would happen if I adjust the angle at which I strike the billiard ball?* Controlled experiments allow scientists to uncover knowledge about the causal mechanisms underpinning processes such as physical laws or drug efficacy. All of these processes require degrees of exploration, hypothesis generation, reasoning, and testing to uncover causal mechanisms. But how to systematically uncover causal knowledge in ways familiar to humans remains elusive for machines.

In artificial intelligence, here are many existing systems that can effectively learn representations to interact with their environment. A question that naturally arises is: why bother with a formal causal representation? For starters, many existing causal systems have derivable guarantees on the existence of a causal relationship [291]. This means the presence of a causal relationship can be verified given a relatively few and general assumptions. More importantly, causal representations provide a task-invariant representation of the world. This means causal knowledge is not specific to a singular task and can naturally be used for multiple tasks. Thus far, most machine learning representations learn representations specific to the training data and task at hand and exhibit very poor performance when transferred to similar but different circumstances. In contrast, models that learn causal relationships as form of model-building offer a representation that is constant assuming constant environment dynamics (such as the dynamics we experience on the surface of Earth). Thus causal knowledge is transferable to new tasks in domains governed by the same causal mechanisms.

The key research question in the field of causal learning is how various intelligent systems, ranging from rats to humans and machines, can learn cause-effect relations in novel situations. Decades ago, a number of researchers (*e.g.*, [292, 293]) suggested that causal knowledge can be acquired by a basic learning mechanism, associative learning, that non-human animals commonly employ in classical conditioning paradigms to learn the relationship between stimuli and responses. A major theoretical account of associative learning is the Rescorla-Wagner model, guided by prediction error in updated associative weights on cue-effect links [294].

However, subsequent research has produced extensive evidence that human causal learning

depends on more sophisticated processes than associative learning of cue-effect links [295]. Human learning and reasoning involves the acquisition of abstract causal structure [296] and strength values for cause-effect relations [297]. Causal graphical models [291] have been integrated with Bayesian statistical inference [298, 299, 300] to provide a general representational framework for human causal learning [295].

Nevertheless, most models of human causal learning assume that the hypothesis space of causal variables and causal structures is given and that inference focuses on selecting the best causal structure to explain the observed contingency information relating causal cues to effects. It is unclear how an agent could *actively* explore a completely novel situation in an online fashion and narrow down the set of potential causal structures to enable efficient inference.

In the context of robotics, causal relationships endow knowledge a task-invariant model of the world. Task-invariant representations are useful for two reasons: (1) they naturally enable transferable knowledge between tasks and (2) aid in efficient exploration to acquire new knowledge. For instance, suppose a naive baby is attempting to open a bottle. Initially, they may observe a parent opening the bottle and wonder how they can achieve the same effect (opening the bottle). If given the bottle, the baby may initial smash, drop, or roll the bottle to see if any of these actions cause the bottle to open. Through this exploration process, the baby is generating and testing causal hypotheses that govern the bottle's operation. Eventually, the baby discovers that twisting the cap opens the top of the bottle, to their parents dismay as the baby has a newfound skill capable of creating an enormous amount of messes.

But the baby continues to explore, hypothesize and test other relationships and is capable of generalizing their knowledge about twisting the cap of the first bottle to bottles of various shapes and sizes. However, one day the baby observes a parent opening a medicine bottle that requires *pushing* in addition to twisting. The baby mischievously acquires the medicine and attempts to open it using its well-founded, generalizable understanding of bottles—twisting the lid and pulling. To the baby's dismay, this seemingly bulletproof approach does not succeed. The baby tries and tries, but no strategy works. The baby observed their parent twisting, so that must be part of the solution, but something is missing. Time passes, and eventually the young child develops the strength to push on the lid to unlock it. Now they've explored the proper extension and adapted their basic twisting strategy to special cases, pushing and twisting on medicine bottles.

The purpose of this baby exploration and learning example is to highlight how learning a causal relationship, when properly coupled with abstraction mechanisms, enables agents to learn compact and effective knowledge for a specific environment and generalize that knowledge to other similar but different cases. Humans have a remarkable ability to learn causal relationships and form abstractions of observed data. This ability allows humans to interact and generalize across a wide range of environments, circumstances, and tasks, using a compact set of causal rules that govern most of everyday life.

4.1.2 What is Causality?

Causality is the formal study of cause-effect relationships. The practical purpose of studying such relationships is to identify how the environment changes after interacting with the environment. Thus far, causality has been formally studying in fields where interventions are typically difficult or impossible—such as economics or epidemiology. However, there is a growing interest in applying formal causal representations to computer vision and robotics. Causal relationships govern our universe: from chemistry to physics to cooking; understanding how to manipulate the environment to produce a desired effect is a critical component of any advanced robotic system.

There are many domains within the field of causality. In psychology, researchers have examined

causal perception to show that humans have an irresistible urge to assign causal relationships to certain types of stimuli, such as a launch event between two balls [301]. Psychologists have also examined how people reason about and assign strengths to causal relationships. Such reasoning has been argued to be based on covariation [297, 302], mechanisms [303], and dynamics [304]. These approaches were all developed from a cognitive science perspective; instead of seeking to establish the ground truth causal relationship, they seek to match human performance.

Within the fields of statistical inference and AI, causal inference, pioneered by Judea Pearl [291], deals what inferences can be made given a particular causal model. This includes things like counterfactual or "what-if" questions. Pearl introduced the *do*-operator that represents an intervention in the causal model, allowing modelers to assess the state of a causal model under different interventions. Causal inference revolution fields where causal structure can be provided by domain experts but interventions in the real world are difficult if not impossible (*e.g.*, economics or epidemiology).

Casual learning from observational data is considered to be an extremely challenging problem in the statistical and machine learning communities. The space of possible causal relations is exponential in the number of variables, and even in cases with two variables, determining the presence and direction of a causal relationship requires multiple assumptions about the underlying data [305]. To tackle these challenges, the space of variables to consider is generally kept low, and conditional independencies or score-based methods are used to determine the presence of a potential cause [306].

Causal learning using experimental data, under randomized controlled experiments, is the gold standard for causal learning. Fisher's randomized controlled experiments [307] provided the modern paradigm for experimental design across statistics, medicine, psychology. To this day, the experimental paradigm outlined by Fisher provides the only scientifically proven way to discover causal relationships from data. To isolate cause and effect, randomized controlled experiments work by controlling for as many confounding variables as possible. For AI, this approach is also the gold standard, and therefore should be leveraged by artificial agents in some capacity, but the setting up a randomized controlled experiment for every possible causal relationship is infeasible - the space of causal relations is exponential in the number of variables and therefore the number of randomized controlled experiments required is exponential. We must find ways for artificial agents to leverage causal knowledge learned from observational data and experimental data to efficiently and robustly learn causal relationships in the world.

We also outline a distinction between quantitative and qualitative causality. *Quantitative causality* deals with physical laws, such as acceleration, friction, electromagnetism, *etc.*, that govern our universe. Quantitative causality is therefore the bedrock of causality, and discovering the functional form of these relationships is the core of the scientific method (see Section 4.2). However, for cognitive agents, this resolution is typically too pedantic and detailed for effective reasoning. To this end, *qualitative causality* captures the more intuitive, high-level causal relationships that humans interact with in daily life. In Pearl's famous "smoking causes cancer" example [291], no processes describe how tobacco causes cellular changes that lead to cancer (a quantitative description of how smoking causes cancer), but instead, the high-level relationship is summarized in the structure of the Directed Acyclic Graph (DAG) and the corresponding probabilities in the Conditional Probability Table (CPT). We believe the lack of effort on merging the quantitative and qualitative causal domains presents a significant hurdle to causal learning; without both, one will not be able to learn grounded representations (quantitative) that allow for efficient inference (qualitative). The representational framework presented in this book, namely the Spatial, Temporal, and Causal And-Or Graph (STC-AOG), is capable of capturing both quantitative and qualitative causality.

In this chapter, we focus on causal learning as we believe it is the most important and least studied domain for artificial intelligence. While we can hand-code known causal knowledge to autonomous agents and use causal inference to aid in their decision-making process, agents without

the ability to learn new causal information will be limited in their generalizability. The world is a complex and ever-changing environment, and being able to learn causal relationships, not just make inferences about them, is a core component of generalized intelligence. The challenge for causal learning is how to efficiency find causal relationships from as little data as possible. We believe an active, "learn as you go" approach is the correct path forward, instead of opting for provable causal discovery. We posit this is how humans learn causal relationships both in daily life and through more rigorous methodologies such as the scientific method.

4.2 Causal Learning as Scientific Exploration

The scientific method is the fundamental mechanism for the acquisition of new knowledge and understanding of our universe. The basic principles of the scientific method have bee used for millennia in some form by great minds such as Copernicus, Kepler, Galileo, and Newton, but Sir Francis Bacon began formalizing the approach in the 19th century [308]. The scientific method gives humankind a systematic way to learn causal relationships. The method consists of the following basic steps: (i) ask a question, (ii) collect background information, (iii) construct a hypothesis, (iv) test hypothesis with controlled experiments (v) analyze results to confirm or refute hypothesis. Our hypothesis in this chapter is that artificial agents need to leverage a similar process to achieve general artificial intelligence.

Agents need mechanisms that enable learning causal relationships from observations *and* interventional information. Simultaneously, agents only endowed with interventional learning must explore the space of all causal relationships, which is exponential in the number of actions and states, to test all possible causal relationships. Thus, these two learning paradigms must be combined for agents to learn causal relationships in a tractable and robust fashion.

Artificial agents should be endowed with processes that emulate the scientific process we use to uncover new physical laws, test new drugs, and learn as children. This process can be summarized into a sequential process that executes in a loop: (i) ask a question (information gain), (ii) construct a hypothesis (experimental setup), (iii) test hypothesis (intervention execution), (iv) update model according to outcome (model update). Robots endow with procedures to facilitate this loop are able to learn about their world and construct a consistent, but always updating model of their reality.

We identify two major components of this scientific process in a computer system: observations and interventions. We focus on these two paradigms as the former enables pruning the space of possible relations to experimentally verify while the latter ensures the robot is capable of eliminating potential confounding factors present in observational settings.

4.3 Necessity of Observations

Learning causal relationships from observations is important to provide a starting point for causal learning. Entertaining the full space of possible causal relationships and verifying or refuting a relationship with a controlled experiment is intractable and exponential in the number of variables. Fortunately, learning causal relationships from observations allows agents (including humans) to prune the space of interventional experiments needed, but observations are incapable of presenting a complete story as the presence of confounders can rarely be eliminated [306]. Constructing causal models from observations offers a crucial starting point for agents to explore and verify hypotheses. Throughout daily life, agents experience the environment around them in temporal order, meaning agents can make the (safe) assumption that causes precede effects. Most causal events in human

society occur with a short time delay between cause and effect, and co-occurrence of cause and effects can be used to identify perceptually causal relationships.

Within the traditional causality community, causal discovery from observational data is utilized when interventions are difficult or impossible to perform. Causal discovery methods are divided into two broad categories: constraint-based and score-based. Constraint-based methods are based on conditional independence (CI) tests; seminal implementations include the PC and FCI algorithms [306]. Score-based methods optimize some score function to rank possible causal structures; seminal methods include K2 [309] and GES [310]. The key difference between the two approaches is that constraint-based methods can handle arbitrary input data and produce DAGs that adhere to the CI constraints present in the data while score-based methods can combine structural confidence using multiple possible DAGs. While these approaches seek recover the ground-truth causal structure, they typically struggle to do so on real-world data; however, they serve as an excellent starting point for understanding the causal structure of the underlying data.

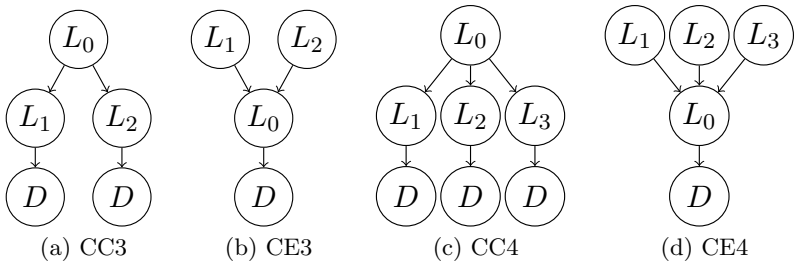
Following the cognitive science work of perceptual causality [301], perceptual causality [311] learns causal relations using an information-theoretic approach that is based on the minimax entropy principle [312]. Perceptual causality constructs a shallow graphical model that links causes to their effects based on contingency data by greedily pursuing causal relations according to their information gain. A graphical structure is constructed iteratively according to this information gain; initially, all actions and effects are assumed to be independent, and links between causes (actions) and their effects (fluent changes) are added with each iteration. The method relies on the assumptions of: (i) cause preceding effect, (ii) a short temporal delay between cause and effect, and (iii) co-occurrence measures strength of causal relation. This method selects causal relations in an order that matches human performance data; causal relations with high information gain (added in earlier iterations) correspond to human perceived causes.

Learning causal relationships from observational (perceptual) data is critical for agents to establish knowledge and prune exploration needed. While any method learning from observational data is prone to potential confounders, leveraging causal information derived from observations prevents agents from needing to establish every causal relation from experimentation.

4.4 Necessity of Experimental Data

While learning causal relations from observations is critical to leverage information from other agents and accelerate learning, any causal learning system incapable of learning through active intervention is prone to confounders and therefore potentially an inaccurate causal model of the environment. Only under highly constrained settings can the ground-truth causal model be recovered from observational data [306]. Thus, any agent expected to learn ground-truth causal models must be endowed with the ability to actively intervene in the environment. Fortunately, as our focus in this chapter is on causal learning for daily activities, allowing the agent to interact with the environment is natural and expected in many applications. But how should an agent choose interventions? How should the agent update its model based on the outcomes?

Selecting interventions should be based on two components: (i) how much the agent stands to gain from performing a particular intervention and (ii) how well the intervention fits into the agent's current goal. How much the agent benefits from a particular intervention can be quantified using information gain, where the information gain corresponds to the amount of model improvement expected by performing a particular intervention. Agents can then make a final action selection by fusing this information gain with an inference about which action is most likely to aid in completing the task.



(a) CC3 (b) CE3 (c) CC4 (d) CE4

Figure 4.1: Common cause (CC) and common effect (CE) structures used in the present study. D indicates the effect of opening the door. (a) CC3 condition, three lock cues; (b) CE3 condition, three lock cues. (c) CC4 condition, four lock cues; (d) CE4 condition, four lock cues.

Agents then must execute the selected intervention, observe the outcome, and then update the causal model of the world according to the outcome. This actively learning procedure can be implemented in many frameworks. We note the work of Bramley *et al*. [313] as an example of active causal learning. In their work, they use a Gibbs sampler to sequentially improve a single hypothesis of the causal model of the environment. The results show the method is able to capture learning properties similar to human learners. This work is one amongst very few works that look at *active* causal learning. In the following subsections, we'll look at a specific case study for active causal learning.

4.4.1 Case Study: OpenLock Task

The OpenLock environment is designed to capture challenges in causal learning for autonomous agents. Specifically, we want to examine how well humans and agents form causal abstractions necessary for task transfer. An essential motivation for studying causal relationships is that they generalize assuming constant environment dynamics. We seek to answer if it possible for learning models to acquire human-like causal knowledge in the form of abstract causal descriptions of tasks. To address this question, we designed a novel task to examine learning of action sequences governed by different causal structures, allowing us to determine in what situations humans can transfer their learned causal knowledge. Our design involves two types of basic causal structures (common cause (CC) and common effect (CE); see Fig. 4.1). When multiple causal chains are consolidated into a single structure, they can form either CC or CE schemas. Previous studies using an observational paradigm have found an asymmetry in human learning for common-cause and common-effect structures [296].

To design a novel environment for humans, we developed a virtual "escape room." Imagine that you find yourself trapped in an empty room where the only means of escape is through a door that will not open. Although there is no visible keyhole on the door—nor do you see any keys lying around—there are some conspicuous levers sticking out of the walls. Your first instinct might be to pull the levers at random to see what happens, and given the outcome, you might revise your theory about how lever interactions relate to the opening of the door. We refer to this underlying theory as a causal schema: *i.e*., a conceptual organization of events identified as cause and effect [314]. These schemas are discovered with experience and can potentially be transferred to novel target problems to infer their characteristics [315].

In the escape room example, one method of unlocking the door is to induce the causal schema connecting lever interactions to the door's locking mechanism. However, it remains unclear whether people are equally proficient in uncovering CC and CE schemas in novel situations. In the current study, we first assessed whether human causal learning can be impacted by the underlying structure,

comparing learning of a CC structure with learning of a CE structure. We then examined whether learning one type of causal structure can facilitate subsequent learning of a more complex version of the same schema involving a greater number of causal variables.

In the remainder of the chapter, we first describe the design of an experiment and report human results. Next, we describe our hierarchical Bayesian model and model results. Finally, we discuss the implications of our findings for causal learning.

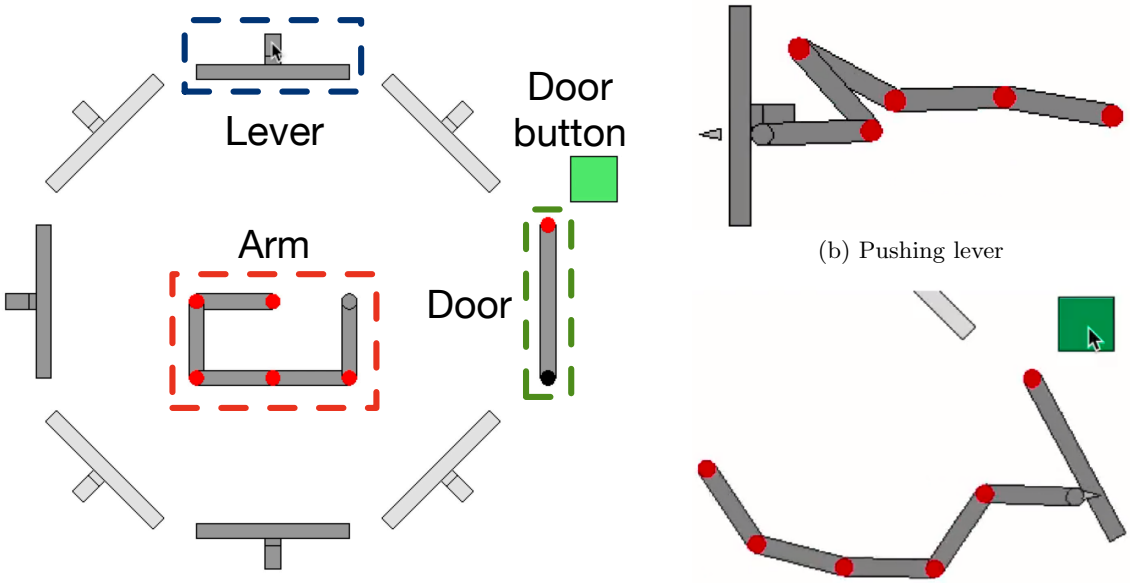
Human Experiments

In the OpenLock task, participants were asked to "escape" from a virtual room by opening a locked door that was controlled by a lever mechanism (see Fig. 4.2). The task was to figure out what level mechanisms can open the door. Each lock situation consisted of seven levers surrounding a robot arm and a door which began in a locked state. The levers pertinent to the locking mechanism (*i.e.*, active levers) were colored grey, and levers irrelevant to the locking mechanism (*i.e.*, inactive levers) were colored white. Participants were not explicitly told which levers were active or inactive but were instead required to learn the distinction through trial and error. This was not generally difficult, however, as the inactive levers could never be moved. The order in which the active levers needed to be moved followed either a common cause (CC) or common effect (CE) schema (see Fig. 4.1), and participants were given 30 attempts to discover *every* solution in each situation. Participants were instructed to consider solutions as "combinations" to each lock, and discovery of every solution/combination was required to ensure that participants understood the underlying causal schema in each situation. Participants also operated under a movement-limit constraint whereby only three movements could be used to both (1) interact with the levers (two movements) and (2) push open the door (one movement). If a participant tried to move an active lever in an incorrect order, the lever would remain stationary and a movement would be expended. Each trial reverted to its initial state once the three movements were expended, and the experiment automatically proceeded to the next trial after 30 attempts. The number of remaining solutions and attempts were provided in a console window located on the same screen as the OpenLock application.

In the environment, users commanded the movement of a simulated robot arm by clicking on desired elements in a 2D display. Levers could either be pushed or pulled by clicking on their inner or outer tracks, although pulling on a lever was never required to unlock the door. There were either 3 or 4 active levers in each lock situation. We refer to the 3- and 4-lever common cause situations as CC3 and CC4 (Fig. 4.1a, Fig. 4.1c), respectively, and the 3- and 4-lever common effect situations as CE3 and CE4 (Fig. 4.1b, Fig. 4.1d), respectively. Note that these numbers correspond with the number of *active* levers. The status of the door (*i.e.*, either locked or unlocked) was indicated by the presence or absence of a black circle located opposite the door's hinge. Once the door was unlocked and the black circle disappeared, participants could command the robot arm to push the door open by clicking on a green *push* button. The robot arm consisted of five segments that were free to rotate such that all elements in the display were easily reached by the arm's free end; the arm position control was implemented using inverse kinematics. Box2D [316] was used to handle collision, and the underlying simulation environment uses OpenAI Gym [317] as the virtual playground to train agents and enforce causal schemas through a finite state machine.

Participants were randomly assigned to one of six conditions in a between-subjects experimental design (40 participants per condition) and began the experiment by viewing a set of instructions outlining important components and details in the lock environment[1]. Fifteen additional participants were recruited but subsequently removed from the analysis due to their inability to complete

[1]The instructional video can be viewed at `https://vimeo.com/265302423`



(a) Room layout

(b) Pushing lever

(c) Pushing door

Figure 4.2: (a) Starting configuration of a 3-lever trial. All levers begin pulled towards the robot arm, whose base is anchored to the center of the display. The arm interacts with levers by either *pushing* outward or *pulling* inward. This is achieved by clicking either the outer or inner regions of the levers' radial tracks, respectively. Only push actions are needed to unlock the door in each lock situation. Light gray levers are always locked, which is unknown to both human subjects and RL at the beginning of training. Once the door is unlocked, the green button can be clicked to command the arm to push the door open. The black circle located opposite the door's red hinge represents the door lock indicator: present if locked, absent if unlocked. (b) Push to open a lever. (c) Open the door by clicking the green button.

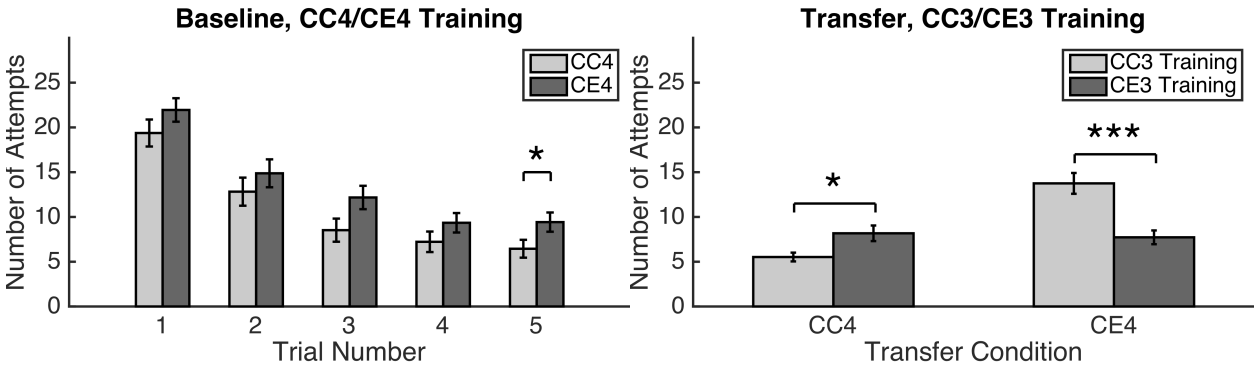


Figure 4.3: (a) Average number of attempts needed to find all unique solutions in the 4-lever common cause (CC4) and common effect (CE4) baseline conditions. Error bars indicate standard error of the mean. (b) Transfer trial results. Average number of attempts needed to find all unique solutions in the 4-lever common cause (CC4; left) and common effect (CE4; right) conditions. Light and dark grey bars indicate CC3 and CE3 training, respectively. Error bars indicate standard error of the mean.

any trial in the allotted number of attempts. The first two experimental conditions were baselines that contained five different lock situations comprised of either CC4 or CE4 trials, exclusively. These baseline conditions for the two control groups, denoted as CC4 and CE4, were included to assess whether human causal learning can be impacted by the underlying structure, comparing learning of a common-cause structure with learning of a common-effect structure. For the remaining four

conditions, we examined whether learning one type of causal structure can facilitate subsequent learning of a more complex version of the same schema involving a greater number of causal variables (*i.e.*, active levers). The four conditions contained six training trials with 3-lever situations, followed by one transfer trial with a 4-lever situation. The schema underlying the 3- and 4-lever situations was either congruent (CC3-CC4, CE3-CE4) or incongruent (CC3-CE4, CE3-CC4) and always remained the same throughout the 3-lever training trials. Participants required approximately 17.4 min to complete the baseline trials and 17.3 min to complete the training and transfer trials.

We first compared performance across the two baseline conditions where participants only completed the CC4 and CE4 trials. The average number of attempts to solve a 4-lever task in each of the baseline trials is shown in Section 4.4.1. Participants showed a clear learning effect as fewer attempts were needed for later trials, $F(4, 75) = 40.16, p < .001$. The main effect of causal structure was trending towards significance, $F(1, 78) = 3.63, p = .06$, and results from a two-sample t-test at the final trial (*i.e.*, Trial 5) indicate that the task with the CE structure took significantly more attempts to solve than the CC structure, $t(78) = 2.00$, $p < .05$. This result suggests that when a situation involved relatively high structural complexity, the CE structure was more difficult to discern than the CC structure.

Next, we examined the training performance in the four groups who completed both the training trials with 3-levers and the transfer trial with 4-levers. A clear learning improvement was found, indicated by a significant main effect of training trials, $F(5, 152) = 56.02, p < .001$. There was no difference in training performance between the CC3 and CE3 groups, $F(1, 158) = 0.11$. Compared with the two control groups in the four-lever situations, participants showed similar performance in the three-lever situations, suggesting that structural complexity impacts the comparative difficulty between CC and CE trials. For simple structures with fewer causal variables, people appear to learn different types of causal structures equally well. However, as complexity increases, some causal structures appear easier to learn than others. To further investigate whether the four training groups achieved the same level of learning, we compared the performance at the final training trial in the three-lever task. There were no differences in performance between the CC3-CC4 and CC3-CE4 groups, $t(78) = 0.87$, or the CE3-CC4 and CE3-CE4 groups, $t(78) = 0.48$. This suggests that participants in each training group had approximately the same level of understanding of the underlying causal schema before moving to their respective transfer trials.

Finally, we examined participants' transfer performance. The average number of attempts needed to solve the transfer trials are depicted in Section 4.4.1. A two-way ANOVA revealed a significant interaction effect between the training structure and the testing structure, $F(1, 156) = 24.94, p < .001$, indicating superior transfer when the same type of causal structures were used in the training and transfer trials. The resulting plot shows that participants trained under a CC3 structure performed better in the CC4 condition than those trained under a CE3 causal structure, $t(78) = 2.62, p = .01$. Similarly, participants trained under a CE3 structure performed better in the CE4 test trials than did those who trained under a CC3 structure, $t(78) = 4.27, p < .001$. Consistent with the baseline groups, there was also a significant main effect of causal structure in the transfer test, as the CE4 condition required more attempts than the CC4 condition, $F(1, 158) = 17.14, p < .001$.

Causal Theory Induction

Next, we examine a hierarchical Bayesian approach to learning causal abstractions. The hierarchy is defined by Spatial, Temporal, and Causal And-Or Graphs (STC-AOGs), where the topmost layer of the hierarchy defines the most abstract knowledge in the model. At each successive layer, the

hierarchy moves from more abstract information to more specific information about this task and environment. The bottommost layer is comprised of individual actions and their effects.

Causal theory induction provides a Bayesian account of how hierarchical causal theories can be induced from data [298, 299, 318]. The key insight is: *hierarchy enables abstraction.* At the highest level, a theory provides general background knowledge about a task or environment. Theories consist of principles, principles lead to structure, and structure leads to data.

Our agent utilizes two theories to learn a model of the OpenLock environment: (i) an instance-level associative theory regarding which attributes and actions induce state changes in the environment, denoted as the bottom-up β theory, and (ii) an abstract-level causal structure theory about which atomic causal structures are useful for the task, denoted as the top-down γ theory.

Instance-level Inductive Learning

A hypothesis space, Ω_C, is defined over possible causal chains, $c \in \Omega_C$. Each chain is defined as a tuple of subchains: $c = (c_0, \ldots, c_k)$, where k is the length of the chain and each subchain is defined as a tuple $c_i = (a_i, s_i, cr_i^a, cr_i^s)$. Each chain comprises an STC-AOG and each subchain represents an STC fragment. Each a_i is an action node that the agent can execute, s_i is a state node, cr_i^a is a causal relation that defines how a state s_i transitions under an action a_i, and cr_i^s is a causal relation that defines how state s_i is affected by changes to the previous state, s_{i-1}. Each s_i is defined by a set of time-invariant *attributes*, ϕ_i and time-varying *fluents*, f_i [319, 96, 95]; *i.e.*, $s_i = (\phi_i, f_i)$. Action nodes can be directly intervened on, but state nodes cannot. This means an agent can directly influence (*i.e.*, execute) an action, but how those actions affect the world must be *actively* learned. The structure of the general causal chain is shown in Fig. 4.4. As an example using Fig. 4.2a, if the agent executes *push* on the *upper* lever, the *lower* lever may transition from *pulled* to *pushed*, and the *left* lever may transition from *locked* to *unlocked*.

The space of states is defined as $\Omega_S = \Omega_\phi \times \Omega_F$, where the space of attributes Ω_ϕ consists of position and color, and the space of fluents Ω_F consists of binary values for lever status (*pushed* or *pulled*) and lever lock status (*locked* or *unlocked*). The space of causal relations is defined as $\Omega_{CR} = \Omega_F \times \Omega_F$, capturing the possibly binary transitions between previous fluent values and the next fluent values.

Our agent induces instance-level knowledge regarding which objects (*i.e.*, instances) can produce causal state changes through interaction (see Section 4.4.1) and simultaneously learns an abstract structural understanding of the task (*i.e.*, schemas; see Section 4.4.1). The two learning mechanisms are combined to form a causal theory of the environment, and the agent uses this theory to reason about the optimal action to select based on past experiences (*i.e.*, interventions; see Section 4.4.1). After taking an action, the agent observes the effects and updates its model of both the instance-level knowledge and the abstract structural knowledge.

The agent seeks to learn which instance-level components of the scene are associated with causal events; *i.e.*, we wish to learn a likelihood term to encode the probability that a causal event will occur. We adhere to a basic yet general associative learning theory: *causal relations induce state changes in the environment, and non-causal relations do not*, referred to as the bottom-up β theory. We learn two independent components: attributes and actions, and we assume they are independent to learn a general associative theory, rather than specific knowledge regarding an exact causal circumstance.

Our agent learns a likelihood term for each attribute ϕ_{ij} and action a_i using Dirichlet distributions because they serve as a conjugate prior to the multinomial distribution. First, a global Dirichlet parameterized by α^G is used across all trials to encode long-term beliefs about various environments. Upon entering a new trial, a local Dirichlet parameterized by $\alpha^L \in [1, 10]$ is initialized

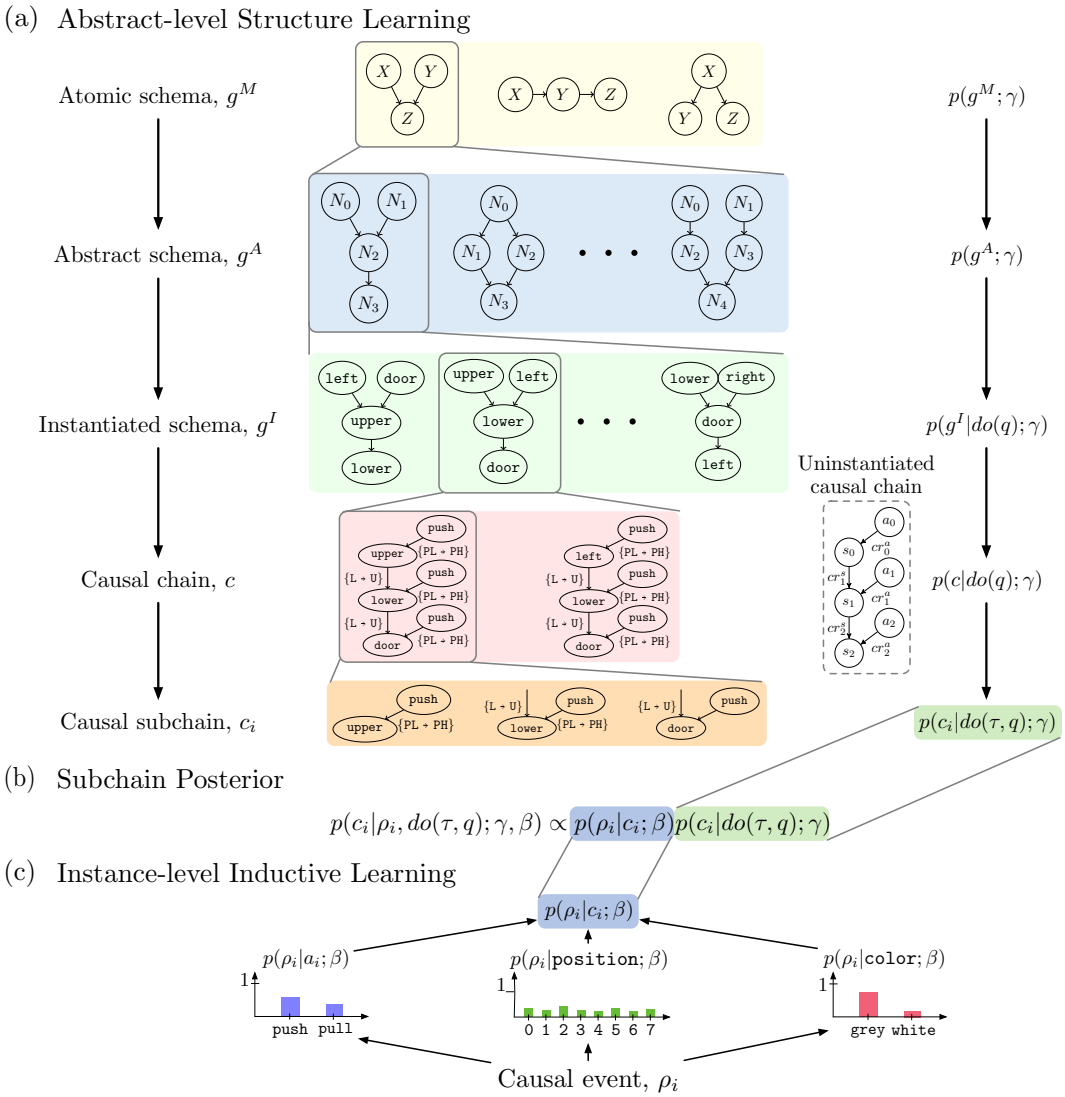


Figure 4.4: Illustration of top-down and bottom-up processes. (a) Abstract-level structure learning hierarchy. At the top, atomic schemas provide the agent with environment-invariant task structures. At the bottom, causal subchains represent a single time-step in the environment. The agent constructs the hierarchy and makes decisions at the causal subchain resolution. Atomic schemas g^M provide the top-level structural knowledge. Abstract schemas g^A are structures specific to a task, but not a particular environment. Instantiated schemas g^I are structures specific to a task and a particular environment. Causal chains c are structures representing a single attempt; an abstract, uninstantiated causal chain is also shown for notation. Each subchain c_i is a structure corresponding to a single action. `PL`, `PH`, `L`, `U` denote fluents *pulled*, *pushed*, *locked*, and *unlocked*, respectively. (b) The subchain posterior computed using the abstract-level structure learning and instance-level inductive learning. (c) Instance-level inductive learning. Each likelihood term is learned from causal events, ρ_i. Likelihood terms are combined for actions, positions, and colors.

to $k\alpha^G$, where k is a normalizing factor. Such design of using a scaled local distribution is necessary to allow α^L to adapt faster than α^G within one trial; *i.e.*, agents must adapt more rapidly to the current trial compared to across all trials. Thus, we have a set of Dirichlet distributions to maintain beliefs: a Dirichlet for each attribute (*e.g.*, position and color) as well as a Dirichlet for actions.

We introduce ρ to represent a casual event or observation occurring in the environment. Our agent wishes to assess the likelihood of a particular causal chain producing a casual event. The

agent computes this likelihood by decomposing the chain into subchains

$$p(\rho|c;\beta) = \prod_{c_i \in c} p(\rho_i|c_i;\beta), \tag{4.1}$$

where $p(\rho_i|c_i;\beta)$ is formulated as

$$p(\rho_i|c_i;\beta) \propto p(\rho_i|a_i;\beta) \prod_{\substack{\phi_{ij} \in s_i \\ s_i \in c_i}} p(\rho_i|\phi_{ij};\beta), \tag{4.2}$$

where $p(\rho_i|\phi_{ij};\beta)$ and $p(\rho_i|a_i;\beta)$ follow multinomial distributions parameterized by a sample from the attribute and action Dirichlet distribution, respectively.[2] Intuitively, this bottom-up associative likelihood encodes a naive Bayesian prediction of how likely a particular subchain is to be involved with any causal event by considering how frequently the attributes and actions have been in causal events in the past, without regard for task structure. For example, we would expect an agent in OpenLock to learn that grey levers always move and white levers never move.

Abstract-level Structure Learning

Given this understanding of how low-level attributes transfer across environments, the agent also needs to learn abstract causal structures that govern a task. We refer to these structures as *schemas*; these schemas are used to encode generalized knowledge about task structure that is invariant to a specific observational environment.

A space of atomic causal schemas, Ω_{g^M}, of causal chain, Common Cause (CC), and Common Effect (CE), serve as categories for the Bayesian prior. The belief in each atomic schema is modeled as a multinomial distribution, whose parameters are defined by a Dirichlet distribution. This root Dirichlet distribution's parameters are updated after every trial according to the top-down causal theory γ, computed as the minimal graph edit distance between an atomic schema and the trial's solution structure. This process yields a prior over atomic schemas, denoted as $p(g^M;\gamma)$, and provides the prior for the top-down inference process. Such abstraction allows agents to transfer beliefs between the abstract notions of CC and CE without considering task-specific requirements; *e.g.*, 3- or 4-lever configurations.

Next, we compute the belief in abstract instantiations of the atomic schemas. These abstract schemas share structural properties with atomic schemas but have a structure that matches the task definition. For instance, each schema must have three subchains to account for the 3-action limit imposed by the environment and should have N trajectories, where N is the number of solutions in the trial. Each abstract schema is denoted as g^A, and the space of abstract schemas, denoted Ω_{g^A}, is enumerated. The belief in an abstract causal schema is computed as

$$p(g^A;\gamma) = \sum_{g^M \in \Omega_{g^M}} p(g^A|g^M) p(g^M;\gamma). \tag{4.3}$$

The abstract structural space can be used to transfer beliefs between rooms; however, we need to perform inference over settings of positions and colors *in this trial* as the agent executes. Thus, the agent enumerates a space of instantiated schemas Ω_{g^I}, where each g^I is an instantiated schema. The agent then computes the belief in an instantiated schema as

$$p(g^I|do(q);\gamma) = \sum_{g^A \in \Omega_{g^A}} p(g^I|g^A, do(q)) p(g^A;\gamma), \tag{4.4}$$

[2]See supplementary materials for additional details.

where $do(q)$ represents the do operator [291], and q represents the solutions already executed. Conditioning on $do(q)$ constrains the space to have instantiated solutions that contain the solutions already discovered by the agent in this trial. Causal chains c define the next lower level in the hierarchy, where each chain corresponds to a single attempt. The belief in a causal chain is computed as

$$p(c|do(q);\gamma) = \sum_{g^I \in \Omega_{gI}} p(c|g^I, do(q))p(g^I|do(q);\gamma). \tag{4.5}$$

Finally, the agent computes the belief in each possible subchain as

$$p(c_i|do(\tau,q);\gamma) = \sum_{c \in \Omega_C} p(c_i|c, do(\tau,q))p(c|do(q);\gamma), \tag{4.6}$$

where $do(\tau,q)$ represents the intervention of performing the action sequence executed thus far in this attempt τ, and performing all solutions found thus far q. This hierarchical process allows the agent to learn and reason about abstract task structure, taking into consideration the specific instantiation of the trial, as well as the agent's own history within this trial.[2]

Intervention Selection

Our agent's goal is to pick the action it believes has the highest chance of (i) being causally plausible in the environment *and* (ii) being part of the solution to the task. We decompose each subchain c_i into its respective parts, $c_i = (a_i, s_i, cr_i)$. The agent combines the top-down and bottom-up processes into a final subchain posterior:

$$p(c_i|\rho_i, do(\tau,q);\gamma,\beta) \propto p(c_i|do(\tau,q);\gamma)p(\rho_i|c_i;\beta). \tag{4.7}$$

Next, the agent marginalizes over the causal relations cr_i and states s_i to obtain a final, action-level term to select interventions

$$p(a_i|\rho_i, do(\tau,q);\gamma,\beta) = \sum_{s_i \in \Omega_S} \sum_{cr_i^a \in \Omega_{CR}} \sum_{cr_i^s \in \Omega_{CR}} p(a_i, s_i, cr_i^a, cr_i^s|\rho_i, do(\tau,q);\gamma,\beta). \tag{4.8}$$

The agent uses a model-based planner to produce action sequences capable of opening the door (following human participant instructions in [320]). The goal is defined as reaching a particular state s^*, and the agent seeks to execute the action a_t to maximize the posterior subject to the constraints that the action appears in the set of chains that satisfy the goal, $\Omega_{C*} = \{c \in \Omega_C \mid s^* \in c\}$. We define the set of actions that appear in chains satisfying the goal as $\Omega_{A*} = \{a \in \Omega_A | \exists c \in \Omega_{C*}, \exists\ s, cr^a, cr^s\ |(a, s, cr^a, cr^s) \in c\}$. The agent's final planning goal is

$$a_t^* = \underset{a_i \in \Omega_{A*}}{\arg\max}\ p(a_i|\rho_i, do(\tau,q);\gamma,\beta). \tag{4.9}$$

At each time-step, the agent selects the action that maximizes this planning objective and updates its beliefs about the world using the processes described in Section 4.4.1 and Section 4.4.1. This iterative process consists of optimal decision-making based on the agent's current understanding of the world, followed by the agent updating their understanding based on the observed outcome.

In this section, we compare results between predominate reinforcement learning (RL) algorithms with the proposed theory-based causal transfer model. Specifically, we compare the proposed method against Deep Q-Network (DQN) [321], DQN with prioritized experience replay (DQN (PE)) [322], Advantage Actor-Critic (A2C) [323], Trust Region Policy Optimization (TRPO) [324], Proximal Policy Optimization (PPO) [325], and Model-Agnostic Meta-Learning (MAML) [326] agents. Below, we use the term *positive transfer* and *negative transfer* to indicate that agent performance benefits from or is hindered from the training phase, respectively.

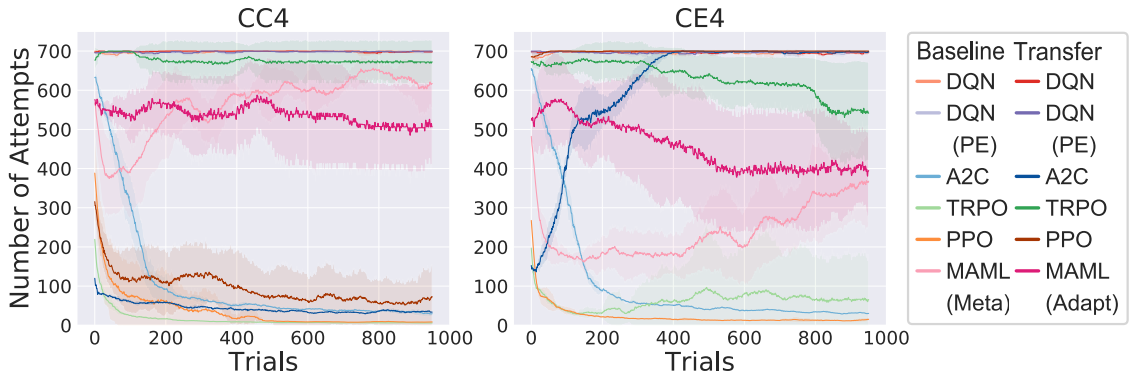


Figure 4.5: RL results for baseline and transfer conditions. Baseline (no transfer) results show the best-performing algorithms (PPO, TRPO) achieving approximately 10 and 25 attempts by the end of the baseline training for CC4 and CE4, respectively. A2C is the only algorithm to show positive transfer; A2C performed better with training for the CC4 condition. The last 50 iterations are not shown due to the use of a smoothing function.

Experimental Setup

The proposed model follows the same procedure as the one used for human studies presented in [320]. Baseline (no transfer) agents are placed in 4-lever scenarios for all trials. Transfer agents are evaluated in two phases: training and transfer. For every training trial, the agent is placed into a 3-lever trial and allowed 30 attempts to find *all* solutions. In the transfer phase, the agent is tasked with a 4-lever trial. Critically, the agent only sees each trial (room) one time, so generalizations must be formed quickly to transfer between trials successfully. See Section 4.4.1 for more details.

When executing various RL agents under this experimental setup, no meaningful learning takes place. Instead, we train RL agents by looping through all rooms repeatedly (thereby seeing each room multiple times). Agents are also allowed 700 attempts in each trial to find all solutions. During training, agents execute for 200 training iterations, where each iteration consists of looping through all six 3-lever trials. During transfer, agents execute for 200 transfer iterations, where each iteration consists of looping through all five 4-lever trials. Note that the setup for RL agents is advantageous; in comparison, both the proposed model and human subjects are only allowed 30 attempts (versus 700) during the training and 1 (versus 200) iteration for testing.

Reinforcement Learning Results

The RL results, shown in Fig. 4.5, demonstrate that A2C, TRPO, and PPO are capable of learning how to solve the OpenLock task from scratch. However, A2C in the Common Cause 4 (CC4) condition is the only agent showing positive transfer; every other agent in every condition shows negative transfer.

These results indicate that model-free RL algorithms are capable of learning how to achieve this task; however, the capability to transfer the learned abstract knowledge is markedly different from the human performance shown in [320]. Due to the overall negative transfer trends shown by nearly every RL agent, we conclude that these RL algorithms cannot capture the correct abstractions to transfer knowledge between the 3-lever training phase and the 4-lever transfer phase. It is worth noting that the RL algorithms found the Common Effect 4 (CE4) condition more difficult than CC4, a result also shown in our proposed model results and human participants.

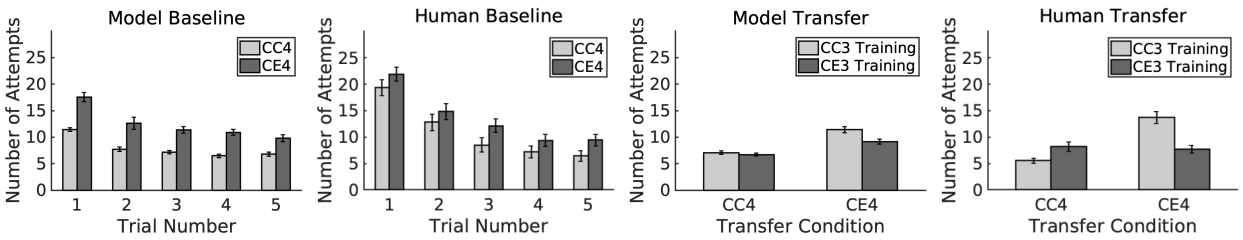


Figure 4.6: Results using the proposed theory-based causal transfer. (a) Proposed model baseline results for CC4/CE4. We see an asymmetry between the difficulty of CC and CE. (b) Human baseline performance [320]. (c) Proposed model transfer results for training in Common Cause 3 (CC3)/Common Effect 3 (CE3). The transfer results show that transferring to an incongruent condition (*i.e.*, different structure, additional lever; *e.g.*, CC3 to CE4) was always more difficult than transferring to a congruent condition (*i.e.*, same structure, additional lever; *e.g.*, CC3 to CC4). (d) Human transfer performance [320].

Theory-based Causal Transfer Results

The results using the proposed model are shown in Fig. 4.6. These results are qualitatively and quantitatively similar to the human participant results presented in [320], and starkly different from the RL results. We execute 40 agents in each condition, matching the number of human subjects described in [320].

Specifically, our agent does not require looping over trials multiple times—it is capable of learning and generalizing from seeing each trial only one time. In the baseline agents, the CC4 condition was more difficult than CE4; this trend was also observed in human participants. During transfer, we see a similar performance as the baseline results; however, for congruent cases (transferring from the same structure with an additional lever) were easier than incongruent cases (transferring to a different structure with an additional lever; CE4 transfer); this result was statistically significant for CE4: $t(79) = 3.0; p = 0.004$. For CC4 transfer, no significance was observed ($t(79) = 0.63; p = 0.44$), indicating both CC3 and CE3 obtained near-equal performance when transferred to CC4.

These learning results are significantly different from the RL results; the proposed causal theory-based model is capable of learning the correct abstraction using instance and structural learning schemes, showing similar trends as the human participants. It is worth noting that RL agents were trained under highly advantageous settings. RL agents: (i) were given more attempts per trial; and (ii) more importantly, were allowed to learn in the same trial multiple times. In contrast, the present model learns the proper mechanisms to: (i) transfer knowledge to structurally equivalent but observationally different scenarios (baseline experiments); (ii) transfer knowledge to cases with structural differences (transfer experiments); and (iii) do so using the *same experimental setup* as humans. The model accomplishes this by understanding which scene components are capable of inducing state changes in the environment while leveraging overall task structure[3].

4.5 Conclusion

In this chapter, we show how the theory-based causal transfer coupled with an associative learning scheme can be used to learn abstract, transferable structural knowledge under both observationally and structurally varying tasks. We executed a plethora of RL algorithms, none of which were capable of learning a transferable representation of the OpenLock task, even under favorable baseline and

[3]For additional model results and ablations, see supplementary.

transfer conditions. In contrast, the proposed model results are not only capable of successfully completing the task, but also adhere closely to the human participant results in [320].

These results suggest that current RL methods lack the necessary learning mechanisms to learn generalized representations in hierarchical, structured tasks. Our model results indicate human causal transfer follows similar abstractions as those presented in this work, namely learning abstract causal structures and learning instance-specific knowledge that connects this particular environment to abstract structures. The model presented here can be used in any reinforcement learning environment where: (i) the environment is governed by a causal structure, (ii) causal cues can be uncovered from interacting with objects with observable attributes, and (iii) different circumstances share some common causal properties (structure and/or attributes).

How can RL benefit from structured causal knowledge? Model-free RL is apt at learning a representation to maximize a reward within simple, non-hierarchical environments using a greedy process. Thus, current approaches do not restrict or impose learning an abstract structural representation of the environment. RL algorithms should be augmented with mechanisms to learn explicit structural knowledge and jointly optimized to learn both an abstract structural encoding of the task while maximizing rewards. Learning such structural knowledge should not only aid in learning transferable policies but also help RL perform better in hierarchical environments. We will investigate how to combine these fields as future work.

Why is CE more difficult than CC? Human participants, RL, and the proposed model all found CE more difficult than CC. A natural question is: why? We posit that it occurs from a decision-tree perspective. In the CC condition, if the agent makes a mistake on the first action, the environment will not change, and the rest of the attempt is bound to fail. However, should the agent choose the correct grey lever, the agent can choose either remaining grey levers; both of which will unlock the door. Conversely, in the CE condition, the agent has two grey levers to choose from in the first action; both will unlock the lever needed to unlock the door. However, the second action is more ambiguous. The agent could choose the correct lever, but it could also choose the other grey lever. Such complexity leads to more failure paths from a decision-tree planning perspective. The CC condition receives immediate feedback on the first action as to whether or not this plan will fail; the CE condition, on the other hand, has more failure pathways. We plan to investigate this property further, as this asymmetry was unexpected and unexplored in the literature.

Why is this task difficult for model-free RL? The OpenLock environment presented here presents many challenges to traditional RL. First, the variation of the lever configurations of trials requires learning abstractions between configurations; each trial can be thought of as a different "game" with the same causal schema. DDQN was designed to learn singular games at a time rather than transfer knowledge *between* different games [327].

Second, the environment's state and action spaces are low dimensional and discrete. This results in a discrete and sparse reward function, which makes gradient descent difficult for DDQN. In contrast to most Atari games where random actions typically move the player (or perform another typically inconsequential action), exploratory mistakes in OpenLock are very common and almost always result in failing to open the door.

Third, state changes modify the underlying mechanics of the environment; *e.g.*, for CC trials, pushing on $L0$ unlocks $L1$ and $L2$. This is unlike traditional Atari games where the visual dynamics of the environment directly influence the reward function. While this maintains the Markov property assumed in Q-learning, it requires reasoning about the latent state space of the causal schema, which is not present in most Atari games.

Fourth, humans using an optimal policy must remember their previous solutions; *i.e.*, an optimal policy is non-Markovian. If humans were using a Markovian policy, their attempts to find the second and/or third solutions should be evenly distributed with the first solution found. However, many

participants find all solutions within 2-3 attempts (finding two solutions in two attempts requires a lucky guess on the first attempt).

RL assumes the problem is Markovian and is therefore unable to remember the solutions already found. We relaxed this constraint by allowing the state space to be semi-Markovian; the number of solutions found was appended to the state space as a binary vector. However, empirically, this made no difference in performance to the fully-Markovian RL results. In fact, using any combination of the *unique solutions* reward function resulted in essentially no learning; after the agent finds a solution and takes the exact same action sequence again, they are given no reward. This means the agent only has one positive example per trial per solution, making it difficult to learn a meaningful policy during experience replay and gradient descent. However, future work should include an exploration into RL agents explicitly equipped with memory, such as a recurrent neural network (RNN). These agents may be better equipped to handle the long-term temporal constraints of finding all solutions.

Acknowledgment The authors thank Feng Gao, Chi Zhang, Prof. Keith Holyoak, and Prof. Ying Nian Wu at UCLA for their constructive discussions on causal inference and RL. The work reported herein was supported by DARPA XAI grant N66001-17-2-4029, ONR MURI grant N00014-16-1-2007, NSF grant BSC-1655300, and an NSF Graduate Research Fellowship.

Chapter 5

Tool Use

In this work, we present a new framework—task-oriented modeling, learning and recognition which aims at understanding the underlying functions, physics and causality in using objects as "tools." Given a task, such as, cracking a nut or painting a wall, we represent each object, *e.g.*, a hammer or brush, in a generative spatio-temporal representation consisting of four components: i) an affordance basis to be grasped by hand; ii) a functional basis to act on a target object (the nut), iii) the imagined actions with typical motion trajectories; and iv) the underlying physical concepts, *e.g.*, force, pressure, *etc*. In a learning phase, our algorithm observes only one RGB-D video, in which a rational human picks up one object (*i.e.*, tool) among a number of candidates to accomplish the task. From this example, our algorithm learns the essential physical concepts in the task (*e.g.*, forces in cracking nuts). In an inference phase, our algorithm is given a new set of objects (daily objects or stones), and picks the best choice available together with the inferred affordance basis, functional basis, imagined human actions (sequence of poses), and the expected physical quantity that it will produce. From this new perspective, any objects can be viewed as a hammer or a shovel, and object recognition is not merely memorizing typical appearance examples for each category but reasoning the physical mechanisms in various tasks to achieve generalization.

5.1 Introduction

[1]

In this work, we rethink object recognition from the perspective of an agent: how objects are used as "tools" in actions to accomplish a "task." Here a task is defined as changing the physical states of a target object by actions, such as, cracking a nut or painting a wall. A tool is a physical object used in the human action to achieve the task, such as a hammer or brush, and it can be any daily objects and is not restricted to conventional hardware tools. This leads us to a new framework—task-oriented modeling, learning and recognition, which aims at understanding the underlying functions, physics and causality in using objects as tools in various task categories.

Fig. 5.1 illustrates the two phases of this new framework. In a learning phase, our algorithm observes only one RGB-D video as an example, in which a rational human picks up one object, the hammer, among a number of candidates to accomplish the task. From this example, our algorithm reasons about the essential physical concepts in the task (*e.g.*, forces produced at the far end of the hammer), and thus learns the task-oriented model. In an inference phase, our algorithm is given a new set of daily objects (on the desk in (b)), and makes the best choice available (the wooden leg) to accomplish the task.

[1] Yixin Zhu and Yibiao Zhao contribute equally to this work.

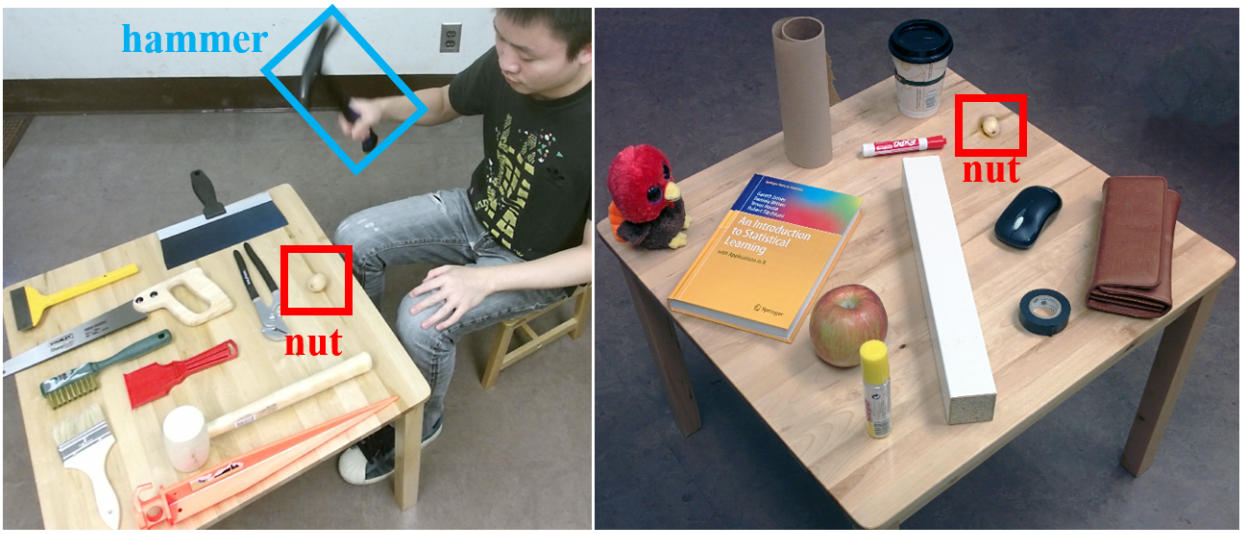


Figure 5.1: Task-oriented object recognition. (a) In a learning phase, a rational human is observed picking a hammer among other tools to crack a nut. (b) In an inference phase, the algorithm is asked to pick the best object (*i.e.*, the wooden leg) on the table for the same task. This generalization entails physical reasoning.

From this new perspective, any objects can be viewed as a hammer or a shovel, and this generative representation allows computer vision algorithms to generalize object recognition to novel functions and situations by reasoning the physical mechanisms in various tasks, and go beyond memorizing typical examples for each object category as the prevailing appearance-based recognition methods do in the literature.

Fig. 5.2 shows some typical results in our experiments to illustrate this new task-oriented object recognition framework.

Given three tasks: chop wood, shovel dirt, and paint wall, and three groups of objects: conventional tools, household objects, and stones, our algorithm ranks the objects in each group for a task. Fig. 5.2 shows the top two choices together with imagined actions using such objects for the tasks.

Our task-oriented object representation is a generative model consisting of four components in a hierarchical spatial-temporal parse graph:

i) An *affordance basis* to be grasped by hand;

ii) A *functional basis* to act on the target object;

iii) An *imagined action* with pose sequence and velocity;

iv) The *physical concepts* produced, *e.g.*, force, pressure.

In the learning phase, our algorithm parses the input RGB-D video by simultaneously reconstructing the 3D meshes of tools and tracking human actions. We assume that the human makes rational decisions in demonstration: picks the best object, grasps the right place, takes the right action (poses, trajectory and velocity), and lands on the target object with the right spots. These decisions are nearly optimal against a large number of compositional alternative choices. Using a ranking-SVM approach, our algorithm will discover the best underlying physical concepts in the human demonstration, and thus the essence of the task.

In the inference stage, our algorithm segments the input RGB-D image into objects as a set of candidates, and computes the task-oriented representation—the optimal parse graph for each candidate and each task by evaluating different combinations. This parse graph includes the best

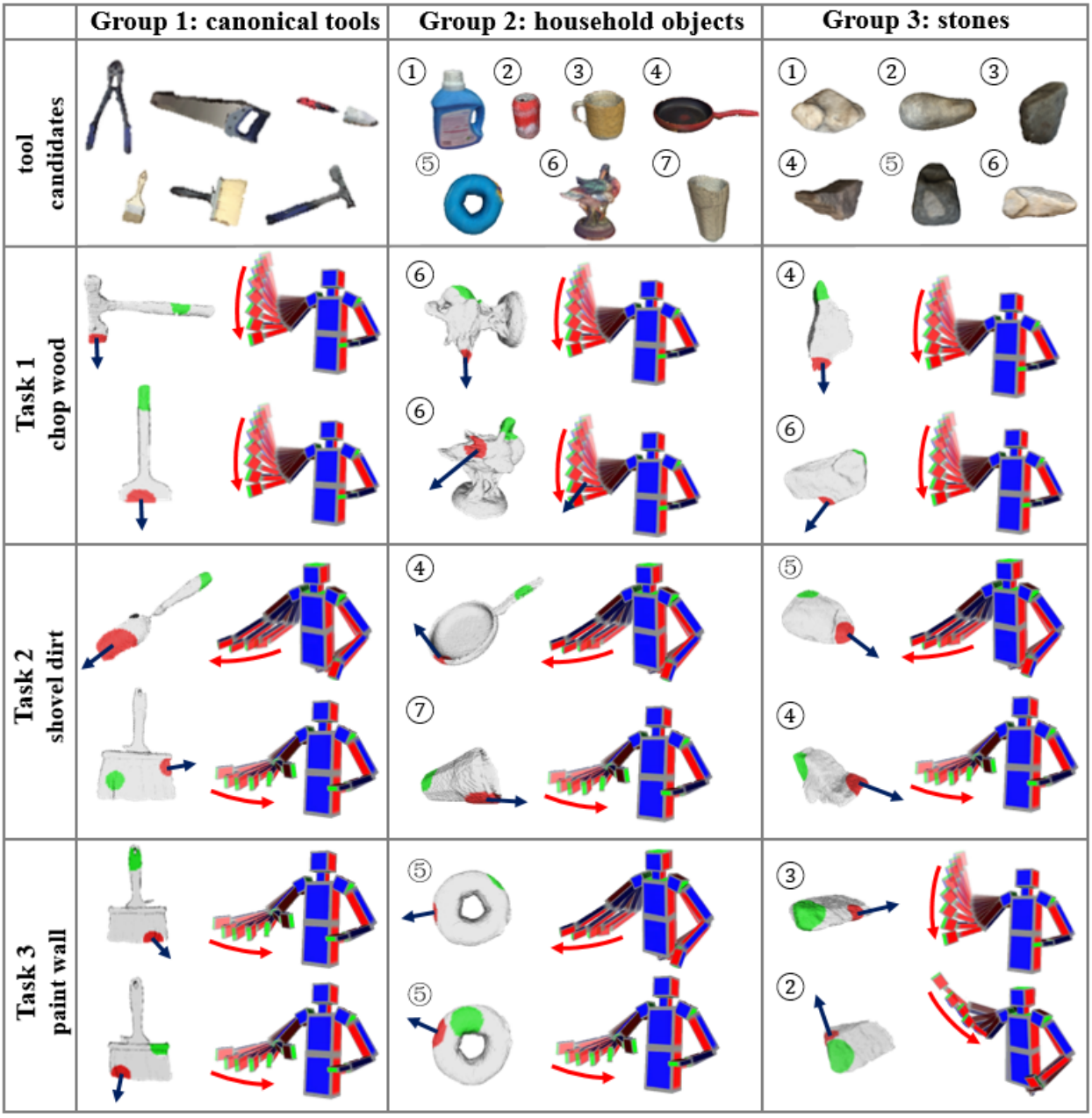


Figure 5.2: Given three tasks: chop wood, shovel dirt, and paint wall. Our algorithm picks and ranks objects for each task among objects in three groups: 1) conventional tools, 2) household objects, and 3) stones, and output the imagined tool-use: affordance basis (the green spot to grasp with hand), functional basis (the red area applied to the target object), and the imagined action pose sequence.

object and its tool-use: affordance basis (green spot), functional basis (red spot), actions (pose sequence), and the quantity of the physical concepts produced by the action.

This work has four major contributions:

1. We propose a novel problem of task-oriented object recognition, which is more general than defining object categories by typical examples, and is of great importance for object manipulation in robotics applications.
2. We propose a task-oriented representation which includes both the visible object and the imag-

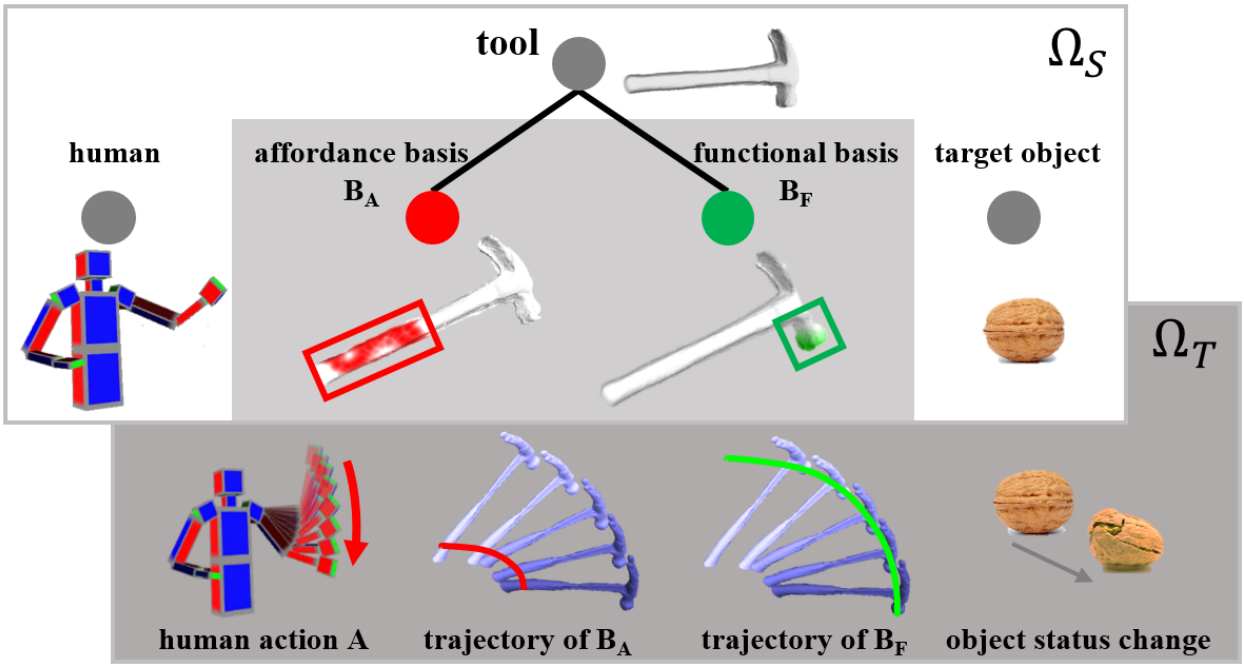


Figure 5.3: The task-oriented representation of a hammer and its use in a task (crack a nut) in a joint spatial, temporal, and causal space. The components in grey area are imagined during inference phase.

ined use (action and physics). The latter is the "dark matter" [157] in computer vision.

3. Given an input object, our method can imagine the plausible tool-use and thus allows vision algorithms to reason innovative use of daily object—a crucial aspect of human and machine intelligence.
4. Our algorithm can learn the physical concepts from a single RGB-D video and reason about the essence of physics for a task.

5.2 Task-oriented object representation

Tools and tool-uses are traditionally studied in cognitive science [328, 329, 330, 331] with verbal definitions and case studies, and an explicit formal representation is missing in the literature.

In our task-oriented modeling and learning framework, an object used for a task is represented in a joint spatial, temporal, and causal parse graph $pg = (pg_s, pg_t, pg_c)$ including three aspects shown in Fig. 5.3:

i) A spatial parse graph pg_s represents object decomposition and 3D relations with the imagined pose;

ii) A *temporal parse graph* pg_t represents the pose sequence in actions; and

iii) A *causal parse graph* pg_c represents the physical quantities produced by the action on the target object.

In this representation, only the object is visible as input, all other components are imagined.

5.2.1 Tool in 3D space

An object (or tool) is observed in a RGB-D image in the inference stage, which is then segmented from the background and filled-in to become a 3D solid object denoted by $\mathbf{X}$. The 3D object is then decomposed into two key parts in the spatial parse graph pg_s:

1) Affordance basis $\mathbf{B_A}$, where the imagined human hand grasps the object with certain pose. Through offline training, we have collected a small set of hand poses for grasping. The parse graph pg_s encodes the 3D positions and 3D orientations between the hand poses and the affordance basis during the tool-use, using 3D geometric relations between the hand pose and the affordance basis, as it is done in [153].

The parse graph pg_s will have lower energy or high probability when the hand hold the object comfortably (see the trajectory of affordance basis $\mathbf{B_A}$ in Fig. 5.3).

2) Functional basis $\mathbf{B_F}$, where the object (or tool) is applied to a target object (the nut) to change its physical state (*i.e.*, fluent). The spatial parse graph pg_s also encodes the 3D relations between the functional basis $\mathbf{B_F}$ and the 3D shape of the target object during the action. We consider three types of the functional basis:

(a) a single contact spot (*e.g.*, hammer); (b) a sharp contacting line segment or edge (*e.g.*, axe and saw); and (c) flat contacting area (*e.g.*, shovel).

We define a space $\Omega_S = \{pg_s\}$ as the set of all possible spatial parse graph pg_s which is a product space of all the possible objects, their affordance bases, functional bases, hand poses, and 3D relations above.

5.2.2 Tool-use in time

A tool-use is a specific action sequence that engages the tool in a task, and is represented by a temporal parse graph pg_t. pg_t represents the human action $\mathbf{A}$ as a sequence of 3D poses. In this work, since we only consider hand-hold objects, we collect some typical action sequences for the arm and hand movements using tools by RGB-D sensors, such as, hammering, shoveling, *etc.* These actions are then clustered into average pose sequences. For each of the sequence, we record the trajectories of the hand pose (or affordnace basis) and the functional basis.

We define a space $\Omega_T = \{pg_t\}$ as the set of possible pose sequences and their associated trajectories of the affordance basis B_A and functional basis B_F.

5.2.3 Physical concept and causality

We consider of thirteen basic physical concepts involved in tool-use, which can be extracted or derived from the spatial and temporal parse graphs as Fig. 5.4 illustrates.

Firstly, as the blue dots and lines in Fig. 5.4 illustrates, we reconstruct the 3D mesh from the input 3D object and thus calculate its volume, and by estimating its material category, we get its density. From volume and density we further calculate the mass of the objects and its parts (when different materials are used).

Secondly, as the green dots and lines Fig. 5.4 illustrates, we can derive the displacement from the 3D trajectory of affordance basis and functional basis, and then calculate the velocity and acceleration of the two bases.

Thirdly, as red dots and line shows, we can estimate the contact spot, line and area from the functional basis and target object, and further compute the momentum, and impulse. We can then also compute basic physical concepts, such as forces, pressure, work, *etc.*

Physical concept operators ∇. We define a set of operators, including addition $\nabla_+(\cdot,\cdot)$, subtraction $\nabla_-(\cdot,\cdot)$, multiplication $\nabla_\times(\cdot,\cdot)$, division $\nabla_/(\cdot,\cdot)$, negation $\nabla_{\text{neg}}(\cdot)$, space integration $\nabla_{\int_S}(\cdot)$, time integration $\nabla_{\int_T}(\cdot)$, space derivation $\nabla_{\partial_S}(\cdot)$ and time derivation $\nabla_{\partial_T}(\cdot)$. For example, the concept of the force and acceleration are defined as: force $= \nabla_\times(\text{mass}, \text{acceleration})$, acceleration $= \nabla_{\partial_t}(\text{velocity})$

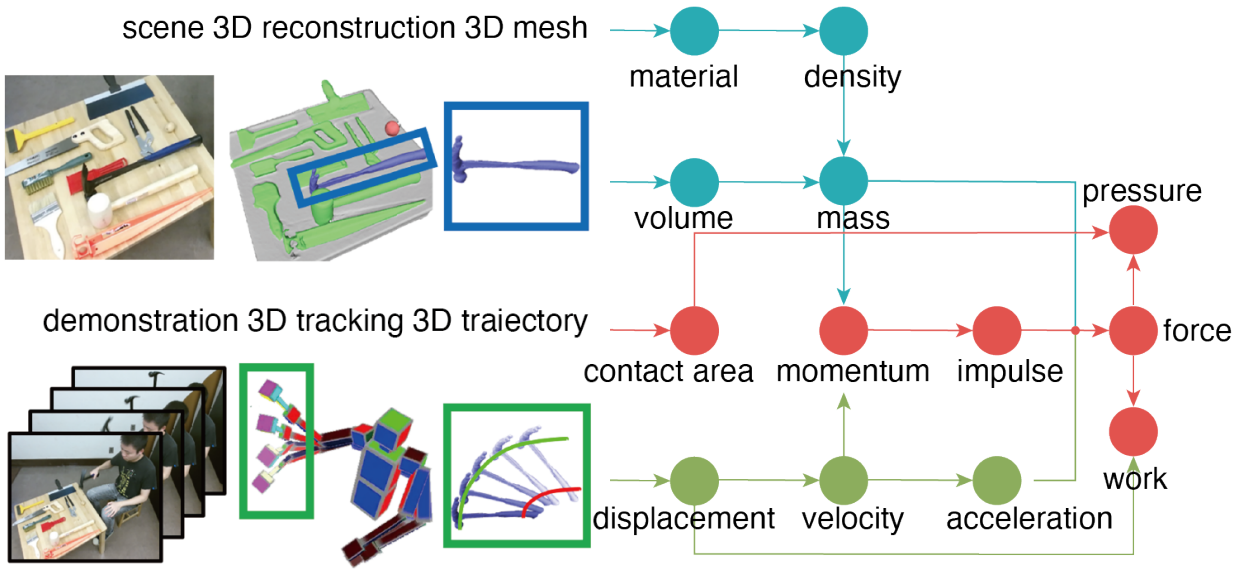


Figure 5.4: Thirteen physical concepts involved in tool-use and their compositional relations. By parsing human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool (blue), trajectories of tool-use (green) or jointly (red). The higher-level physical concepts can be further derived recursively.

The causal parse graph pg_c includes the specific physical concepts used in a tool-use which is often an instantiated sub-graph of the concept graph in Fig. 5.4.

Since the law of physics is universally applicable, the major advantage of using physical concepts is the ability to generalize to novel situations.

5.3 Problem definition

5.3.1 Learning physical concept

Given a task, the goal of the learning algorithm is to find the essential physical concept that best explains why a selected tool and tool-use is optimal.

Rational choice assumption states that human choices are rational and near-optimal. As shown in Fig. 5.5 (a-d), we assume that human chooses the optimal tool and tool-use pg^* (in blue box) based on the essential physical concept, so that most of other tools and tool-uses in the hypothesis spaces should not outperform the demonstration.

For instance, let us assume the essential physical concept to explain the choice of a tool is to maximize "mass," then other tools should not offer more "mass" than the selected one. If there is a heavier tool not picked by human, it implies that "mass" is not the essential physical concept.

During learning stage, we consider the selected tool and tool-use as the only positive training example, and we randomly sample n different combinations of tools and tool-uses pg_i, $i = 1 \cdots n$ in the hypothesis spaces as negative training samples.

Ranking function. Based on the rational choice assumption, we pose the tool recognition as a ranking problem [332], so that the human demonstration should be better than other tools and tool-uses with respect to the learned ranking function.

The goal of the learning is to find a ranking function indicating the essential purposes of tool-use

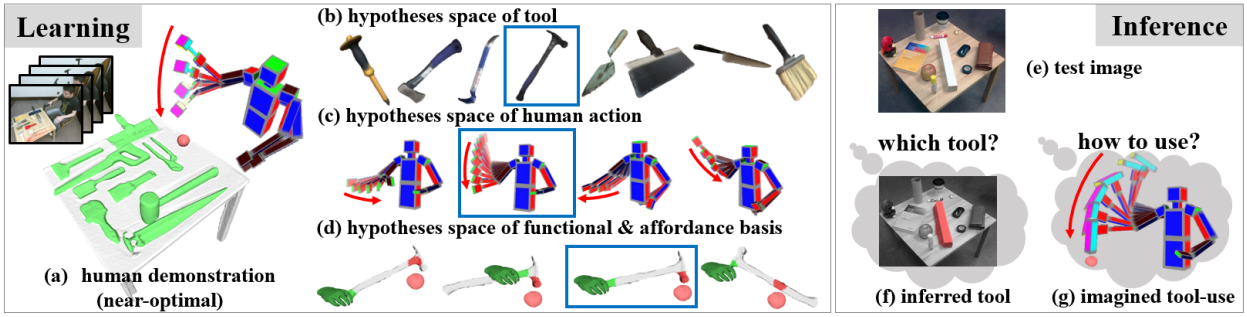


Figure 5.5: An illustration of learning and inference. (a)–(d) We assume the human choice (shown in blue bounding box) of tool and tool-use (action and affordance / functional bases) is near-optimal, thus most of other combinations of tool and tool-use (action, affordance / functional bases) in the hypotheses spaces should not outperform human demonstration. Based on this assumption, we treat the human demonstration as positive example, and random sample other tools and tool-uses in the hypothesis spaces as negative examples. (e) During the inference, given an image of static scene in a novel situation, (f) the algorithm infers the best tool and imagines the optimal tool-use.

in a given task.

$$R(pg) = \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg), \tag{5.1}$$

where $\boldsymbol{\omega}$ are the weighting coefficients of the physical concepts. Intuitively, each coefficient reflects the importance of its corresponding physical concept for the task.

Learning ranking function is equivalent to find the weight coefficients so that the maximum number of pairwise constraints is fulfilled.

$$\forall i \in \{1, \cdots, n\} : \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg^*) > \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg_i) \tag{5.2}$$

In this way, these constraints enforce the human demonstration pg^* has the highest ranking score compared with the other negative samples pg_i under the essential physical concept.

We approximate the solution by introducing nonnegative slack variables, similar to SVM classification [332]. This leads to the following optimization problem

$$\min \quad \frac{1}{2}\, \boldsymbol{\omega} \cdot \boldsymbol{\omega} + \lambda \sum_{i}^{n} \xi_i^2 \tag{5.3}$$

$$\text{s.t.} \quad \forall i \in \{1, \cdots, n\} :$$

$$\boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg^*) - \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg_i) > 1 - \xi_i^2 \tag{5.4}$$

$$\xi_i \geqslant 0, \tag{5.5}$$

where ξ_i is a slack variable for each constraint, and λ is the trade-off parameter between maximizing the margin and satisfying the rational choice constraints.

This is a general formulation for the task-oriented modeling and learning problem, where the parse graph pg includes objects $\mathbf{X}$, human action $\mathbf{A}$ and affordance / functional basis B_A / B_F. In this way, this framework subsumes following special cases: i) object recognition based on appearance and geometry $\boldsymbol{\phi}(\mathbf{X})$, ii) action recognition $\boldsymbol{\phi}(\mathbf{A})$, iii) detecting furniture by their affordance $\boldsymbol{\phi}(B_A)$, and iv) physical concept $\boldsymbol{\phi}(pg_c)$. In this work, we only focus on learning physical concepts.

In our experiment, we only consider the scenario that the learner only observes one demonstration of the teacher choosing one tool from a few candidates. Instead of feeding a large dataset for training, we are more interested in how much the algorithm can learn from such a small sample

learning problem. Therefore, we only infer a single physical concept for functional and affordance basis respectively by iterating over the concept space, while this formulation can be naturally generalized to more sophisticated scenarios for future study.

5.3.2 Recognizing tools by imagining tool-uses

Traditional object recognition methods assume that visual patterns of the objects in both training and testing sets share the same distribution. However, such assumption does not hold in tool recognition problem. The visual appearances of tools at different situations have fundamental differences. For instance, a hammer and a stone can be used to crack a nut, despite the fact the their appearances are quite different.

In order to address this challenge, we propose this algorithm to recognize tools by essential physical concepts and imagine tool-uses during the inference.

Recognize tools by essential physical concepts. Fortunately, as domain general mechanisms, the essential physical concepts in a given task are invariant across different situations. For instance, a hammer and a stone can be categorized as the same tool to crack a nut due to the similar ability to provide enough "force." In the inference, we use the learned ranking function to recognize the best tool.

$$pg^* = \arg\max \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg), \tag{5.6}$$

Imagine tool-use beyond observations. Given an observed image of tool without actually seeing the tool-use, our algorithm first imagines different tool-uses (human action and affordance / functional bases), and then combines the imagined tool-uses with observed tools to recognize the best tool by evaluating the ranking function.

The imagined tool-uses are generated by sampling human action and affordance/functional bases from the hypothesis spaces as shown in Fig. 5.5 (c-d). We first assign the trajectories of imaged human hand movement to the affordance basis, then compute the trajectory of functional basis by applying the relative 3D transformation between the two bases. Lastly, we calculate the physical concepts recursively as discussed in Section 5.2.3, and evaluate the ranking function accordingly.

The ability of imagining tool-use is particularly important for an agent to predict how they can use a tool, and physically interact with their environment.

Moreover, such ability of imagining tool-use enables the agent to actively explore different kinds of tool-uses instead of to simply mimic the observed tool-use in human demonstration. Although the tool-use in human demonstration is assumed to be optimal, other tool-uses may be better in different situations. For example, the way you use a stone to crack a nut may be quite different from the way you use a hammer.

5.3.3 Parsing human demonstration

In this section we show how we use the off-the-shelf computer vision algorithms to parse the input RGB-D video of human demonstration.

3D reconstruction. We apply the KinectFusion algorithm [271] to generate a 3D reconstruction of the static scene, including a tool and an object. KinectFusion is GPU optimized such that it can run at interactive rates. Each frame of depth image captured by RGB-D sensors has a lot of missing data. By moving the sensor around, the KinectFusion algorithm fills these holes by combining temporal frames into a smooth 3D point cloud / mesh (Fig. 5.6 (a)). In this work, we only focus on medium sized tool that can be held in one hand, and can be well reconstructed by

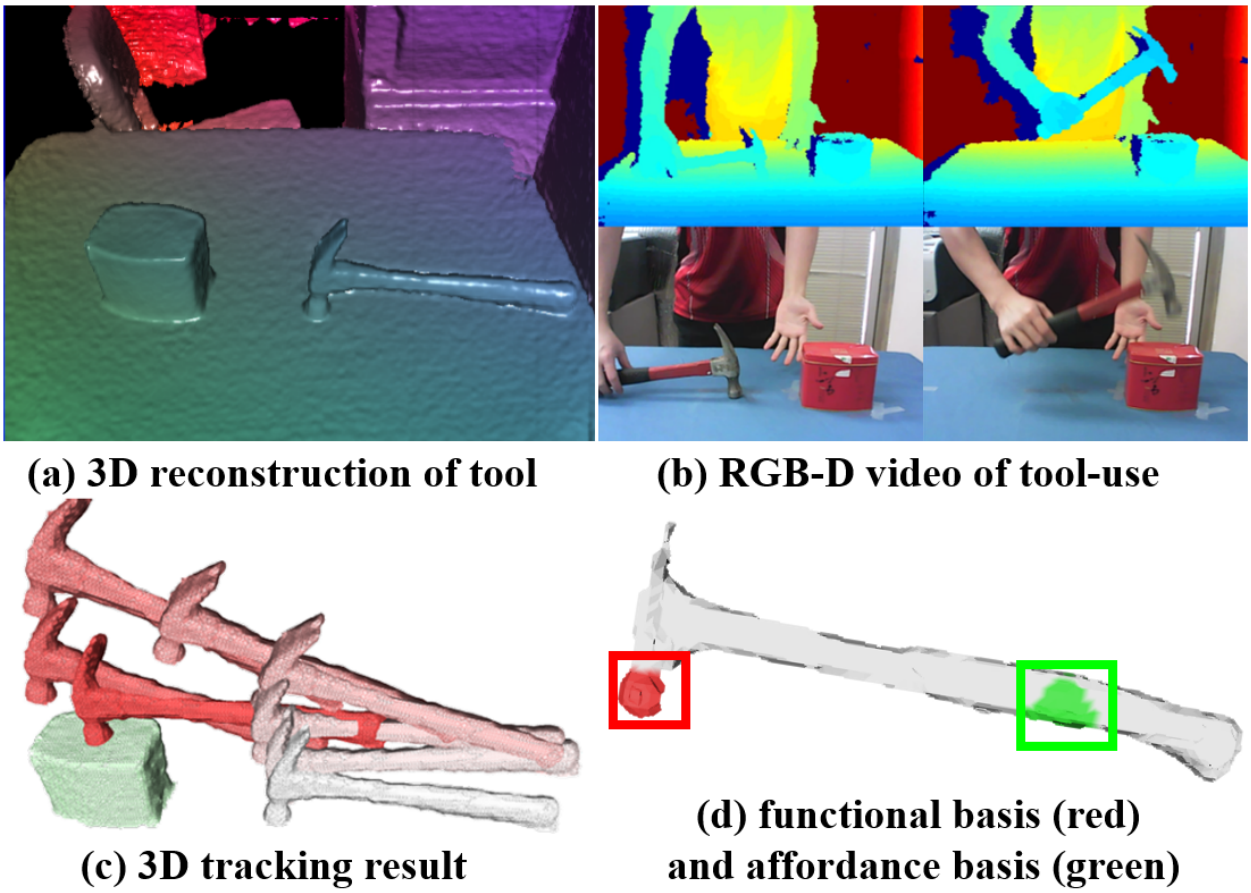


Figure 5.6: Spatial-temporal parsing of human demonstration. (a) Using KinectFusion, we first reconstruct 3D scene, including the tool and the target object. (b) Given a RGB-D video of tool-use by human demonstration, (d) affordance / functional bases can be detected by (c) 3D tracking.

a consumer-level RGB-D sensor. By fitting the plane of the table, the tool and the target object then can be extracted from background.

3D tracking of tool and target object. Tracking the 3D mesh of tool and target object allows the algorithm to perceive the interactions and detect status changes. In this work, we use an off-the-shelf 3D tracking algorithm based on Point Cloud Library [333]. The algorithm first performs object segmentation using the first depth frame of the RGB-D video, and then invokes particle filtering [334] to track each object segment as well as estimating the 3D orientation frame by frame (Fig. 5.6 (c)).

3D hand tracking. 3D tracking of hand positions and orientations are achieved by 3D skeleton tracking [155]. The skeleton tracking outputs a full body skeleton, including 3D position and orientation of each joint. Without loss of generality, we assume the interacting hand to be the right hand.

Contact detection. Given the tracked 3D hand pose / tool / target object, we perform touch detection (Fig. 5.6 (d)) by measuring the euclidean distance among them. The touch detection between the human hand and the tool localizes the 3D location of the affordance basis, while the touch detection between the tool and the target object yields the 3D location of the functional basis.

5.4 Experiment

In this section, we first introduce our dataset, and evaluate our algorithm in three aspects: (i) learning physical concepts; (ii) recognizing tools; and (iii) imagining tool-uses.

5.4.1 Dataset

We designed a new Tool & Tool-Use (TTU) dataset for evaluating the recognition of tools and task-oriented objects. The dataset contains a collection of static 3D object instances, together with a set of human demonstrations of tool-use.

The 3D object instances include 452 static 3D meshes, ranging from typical tools, household objects and stones. Some of these object instances are shown in Fig. 5.7. Some typical actions are illustrated in Fig. 5.5. Each action contains a sequence (3-4 seconds) of full body skeletons.

5.4.2 Learning physical concept

We first evaluate our learning algorithm by comparing with human judgments. Forty human subjects annotated the essential physical concepts for four different tasks, the distribution of annotated the essential physical concepts is shown as the blue bars in Fig. 5.8. Interestingly, human subjects have relative consistent common knowledge that force and momentum are useful for cracking nuts, and pressure is important for chopping wood. Our algorithm learned very similar physical concepts as the red bars shown in Fig. 5.8. For the other two tasks *i.e.*, shovel dirt and paint wall, although the human judgments are relatively ambiguous, our algorithm still produces relative similar results of learned physical concepts.

Fig. 5.9 shows an example of learning physical concept for cracking a nut. Given a set of RGB-D images of ten tool candidates in Fig. 5.9 (a) and a human demonstration of tool-use in Fig. 5.9 (b), our algorithm imagines different kinds of tool-use as shown in Fig. 5.9 (c), and ranks them with respect to different physical concepts. By assuming human demonstration is rational and near-optimal, our learning algorithm selects physical concepts by minimizing the number of violations as the red area on the left of Fig. 5.9 (c). For instance, the plot of "force" shows ranked pairs of tool and tool-use with respect to the forces applied on the functional basis. The force produced by human demonstration (the black vertical line) is larger than most of the generated tool-uses, thus it is near-optimal. The instances on the right of Fig. 5.9 (c) are sampled tools and tool-uses. The red ones are the cases outperform human demonstration, while the gray ones are the cases underperform human demonstration.

5.4.3 Inferring tools and tool-uses

In the Fig. 5.2, we illustrate qualitative results of inferred tool and tool-use for three tasks, *i.e.* chop wood, shovel dirt, and paint wall. By evaluating in three scenarios: (a) typical tools, (b) household objects, (c) natural stones, we are interested in the generalization ability of the learned model.

Recognizing tools

We asked four human subjects to rank tool candidates shown in Fig. 5.2. For the task of chopping wood in Fig. 5.10, we plot tool candidates in terms of their average ranking by human subjects (x-axis) and their ranking generated by our algorithm (y-axis).

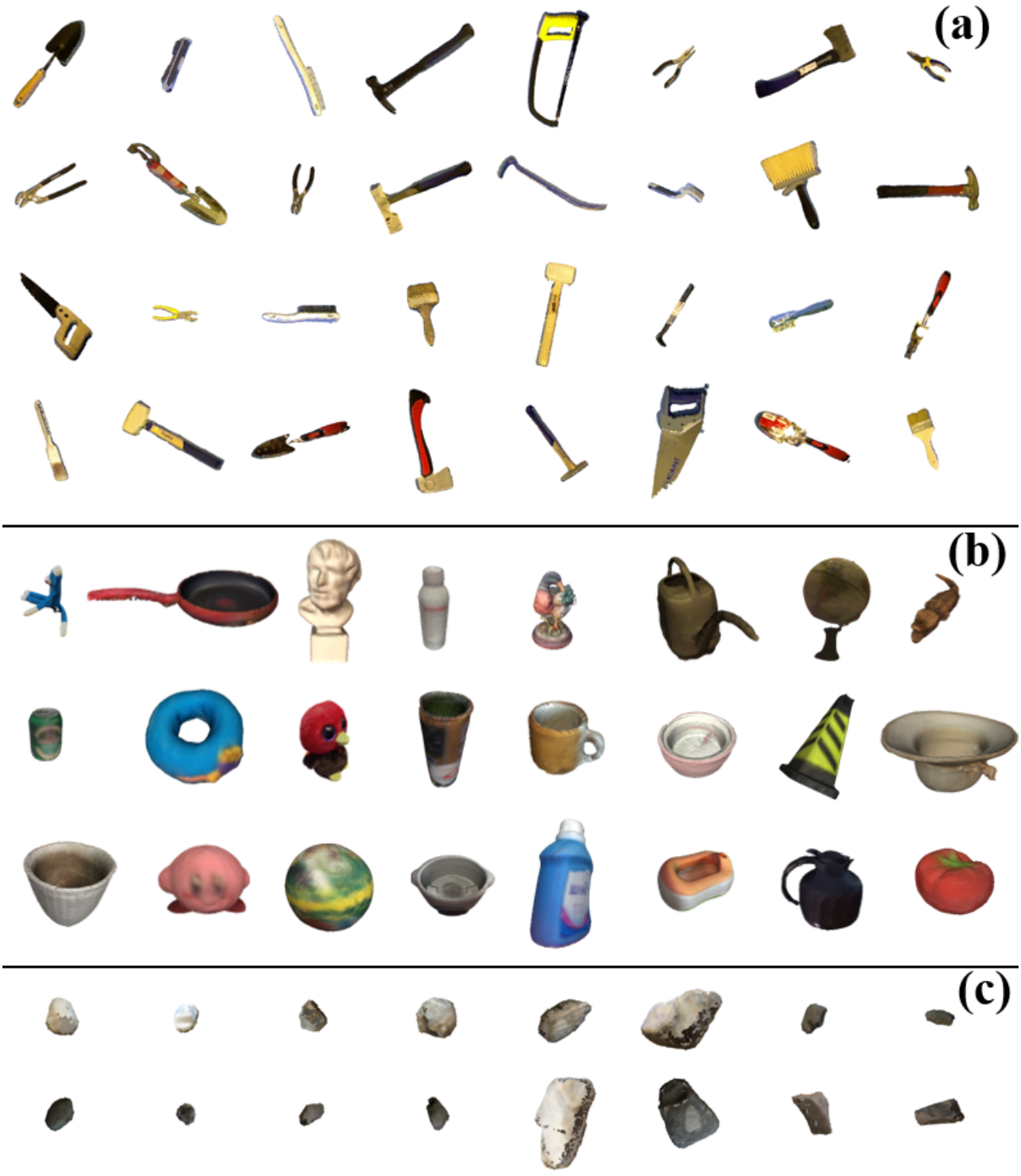


Figure 5.7: Sample tool instances in dataset. (a) typical tools (b) household objects (c) natural stones.

The three columns show different testing scenarios. We can see that our model learned from canonical cases of tool-use can be easily generalized to recognize tools in novel situation, *i.e.*, household objects and natural stones. The correlation between algorithm ranking and human ranking is consistent across these three scenarios. Sometimes, the algorithm works even better on the stone scenarios.

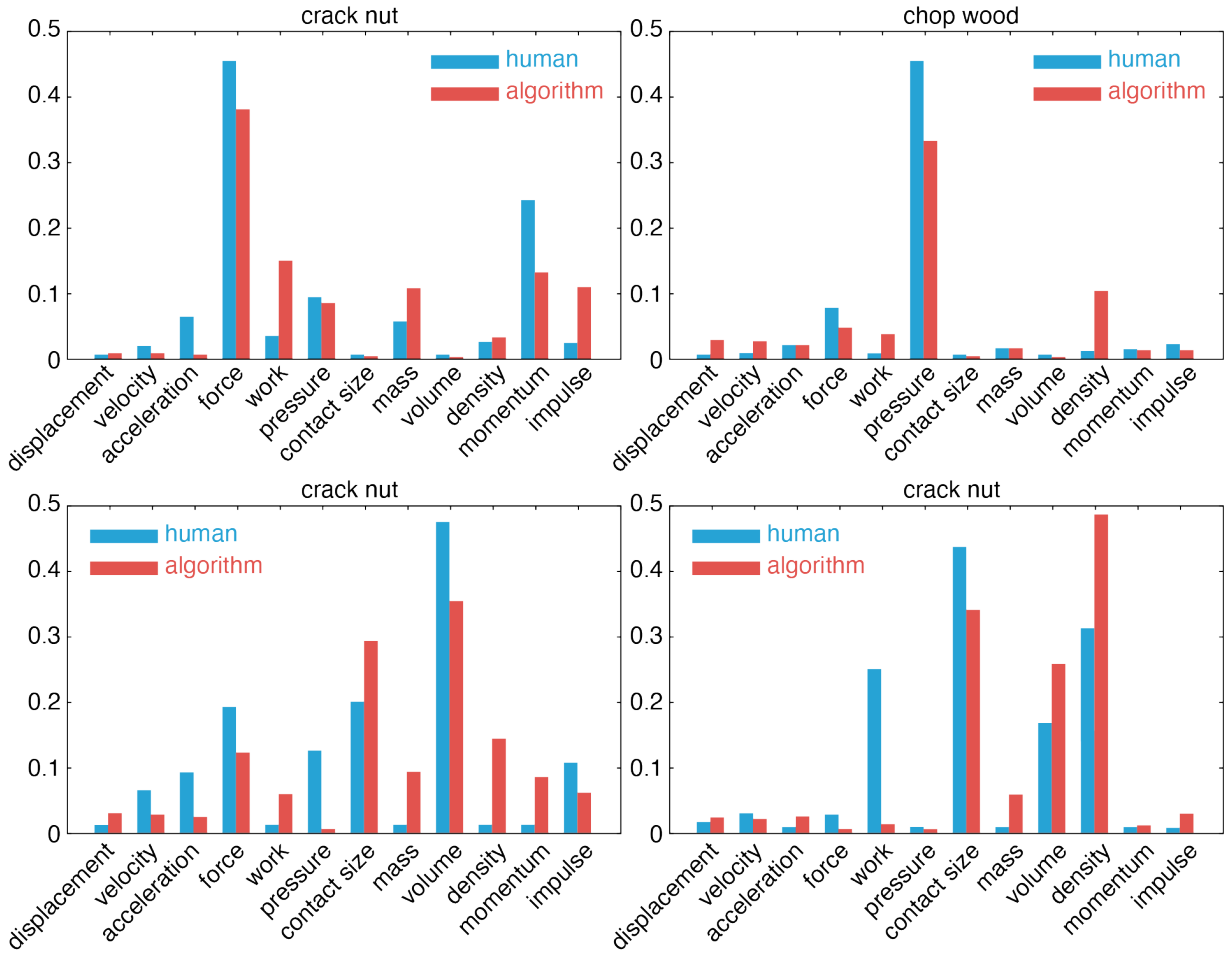


Figure 5.8: Learning essential physical concepts of tool-use. The red bars represent human judgments about what the essential physical concepts are for each task. The blue bars represent weight coefficients of different physical concepts learned by our algorithm.

The three rows represent different levels of tool-use: (a) the "tool-ranking with random use" evaluates the ranking of tools by calculating the expected scores of random tool-use; (b) the "tool-ranking with inferred use" evaluates the ranking of tools by calculating their optimal tool-use inferred by our algorithm; (c) the "tool-ranking with best use" evaluates the ranking of tools by their best uses given by human subjects. The Table 5.1 summarizes the correlations between human rankings and algorithm rankings on three tasks.

Table 5.1: Accuracy of tool recognition. This table shows the correlation between the ranking generated by our algorithm and the average ranking annotated by human subjects. The three rows represent different levels of tool-use imagined by our inference algorithm. The qualitative and quantitative ranking results of tool candidates are illustrated in Fig. 5.2 and Fig. 5.10 respectively.

correlation of ranking	chop wood			shovel dirt			paint wall		
algorithm vs. human	tool	object	stone	tool	object	stone	tool	object	stone
tool + random use	0.07	0.14	0.20	0.52	0.32	0.09	0.12	0.11	0.31
tool + inferred use	0.48	0.25	0.89	0.64	0.89	0.14	0.10	0.64	0.20
tool + best use	0.83	0.43	0.89	0.64	0.89	0.14	0.10	0.64	0.20

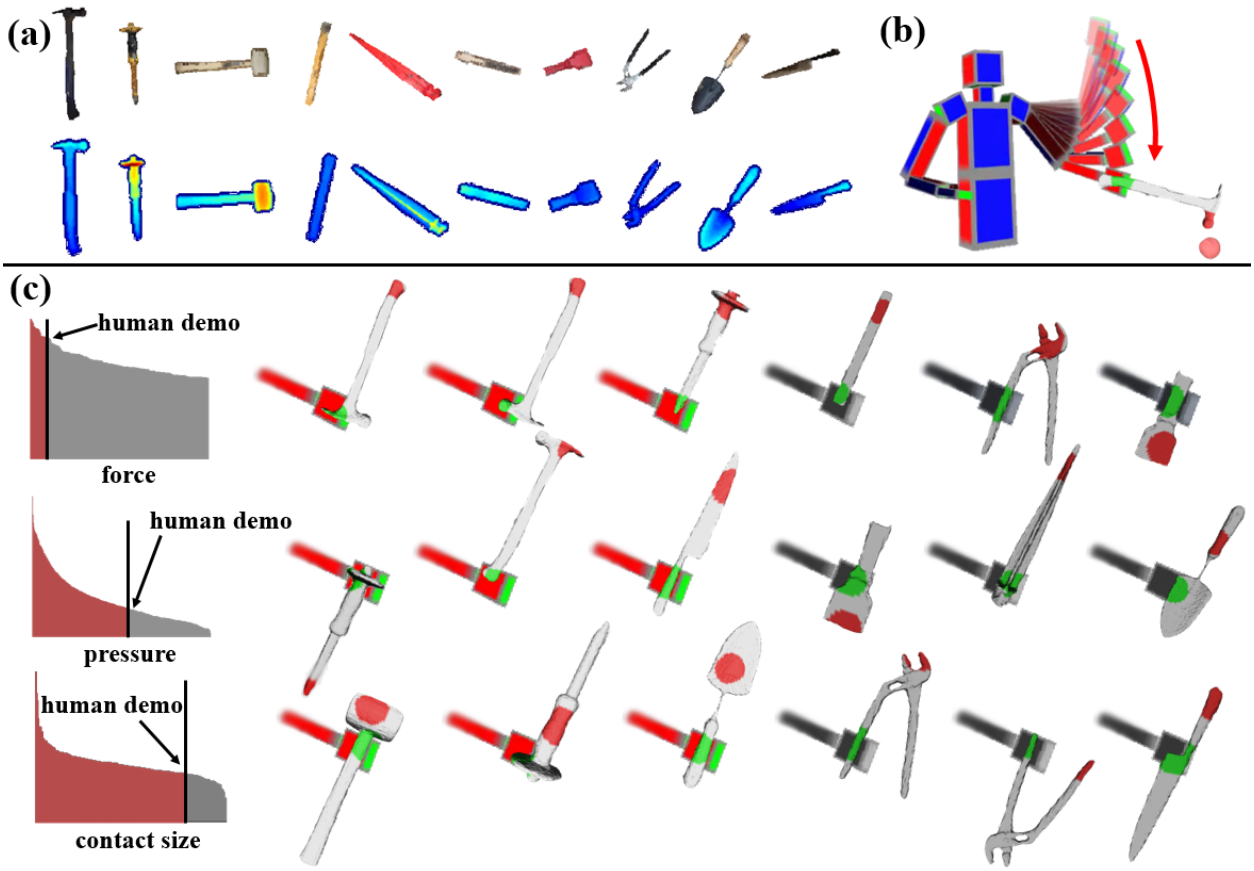


Figure 5.9: Learning physical concept from single human demonstration for cracking a nut. (a) A set of tool candidates are given by RGB-D images. (b) The human demonstration of tool-use is assumed to be near-optimal. (c) The algorithm sorts all the samples of tool-uses with respect to different physical concepts. The black vertical bar represents the human demonstration of tool-use, while the red area and gray area represent samples that outperform and underperform human demonstration receptively. We showed six sampled tool and tool-use, three of which outperform human demonstration, and the others underperform human demonstration. In this cracking nut example, the "forces" is selected as the essential physical concept because there are minimum number of samples that violate the "rational choice assumption."

Imagining tool-uses

We also evaluated the imagined tool-uses in three aspects: human action A, affordance basis B_A, functional basis B_F.

The evaluation of human action is based on the classification of action directions, which are "up," "down," "forward," "backward," "left" and "right." The classification accuracy for this problem over all the experiments is 89.3%. The algorithm can reliably classify the action of cracking a nut as "down." But there are some ambiguities in classifying the action of shoveling dirt, because "left" and "right" are physically similar.

The Fig. 5.11 illustrates three example of imagined affordance basis B_A and functional basis B_F. Comparing to human annotations, the algorithm finds very similar positions of affordance basis B_A and functional basis B_F respectively. In Table 5.2 we show the 3D distances between the positions imagined by our algorithm and the positions annotated by human subjects in centimeter.

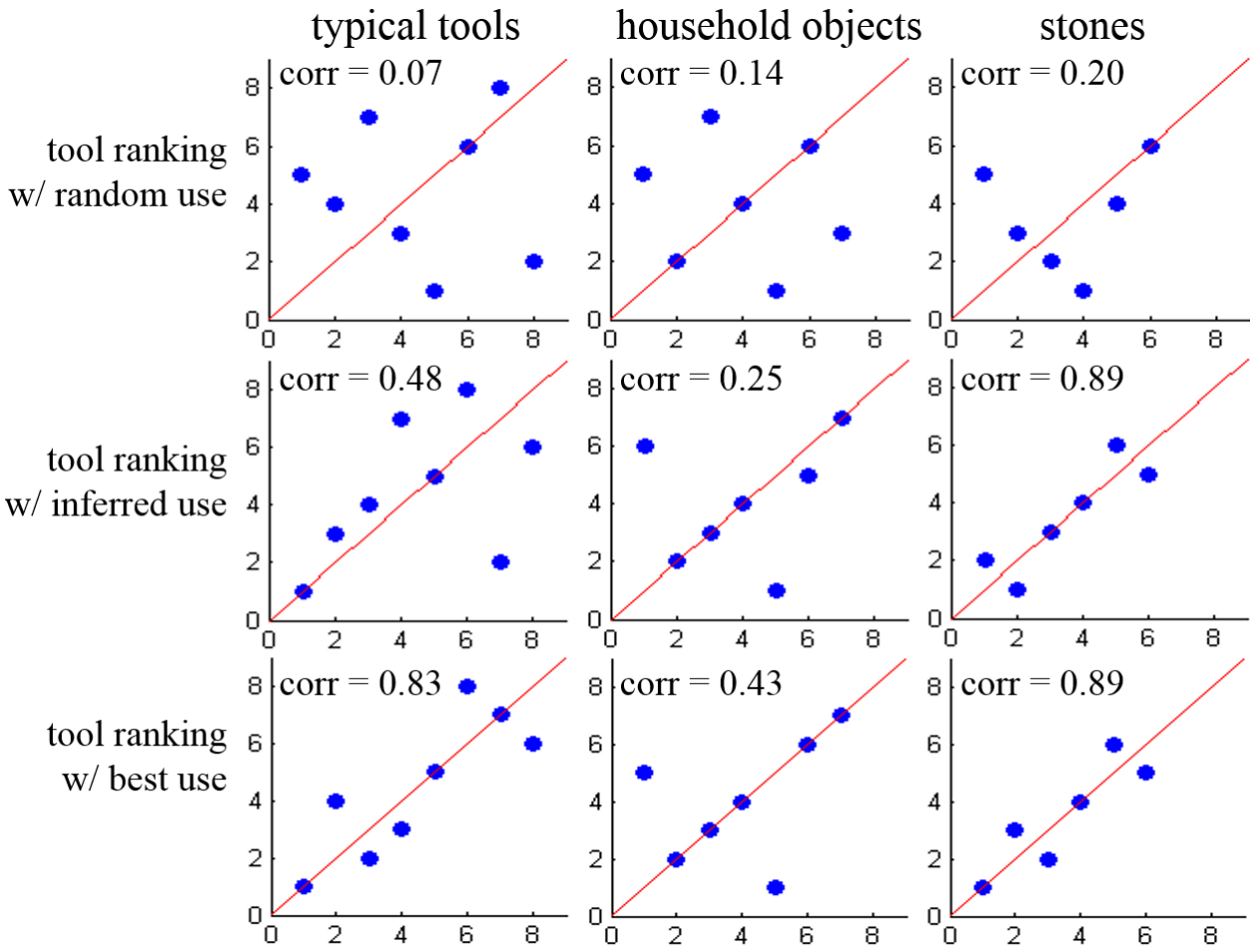


Figure 5.10: Recognizing tools for chopping wood. The scatters show tool candidates ranked by our algorithm (y-axis) with respect to the average ranking by human subjects (x-axis). The three columns show different testing scenarios, while the three rows represent different levels of tool-use imagined by inference algorithm.

Table 5.2: Errors of imagining tool-use for affordance / functional bases (B_A and B_F) . The table shows the 3D distances between their positions imagined by our algorithm and the positions annotated by human subjects. The specific positions for sample tool candidates are shown in Fig. 5.11.

3D distance (cm)	chop wood			shovel dirt			paint wall		
algorithm vs. human	tool	object	stone	tool	object	stone	tool	object	stone
B_A - top 1	1.75	3.02	3.19	1.17	2.03	3.28	0.43	2.48	2.86
B_A - top 3	1.04	2.17	2.81	0.97	0.52	2.21	0.31	2.32	2.67
B_F - top 1	0.48	5.97	3.91	6.98	6.38	0.23	2.35	2.74	2.65
B_F - top 3	0.27	5.92	3.95	2.85	3.29	0.31	1.43	2.64	2.71

5.5 Discussions

In this work, we present a new framework for task-oriented object modeling, learning and recognition.

An object for a task is represented in a spatial, temporal, and causal parse graph including:
i) spatial decomposition of the object and 3D relations with the imagined human pose;
ii) temporal pose sequences of human actions; and

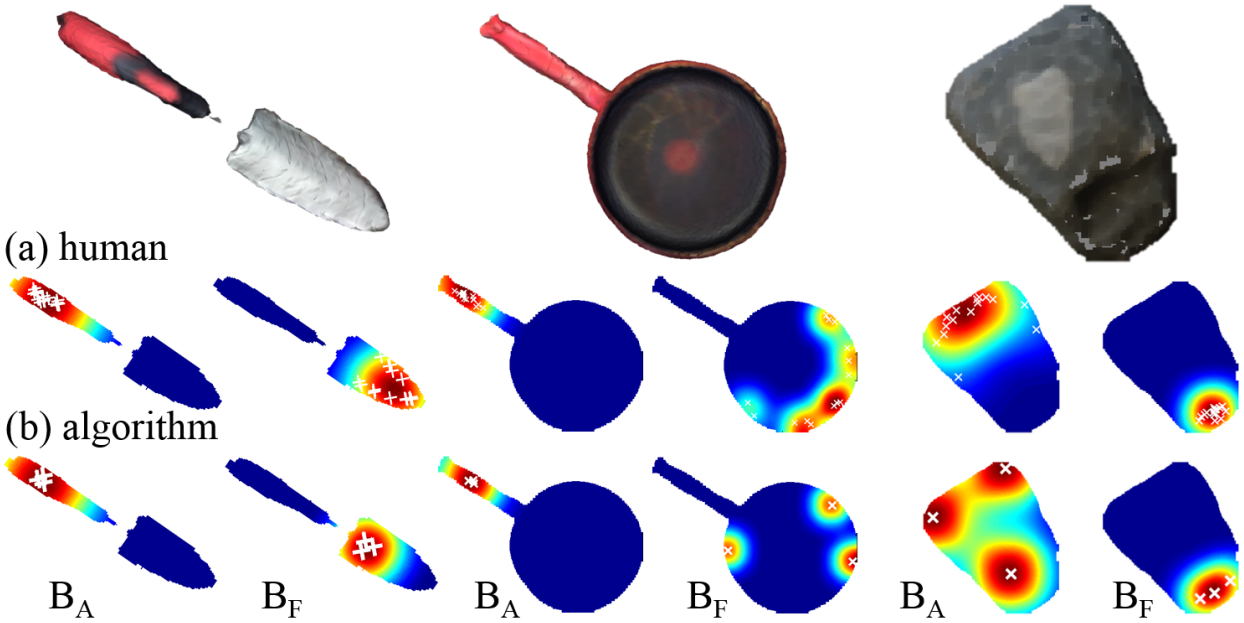


Figure 5.11: Comparison of human predicted tool-use (a) and algorithm imagined tool-use (b) for shoveling dirt.

iii) causal effects (physical quantities on the target object) produced by the object and action. In this inferred representation, only the object is visible, while all other components are imagined "dark" matters. This framework subsumes other traditional problems, such as:

(a) object recognition based on appearance and geometry; (b) action recognition based on poses; (c) object manipulation and affordance in robotics. We argue that objects, especially man-made objects, are designed for various tasks in a broad sense [328, 329, 330, 331], and therefore it is natural to study them in a task-oriented framework.

In the following, we briefly review related work in the literature of cognitive science, neuroscience, vision and robotics.

5.5.1 Related work

1) *Cognitiove Science and psychology.* The perception of tools and tool-uses has been extensively studied in cognitive science and psychology. Our work is motivated by the astonishing ability of animal tool-uses [335, 336, 337, 329, 330, 338]. For example, Santos *et al.* [339] trained two species of monkeys on a task to choose one of the two canes to reach food under various conditions that involve physical concepts. Weir *et al.* [340] reported that New Caledonian crows can bend a piece of straight wire into a hook and successfully used it to lift a bucket containing food from a vertical pipe. These discoveries suggest that animals can reason about the functional properties, physical forces and causal relations of tools using domain general mechanisms. Meanwhile, the history of human tool designing reflects the history of human intelligence development [341, 342, 343, 344]. One argument in cognitive science is that an intuitive physics simulation engine may have been wired in the brain through evolution [90, 345, 93], which is crucial for our capabilities of understanding objects and scenes.

2) *Neuroscience.* Studies in neuroscience [346, 25, 26] found in fMRI experiments that cortical areas in the doral pathway are selectively activated by tools in contrast to faces, indicating a very different pathway and mechanism for object manipulation from that of object recognition. Therefore

studying this mechanism will lead us to new directions for computer vision research.

3) *Robotics and AI.* There is also a large body of work studying tool manipulation in robotics and AI. Some related work focus on learning affordance parts or functional object detectors, *e.g.* [347, 348, 349, 350, 351, 352, 353, 354, 355]. They, however, are still learning high level appearance features, either selected by affordance / functional cues, or through human demonstrations [356], not to reason the underlying physical concepts.

4) *Computer vision.* The most related work in computer vision is a recent stream that recognizes functional objects (*e.g.*, chairs) [357, 358, 128, 359, 153, 360, 361, 362] and functional scene (*e.g.*, bedroom) [54, 121, 53, 130] by fitting imagined human poses. The idea of integrating physical-based models has been used for object tracking [227, 363] and scene understanding [116, 117] in computer vision. But our work goes beyond affordance.

5.5.2 Limitation and future work

In this work, we only consider handhold physical objects as tools. We do not consider other tools, such as, electrical, digital, virtual or mental tools. Our current object model is also limited by rigid bodies, and can not handle deformable or articulated objects, like scissors, which requires fine-grained hand pose and motion. All these request richer and finer representations which we will study in the future work.

Chapter 6

Mirroring and Imitation

A hallmark of machine intelligence is the capability to adapt to new tasks rapidly and "achieve goals in a wide range of environments" [364]. In comparison, a human can quickly learn new skills by observing other individuals, expanding their swiftly to adapt to the ever-changing environment. To emulate the similar learning process, the robotics community has been developing the framework of *Learning from Demonstration* (LfD) [356, 365]. This framework aims to have the robot learn human's demonstrated skills, manipulation in particular, naturally and quickly.

6.1 Robot Learning from Demonstration: Methods and Challenges

The main approaches for LfD can roughly be divided into three main categories. The first category of work uses kinesthetic teaching, where human demonstrators physically manipulate the robot to guide its task performance [366, 367, 368, 369], or uses teleoperation to collect demonstrations [370]. These methods is capable of incorporating forces into the demonstrations for in-contact tasks. Whereas the disadvantage is that the robot, *e.g.*, a manipulator, is bulky for humans to administrate fine manipulation tasks or those require precious trajectories.

Imitation learning is the second category. Work in this category either supervises demonstrations that directly mimic the demonstrator's behaviors [371, 372, 373, 374, 375, 376] or uses demonstrations as the initial policy to constrain the search space [377] and usually applies reinforcement learning to derive a control policy. These latter methods receive considerable attention recently and have succeeded in robot's constrained reaching [378], locomotion [379], grasping [380] and soft hand controlling [381]. To avoid being confined by the human demonstration, [382, 383] uses guided policy search for robot manipulations. Vision and other sensing techniques are usually used to track and record demonstrator's motion. However, policy search methods have not yet demonstrated successful applications in very complex tasks.

Finally, inverse reinforcement learning, or inverse optimal control can be classified as the third category. [384, 385, 386, 387] gains increasing interests in robotics community. Although it alleviates the need for reward engineering by inferring the reward/objective function from demonstrations, IRL has not been shown to scale to the same complexity of tasks as direct imitation learning, since there may exist many optimal policies that can explain a set of given demonstrations [388]. This challenge is often magnified by task complexity, making it computationally highly expensive [389].

Although these three types of main approaches have their own advantages and many of them have shown promising results, one common problem in LfD remains unsolved is the "correspondence problem" [390], *i.e.*, the difference of embodiments between a human and a robot. As the example illustrated in Fig. 6.1, a human hand with five fingers can firmly grasp a hammer, but a robot

(a) Demonstration (b) Failure by pure imitation

Figure 6.1: (a) Given a successful human demonstration, (b) the robot may fail to accomplish the same task by imitating the human demonstration due to different embodiments. In this case, a two-finger gripper cannot firmly hold a hammer while swinging; the hammer slips, and the execution fails. Reprinted, with permission, from [22].

gripper with the typical two or three fingers might struggle to wield. In this case, a one-to-one mapping that is usually handcrafted between the human demonstration and the robot execution, restricting the LfD only to mimic the demonstrator's low-level motor controls and replicate the (almost) identical procedure to achieve the goal, results in a failure. A system must reason about the underlying mechanisms of imitation, rather than simply mimicking the motions of a human demonstration, to allow the acquired skills to be adapted to new robots or new situations.

6.2 An Introduction to Mirror Neurons

In order to address the above difficulty in LfD for robots, let's take a step back to see how humans, or primates, understand and imitate actions.

6.2.1 Mirror Neurons in Monkeys and Humans

Neuroscientists discovered a special type of neurons, which are termed *Mirror Neurons*, in macaque monkeys' rostral part of inferior area 6 (area F5). The neurons fire (discharge) when the monkey performs a *goal-directed* action or sees others performing the same action [391]. Within the same area, there are motor neurons that activate under the presence of visual stimuli while some other activate for three-dimensional objects, but mirror neurons require the presentation of both the action (hand movement in particular) and the target object to be activated [392].

There is less direct evidence on mirror neurons in humans until recent brain-imaging techniques, Functional Magnetic Resonance Imaging (fMRI) in particular, reveal the existence of mirror neurons or mirror system in humans by observing similar neuron activities when observing or performing an action in the Broca's area of the frontal lobe, which is a neural center important for language, and in the parietal lobe that is related to perception and action [393]. In fact, the mirror system's activation appears not only when observing hand-related actions, but mouth and foot actions as well [392]. The activation also appears for non-human animals with different embodiments, such as monkeys and dogs [394], and even non-animals, *i.e.*, robots [395, 396].

While mirror neurons were initially regarded as providing an abstraction of actions, later experiments indicate that rather than low-level motor controls, the mirror neurons encode the goal or the intention of an action [397]. In the monkey side, Umilta *et al.* introduced normal pliers and reverse pliers for monkeys to use them to grip an object. While the goal—gripping the object—is the same, the actions of using the pliers are different. The monkeys needed to close their hand to operate a normal plier, but to open their hands for a reverse one [398]. The (mirror) neurons in area F5 discharged when the hands was opening with a normal pliers would discharge when the hands

was closing with a reverse pliers, or vice versa. The similarity in the temporal discharge pattern suggested that that mirror neurons code the goal or intention behind actions rather than low-level motor act.

In the human side, a seminal *tea party* experiment conducted by Iacoboni *et al*. reached a similar conclusion [122]. The human subjects observed three experimental conditions: Context, Action, and Intention. In the Context condition, there was a "before tea" context where items were properly arranged and an "after tea" context where those were not. No action is involved in this condition. Two types of actions, a whole-hand prehension grasp and a precision grip action of the tea cup, were displayed an equal number of times. The Intention condition included both the grasping actions and the two scenes used in the Context condition, making the intentions, drinking tea or cleaning up, less ambiguous compared to the Context condition or the Action condition alone.

A significant signal increase observed in the right inferior frontal cortex in the Intention condition suggested that mirror neurons are involved in understanding the intentions of others. This study hints a deeper cognitive process supported by mirror neurons that connects the action observation and action imitation.

Moreover, the functions of mirror neurons have been studied in various other cognitive processes, such as affordance, goal or intend prediction, action understanding, theory of mind *etc*. [394, 399]. They are also believed to support human's social interactions by allowing humans to feel certain emotions of each other, such as empathy [400].

Despite the vast amount of work related to mirror neurons, it is still a controversial topic and questions and doubts are cast to the theory that worth notices. In summary, there are two categories of questionings. Firstly, some researchers show their reservations on the existence of mirror neurons. Meanwhile, some argue about what role exactly mirror neurons play in our cognitive processes [401]. Nevertheless, mirror neurons provide a profounding basis of action understanding and imitation.

6.3 Mirroring with Functional Equivalence

Inspired by the mirror neurons, a *mirroring* approach that extends the current LfD in Robotics, through the physics-based simulation, is proposed to address the correspondence problem in robot imitation. Rather than overimitating the motion controls from the demonstration, it is advantageous for the robot to seek *functionally equivalent* but possibly visually different actions that can produce the same effect and achieve the same goal as those in the demonstration. The proposed *mirroring* approach emphasizes the intent of the demonstration as changing the target object to desired states regardless of the embodiment

6.3.1 Force-based Goal-oriented Mirroring

Consider the task of opening medicine bottles that have child-safety locking mechanisms. These bottles require the user to push or squeeze in various places to unlock the cap. By design, attempts to open these bottles using a standard procedure will result in failure. Even if the agent visually observes a successful demonstration, imitation of this procedure will likely omit critical steps in the procedure. The visual procedure for opening both medicine and traditional bottles are typically identical.

To achieve this, more explicit modeling knowledge about physical objects and forces is required as the ability of imitating and replicating contact forces could be a key in imitating manipulation. However, measuring human manipulation forces is difficult due to the lack of proper instruments that is accurate and imposes little constrain to natural hand motions. For example, vision-based manipulation force sensing method [227] is under too many constraints and the results are not

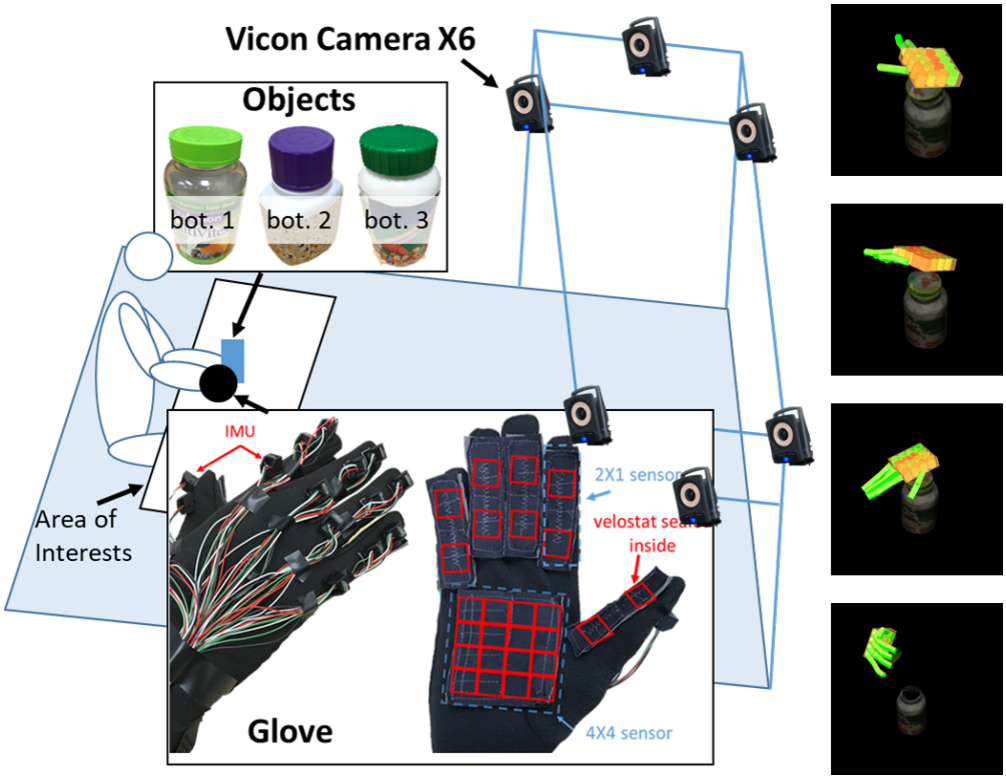


Figure 6.2: Data collection environment. A tactile glove is utilized to collect hand poses and forces, and the Vicon MoCap system for relative poses of hand and objects. Reprinted, with permission, from [406].

always reliable. [402] modifies target object to embed force/torque sensors inside. Although this method produces accurate force measurements, it only affords one grasp type for one object. Other force sensing devices such as strain gauge, FlexForce [403], or liquid-metal embedded elastomer sensor [404] can be implemented to hand using glove-based systems. But they can be too rigid to conform to the contours of the hand, resulting in limitations on natural hand motion during fine manipulative actions. Recently, [405] introduces Velostat, a soft piezoresistive conductive film whose resistance changes under pressure, to a IMU-based pose sensing glove to reliably record manipulation demonstrations with fine-grained force information, as shown in Fig. 6.2. This kind of demonstration is particularly important for the tasks with visually latent changes, supporting three characteristics in the mirroring approach compared to the standard LfD:

- *Force-based*: A low-cost tactile glove is deployed to collect human demonstration with fine-grained manipulation forces. Beyond visually observable space, these tactile-enabled demonstrations capture a deeper understanding of the physical world that a robot interacts with, providing an extra dimension to address the correspondence problem.
- *Goal-oriented*: A "goal" is defined as the desired state of the target object and is encoded in a grammar model. The terminal node of the grammar model is the state changes caused by the forces, independent of the embodiments.
- *Mirroring*: Different from the classic LfD, a robot does not necessarily mimic every action in the human demonstration. Instead, the robot reasons about the action to achieve the goal states based on the learned grammar and the simulated forces.

To validate the proposed approach, this study *mirror*s the human manipulation actions of opening medicine bottles with a child-safety lock to a real Baxter robot. The challenge in this task lies in the fact that opening such bottles requires to push or squeeze various parts, which is visually

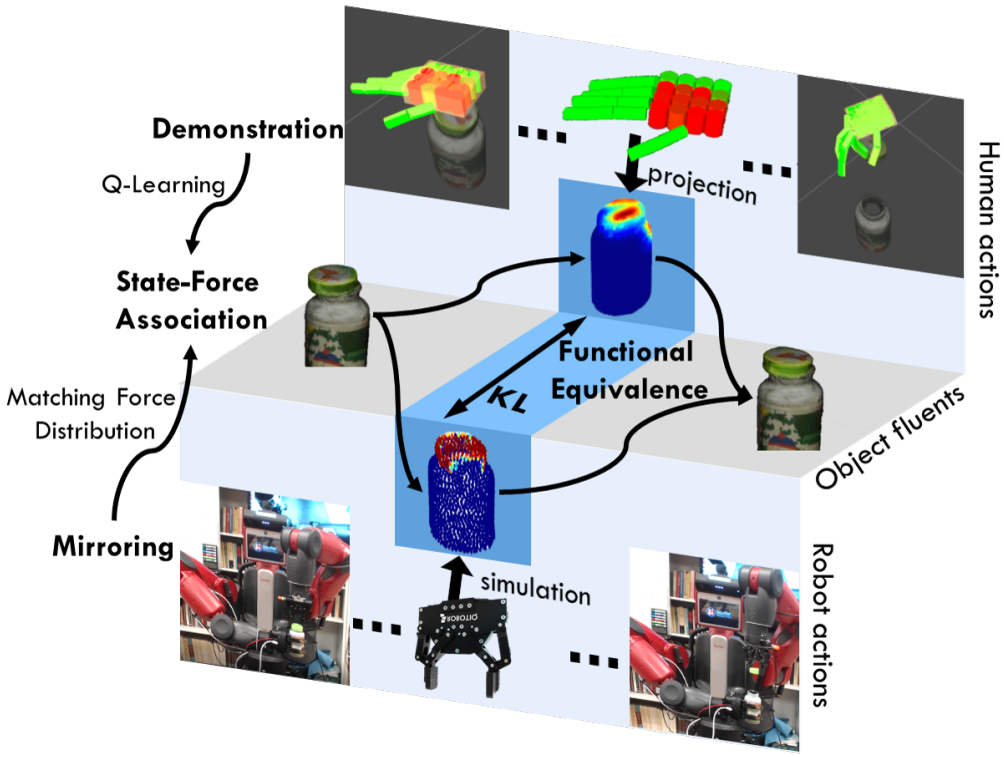


Figure 6.3: A robot mirrors human demonstrations with functional equivalence by inferring the action that produces similar force, resulting in similar changes of the physical states. Q-Learning is applied to associate types of forces with the categories of the object state changes to produce human-object-interaction (*hoi*) units. Reprinted, with permission, from [406].

similar to opening one without a child-safe lock. Fig. 6.3 outlines the *mirroring* approach with *functional equivalence*. Specifically, the forces on the object exerted by the hand in the demonstration is explicitly modeled with a pose and force sensing tactile glove. The collected distribution of the forces on the object is compared to a set of the force distributions exerted by the robot gripper on the same object in a physics-based simulator. Simulated actions with sufficiently small Kullback-Leibler (KL) divergence with respect to the demonstration are considered *functionally equivalent*, thus hinting this action would be the best robot action to accomplish the task.

Representation

The action sequence to execute a task is represented by a structural grammar model *Temporal And-Or Graph (T-AOG)* [154] (see Fig. 6.4). A T-AOG is a directed graph which describes a stochastic context-free grammar (SCFG), encoding both a hierarchical and a compositional representation. Formally, a T-AOG is defined as a five-tuple $G = (S, V, R, P, \Sigma)$. Specifically,

- S is the start symbol that represents an event category (*e.g.*, opening a bottle).
- V is a set of nodes including non-terminal nodes V^{NT} and terminal nodes V^{T}: $V = V^{NT} \cup V^{T}$.
- The **non-terminal** nodes can be divided into And-nodes and Or-nodes: $V^{NT} = V^{AND} \cup V^{OR}$. And-nodes V^{AND} represent the compositional relations: a node v is an And-node if the entity represented by v can be decomposed into multiple parts represented by its child nodes. Or-nodes V^{OR} indicate the alternative configuration among its child nodes: a node v is an Or-node if the entity represented by v has multiple mutually exclusive configurations represented by its child nodes.

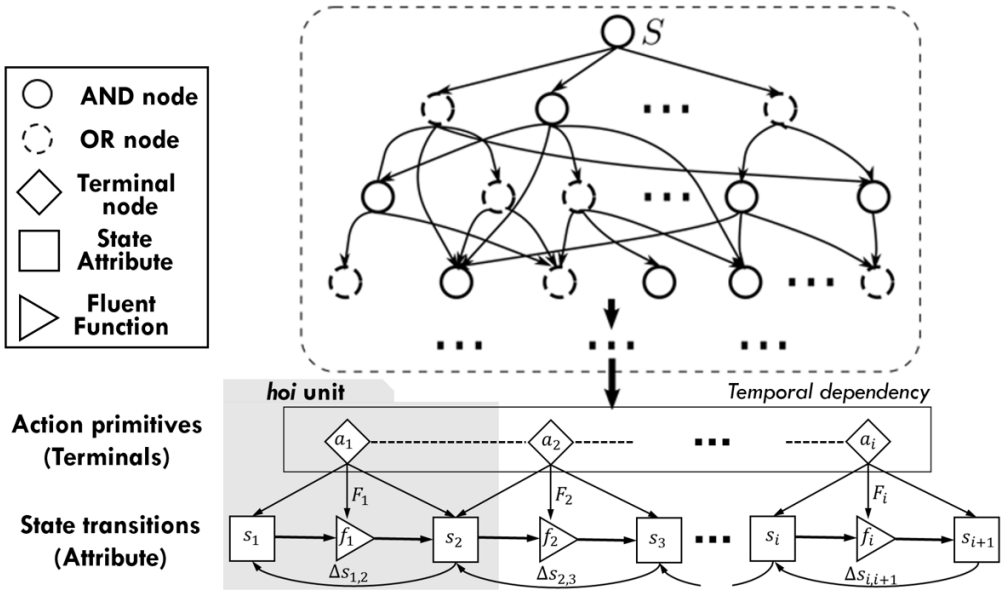


Figure 6.4: Illustration of a T-AOG. The T-AOG is a temporal grammar in which the terminal nodes are the *hoi* units. An *hoi* unit (shown in the grey area) contains a single action a_i that transits the state from the pre-condition s_i to the post-condition s_{i+1}. The fluents function f_i represents the changes of the physical state s_i on object caused by the forces F_i exerted by the action a_i: $s_{i+1} = f_i(s_i, a_i; F_i)$. Reprinted, with permission, from [406].

- The **terminal** nodes V^T are the entities that cannot be further decomposed or do not have different configurations. For a T-AOG, the terminal nodes represent the *human-object-interaction (hoi)* units [407]. An *hoi* unit encodes actions a_i that an agent can perform (*e.g.*, grasp, twist), the spatiotemporal relations between the object and the agent's hand, and how the force F_i produced by such primitive causes the changes of physical states on the object.
- $R = \{r : \alpha \rightarrow \beta\}$ is a set of production rules that represent the top-down sampling process from a parent node α to its child nodes β.
- $P : p(r) = p(\beta|\alpha)$ is the probability associated with each production rule.
- Σ is the language defined by the grammar, *i.e.*, the set of all valid sentences that can be generated by the grammar.

A **parse tree** *pt* is an instance of the T-AOG, where one of the child nodes is selected for each Or-node. The terminal nodes of a *pt* form a valid sentence; in this case, terminal nodes are a set of *hoi* units consisting of the actions for an agent to execute in a fixed order, as well as the state changes after performing such an action sequence.

Learning Force and State Associations as *hoi*

To transfer across different embodiment, we need to know the effect of a particular type of forces so that the desired action can be planned, requiring to investigate the state changes caused by the forces. We cast this problem in a reinforcement learning framework to learn a policy that associates forces and state changes. The state space and the action (force) space from human demonstrations are discretized and quantized, and an iterative Q-Learning scheme is applied. We believe the proposed learning framework does not lose generality since one can scale up the process to continuous state space or action space by using DQN [321] or other advanced policy gradient methods.

Categorize Force The pose and force data of human demonstrations were collected using a tactile glove. The forces exerted by a human hand, together with the poses, are projected onto the mesh of the object. Formally,

$$F_t^o = g(a_t^h(F_t^h, p_t^h)), \qquad t \in \{1, 2, \ldots, n\} \tag{6.1}$$

where t is the frame index, and n is the total number of frames. g is an implicit projection function that maps a human action a_t^h, parameterized by the force exerted F_t^h and the pose p_t^h, to F_t^o the force projected on the object mesh.

Each element in the resulting force F_t^o is a 4-dimensional vector, where the first three dimensions represent the position of one object surface vertex and the fourth dimension the force magnitude on this vertex.

K-means clustering [408] is adopted to categorize the force F_i^o into N types, *i.e.*,

$$l_k = c(F_t^o), \qquad t \in \{1, 2, \ldots, n\}, k \in \{1, \ldots, N\} \tag{6.2}$$

where $c(\cdot)$ denotes the clustering function and l_k is the label of the k-th cluster type. After assigning labels to each frame, the algorithm aggregates the frames with the same label into a segment and take the average,

$$F_k = \text{avg}(F_t^o), \qquad \forall t, c(F_t^o) = l_k. \tag{6.3}$$

The segments form a discretized action (force) sequence (Fig. 6.5c) to complete the given task.

Quantize State The relative poses can describe the states of a rigid target object under manipulation actions among object's parts, *e.g.*, bottle and lid, multiple Lego blocks, *etc*. Relative distance and relative rotation angle between the lid and the bottle, which are derived from their relative poses, are used as the state space. As shown in Fig. 6.5b, within each segment of the force (shown in color bars), the algorithm takes the average of the corresponding angle and distance and normalize their magnitude to unit size,

$$s_i = \langle d_i, \theta_i \rangle \in [0, 1]^2, \forall i \in \{1, \ldots, M\} \tag{6.4}$$

where M is the total number of states, and d_i and θ_i denotes the relative distance and angle, respectively.

Associate Force and State as *hoi* Units by Q-Learning By replacing the actions in Q-learning with the labels of the force l_k, we adopt the tabular Q-Learning that associates the current state s_i to a force type using the iterative Q-Learning update rule in a temporal difference fashion,

$$Q(s_i, l_k) = (1 - \alpha) \cdot Q(s_i, l_k) + \alpha \cdot \left[r(s_i, l_k) + \gamma \cdot \max_k Q(s_{i+1}, l_k) \right], \tag{6.5}$$

where r denotes the reward, Q the Q-function, α the learning rate, and γ the discount factor, assuming a deterministic system dynamics.

Inference We pick the best action according to the Q-function $l_* = argmax_k Q(s_i, l_k)$. The association among s_i, s_{i+1} and corresponding F_k naturally forms an *hoi* unit (see Fig. 6.4) and will be used for learning a goal-oriented grammar discussed in the next section.

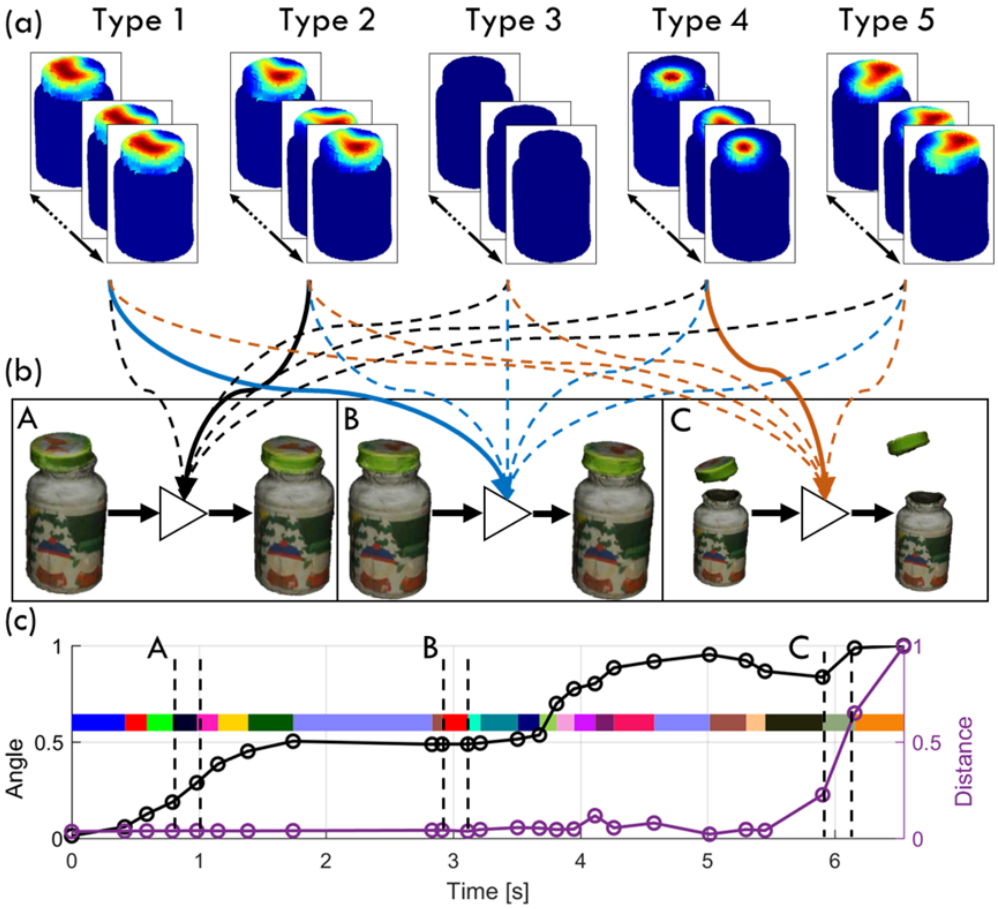


Figure 6.5: Force and state associations as *hoi* units. The manipulation force is clustered into 21 types. (a) Five examples of force types, in which Type 3 has no force. (c) Given the categorized force and quantized states based on the forces, (b) the Q-learning algorithm associates a force to a specific state change (A: lid is twisted; B: initiate contact; C: pull off the lid) shown by the solid lines. The dash lines indicate the forces that are incompatible to the given fluents functions, represented by the triangles. Reprinted, with permission, from [406].

Learning Goal-Oriented *hoi* Grammar

Grammar Induction Each successful demonstration contributes a sequence of *hoi* units that encode the types of forces and the state evolvement. A T-AOG $\mathcal{G}$ is induced from multiple demonstrations using a modified version of Automatic Distillation of Structure (ADIOS) algorithm presented in [45]. The objective function is the posterior probability of the grammar given the training data X,

$$p(G|X) \propto p(G)p(X|G) = \frac{1}{Z} e^{-\alpha||G||} \prod_{pt_i \in X} p(pt_i|G), \tag{6.6}$$

where $pt_i = (hoi_1, hoi_2, \ldots, hoi_m) \in X$ represents a valid parse graph of *hoi* units with length m.

Action Sequence Sampling To generate a valid sentence, *i.e.*, a parse tree $pt = (hoi_0, \ldots, hoi_K)$, we sample T-AOG $\mathcal{G}$ by decomposing all the And-nodes and selecting one branch at each Or-node. This pt is *goal-oriented* in the sense that its terminal nodes $hoi_k \in pt$ encode the forces of reaching sub-goal states that are invariant across embodiments for the given task. Note that this process is non-Markovian, while the force-state association using Q-Learning is Markovian.

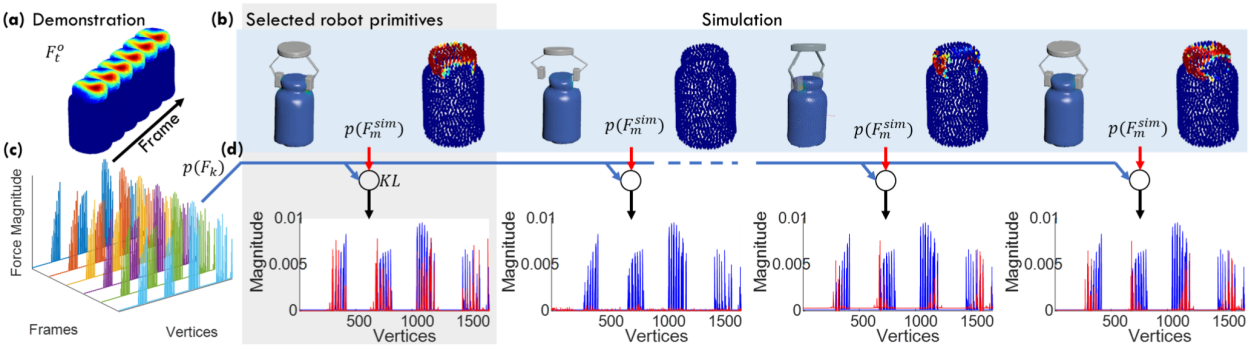


Figure 6.6: Based on the demonstrations, the force in the same cluster l_k produces a force distribution on the object F_t^o, and the average is the distribution of the force category F_k. Among the simulated force responses F_m^{sim} obtained from a physics-based simulator, the corresponding primitive of the most similar force, measured by the KL distance, is selected for the robot execution. (a) The forces in the same cluster. (b) The simulated robot primitives (downward, no contact, contact, and twist) and their force responses. (c) The force distributions of the same cluster in each frame. (d) The distributions of F_k against each simulated force distribution F_m^{sim}, denoted by blue and red, respectively. Reprinted, with permission, from [406].

Simulation

Simulation-based Action Synthesis Discrete robot action primitives are given by a dictionary $\Omega_{a^r} = \{a_1^r, \ldots, a_M^r\}$, $M = 10$, parameterized by the change of end-effector poses, including moves in all six canonical directions, rotations in both clockwise and counter-clockwise directions, and opening/closing the gripper. The task of opening a medicine bottle can be accomplished by the combinatorics of the actions. Given a pt, we seek to generate a sequence of robot actions $\{a_i^r, i = 1, \ldots, m\}$ that produce forces sufficient to cause the same changes of states as encoded in the sampled pt. In this sense, we say the robot action a_i^r is *functionally equivalent* to the demonstration action sequence a_i^h. Additionally, since the goal of the generated action sequences is to achieve the same effects, such generated action sequences can be different from the observed demonstrations and will not overimitate the observed ones.

A physics-based simulator (see Fig. 6.6) is introduced to estimate the force exerted by the robot gripper on the bottle. We denote the force obtained from the simulator as F_m^{sim}, where m is the index of the robot primitives, and compare it to the corresponding F_k, the average force exerted by human demonstrations with label l_k. Formally, F_k and F_m^{sim} are formalized as distributions,

$$P(F_k) = \frac{1}{Z_k} F_k, \quad \text{and} \quad P(F_m^{sim}) = \frac{1}{Z_m^{sim}} F_m^{sim}, \tag{6.7}$$

where Z_k and Z_m^{sim} are the normalization factors, obtained by summing over the force magnitudes on all vertices of the object. The similarity of the two forces can be measured by the KL divergence, and the robot action is selected by

$$\begin{aligned} F_*^{sim} &= argmin_m \mathrm{KL}\left(P(F_k) \parallel P(F_m^{sim})\right) \\ &= argmin_m \sum_v \left[P_{F_k}(v) \log \frac{P_{F_k}(v)}{P_{F_m^{sim}}(v)} \right], \end{aligned} \tag{6.8}$$

where v is the vertex index on the object mesh. Once F_*^{sim} is selected, the robot would choose the corresponding primitive a_*^r that produces F_*^{sim}.

Physics-based Simulation The physics-based simulation needs to be able to capture intricate frictional contact between the robot gripper and the bottle. The total force applied at each point located at the surface of the bottle consists of several terms: the normal component of squeezing force from the gripper, the tangential component of static friction force from the gripper, the internal elastic force from the rest of the continuous bottle material and gravity.

The key to achieving such a force balance in the simulator is to model the deformation of the bottle. Various physical constitutive models and stress-strain relationships exist for polymers, and it is impractical for us to find the exact material parameters through mechanical tension or compression tests. Thus, we assume the deformation of the bottle is sufficiently far away from the plastic regime, and adopt a standard hyperelastic model: the Neo-Hookean model [409] to describe the mechanical stress under deformation

$$\mathbf{P} = \mu(\mathbf{F} - \mathbf{F}^{-T}) + \lambda \log(\det(\mathbf{F}))\mathbf{F}^{-T}, \tag{6.9}$$

where $\mathbf{F}$ is the deformation gradient tensor encoding the strain at each point, $\mathbf{P}$ is the first Piola-Kirchoff stress tensor describing its elastic mechanical stress, and μ, λ are material parameters describing the stiffness and incompressibility of the bottle, respectively. The governing equation describing the force balance of the bottle is given by

$$\nabla \cdot \mathbf{P} = \mathbf{f}^{ext}, \tag{6.10}$$

where $\mathbf{f}^{ext}$ denotes the total external force on the bottle.

We solve Eq. (6.10) using the Finite Element Method [410]. The input bottle geometry is first converted from a triangulated surface to a tetrahedralized volume using TetGen [411]. The robot gripper mesh is converted into a watertight level set represented by OpenVDB [258], which allows natural treatment of frictional contact under arbitrary kinematic rigid motion. The additional parameters including friction coefficient, μ, and λ are set empirically. Once the discretized equation system is solved to convergence, we evaluate the force magnitude at each discrete point of the object surface mesh and store them in F_m^{sim}.

6.3.2 Mirroring to Robot without Overimitation

Preliminary

Robot Platform We exercise the proposed framework in a robot platform with a dual-armed 7-DoF Baxter robot mounted on a DataSpeed mobility base. The robot is equipped with a ReFlex TakkTile gripper on the right wrist and a Robotiq S85 parallel gripper on the left. The entire system runs on ROS, and the arm motion is planned by *MoveIt!*.

Dataset The hand pose and force data is collected using an open-sourced tactile glove [405] that is equipped with i) a network of 15 IMUs to measure the rotations between individual phalanxes, and ii) 6 customized force sensors using Velostat, a piezoresistive material, to record the force in two regions (proximal and distal) on each phalange and a 4×4 regions on palm. Fig. 6.2 depicts the tactile glove and the data collection environment. The relative poses between the wrist of hand and object parts (*i.e.*, bottle, and lid) are obtained from Vicon. The data of 10 human manipulation sequences is collected, processed, and visualized using ROS.

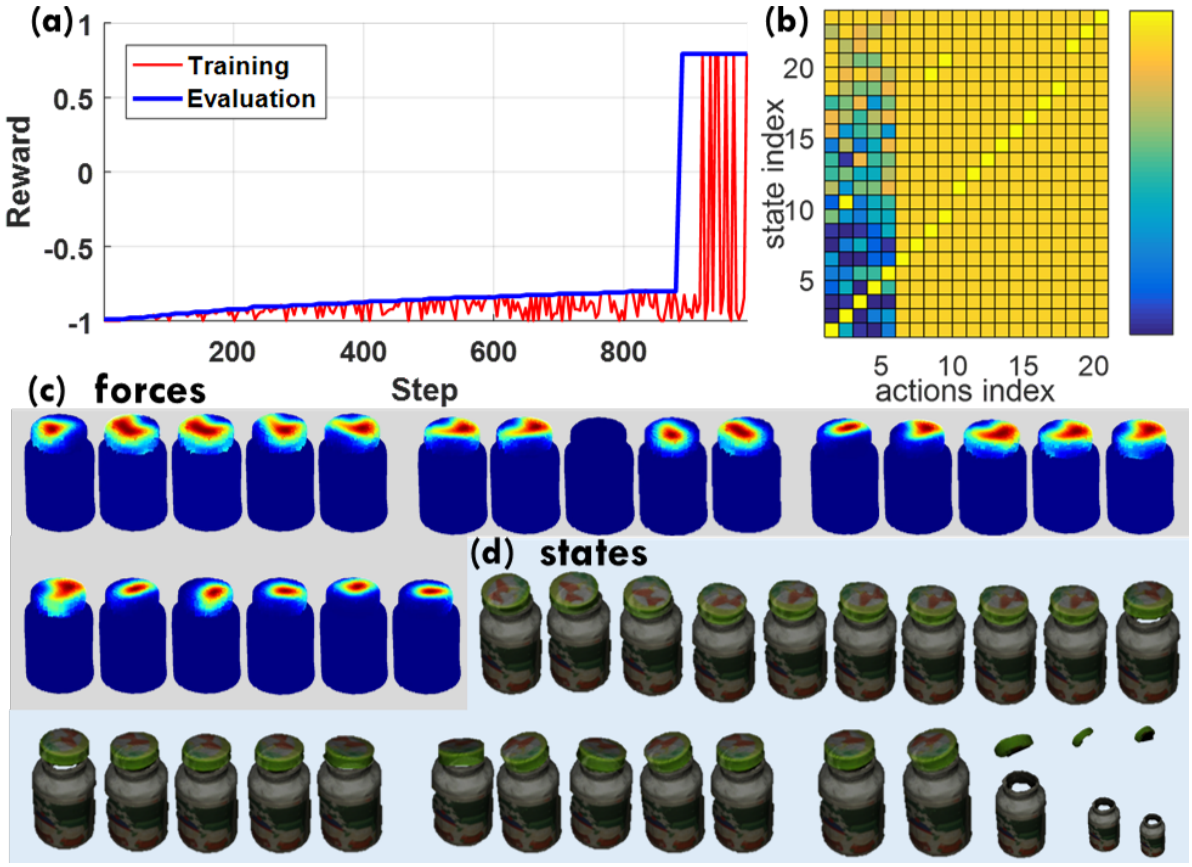


Figure 6.7: (a) The cumulative rewards during training and evaluation. (b) The landscape of the learned Q table, where yellow indicates high values and blue low. (c) The 21 types of actions (forces) by clustering in one exemplary demonstration. (d) The 25 discretized states based on the forces (some force types appear more than once). Reprinted, with permission, from [406].

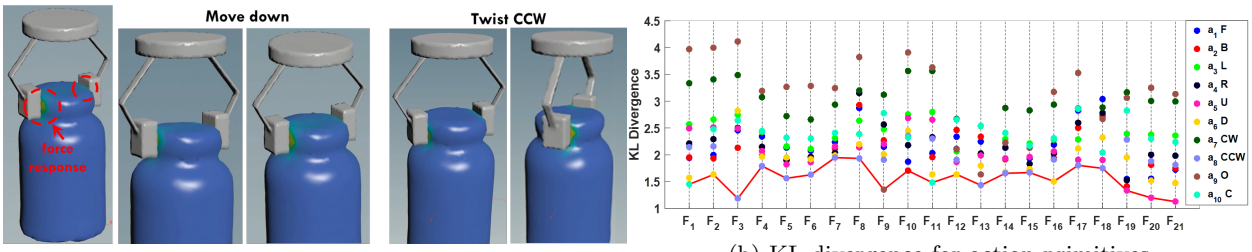
Learning

Fig. 6.7a-b shows the Q-learning results, with a discount factor 0.99, reward for success +1, reward for failure −1, and reward for all others 0. We use ϵ-greedy exploration with exponential decay to obtain the state-force associations.

Fig. 6.7a shows the cumulative reward during each training episode in red, and the average cumulative reward during evaluation in blue. During training, the cumulative reward generally increases until finding a path that leads to the maximum reward and begins fluctuating. This fluctuation happens due to the marginal probability of a non-optimal action being chosen at each step in ϵ-greedy exploration policy, even though an optimal path has been found. The evaluation is performed every ten episodes during training with a policy induced by the Q-table. During the evaluation, the reward monotonically increases slowly at first and jumps to the maximum, due to the optimal path found during training and the learning signals propagated into the Q-table. The policy induced from the Q-table converges to the optimum after approximately 900 episodes in training. The resulting Q-table is shown in Fig. 6.7b.

Robot Execution with Functional Equivalence

A *pt* is first sampled from the T-AOG induced from the learned policy to obtain a sequence of force types the robot should imitate in order to cause the same changes of object states. Our

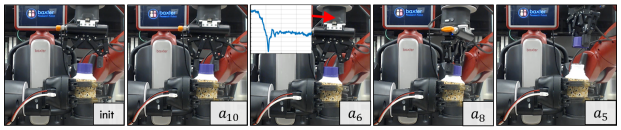


(a) Force simulation

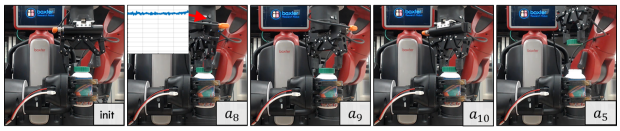
(b) KL divergence for action primitives

Figure 6.8: (a) Simulations of the robot actions' force responses. (b) The KL divergence for all action primitives in a *pt*. In this case, the primitives are a_1 *move forward*, a_2 *move backward*, a_3 *move left*, a_4 *move right*, a_5 *move up*, a_6 *move down*, a_7 *rotate clockwise*, a_8 *rotate counter-clockwise*, a_9 *open gripper*, and a_{10} *close gripper*. The solid red line is the sequence of actions for a robot to execute. Reprinted, with permission, from [406].

(a) Robot execution to open Bottle 1



(b) Robot execution to open Bottle 2



(c) Robot execution to open Bottle 3

Figure 6.9: Starting from the initial pose, the primitives (in grey) are performed sequentially. The robot "pushes" by a_6 (*downward*)(see force plot) and opens the medicine bottle by a_5 (*upward*). Reprinted, with permission, from [406].

physics-based simulation then emulates a set of robot actions to obtain their force responses; some examples are shown in Fig. 6.8a.

Fig. 6.8b shows an example of a *pt* consisting of 21 *hoi* units (x-axis). The force responses of the ten robot primitives are simulated, and the similarities (y-axis) to the corresponding F_k are measured in each stage. The primitives with the lowest KL divergence (connected by the red line) are selected for robot execution.

The execution of a Baxter robot is shown in Fig. 6.9a. It starts from an initial position and sequentially performs the corresponding primitives indicated in the grey area in the lower right corner. The a_6 *downward* primitive indeed generates forces which are captured by the force sensor (top left) in the robot wrist, which demonstrate that the *mirroring* approach indeed allows the robot to fulfill the challenging task of opening medicine bottles with a set of actions that are different from demonstrations.

The result also shows that the similarity between forces can be adequately measured by KL divergence to determine whether two actions are *functionally equivalent*. For instance, the primitive *opening the gripper* has the largest divergence in most of the cases as it produces no force to the object, except in F_9 when the demonstrator releases the lid after one rotation. The pressing force critical to our task is also captured and mirrored to robot well (see F_2 and F_{16} where a *downward* primitive is planned). Finally, *upward* primitives are selected to finish the task by pulling the lid.

6.4 Mirroring and Planning

The demonstrated task of opening medicine bottles shows the efficacy of mirroring, that is understanding the goal of an action and the reason, *i.e.* the force exerted, why the goal is achieved, in in-hand fine manipulation. In order to extend mirroring to more general manipulation tasks, an additional step in motion planning is needed.

6.4.1 Motion Planning for Mobile Manipulation

Manipulation is a core capability of both humans and robots, but human manipulation skills differ from machines in many ways [412]. Machines can nowadays easily outperform humans in some select tasks with extreme precision and efficiency, whereas human manipulation demonstrates robustness and adaptability at levels well beyond robots. In particular, humans possess excellent foot-arm coordination, and their manipulation strategy can be easily adapted to objects with similar underlying kinematic structures but with dramatically different appearance and geometric shapes. Such exceptional coordination and adaptability enable humans to accomplish a variety of manipulation tasks across a wide range of objects and environments.

These astonishing capabilities possessed by humans, however, are still extremely challenging to be replicated or implemented in a mobile manipulator due to the following two major difficulties: (i) How to smoothly coordinate between the mobile manipulator's locomotion, manipulation, and manipulated structures or objects? (ii) What is a proper representation that can abstract a physical structure and facilitate a general manipulation strategy?

Motion planning for a mobile manipulator has been extensively studied in past decades. Most of the approaches are based on probabilistic sampling, such as PRM [413], RRT [414], and its variants RRTConnect [415] and RRT* [416]. Although recent studies demonstrated the efficacy of sampling-based methods in high-dimensional motion planning [417, 418, 419], there are still some serious issues including (i) The generated solutions may violate physical constraints imposed by the robot hardware. Although some recent variants take simple kinematic models into account [420], they require additional modeling and computational efforts and are difficult to handle complicated kinematic chains. (ii) Many sampling-based methods require a pre-defined final pose as a goal to search the configuration space (except a recent study [421]). Such a procedure prohibits the algorithm from exploring better robot poses that satisfy the manipulation task.

Trajectory optimization is crucial in robotic motion planning to produce smooth trajectories by imposing constraints (*e.g.*, CHOMP [422], STOMP [423], and TrajOpt [424]) based on robot kinematics models and hardware specifications. Although some algorithms provide promising results with collision-free trajectories in mobile manipulators [425, 426, 427], these approaches *individually* optimize the base and arm (manipulator) trajectory and require an extra step to refine the base-arm coordination, lacking generalizability.

Although a number of full-body motion planning methods for a mobile manipulator have been proposed [428, 425, 429, 418], they either individually plan the trajectory of manipulator and mobile base or only support specific manipulation tasks, requiring additional modeling and computational efforts.

6.4.2 Virtual Kinematic Chain

Inspired by the virtual mechanism [430], we propose an optimization-based approach to tackle the aforementioned difficulties in motion planning for a mobile manipulator to interact with *articulated* structures and objects. The proposed approach utilizes Virtual Kinematic Chain (VKC), consisting of two steps: (i) VKC modeling, and (ii) an optimization-based motion planning.

Virtual mechanisms [431] is not a new idea in robotics; it has been used to chain serial manipulators [430] and parallel structures [432] via rigid-body objects. However, prior attempts do not support articulated objects.

Specifically, to build a VKC, the manipulated object needs to be abstracted as a *virtual* articulated object, and a virtual transformation between the mobile base and the virtual base link is augmented by incorporating the locomotion information into the VKC. Finally, a VKC could be formed by connecting the robot end-effector to an attachable location of the manipulated object. Once the VKC is formed, the motion planning solver performed on the VKC could be treated as a *single* optimization problem, *i.e.*, *jointly* optimizing both the robot locomotion and manipulation trajectories according to the desired goal state of the manipulated object *without* explicitly defining the final pose of the mobile manipulator.

In short, introducing VKC bridges the mobile manipulator and the manipulated (articulated) object using virtual serial chains, enabling robots to easily adapt to a variety of manipulation tasks and objects with a better generalization.

Problem definition

Prior to discussing modeling and motion planning for the VKC, we first define notations for robot and object models, as well as the formulation of a manipulation task.

In this chapter, the definition of joints and links generally follows the Unified Robot Description Format (URDF) standard, allowing easy implementation of the kinematic model in Robot Operating System (ROS); we use a tree for physical properties and kinematic constraints of links and joints. Fig. 6.10 (b) shows a conceptual model of kinematic chains, and Table 6.1 lists the notations:

- The group *Robot* refers to a mobile manipulator, which consists of three components: a mobile base, a robot manipulator, and an end-effector.
- The group *Object* represents the articulated object, which has a base link and other links anchored by this base link; a joint in an articulated object could be either prismatic, revolute, or fixed. Compared to the group *Robot*, instead of having an end-effector, an object can have multiple attachable links (*e.g.*, doorknob, drawer handle). An *attachment* is defined as a *virtual* joint, which represents a local transformation from an attachable link to an end-effector link. Note that both $\mathcal{F}_{ee}^{R}$ and $\mathcal{F}_{at}^{O}$ do not have to be terminal nodes in a $\mathcal{TG}$, and one attachable link is capable of having multiple attachments.
- The group *Problem Configuration* defines the configuration for a manipulation task. A mapping ${}^{a}_{b}\mathbf{T}$ between two adjacent link frames $\mathcal{F}_a$ and $\mathcal{F}_b$ implicitly encodes the joint state, and the joint type constraints the codomain of the mapping. Hence, we also use ${}^{a}_{b}\mathbf{T}$ to represents the joint connecting the parent link a and child link b.

Our modeling algorithm requires the knowledge of $\mathcal{TG}^{R}$, $\mathcal{TG}^{O}$, and ${}^{at}_{ee}\mathbf{T}^{O}$ as inputs to construct a *serial* chain $\mathcal{C}^{V}$. A serial kinematic chain's forward kinematics (FK), inverse kinematics (IK), and Jacobians can be effectively solved by most of the existing kinematic solvers. Thus, the constructed kinematics model can be easily adapted to various trajectory optimization frameworks; in this work, we use TrajOpt [433] to optimize the trajectory. The motion planning problem is to find a collision-free path for $\mathcal{C}^{V}$ through the trajectory optimization that satisfies either $\mathbf{q}_G$ or ${}^{w}_{ee}\mathbf{T}_G$ starting from $\mathbf{q}_I$. We assume the connection established between the end-effector and the attachable link is invariant during optimization or execution. The objectives of a manipulation task is implicitly encoded by $\mathbf{q}_G$ or ${}^{w}_{ee}\mathbf{T}_G$; for instance, the *Goal* of grasping a door handle could be set as ${}^{w}_{ee}\mathbf{T}_G = {}^{w}_{dh}\mathbf{T}^{O}\,{}^{dh}_{ee}\mathbf{T}^{O}$. We also define a set of symbolic *Actions*, which bridges symbolic task sequences with limited *Goal* information for motion planning in a manipulation task.

Table 6.1: Notation for constructing VKCs.

Group	Notation	Description
Robot	$\mathcal{TG}^R$	A tree represents the robot kinematic model
	$\mathcal{F}_b^R$	Robot base link's frame; the root of $\mathcal{C}^R$
	$\mathcal{F}_{ee}^R$	Robot end-effector link's frame
	$\mathcal{C}^R$	$\subset \mathcal{TG}^R$, a kinematic chain from $\mathcal{F}_b^R$ to $\mathcal{F}_{ee}^R$
	$\mathcal{F}_i^R$	Frame of link i in the kinematic chain $\mathcal{C}^R$
Object	$\mathcal{TG}^O$	A tree represents the object kinematic model
	$\mathcal{F}_b^O$	Object base link's frame; the root of $\mathcal{TG}^O$
	$\mathcal{F}_{at}^O$	Object attachable link's frame
	$\mathcal{C}^O$	$\subset \mathcal{TG}^O$, a kinematic chain from $\mathcal{F}_b^O$ to $\mathcal{F}_{at}^O$
	$\mathcal{F}_i^O$	Frame of link i in the kinematic chain $\mathcal{C}^O$
Others	$\mathcal{C}_n^V$	A serial VKC with n DoF
	$\mathbf{q}$	$\in \mathbb{R}^n$, the state of VKC in joint space
	$\mathbf{g}$	$\in \mathbb{R}^k$ $(k \leqslant n)$, the joint goal state
	${}_b^a T$	A homogeneous transformation from $\mathcal{F}_a$ to $\mathcal{F}_b$
	${}_{ee}^w T_g$	The end-effector's goal pose in world frame

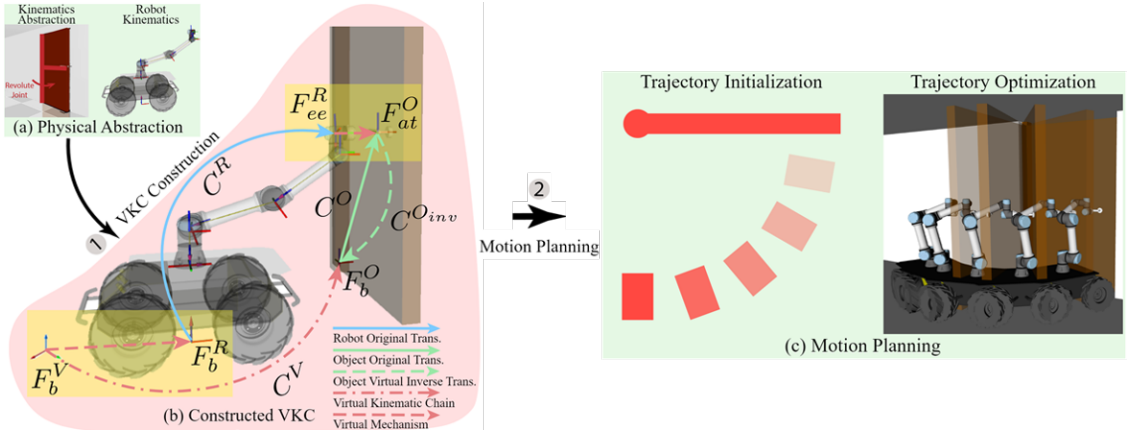
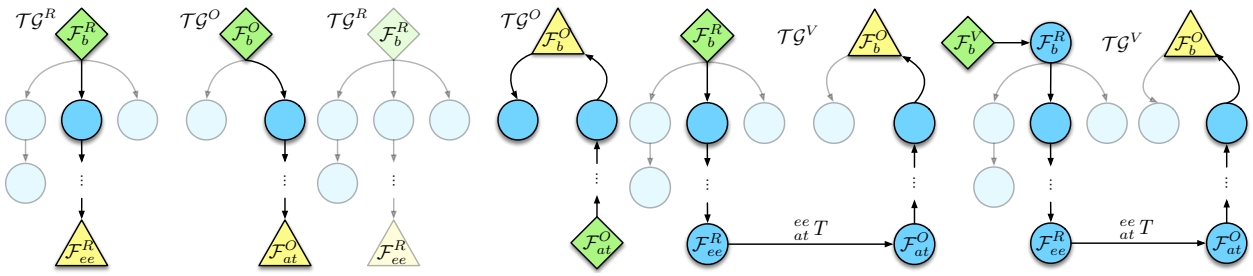


Figure 6.10: Overview of the proposed motion planner using VKC. (a) Perceiving and abstracting the underlying kinematics of the manipulated object, wherein (b) a VKC is constructed, yellow boxes denote where the virtual mechanism is formed. $\mathcal{F}_b^V$ is the virtual base frame. $\mathcal{F}_b^R$ and $\mathcal{F}_{ee}^R$ are the robot base frame and the end-effector frame, respectively. $\mathcal{F}_b^O$ and $\mathcal{F}_{at}^O$ are the manipulated object base frame and the attachable frame, respectively. Two augmented virtual connections are established: one between $\mathcal{F}_b^V$ and $\mathcal{F}_b^R$ to reflect the pose changes of the mobile base to the world, and another between $\mathcal{F}_{ee}^R$ and $\mathcal{F}_{at}^O$ to transfer effects of the manipulator to the manipulated object. (c) presents the motion planning procedure. A trajectory initialization, which utilizes A* algorithm and inverse kinematics, is adapted before the trajectory optimization, which produces the final optimized motion trajectory.

VKC Modeling

In this section, we discuss the proposed modeling and motion planning methods for VKC. We assume that the underlying kinematics of the manipulated object are already available; Figure



(a) Original Structure (b) Kinematic Inversion (c) VKC Construction (d) Virtual Mobile Base

Figure 6.11: The construction process of VKC. The green diamond denotes the *root frame*, the yellow triangle denotes a *robot end-effector frame* or an *object attachable frame* as a *terminal frame*, and the blue circles denote *other frames*.

illustrates an example of a constructed VKC in which both robot and object kinematics is incorporated, such that trajectories of robot locomotion and manipulation could be jointly optimized as a *single* optimization problem.

The objective of the modeling is to construct a virtual serial kinematic chain $\mathcal{C}^V$ by composing the robot and the object kinematics model.

Original Structure As highlighted in Fig. 6.11a, two trees represent the original kinematic chains of the robot $\mathcal{C}^R$ and the articulated object $\mathcal{C}^O$, respectively.

Kinematic Inversion To insert a virtual joint (*i.e.*, an *attachment*) between the robot's end-effector frame $\mathcal{F}^R_{ee}$ and the articulated object's attachable frame $\mathcal{F}^O_{at}$, one has to invert the kinematic relationship of the articulated object $\mathcal{C}^O$. Note such an inversion is not as simple as taking an inverse of the transformation between two adjacent links, since the joint frame might be falsely aligned to the new child frame, resulting in a wrong kinematics model.

Consider an articulated object with a terminal link t and a root/base link b fixed to the world frame w with a world joint ${}^w_b\mathbf{T}$. A mapping ${}^b_t\mathbf{T}$ represents the joint connecting the base and the terminal links. Then the joint frame in the world can be derived as ${}^w_t\mathbf{T} = {}^w_b\mathbf{T}\,{}^b_t\mathbf{T}$. In an inverted model, b becomes terminal link, and the new joint frame in the world ${}^w_b\mathbf{T}_{\text{inv}}$ has to be identical to ${}^w_t\mathbf{T}$ to ensure the same kinematics relation remains as prior to the inversion. Hence, the new joint ${}^t_b\mathbf{T}_{\text{inv}}$ within articulated object should be:

$$ {}^w_b\mathbf{T}_{inv} = {}^w_t\mathbf{T},\ {}^w_t\mathbf{T},\ {}^t_b\mathbf{T}_{inv} = {}^w_t\mathbf{T},\ {}^t_b\mathbf{T}_{inv} = {}^w_t\mathbf{T}^{-1}\ {}^w_t\mathbf{T} = \mathcal{I}_4. \tag{6.11} $$

A general formulation of kinematic chains with multiple links could be easily extended and derived from the above example with two links:

$$ {}^i_{i-1}\mathbf{T}_{\text{inv}} = {}^{i+1}_{i}\mathbf{T}, \tag{6.12} $$

where $i-1$ represents the parent link of link i, and $i+1$ represents the child link of link i *along* the kinematic chain *before* the inversion.

VKC Construction with Virtual Mobile Base After inverting the object model $\mathcal{C}^O$, a virtual kinematic chain can be constructed by adding a virtual joint between $\mathcal{F}^{O_{\text{inv}}}_b$ and $\mathcal{F}^R_{ee}$. The virtual base is further added to link b to enable a joint optimization of the locomotion and manipulation. It simply adds two perpendicular prismatic virtual joints to imitate a planar motion between the mobile base and the ground, while ensuring the virtual kinematic chain remains serial.

6.4.3 Optimization-based Motion Planning

Optimization Framework

Given an environment with obstacles, the motion planning of a mobile manipulator using VKC could be regarded as finding a collision-free trajectory with the newly constructed VKC, solvable by trajectory optimization that minimizes the given objective functions. To simplify the problem, we only consider all the feasible state reachable by VKCs; *i.e.*, we do not consider manipulating trivially underactuated articulated objects, such as a double pendulum. This assumption is generally reasonable for a mobile manipulator since the constrained mechanisms of human-made indoor environments are designed to be fully-actuated. We also assume holonomic constraints for the robot mobile base (*i.e.*, an omnidirectional mobile base).

The optimization problem for a mobile manipulation task [433] using VKC can be formally expressed as:

$$\begin{aligned}\underset{\mathbf{q}_{1:T}}{\text{minimize}} \quad & \sum_{t=1}^{T-1} ||W_{vel}^{1/2}(\mathbf{q}_{t+1}-\mathbf{q}_t)||_2^2 \\ & + \sum_{t=2}^{T-1} ||W_{acc}^{1/2}(\mathbf{q}_{t+1}-2\mathbf{q}_t-\mathbf{q}_{t-1})||_2^2\end{aligned} \tag{6.13}$$

$$\text{subject to} \quad h_{\text{chain}}(\mathbf{q}_t)=0, \forall t=1,2,\ldots,T \tag{6.14}$$

$$\mathbf{q}^{\min} \leqslant \mathbf{q}_t \leqslant \mathbf{q}^{\max}, \forall t=1,2,\ldots,T \tag{6.15}$$

$$||\ddot{\mathbf{q}}_t||_\infty \leqslant \xi_{acc}, \forall t=2,3,\ldots,T-1 \tag{6.16}$$

$$||f_{\text{task}}(\mathbf{q}_T)-\mathbf{g}||_2^2 \leqslant \xi_{goal},\ ||f_{\text{fk}}(\mathbf{q}_T)-{}_{ee}^{w}T_g||_2^2 \leqslant \xi_{goal} \tag{6.17}$$

$$\sum_{i=1}^{N_{link}}\sum_{j=1}^{N_{obj}} |\text{dist}_{\text{safe}}-f_{\text{dist}}(L_i,O_j)|^+ \leqslant \xi_{dist} \tag{6.18}$$

$$\sum_{i=1}^{N_{link}}\sum_{j=1}^{N_{link}} |\text{dist}_{\text{safe}}-f_{\text{dist}}(L_i,L_j)|^+ \leqslant \xi_{dist}\ . \tag{6.19}$$

Objective Eq. (6.13) is the objective function, where we penalize the overall velocities of every joint with the approximation $\dot{\mathbf{q}}_t \approx \mathbf{q}_{t+1}-\mathbf{q}_t$ and overall acceleration of every joint with the approximation $\ddot{\mathbf{q}}_t \approx \mathbf{q}_{t-1}-2\mathbf{q}_t+\mathbf{q}_{t+1}$. W_{vel} and W_{acc} are diagonal weight matrices for each joint, respectively. $||\cdot||_2$ denotes the $l2$ norm, and $\mathbf{q}_{1:T}$ represents the trajectory sequence $\{q_1,q_2,\ldots,q_T\}$, where $\mathbf{q}_t$ denotes the VKC state at the t^{th} time step.

Constraints Eq. (6.14) is an equality constraint that specifies the kinematics of the VKC, which includes the forward kinematics of the VKC, as well as other physical constraints of the manipulated object; *e.g.*, the base link of door is fixed to the ground: ${}_{b}^{w}T_{1:T}^{O}-{}_{b}^{w}T_{1}^{O}=0$.

Eq. (6.15) is an inequality constraint that defines joint limits, in which $\mathbf{q}^{\min}$ and $\mathbf{q}^{\max}$ specify the lower and upper bound of every joint, respectively.

Eq. (6.16) is an inequality constraint that bounds the joint acceleration by ξ_{acc} in order to obtain a feasible trajectory that can be executed without saturation. $||\cdot||_\infty$ denotes the infinity norm.

The first equation in Eq. (6.17) bounds the squared $l2$ norm between the final state in the goal space $f_{\text{task}}(\mathbf{q}_T)$ and the goal state $\mathbf{g}$ with a tolerance ξ_{goal}. The second equation in Eq. (6.17)

bounds the squared $l2$ norm between the final end-effector pose $f_{\text{fk}}(\mathbf{q}_T)$ and desired end-effector pose ${}^{w}_{ee}T_g$ via an inequality constraint in $SE(3)$. Note that these two constraints do not need to be specified in every task. The function $f_{task}(\cdot)$ is a task-dependent function that maps the joint space of a VKC to the goal space as the goal space could be different from task to task. For example, in a door opening task, $f_{task}(\cdot)$ will map the joint space of a VKC to the joint of the door revolute axis. Hence, we can use an angle θ to describe the task goal (*i.e.*, the desired pose of the door), instead of explicitly specifying the final pose of every joint in the VKC. In this way, the end-effector's and the mobile base's paths are implicitly optimized in the joint space, together with obstacle avoidance and trajectory smoothing. It is also straightforward to add additional task constraints to the very same optimization problem, depending on other requirements.

Eq. (6.18) and Eq. (6.19) are inequality constraints that check link-object collisions and link-link collisions, respectively, where N_{link} and N_{obj} are the number of links and the number of objects, respectively. dist_{safe} is a pre-define safety distance, and $f_{dist}(\cdot)$ is a function that calculates the signed distance [433] between i-th link L_i and j-th object O_j. The function $|\cdot|^+$ is defined as $|x|^+ = \max(x, 0)$. The inequality constraints Eq. (6.18) and Eq. (6.19) make the preceding optimization problem highly non-convex and cannot be solved by a general convex solver. To address this issue, we adopt the algorithm proposed by Shulman *et al.* [433] to solve the optimization problem by approximating the non-convex problem to a sequence of convex problems and utilizing a sequential convex optimization method to solve them.

In addition to the space constraints, we also incorporate:

- *Manipulability*, which could be easily evaluated by the Jacobian of the end-effector J_{ee} for a serial chain: $w = \sqrt{\det(J_{ee}J_{ee}^T)}$ [425, 434].
- *Stability*, achieved by adopting the center of mass position as inequality constraints; see [432, 428].

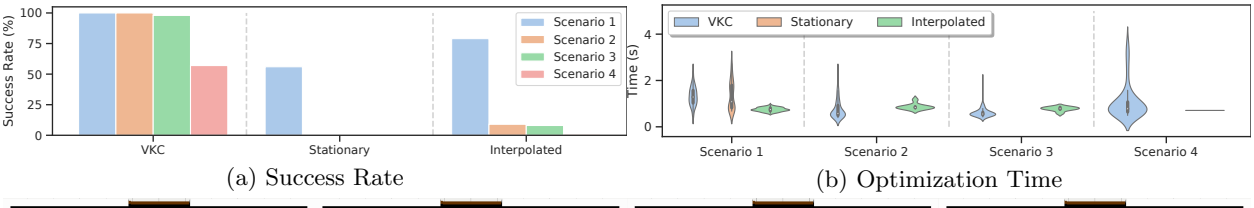
Trajectory Initialization

Trajectory optimization could easily get trapped in local minima using pure gradient descent approaches; a properly initialized trajectory would significantly improve the generated trajectory. In particular, in an indoor environment with cluttered obstacles, a mobile manipulator often gets stuck when obstacles block the path to the target. To alleviate this issue without sacrificing planning time, we adopt a native A* algorithm to search a feasible path around the obstacles for the mobile base to traverse the space. In some tasks where the final pose of the mobile base is not specified, an initial guess of final base position for trajectory initialization could be obtained by solving the inverse kinematics of $\mathcal{C}^V$ numerically with given joint goal state $\mathbf{g}$ and/or end-effector goal pose ${}^{w}_{ee}T_g$.

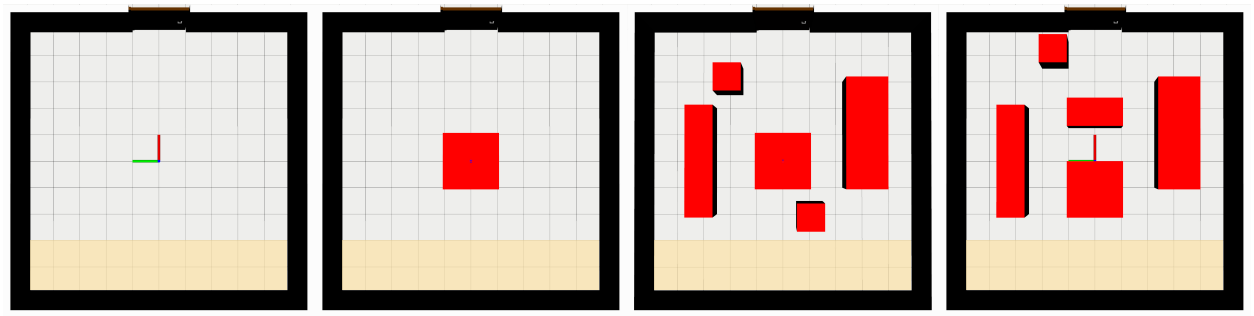
This experiment demonstrates how the aforementioned motion planning framework is applied on a mobile manipulator to approach and open a door, where the locomotion and manipulation are modeled using VKC. We also evaluate the performance of our A*-based trajectory initialization method by comparing the success rate and computation time with two trajectory initialization methods as baselines:

- **Stationary:** The trajectory $\mathbf{q}_{1:T}$ is initialized by waypoints $\mathbf{q}_t$ that are identical to the initial pose $\mathbf{q}_{init}$.
- **Interpolated:** The trajectory $\mathbf{q}_{1:T}$ is initialized by waypoints that are linearly interpolated between initial and goal pose.

Fig. 6.12c-Fig. 6.12f show four scenarios for evaluations. The task is for the mobile manipulator to navigate from the shaded yellow region (bottom) to the door (top) and pull to open it. The scenarios range from having no obstacle, one obstacle in the center, five randomly generated obstacles, and four randomly generated obstacles with an additional obstacle purposely placed next to the



(a) Success Rate (b) Optimization Time



(c) Scenario 1 (d) Scenario 2 (e) Scenario 3 (f) Scenario 4

Figure 6.12: Quantitative results in motion planning using VKC modeling. (a) The success rates in generating a feasible plan using A*-based approach (Ours), Stationary, and Interpolated trajectory initialization methods. (b) The violin plots [435] (a hybrid of a box plot and a kernel density plot) of the optimization time for different methods, wherein the white dot represents the median, the thick gray bar in the center represents the interquartile range, and the thin gray line represents the rest of the distribution, except for points that are determined to be "outliers." (c)–(f) The experimental scenarios with an increasing complexity, where the initial condition was highlighted in yellow.

door to block the mobile manipulator. These scenarios are in increasing complexity and difficulty for generating a feasible trajectory.

The experiment runs 100 times for 100 different initial robot pose uniformly sampled within the shaded region for each of the four scenarios; Fig. 6.12a shows the success rate. A successfully optimized trajectory has to be a converged result without violating any constraints (*e.g.*, collisions).

The proposed VKC-based method has the highest success rates among all four scenarios. Specifically, it achieves above 95% for the simpler three scenarios. In the fourth one, most of the failures are due to sharp turns near the obstacle. In comparison, we find that the mobile manipulator to be commonly trapped into a local minimal surrounded by obstacles using the *Stationary* trajectory initialization method, resulting in an almost 0% success rate in the scenarios expect the one with no obstacle. Although the initial waypoints could drive the mobile manipulator out of such a local minimal using the *Interpolated* trajectory initialization method, it is very typical [436] that the mobile manipulator would prefer to go through thin and long obstacles despite the imposed penalties. Although the *Interpolated* method performs better than the *Stationary* method, it can barely succeed in scenario 2 and 3 with about 8% success rates. The proposed method significantly outperforms two baselines; it is the only method that can successfully solve scenario 4, which requires excellent arm-base coordination in pulling to open the door.

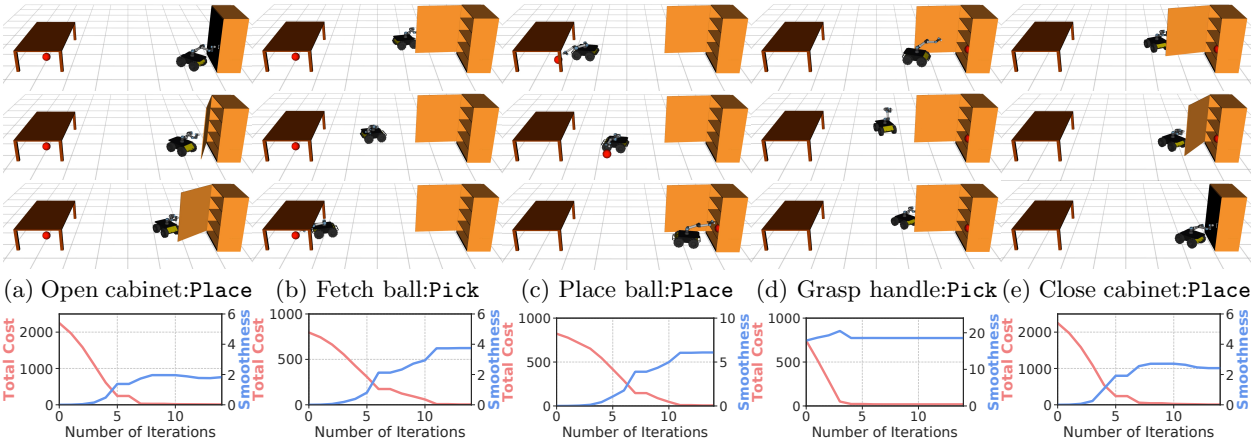
Fig. 6.12b further quantitatively compares the optimization times of the three trajectory initialization methods among the successful trials. The time needed (if any) is comparable for all cases, indicating the proposed method can be scaled up to more challenging mobile manipulation tasks.

6.4.4 Symbolic Task Predicates

Using high-level predicates, we develop an interface to specify manipulating goals. These predicates are tabulated in Table 6.2 with descriptions of how VKC are constructed to satisfy the goals;

Table 6.2: Predicates, robot actions, and VKC modifications.

Predicates	Description	VKC Modification
`Goto` ($\mathbf{g}_b$)	Move base to position $\mathbf{g}_b$	-
`Pick` ($O, {}^{w}_{at}T^{O}$)	Pick object O at location ${}^{w}_{at}T^{O}$	Connect O and end-effector via a virtual joint
`Place` ($O, \mathbf{g}_O$)	Place the object O in the pose of $\mathbf{g}_O$	Break virtual connection between O and end-effector
`Use` ($O_a, O_b, {}^{w}_{at}T^{O_b}$)	Use O_a manipulates O_b	Connect O_a and O_b via a virtual joint at ${}^{w}_{at}T^{O_b}$



(a) Open cabinet:`Place` (b) Fetch ball:`Pick` (c) Place ball:`Place` (d) Grasp handle:`Pick` (e) Close cabinet:`Place`

(f) Cost over iterations of each action

Figure 6.13: Motion planning results using VKC modeling in a simulated environment. (a)–(e) Different predicate actions. (f) The cost over optimization iterations of each action, in which total costs (blue lines) are effectively optimized, and joint velocities and accelerations (red lines) are bounded at the same time.

they allow a mobile manipulator to accomplish more complex tasks by properly sequencing the predicates. Once the VKCs are constructed according to the specified predicates, corresponding optimization problems are formulated and solved automatically without the need for manual designs or modifications of intermediate steps.

Regardless of the number of predicates involved in a task, they all share the same objective (Eq. (6.13)) in the optimization problems, and at most two constraints (Eq. (6.14) and Eq. (6.17)) are altered. In particular, Eq. (6.14) is updated based on the newly constructed VKC to satisfy kinematics constraints, and Eq. (6.17) is updated with a different task function f_{task}. For examples, `Goto` specifies the desired coordination of the mobile base, `Pick` specifies the desired pose of the robot end-effector, and `Place` specifies the final pose of the manipulated object.

Fig. 6.13 qualitatively shows the motion planning results using VKCs. The mobile manipulator needs to: (a) open the door of a cabinet, (b) fetch a ball under the table, (c) pick up the ball and place it on a shelf of the cabinet, (d) grasp the handle of the cabinet, and (e) close the cabinet door. This task requires the motion planner to deliver proper locomotion and manipulation with excellent arm-base coordination, demonstrating the efficacy and practicality of the proposed VKC modeling method.

Chapter 7

Utility

7.1 Learning Human Utility from Demonstration

7.1.1 Introduction

Explicitly programming service robots to accomplish new tasks in uncontrolled environments is time-consuming, error-prone, and sometimes even infeasible. In Learning from Demonstration (LfD), many statistical models have been proposed that maximize the likelihood of observations. For example, Bayesian formulations assume a prior model of the goal, and use Bayes' Theorem to explain the relationship between the posterior and likelihood. These Bayesian formulations learn a model of the demonstrated task most consistent with training data. Such approaches are often referred to as inductive learning.

In contrast, robot autonomy was originally studied as a rule-based deductive learning system. There is a paradigm shift in applying inductive models to deduction based inference. In this work, we explore a middle-ground, where deductive rules are learned through statistical techniques. Specifically, we teach a robot how to fold shirts through human demonstrations, and have it reproduce the skill under both different articles of clothing and different sets of available actions. Our experimental results show good performance on a two-armed industrial robot following causal chains that maximize a learned latent utility function. Most importantly, the robot's decisions are interpretable, facilitating immediate natural language description of plans. Human preferences are modeled by a latent utility function over the states of the world. To rank preferences, we pursue relevant fluents of a task, and then learn a utility function based on these fluents. For example, Fig. 7.1 shows the utility landscape for a cloth-folding task, obtained through 45 visual demonstrations.

Utility learning equios autonomous agents, such as service robots with a high-level understanding of goals, as well as the ability to adapt that understanding to new situations. By studying a utility function to rank states of the world, we steer away from imitation as a measure of success, thereby bypassing the complications of the correspondence problem. Shukla *et al*. proposed framework simultaneously learns both interpretable features as well as the utility function from few human demonstrations to model non-Markovian behavior. Moreover, our model can explain its motivations and how those motivation translate into actions. In particular, we parse videos if shirt-folding demonstrations to develop a preference model that is capable of generalizing its learned features and utility function across different situations.

The utility landscape shows a global perspective of candidate goal states. To close the loop with autonomous behavior, we further design a dynamics equation to connect high-level reasoning to low-level motion control. The primary contributions of our work include:

- A utility value driven planning framework for non-Markovian tasks.

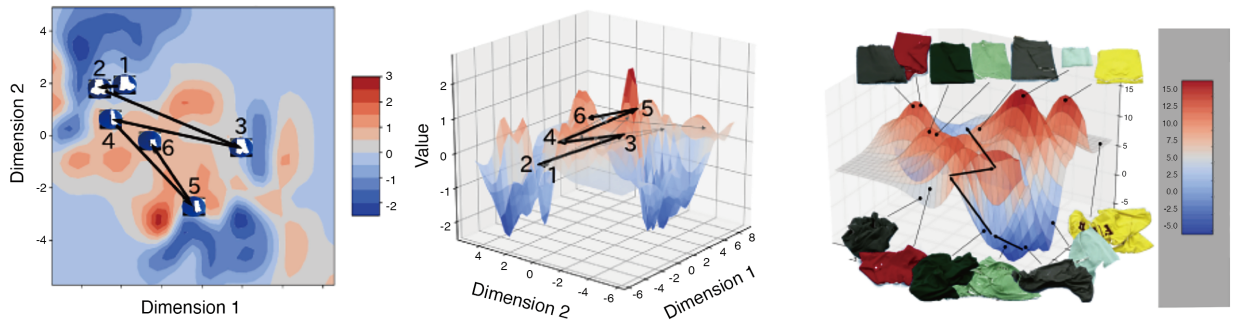


Figure 7.1: The utility landscape identifies desired states. This one, in particular, is trained from 45 cloth-folding video demonstrations. For visualization purposes, we reduce the state-space to two dimensions through multidimensional scaling (MDS). The canyons in this landscape represent wrinkled clothes, whereas the peaks represent well-folded clothes. Given this learned utility function, a robot chooses from an available set of actions to craft a motion trajectory that maximizes its utility.

- Learning an interpretable utility ranking model to explain the goal which is independent of system dynamics
- Derive a dynamics equation that uncouples the utility of a situation from the available set of actions.
- Teaching a robot to fold t-shirts, having it generalize to arbitrary articles of clothing.

7.1.2 Model

Definition 1. *Environment: The world (or environment) is defined by a generative composition model of objects, actions, and changes in conditions. Specifically, we use the stochastic context free And-Or graph (AOG).*

The atomic (terminal) units of this composition grammar are tuples of the form $(F_{start}, u_{[1:t]}, F_{end})$, where F_{start} and F_{end} are pre- and post-fluents of a sequence of interactions $u_{[1:t]}$. Concretely, the sequence of interactions $u_{[1:t]}$ is implemented by spatial and temporal features of human-object interactions (4D HOI).

Definition 2. *State: A state is a configuration of the believed model of the world. In our case, a state is a parse-graph (pg) of the And-Or graph, representing a selection of parameters (θ_{OR}) for each Or-node. The set of all parse-graphs is denoted Ω_{pg}.*

Definition 3. *Fluent: A fluent is a condition of a state that can change over time. It is represented as a real-valued function on the state (indexed by $i \in N$): $f_i : \Omega_{pg} \longrightarrow R$.*

Definition 4. *Fluent-vector: A fluent-vector F is a column-vector of fluents: $F = (f1, f2, ..., fk)^T$*

Definition 5. *Goal: The goal of a task is characterized by a fluent-change Δ_F. The purpose of learning the utility function is to identify reasonable goals.*

Utility Model

We assume human preferences are derived from a utilitarian model, in which a latent utility function assigns a real-number to each configuration of the world. For example, if a state pg^1 has a higher utility than another state pg^2, then the corresponding ranking is denoted $pg^1 \succ pg^2$, implying the utility of pg^1 is greater than the utility of pg^2.

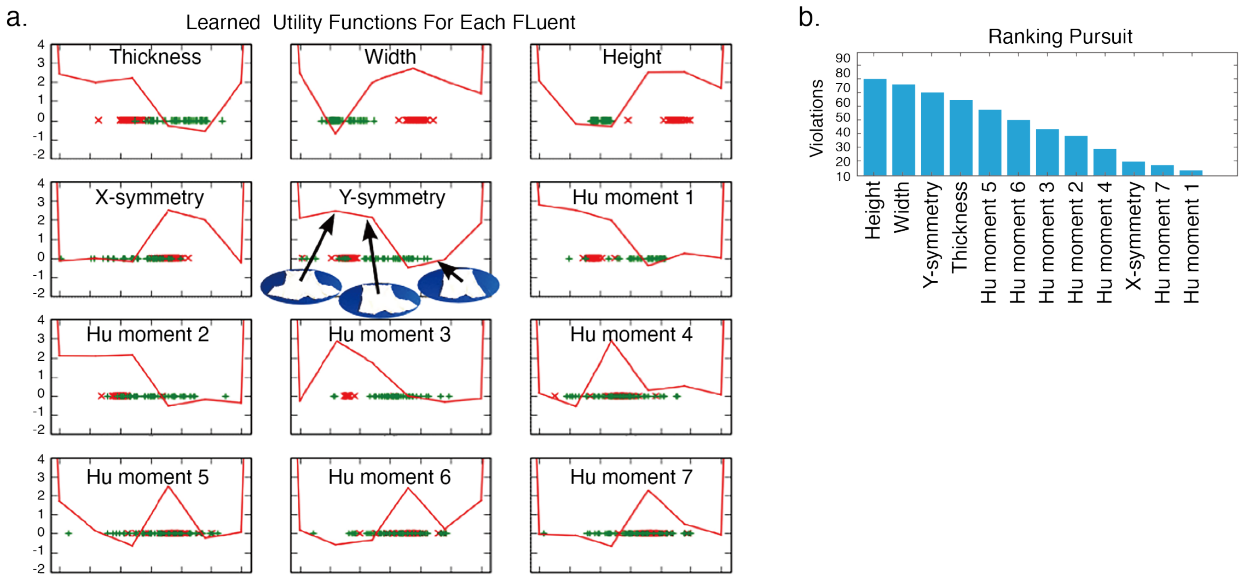


Figure 7.2: (a) The 12 curves represent the negative utility function corresponding to each fluent. The functions are negated to draw parallels with the concept of potential energy. Red marks indicate fluent values of pg^0, which the learned model appears to avoid, and the green marks indicate fluent values of the goal pg^*, which the learned model appears to favor. Notice how the y-symmetry potential energy decreases as the cloth becomes more and more symmetric. By tracing the change in utilities of each individual fluent, the robot can more clearly explain why it favors one state over another. (b) The ranking pursuit algorithm extracts fluents greedily to minimize ranking violations.As shown in the chart, the top 3 most important fluents for the task of cloth-folding are height, width,and y-symmetry.

Each video demonstration contains a sequence of n states $pg^0, pg^1, ..., pg^n$, which offers $\binom{n}{2}$ = n(n - 1)/2 possible ordered pairs (ranking constraints). Given some ranking constraints, we define an energy function by how consistent a utility function is with the constraints.

The energy function described above is used to design its corresponding Gibbs distribution. In the case of Zhu and Mumford [437], a maximum entropy model reproduces the marginal distributions of fluents. Instead of matching statistics of observations, our work attempts to model human preferences. We instead use a maximum margin formulation, and select relevant fluents by minimizing the ranking violations of the model. The specific details of this preference model is described below

Minimum Violations

Let $D = \{f^{(1)}, f^{(2)}, ...\}$ be a dictionary of fluents, each with a latent utility function $\lambda : R \longrightarrow R$. Using a sparse coding model, the utility of a parse-graph pg is estimated by a small subset of relevant fluents $F = \{f^{(1)}, f^{(2)}, ..., f^{(K)}\} \subset D$. Denote $\Lambda = \lambda^{(1)}(), \lambda^{(2)}(), ..., \lambda^{(K)}()$ as the corresponding set of utility functions for each fluent in F. For example, 12 utility functions learned from human preferences are shown in Fig. 7.2, approximated by piecewise linear functions. The total utility function is thus

$$U(pg; \Lambda, F) = \sum_{\alpha=1}^{K} \lambda^{\alpha}(f^{\alpha}(pg)) \tag{7.1}$$

Of all selection of parameters (Λ) and fluent-vectors (F) that satisfy the ranking constraints, we choose the model with minimum ranking violations. In order to learn each utility function in

Λ, we treat the space of fluents as a set of alternatives. Let R denote the set of rankings over the alternatives. Each human demonstration is seen as a ranking $\sigma_i \in R$ over the alternatives. We say $a \succ_{\sigma_i} b$ if person i prefers alternative a to alternative b. The collection of a person's rankings is called their preference profile, denoted $\vec{\sigma}$.

Each video v provides a preference profile $\vec{\sigma_v}$. For example, we assume at least the following ranking: $pg^* \succ_{\sigma_v} pg^0$, where pg^0 is the initial state and pg^* is the final state. The learned utility functions try to satisfy $U(pg^*) > U(pg^0)$.

U is treated as a ranking score: higher values correspond to more favorable states. We want to model the goal of a task using rankings obtained from visual demonstrations. The goal model, or preference model, of a parse-graph pg takes the Gibbs distribution of the form, $p(pg; \Lambda, F) = \frac{1}{Z} e^{U(pg;\Lambda,F)}$, where $U(pg; \Lambda, F)$ is the total utility function that minimizes ranking violations:

$$\begin{aligned} & min \sum_{\alpha=1}^{K} \int_x \lambda^{''(\alpha)} dx + C \sum_v \xi_v \\ & s.t. \sum_\alpha (\lambda^\alpha(f^\alpha(pg_v^*)) - \lambda^\alpha(f^\alpha(pg_v^0))) > 1 - \xi_v, \\ & \xi_v \geqslant 0 \end{aligned} \tag{7.2}$$

Here, ξ_v is a non-negative slack variable analogous to margin maximization. C is a hyper-parameter that balances the violations against smoothness of the utility functions Of all utility functions, we select the one which minimizes the ranking violations. The next section explains how to select the optimal subset of fluents.

Ranking pursuit

The empirical rankings of states $pg^* > pg^0$ in the observations must match the predicted ranking. We start with an empty set of fluents F = , and select from the elements of D that result in the least number of ranking violations.

This process continues greedily until the amount of violations can no longer be substantially reduced. Fig. 7.2 shows empirical results of pursuing relevant fluents for the cloth-folding task. The dictionary of initial fluents may be hand-designed or automatically learned through statistical means, such as from hidden layers of a convolutional neural network.

Ranking Sparsity

The number of ranking pairs we can extract from the training dataset is not immediately obvious. For example, each video demonstration supplies ordered pairs of states that we can use to learn a utility function. A sequence of n states $(pg^0, pg^1, ..., pg^n)$ allows $\binom{n}{2}$ = n(n - 1)/2 ordered pairs.

On one end of the spectrum, which we call sparse ranking, we know at the very least that $pg^n > pg^0$ for each demonstration. This is a safe bet since each video demonstration is assumed to successfully accomplish the goal. However, the utility model throws out useful information when ignoring the intermediate states.

On the other end, in dense ranking, all $\binom{n}{2}$ are used. Despite using all information available, this approach may be prone to introducing many ranking violations.

Fig. 7.3 visualizes performance of both approaches as we increment the number of available video demonstrations.

In recent years, there has been growing interest in studying object affordance in computer vision and graphics. As many object classes, especially man-made objects and scene layouts, are designed

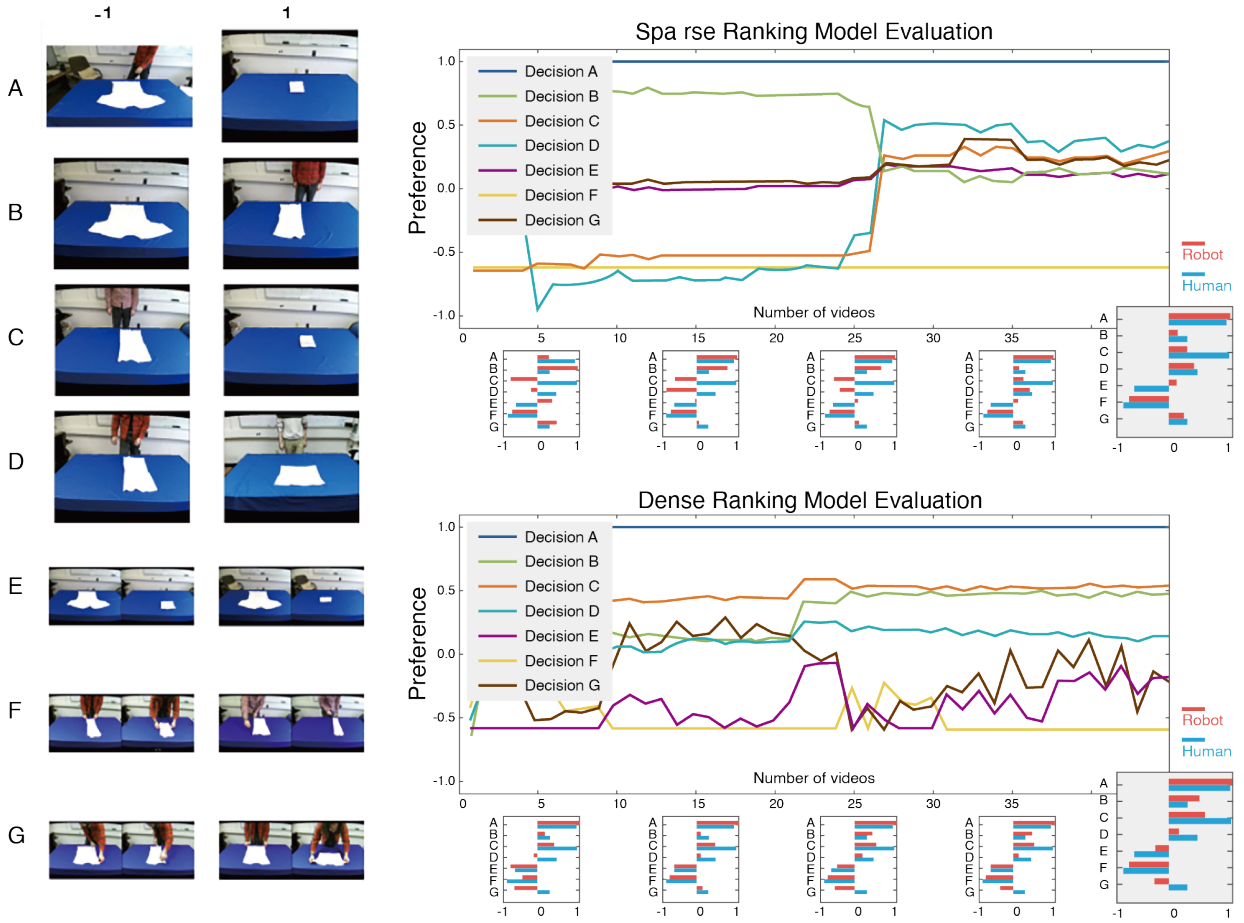


Figure 7.3: The sparse and dense ranking models are evaluated by how quickly they converge and how strongly they match human preferences. The x-axis on each plot indicates the number of unique videos shown to the learning algorithm. The y-axis indicates two alternatives (1 vs. -1) for 7 decisions (A, B, C, D, E, F, and G) of varying difficulty. The horizontal bar-charts below each plot show comparisons between human and robot preferences. As more videos are made available, both models improve performance in convergence as well as alignment to human preferences (from 330 survey results).

primarily to serve human purposes, the latest studies on object affordance include reasoning about geometry and function, thereby achieving better generalizations to unseen instances than conventional appearance-based machine learning approaches. In particular, Grabner *et al.* [128] designed an "affordance detector" for chairs by fitting typical human sitting poses to 3D objects.

Zhu *et al.* propose to go beyond visible *geometric compatibility* to infer, through physics-based simulation, the forces/pressures on various body parts (hip, back, head, neck, arm, leg, *etc.*) as people interact with objects. By observing people's choices in videos—for example, in selecting a specific chair in which to sit among the many chairs available in a scene (Fig. 7.4)—it can learn the *comfort intervals* of the pressures on body parts as well as human preferences in distributing these pressures among body parts. Thus, our system is able to "feel," in numerical terms, discomfort when the forces/pressures on body parts exceed comfort intervals. Zhu *et al.* argue that this is an important step in representing *human utilities*—the pleasure and satisfaction defined in economics and ethics (*e.g.*, by the philosopher Jeremy Benthem) that drives human activities at all levels. In our work, human utilities explain why people choose one chair over others in a scene and how they adjust their poses to sit more comfortably, providing a deeper and finer-grained account not only

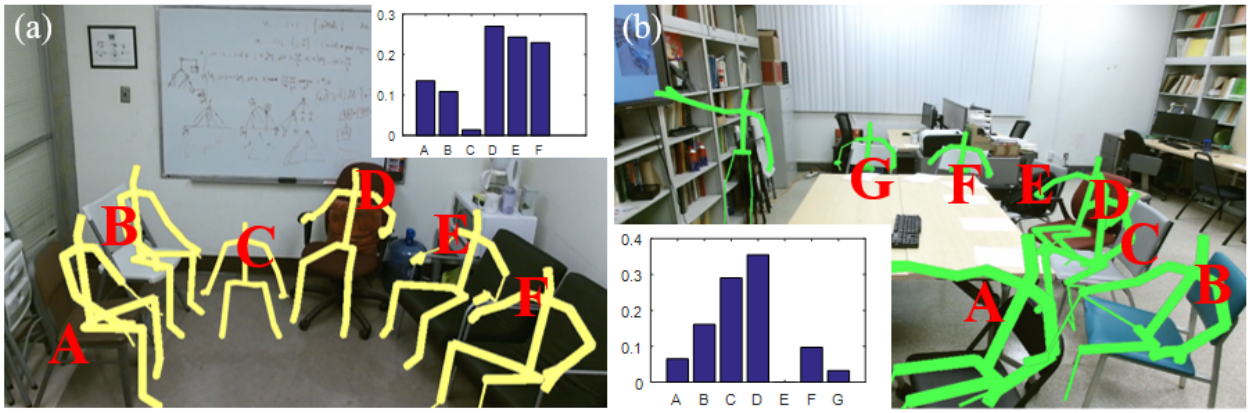


Figure 7.4: Examples of sitting activities in (a) an office and (b) a meeting room. In addition to geometry and appearance, people also consider other important factors including comfortability, reaching cost, and social goals when choosing a chair. The histograms indicate human preferences for different candidate chairs.

of object affordance but also of people's behaviors observed in videos.

In addition to comfort intervals for body pressures, our notion of human utilities also takes into consideration: (i) the tasks observed in a scene—for example, students conversing with a professor in an office (Fig. 7.4 (a)) or participating in a teleconference in a lab (Fig. 7.4 (b))—where people must attend to other objects and humans, and (ii) the space constraints in a planned motion—*e.g.*, the cost to reach a chair at a distance. In a full-blown application, this work demonstrate that human utilities can be used to analyze human activities, such as in the context of robot task planning.

7.1.3 Related Work

Modeling Affordance: The concept of affordance was first introduced by Gibson [438]. Hermans *et al.* [439] and Fritz *et al.* [440] predicted action maps for autonomous robots. Later, researchers incorporated affordance cues in shape recognition by observing people interacting with 3D scenes [281, 280, 153]. Adding geometric constraints, several researchers computed alignments of a small set of discrete poses [128, 121, 282]. By searching a continuous pose parameter space of shapes, Kim *et al.* [362] obtained accurate alignments between shapes and human skeletons. More recently, Savva *et al.* [441] predicted regions in 3D scenes where actions may take place. Applications that use affordance in scene labeling and object placement are reported in [442, 443, 130]. A closely related topic is to infer the stability and the supporting relations in a scene [129, 117, 257].

Inferring Forces from Videos: For pose tracking, Brubaker *et al.* [224, 225, 226] estimate contact forces and internal joint torques using a mass-spring system. More recently, Zhu *et al.* and Pham *et al.* [222, 227] use numerical differentiation methods to estimate hand manipulation forces. These methods are either limited to rigid body problems or employ oversimplified volumetric human models inadequate in simulating detailed human interactions with arbitrary 3D objects in scenes. In computer graphics, soft body simulation has been used to jointly track human hands and calculate contact forces from videos [229, 228].

Contributions

This work makes five major contributions:

- It incorporate physics-based, soft body simulations to infer the *invisible* physical quantities—*e.g.*, forces and pressures—during human-object interactions. To our knowledge, this is the first work to adopt state-of-the-art, physically accurate simulations to scene understanding. A major advantage of our method is its robustness in inferring both the forces and pressures acting on the entire human body as our model, which is comprised of more than 2,000 vertices, deforms in a realistic manner.
- Given a static scene acquired by RGB-D sensors, our proposed framework reasons about the relevant physics in order to synthesize creative, *physically stable* ways of sitting on objects.
- By incorporating a conventional robotics path planner, our proposed framework can generalize a static sitting pose to extend over a *dynamic* moving sequence.
- From human demonstrations, our system learns to generate the force histograms of each human body part, which essentially defines human utilities, such as comfortability, in terms of the force acting on each body part.
- We propose a method to robustly generate *volumetric* human models from the widely-used stick-man models acquired using Kinect sensors [363], and introduce a pipeline to reconstruct *watertight* 3D scenes with well-defined interior and exterior regions, which are critical to the success of physics-based scene understanding using advanced simulations.

Overview

The remainder of this chapter is organized as follows: In Section 7.1.4, we introduce our representation, which incorporates physical quantities into the spatiotemporal spaces of interest. In Section 7.1.5, we describe the pipeline for calculating the relevant physical quantities, which makes use of the Finite Element Method (FEM). In Section 7.1.6, we formulate the problem as a ranking task, and introduce a learning and inference algorithm under the assumption of rational choice. Section 7.1.7 demonstrates that our proposed framework can be easily generalized to challenging new situations. Section 7.1.8 concludes the chapter by discussing limitations and future work.

7.1.4 Representation

Spatial Entities and Relations in 3D Spaces

We represent sitting behaviors and associated relations in a parse graph pg, which includes (i) spatial entities—objects and human poses extracted from 3D scenes—and (ii) spatial relations—object-object and human-object relations.

Spatial Entities: For each frame of the input video, the parse graph pg is first decomposed into a static scene and a human pose. The static scene is further decomposed into a set of 3D objects, including chairs (Fig. 7.5 (b)). In this work, we consider only human poses related to sitting. We collect typical sitting poses using a Kinect sensor, and align and cluster them into 7 average poses (Fig. 7.5 (a)). For each average pose, we first convert the Kinect stick-man models (Fig. 7.6 (a)) into tetrahedralized human models (Fig. 7.6 (b)). These are then discretized into 14 pre-defined human body parts (Fig. 7.6 (c)) for simulations, as shown in Fig. 7.6 (d).

Spatial Relations: Pairs of objects extracted from 3D scenes form object-object relations, and each object and human pose pair forms a human-object relation. Fig. 7.7 (d)(e) show an example of spatial relations. For the purposes of this work, we define these two spatial relations as spatial features $\phi_s(pg)$ that encode the relative spatial distances and orientations. At a higher level, human-object relations also encode visual attention and social goals.

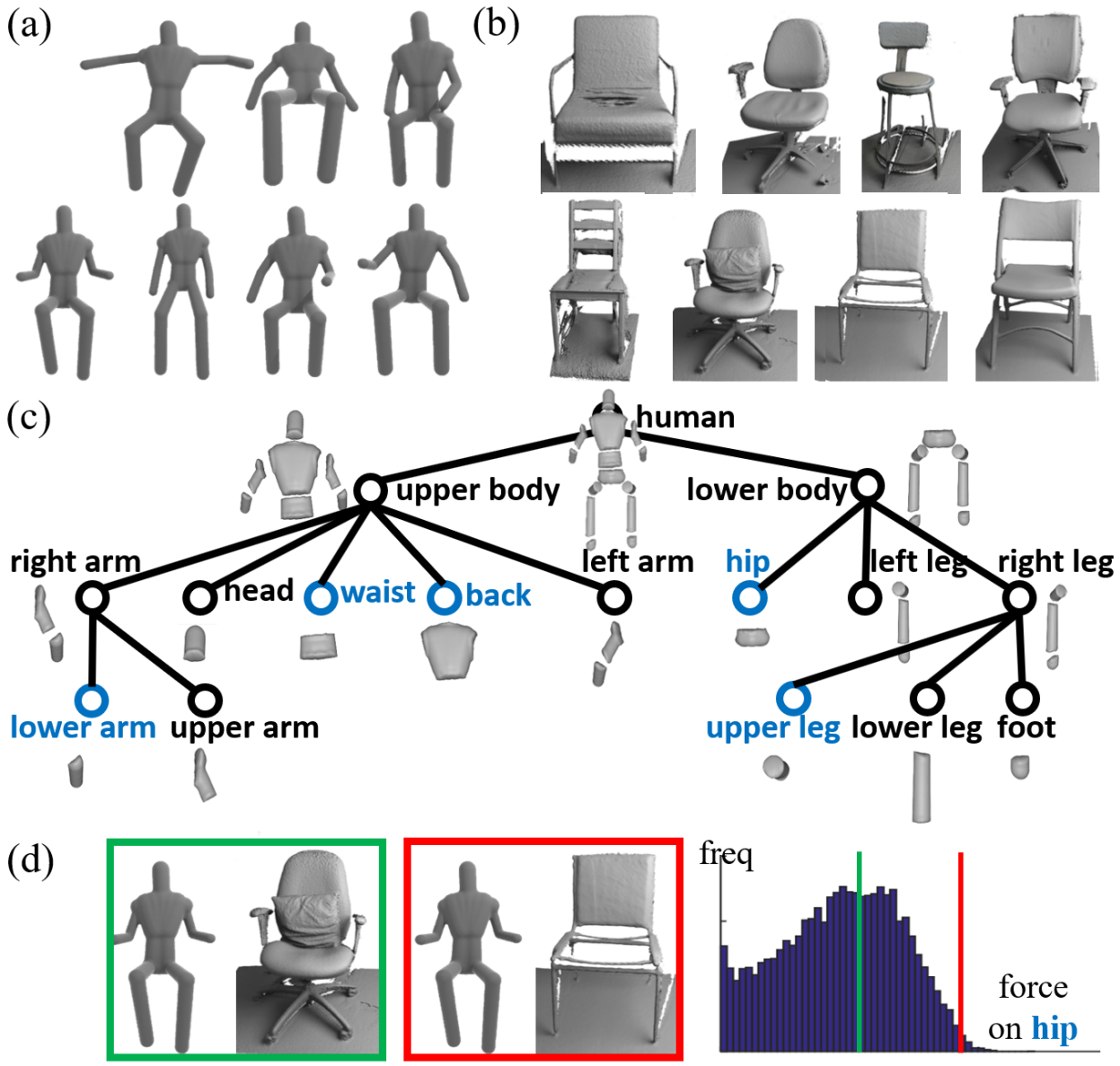


Figure 7.5: (a) We collect a set of human poses and cluster them into 7 average poses. (b) Various chairs extracted from scanned scenes. (c) Each human pose is decomposed into 14 body parts. When a human interacts with a chair, we infer the forces on each body part using FEM simulations. (d) Examples illustrating human preferences; green indicates a comfortable sitting activity, red an uncomfortable one.

Physical Quantities of Human Utilities

To date, researchers have mostly generated affordance maps by evaluating the geometric compatibility between people and objects [362, 443, 280, 130, 441, 153]. We employ a more meaningful and quantifiable metric—forces (including pressures) as physical quantities $\phi_p(pg)$ produced during human-object interactions. The forces acting on each body part essentially determines the *comfortability* of a person interacting with the scene. People tend to choose more comfortable chairs that will apparently provide better distributions of supporting forces at each body part (Fig. 7.5 (d)).

Deploying our physically simulated volumetric human models in the reconstructed scenes, we can estimate fine-grained external forces at each vertex of the human model, as shown in Fig. 7.6

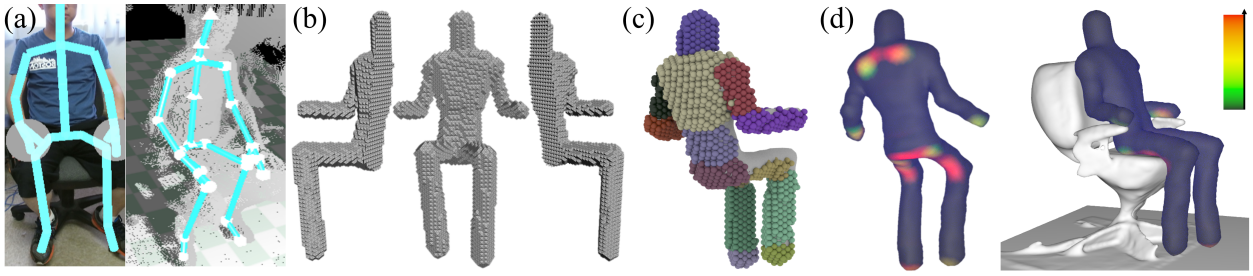


Figure 7.6: The stick-man model (a) captured using a Kinect is converted into a tetrahedralized human model (b) and then segmented into 14 body parts (c). Using FEM simulation the physical quantities $\phi_p(pg)$ are estimated at each vertex of the FEM mesh; the forces at each vertex are visualized in (d).

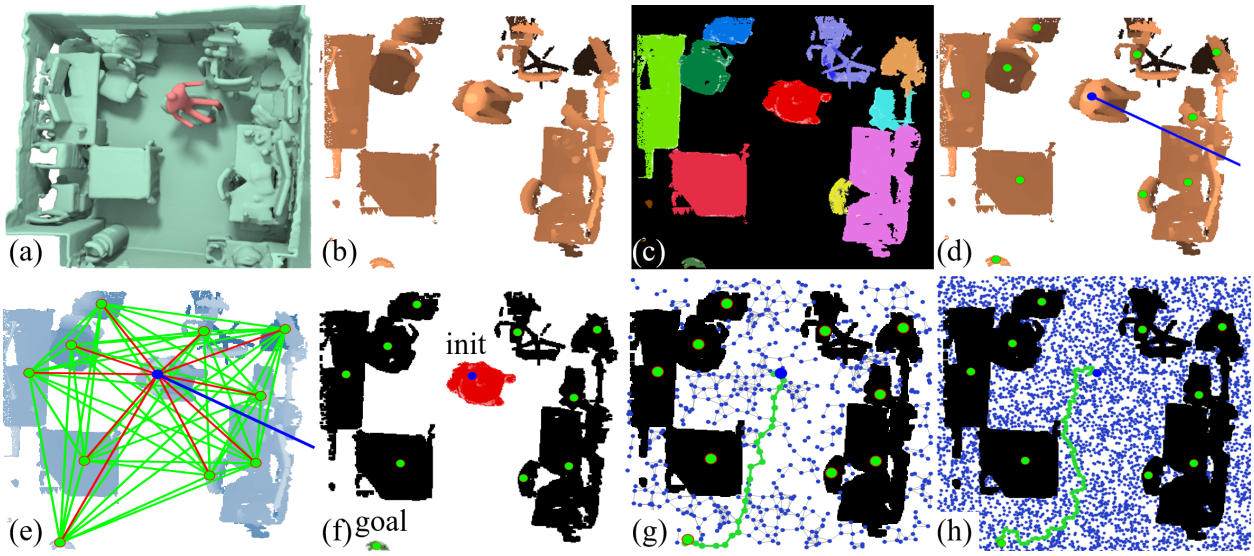


Figure 7.7: **Data pre-processing.** (a) Given a reconstructed 3D scene, (b) we project it down onto a planar map, and (c) segment 3D objects from the scene. (d) visualizes 3D object positions (green dots), human head position (blue dot), and orientation (blue line). (e) **Spatial features** $\phi_s(pg)$ are defined as human-object (red lines) and object-object (green lines) relative distances and orientations. (f) **Temporal features** $\phi_t(pg)$ are defined as the plan cost from a given initial position to a goal position. (g)(h) Two solutions generated by the PRM planner using graphs with different numbers of nodes (more nodes yield finer-grained plans at higher cost).

(d). In this work, we use the FEM to compute forces. The force acting on each body part can be estimated by summing up vertex-wise force contributions. A major advantage of using physical concepts is their ability to generalize to new situations.

Human Utilities in Time

To model the human utility, a plan cost $\phi_t(pg)$ is incorporated into our proposed framework. This is defined as a body pose sequence from a given initial state to a goal state, which encodes people's intentions and task planning through time. Compared to prior work, adding plan cost extends the solution space from a static human pose to *dynamic* pose sequences.

To simplify the problem, we use the Probabilistic Roadmap (PRM) planner [413] to calculate the plan cost. Viewed from above, we project the 3D scene to create a planar map, and use a 2D PRM to calculate the plan cost. However, our proposed framework does not preclude the use of more sophisticated planning methods in 3D space.

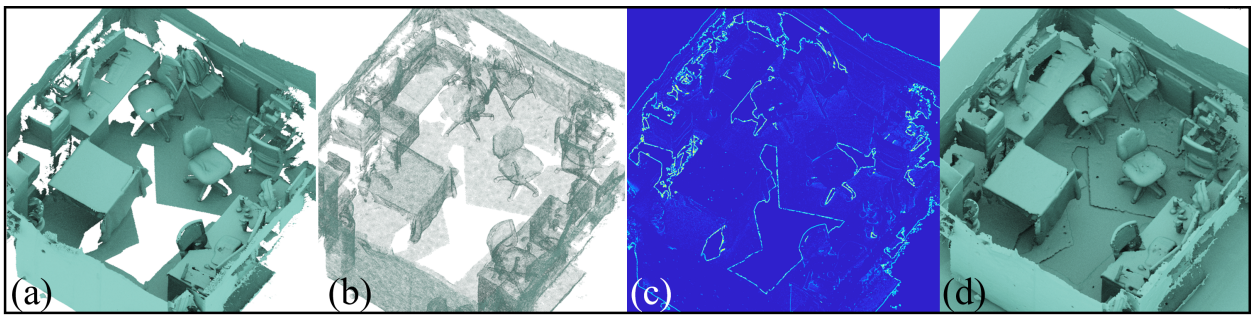


Figure 7.8: (a) From a reconstructed 3D indoor scene [444, 441], (b) we uniformly sample vertices in the input mesh with Poisson disk sampling [450], then convert them into a watertight mesh [451, 258] with well-defined interior and exterior regions. Differences (c) between the input mesh and the converted watertight mesh. By adding a ground geometry, we obtain a detailed, watertight reconstruction (d) of the 3D scene, which is inputted to the simulation.

7.1.5 Estimating the Forces in 3D Scenes

Dataset of 3D Scenes and Human Models

Our dataset includes reconstructed *watertight* 3D scenes, 3D objects (including chairs) extracted from the scenes, tracked human skeletons and *volumetric* human poses. The skeletons and volumetric human poses are registered in the reconstructed scenes.

The most distinguishing feature of our dataset relative to previous ones (*e.g.*, [444, 445, 446, 441]) is the watertight property of our reconstructed scenes. This is crucial for physics-based simulation methods such as the FEM. Furthermore, our dataset includes much larger variations of chair-shaped objects and human poses, as shown in Fig. 7.5 (a)(b), as well as more challenging and cluttered scenes.

Reconstructing Watertight Scenes

Reconstructing Closed-loop Scenes: Reconstruction methods that use purely geometric registration [271, 447, 448, 449] suffer from aliasing of fine geometric details and an inability to disambiguate different locations based on local geometry. Such problems are compounded when attempting to register loop closure fragments with low overlap. In our work, we reconstruct 3D scenes with global optimization based on line processes [444], resulting in detailed reconstructions with loop closures, as shown in Fig. 7.8 (a).

Converting to Watertight Scenes: Collision detection and resolution in the simulation requires a watertight scene mesh. We first use Poisson disk sampling [450] to generate uniformly distributed vertices from the input triangle mesh, as illustrated in Fig. 7.8 (b). Each vertex is then replaced with a fixed-radius sphere level set [258]. Subsequently, the Constructive Solid Geometry (CSG) union operation is applied to this level set and a ground level set to produce a complete scene with a filled-in floor. Finally, the Marching Cubes algorithm [451] is applied to the level set in order to generate the watertight surface, as shown in Fig. 7.8 (d). The resulting scene has the well-defined interior and exterior regions required by the simulation.

Modeling Volumetric Human Pose

Skeleton Alignment and Clustering: The resting poses of human skeletons acquired using the Kinect are aligned by solving the absolute orientation problem using Horn's quaternion-based

method [452]; *i.e.*, finding the optimal rotation and translation that maps one collection of vertices to another in a least squares sense:

$$\min \sum_i ||\mathbf{R}\mathbf{A}(:,i) + \mathbf{t} - \mathbf{B}(:,i)||^2, \tag{7.3}$$

where $\mathbf{A}$ and $\mathbf{B}$ are a $3 \times N$ matrices whose columns comprise the coordinates of the N source vertices and N target vertices, respectively. Presently, we have $N = 3$ (left shoulder, right shoulder, and spine base) for skeleton alignment. The K-means clustering algorithm [453, 454, 455] is then applied to cluster the resting poses into 7 categories, as shown in Fig. 7.5 (a).

Skeleton Skinning: Human skeleton data comprise joints, segments, and their orientations. For simplicity, an analytic geometric primitive is assigned to each body part. The primitives include ellipsoids (including spheres), hexahedra, and cylinders. The parameters of the primitives are chosen such that they best fit the body parts. A high-resolution level set is then applied to wrap around the union of all the primitives [258]; its zero isocontour approximates the skin [451].

Volumetric Discretization: Although the Marching Cubes algorithm suffices to extract a triangulated skin mesh from the level set, our simulation requires a full discretization of the volume bounded by the skin. To achieve this, we embed the skin level set into a body-centered cubic tetrahedral lattice as in [456]. This results in a tetrahedralized human shape geometry as shown in Fig. 7.6 (b).

Simulating Human Interactions With Scenes

As stated earlier, we chose the FEM to simulate human tissue dynamics. Our simulation requires only reconstructed watertight scenes and volumetric human poses as inputs. The outputs of the simulation are the relevant physical quantities $\phi_p(pg)$; *e.g.*, forces and pressures.

Elasticity: The human body is modeled as an elastic material. The total elastic potential energy is defined as

$$\Phi^E(\mathbf{x}) = \int_\Omega \Psi^E(\mathbf{x}) d\mathbf{x} \approx \sum_e V_e^0 \Psi^E(\mathbf{F}(\mathbf{x})), \tag{7.4}$$

where Ω is the simulation domain defined by the tetrahedral body mesh, $\mathbf{x}$ denotes the deformed vertex positions, and V_e^0 is the initial undeformed volume of tetrahedral element e. The hyperelastic energy density function Ψ^E is defined in terms of the deformation gradient $\mathbf{F} = \frac{\partial \mathbf{x}}{\partial \mathbf{X}}$, where $\mathbf{X}$ denotes the undeformed vertex positions. We use the fixed corotated elasticity model [457] for Ψ^E due to its robustness in handling large deformations.

Contact Forces: To model contact forces, we need to penalize penetrations of the human body mesh into the scene mesh. This requires a differentiable volumetric description of the scene geometry. With watertight scenes, the level set reconstruction is performed by directly computing signed distances from level set vertices to the mesh surface. In each simulation timestep, all human mesh vertices are checked against the scene level set. If a penetration is detected for vertex i, a collision energy $\Phi^C(\mathbf{x}_i)$ that penalizes the penetration distance in the normal direction is assigned to the corresponding vertex

$$\Phi^C(\mathbf{x}_i) = \frac{1}{2} k_c (\mathbf{x}_i - \mathcal{P}(\mathbf{x}_i))^2, \tag{7.5}$$

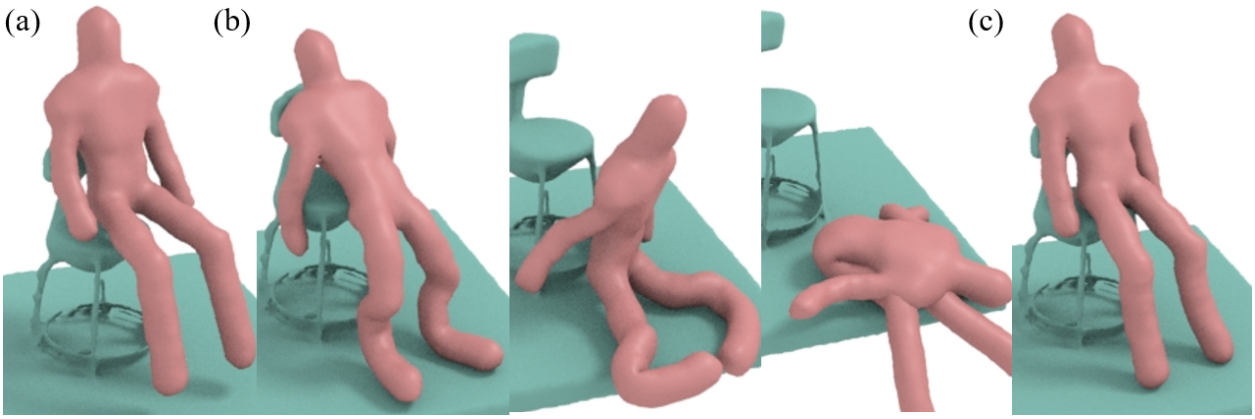


Figure 7.9: (a)Given an initial human pose in a 3D scene subject to gravity, (b) without adequate damping, the human body is too energetic and produces unnaturally bouncy motion. (c) With proper damping, the simulation converges to a physically stable rest pose in a small number of timesteps.

where k_c is a penalty stiffness constant and $\mathcal{P}(\mathbf{x}_i)$ projects $\mathbf{x}_i$ onto the closest point on the level set zero isocontour along its normal direction. To prevent free sliding along the collision geometry, we further introduce a friction force that slightly damps the tangential velocity for vertices in collision.

Dynamics Integration: Backward Euler time integration is used to solve the momentum equation. From time n to $n+1$, the nonlinear system to solve is

$$\mathbf{M}\frac{\mathbf{v}^{n+1}-\mathbf{v}^n}{\Delta t} = \mathbf{f}(\mathbf{x}^{n+1},\mathbf{v}^{n+1}) + \mathbf{M}g, \tag{7.6}$$

$$\mathbf{f}(\mathbf{x}^{n+1},\mathbf{v}^{n+1}) = \mathbf{f}^E(\mathbf{x}^{n+1}) + \mathbf{f}^C(\mathbf{x}^{n+1}) + \mathbf{f}^D(\mathbf{v}^{n+1}), \tag{7.7}$$

$$\mathbf{x}^{n+1}-\mathbf{x}^n = \mathbf{v}^{n+1}\Delta t. \tag{7.8}$$

Here $\mathbf{M}$ is the mass matrix, $\mathbf{x}$ denotes position, $\mathbf{v}$ denotes velocity, $\mathbf{f}^E = -\frac{\partial \Phi^E}{\partial \mathbf{x}}$ is the elastic force, $\mathbf{f}^C = -\frac{\partial \Phi^C}{\partial \mathbf{x}}$ is the contact force, $g = 9.8m/s$ is gravity, and $\mathbf{f}^D = -\nu\mathbf{v}$ is an additional force to dampen the velocities, where ν is the damping coefficient. Fig. 7.9 (b) shows that without the damping force, the deformable human body model is too energetic and may produce unnaturally bouncy motion. While there exist more accurate viscoelastic material models of human tissue, our simple damping force is easy to implement and achieves similar behaviors for the simulation results. We solve the above nonlinear system for positions $\mathbf{x}^{n+1}$ and velocities $\mathbf{v}^{n+1}$ using Newton's method [458].

Table 7.1: Physical simulation parameters

Timestep: $1 \times 10^{-3}s$	Density: $1000kg/m^3$	Young's modulus: $0.15kPa$	Poisson's ratio: 0.3
Collision stiffness: $1 \times 10^4 kg/s^2$	Friction coeff: 1×10^{-3}	Damping coeff: $50kg/s$	Gravity: $9.81m/s^2$

Simulation Outputs: When the simulation comes to rest, $\mathbf{v} = \mathbf{0}$ and the damping forces vanish. The elastic, contact, and gravity forces sum to zero everywhere over the mesh. As the output of the simulation, we export the computed contact forces acting on the skin surface.

7.1.6 Learning and Inferring Human Utilities

Extracting Features

We craft features $\boldsymbol{\phi}(pg)$ of three types: (i) spatial features $\phi_s(pg)$ encoding spatial relations, (ii) temporal features $\phi_t(pg)$ associated with plan cost, and (iii) physical quantities $\phi_p(pg)$ produced during human interactions with scenes.

Data pre-processing is illustrated in Fig. 7.7 (a)-(c). Given a reconstructed watertight scene, we remove the ground plane by setting a 0.05 m depth threshold and projecting it down onto a planar map. 3D objects in the scene are first segmented into primitives [274] and then grouped into object segments as in [118, 116]. Some manual labeling and processing is needed for certain cluttered scenes. Finally, a semantic label is manually assigned to each object; *e.g.*, a desk with a monitor, a door, *etc.*

Spatial features $\phi_s(pg)$ are defined as human-object / object-object relative distances and orientations as shown in Fig. 7.7 (d)(e). For each object, the geometric center is obtained by averaging over all the vertices. The human head position and orientation is acquired with the Kinect.

Temporal features $\phi_t(pg)$ are defined as the plan cost from a given initial position to a goal position. To simplify the problem, we project the 3D scene down onto a planar map. We build a binary obstacle map where the free spaces devoid of objects have unit costs, whereas the spaces occupied by objects have infinite costs. We use a 2D PRM planner to calculate the costs using 2D human positions and head orientations. Thus the planner constructs a probabilistic roadmap to approximate the possible motions. Finally, the optimal path is obtained using Dijkstra's shortest path algorithm [459]. Fig. 7.7 (f)–(h) show two solutions using different numbers of nodes in the planner graph.

Physical quantities $\phi_p(pg)$ produced by people interacting with scenes are computed using the FEM. Currently, we consider only the forces and pressures acting on 14 body parts of the tetrahedralized human model, as shown in Fig. 7.5 (c). The net force on each body part is obtained by summing up the forces at all its vertices. The net force divided by the number of contributing vertices yields the local pressure. Fig. 7.6 (d) illustrates a force heatmap for sitting.

Learning Human Utilities

The goal in the learning phase is to find the proper coefficient vector $\boldsymbol{\omega}$ of the feature space $\boldsymbol{\phi}(pg)$ that best separates the positive examples of people interacting with the scenes from the negative examples.

Rational Choice Assumption: We assume that in interacting with a 3D scene, the *observed* person makes near-optimal choices to minimize the cost of certain tasks. This is known as rational choice theory [460, 461, 462, 463]. More concretely, the person tries to optimize one or more of the following factors: (i) the human-object and object-object orientations and distances defined as $\phi_s(pg)$, (ii) the plan cost from the current position to a goal position $\phi_t(pg)$, and (iii) the physical quantities $\phi_p(pg)$ that quantify the comfortability of interactions with the scenes.

In accordance with rational choice theory, for an observed person choosing an object (*e.g.*, an armchair) on which to sit, their choice $pg^\star$ is assumed to be optimal; hence, this is regarded a positive example. If we *imagine* the same person making random choices $\{pg_i\}$ by randomly sitting on other objects (*e.g.*, the ground), the rational choice assumption implies that the costs of the imagined configurations $\{pg_i\}$ should be higher; hence, these should be regarded negative examples.

Let us consider a simplified scenario as an example: Suppose the ground-truth factors that best explain the observed demonstration are that the object is comfortable to sit on and that it faces the blackboard. Then, other objects in the imagined configurations should fall into one of the following three categories: they (i) may be more comfortable, but have less desirable orientations relative to the blackboard, or (ii) may have better orientations with the blackboard, but be less comfortable, or (iii) may be less comfortable and have worse orientations.

To summarize, under the rational choice assumption, we consider the *observed* rational person interacting with the scenes $pg^\star$ a positive example, and the *imagined* random configurations $\{pg_i\}$ as negative examples. However, the random generated configurations $\{pg_i\}$ may be similar or even identical to the observed optimal configuration $pg^\star$. To avoid this problem, we remove random configurations that are too similar to observed configurations before applying the learning algorithm.

Ranking function: Based on the rational choice assumption, it is natural to formulate the learning phase as a ranking problem [332]—the *observed* rational person interaction $pg^\star$ should have lower cost than any *imagined* random configurations $\{pg_i\}$ with respect to the correct coefficient vector $\boldsymbol{\omega}$ of $\boldsymbol{\phi}(pg)$, which includes spatial relations $\phi_s(pg)$, plan cost $\phi_t(pg)$, and physical quantities $\phi_p(pg)$. Each coefficient ω_i reflects the importance of its corresponding feature. The ranking function is defined as

$$R(pg) = \langle \boldsymbol{\omega}, \boldsymbol{\phi}(pg) \rangle. \tag{7.9}$$

Learning the ranking function is equivalent to finding the coefficient vector $\boldsymbol{\omega}$ such that the maximum number of the following inequalities are satisfied:

$$\langle \boldsymbol{\omega}, \boldsymbol{\phi}(pg^\star) \rangle > \langle \boldsymbol{\omega}, \boldsymbol{\phi}(pg_i) \rangle, \quad \forall i \in \{1, 2, \cdots, n\}, \tag{7.10}$$

which corresponds to the rational choice assumption that the observed person's choice is near-optimal.

To approximate the solution to the above NP-hard problem [464], we introduce non-negative slack variables ξ_i [259]:

$$\min \frac{1}{2}\langle \boldsymbol{\omega}, \boldsymbol{\omega} \rangle + \lambda \sum_i^n \xi_i^2, \quad \forall i \in \{1, \cdots, n\} \tag{7.11}$$

$$\text{s.t. } \xi_i \geqslant 0, \quad \langle \boldsymbol{\omega}, \boldsymbol{\phi}(pg^\star) \rangle - \langle \boldsymbol{\omega}, \boldsymbol{\phi}(pg_i) \rangle > 1 - \xi_i^2, \tag{7.12}$$

where λ is the trade-off parameter between maximizing the margin and satisfying the pairwise relative constraints.

Inferring the Optimal Affordance

Given a static scene, the goal in the inference phase is to find, among all the *imagined* configurations $\{pg_i\}$ in the solution space, the best configuration $pg^\star$ that receives the highest score:

$$pg^\star = \arg\max_{pg_i} \langle \boldsymbol{\omega}, \boldsymbol{\phi}(pg_i) \rangle. \tag{7.13}$$

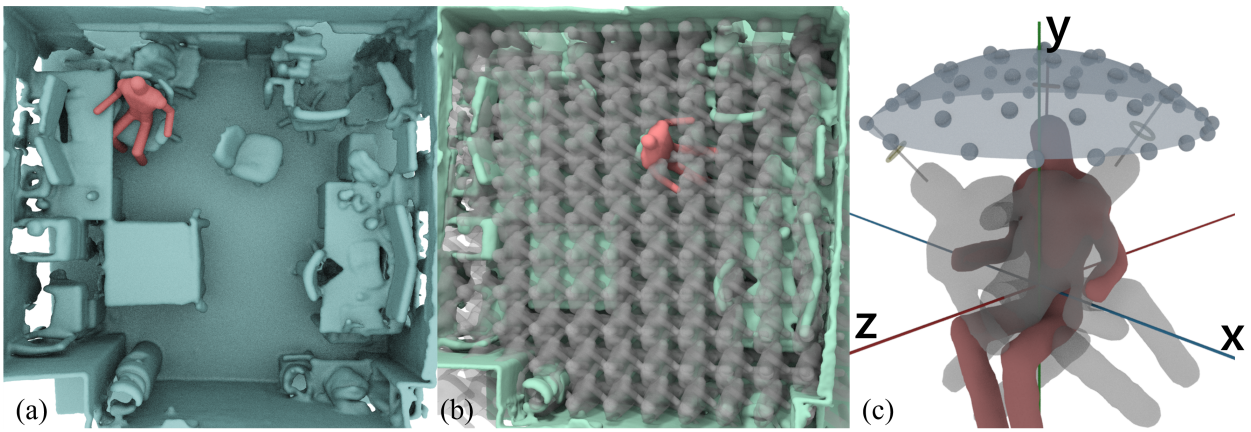


Figure 7.10: In the learning phase, based on rational choice theory, we assume that the observed demonstration is optimal, and therefore regard it a positive example. (a) In this example, a person is sitting on an armchair facing a desk with a monitor. The learning algorithm then imagines different configurations $\{pg_i\}$ in the solution space by initializing with different human poses P_a, (b) translations T_b, and (c) orientations O_c. The imagined randomly generated configurations $\{pg_i\}$ are regarded negative examples. In the inference phase, the inference algorithm performs the same sampling process (b)(c), and finds the optimal configuration $pg^\star$ with the highest score.

Sampling the Solution Space

Without observing a human interacting with the scenes, the inference algorithm must sample the solution space by imagining different configurations $\{pg_i\}$. The same sampling process is also required in the learning phase to generate negative examples.

We first quantize the human poses into the 7 categories shown in Fig. 7.5 (a). The imagined configurations of the human model are initialized with different poses P_a, translations T_b, and orientations O_c, as shown in Fig. 7.10 (b)(c). The tuple (P_a, T_b, O_c) specifies a unique human configuration. Given such a tuple, the simulation will impose gravity and the simulated human model will reach its rest state. The methods described in Section 7.1.6 are then used to extract the features $\boldsymbol{\phi}(pg_i)$.

In the learning phase, the $\boldsymbol{\phi}(pg_i)$ are then used to learn the ranking function (Eq. (7.9)). In the inference phase, the extracted features are then evaluated by Eq. (7.13). The configuration with the highest score is taken as the optimal configuration $pg^\star$.

7.1.7 Experiments

Learning Human Utilities From Demos

A set of demonstrations of people sitting in the scene were collected using RGB-D sensors, as shown in Fig. 7.10 (a). The observed demonstrations were then used as positive training examples. For each 3D scene, we further generated over 4,000 different configurations pg_i by enumerating all poses and randomly sampling different initial human translations and rotations in the solution space, as shown in Fig. 7.10 (b)(c). The synthesized configurations that are similar to the human demonstrations were pruned. The remaining configurations were used as negative examples. The learning algorithm (Eq. (7.9)) learned the coefficient vector $\boldsymbol{\omega}$ of the ranking function under three different settings: (i) physical quantities $\phi_p(pg)$, (ii) with additional spatial relations $\phi_s(pg)$, and (iii) with all features $\phi_p(pg)$, $\phi_s(pg)$, and $\phi_t(pg)$.

Fig. 7.11 (a) shows the final force histograms of 6 (out of 14) body parts. Unsurprisingly when

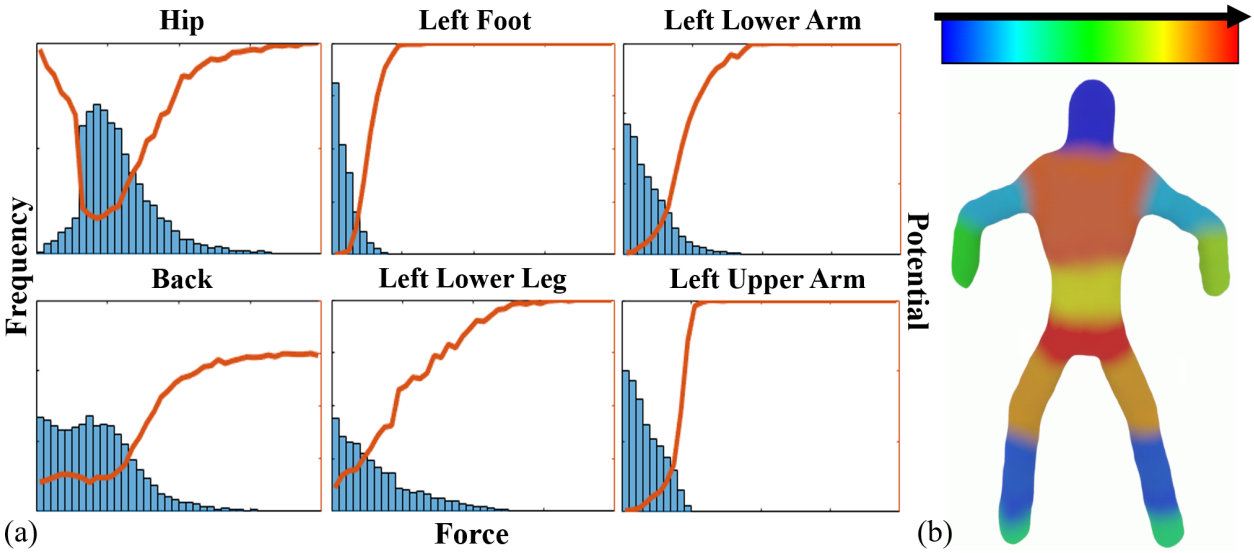


Figure 7.11: (a) The final force histograms of 6 (out of 14) body parts. The x axis indicates the magnitudes of the forces, the y axis their frequencies and potential energy. Histogram areas reflect the number of cases with non-zero forces. (b) The average forces of each body part normalized and remapped to a T pose.

sitting, forces act on the hip in almost all cases, upper legs and lower arms also tend to be subject to relatively large magnitude forces, upper arms and heads are much less likely to interact with the scene, and the feet contact the scene in many cases, but with overall small force magnitudes. The heat map of the average forces acting on each human body part over all the collected human sitting activities is shown in Fig. 7.11 (b).

Inferring Optimal Affordance in Static Scenes

Next, we tested the learned models on our dataset as well as on prior 3D datasets [441, 444] in three different scenarios: (i) canonical scenarios with chair-shaped objects, (ii) cluttered scenarios with severe object overlaps, and (iii) novel scenarios extremely different from the training data.

The first testing was done in the same scene as the training. Fig. 7.12 shows examples of the top ranked human poses. Although using physical quantities $\phi_p(pg)$ produced physically plausible sitting poses (Fig. 7.12 (a)), some of the results do not look like sitting poses (*e.g.*, lying poses and upside-down poses). Such diverse results are caused by the lack of spatial and temporal constraints.

Including the spatial features $\phi_s(pg)$, the relative orientations and distances between the human model and objects in the scene, improved the results, as shown in Fig. 7.12 (b). Intuitively, the top poses become more natural because they share similar human attentions and social goals to those in the observed demonstrations. For the case shown in Fig. 7.12, the relative orientation between the human model and the desk with monitor prunes the configurations for which the human poses are not facing towards the monitor. The laying poses and upside-down poses are also pruned.

Integrating the temporal features $\phi_t(pg)$ also takes into consideration the plan cost, which prunes the poses with large plan cost differences compared to the observed person demonstrations. Note that the plan cost used in temporal features enables our system to output a dynamic moving sequence, which extends the static sitting poses in previous work.

Additional results including canonical, cluttered, and novel scenarios from our dataset and other datasets [441, 446, 444, 465] are shown in Fig. 7.13.

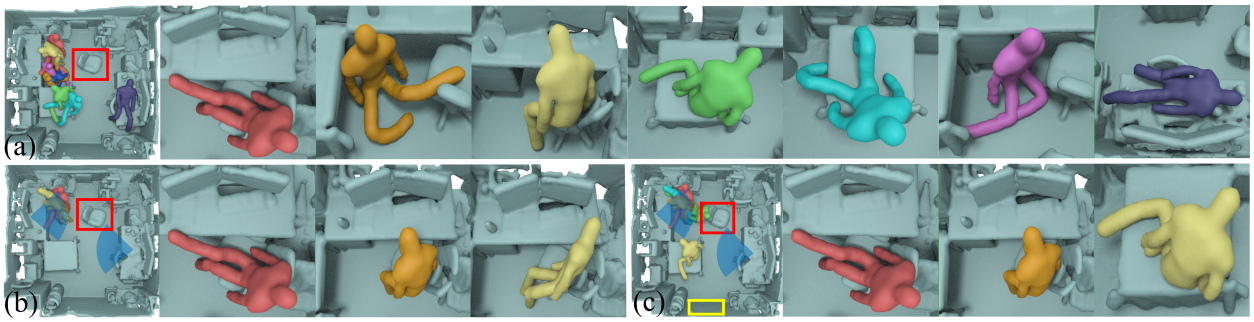


Figure 7.12: (a) The top 7 human poses using physical quantities $\phi_p(pg)$. The algorithm seeks physically comfortable sitting poses, resulting in casual sitting styles; *e.g.*, lying on the desk. (b) Improved results after adding spatial features $\phi_s(pg)$ to restrict the human-object relative orientations and distances. Further including temporal features $\phi_t(pg)$ yields the most natural poses (c). The yellow bounding box indicates the door, the initial position for the path planner. Samples generated near the 3D chair labeled with a red bounding box do not produce high scores as forces apply on the arms of the person in the observed demonstration (Fig. 7.10 (a)). The lack of chair arms leads to low scores.

Evaluations: We asked 4 subjects to rank the highest-scored sitting poses. Fig. 7.14 plots the correlations between their rankings and our system's output.

7.1.8 Discussion and Future Work

The current stream of studies on object affordance [281, 280, 153, 128, 121, 282, 362, 441, 222] have attracted increasing interest on geometry-based methods, which offer more generalization power than the prevailing appearance-based machine learning approach. We have taken a step further by inferring the invisible physical quantities and learning human utilities based on rational human behaviors and choices observed in videos. Physics-based simulation is more general than geometric compatibility, as suggested by the various "lazy/casual seated poses" that are typically not observed in public videos. We argue that human utilities provide a deeper and finer-grained account for object affordance as well as for human behaviors. Incorporating spatial context features, temporal plan costs, and physical quantities computed during simulated human-object interactions, we demonstrated that our framework is general enough to handle novel cases using models trained from canonical cases.

Our current work has several limitations that we will address in future research: First, we have assumed a rigid scene. We shall consider various material properties of objects and allow two-way causal interactions between the objects and human models. This promises to enable deeper scene understanding with the help of more sophisticated hierarchical task planners. Second, currently we model the anatomically complex human body simply as a homogeneous elastodynamic material. We believe that a more realistic biomechanical human model with articulated bones actuated by muscles surrounded by other soft tissues (see, *e.g.*, [466, 467]) could enable our framework to yield more refined solutions. Optimal motor controllers could also be employed within the human simulation to support fine-grained motor planning, thus going beyond task planning, although this will increase computational complexity.

By solving these problems, we will be a step closer to consolidating several different research streams and associated methods in vision, graphics, cognition, and robotics.

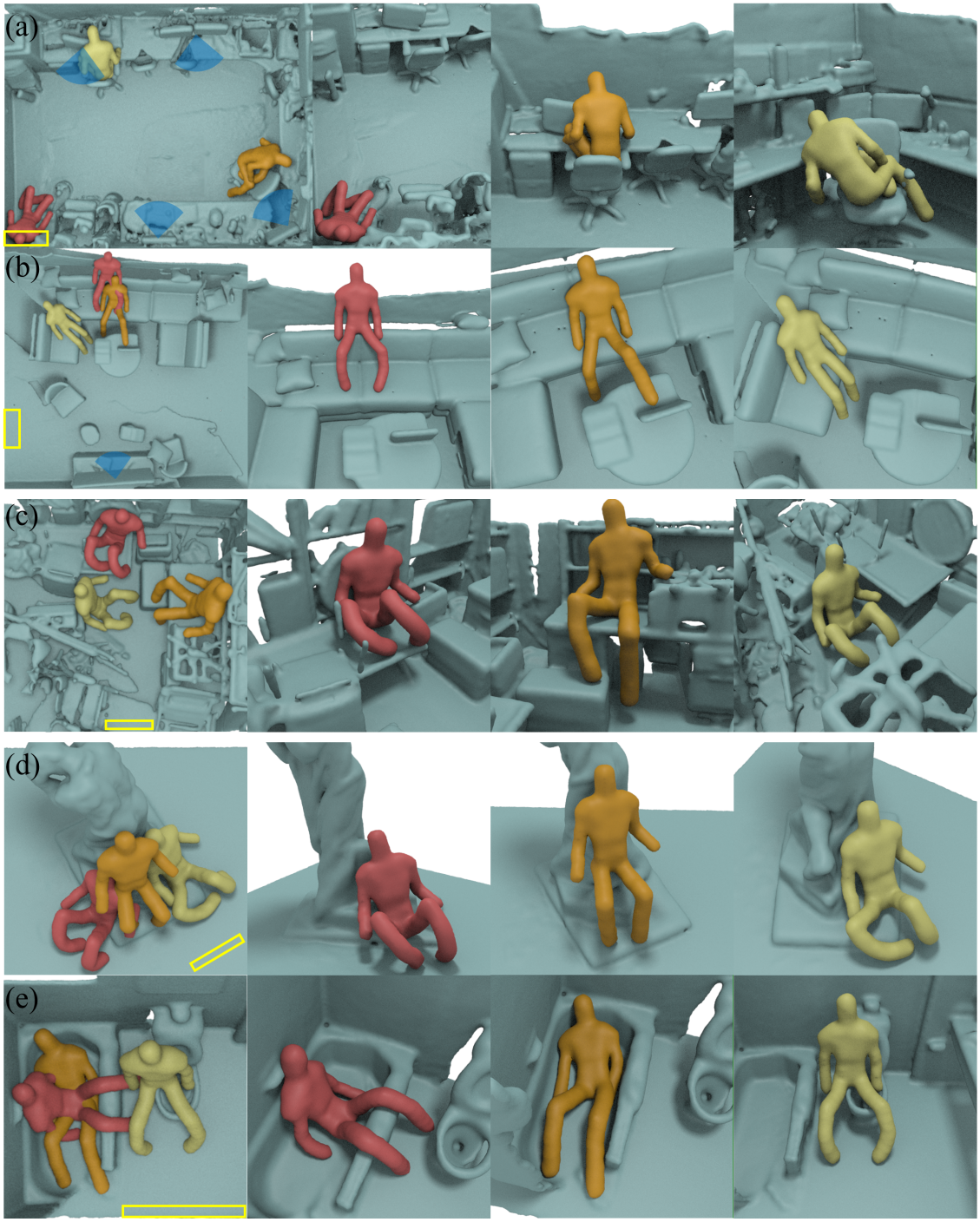


Figure 7.13: Top 3 poses in (a)(b) canonical scenarios, (c) cluttered scenarios, and (d)(e) novel scenarios. All the features $\boldsymbol{\phi}(pg)$ are used in (a) and (b). Both physical quantities $\phi_p(pg)$ and plan costs $\phi_t(pg)$ are used in (c)–(e). The initial position for the path planner is indicated by the yellow bounding box.

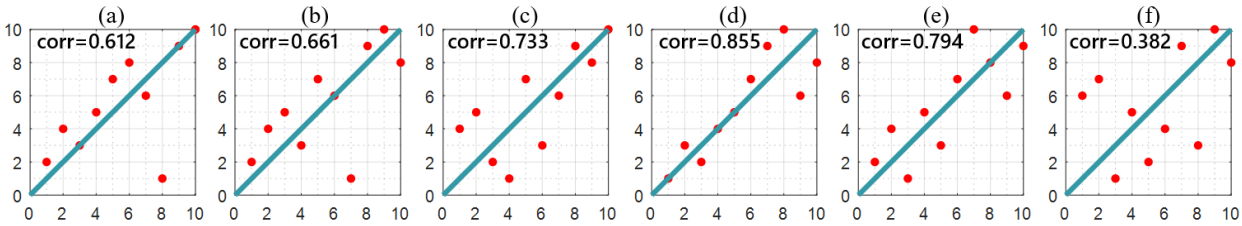


Figure 7.14: Correlations of the ranking by human subjects (x-axis) and our system's output (y-axis). The closer the plotted points fall to the diagonal lines the better our proposed method matches the performance of the human subjects. Plots (a)–(e) correspond to Fig. 7.13 (a)–(e). Plot (f) corresponds to Fig. 7.12 (c).

Chapter 8

Nonverbal Communication: Attention, Gaze, Pointing, and Coattention

8.1 Nonverbal Behavior

What is the role of nonverbal behavior in everyday life? Other than verbal cues, nonverbal behavior seems to provide an indirect means of knowing more about other people. The information gleaned from nonverbal behavior is more representative of "true" characteristics, attitudes, and feelings than that offered verbally; most people assume nonverbal behavior to be spontaneous and sincere than verbal behavior [468]. Much of what social psychologists think about nonverbal behavior derives from a proposal made more than a century ago by Charles Darwin. Darwin argues that we have such nonverbal behaviors primarily because they are the vestiges of serviceable associated habits, *i.e.*, behaviors that had specific and direct functions earlier in our evolutionary history [469, 470]. Over the course of evolution, such behaviors have acquired communicative value, *i.e.*, they provide others with external evidence of an individual's internal state, and persisted even though they no longer serve the original purposes [471].

The functions that nonverbal behaviors serve could be interpersonal or intrapersonal. The interpersonal functions involve information such behaviors convey to others, regardless of whether they are employed intentionally or serve as the basis of an inference the listener makes about the speaker [469]. The intrapersonal functions involve noncommunicative purposes the behaviors serve.

How do they contribute to our understanding of the speaker's message? Some evidence that gestures can convey nonsemantic information, and it is not too difficult to think of circumstances in which such information could be useful. Here, the study of speech and gestures overlaps with the study of person perception and attribution processes, because gestures, in their cultural and social context, may enter into the process by which we draw conclusions about people—their backgrounds, their personalities, their motives and intentions, their moods and emotions, *etc.* Further, since the significance of gestures can be ambiguous, it is likely that our beliefs and expectations about the speaker-gesturer will affect the meanings and consequences we attribute to the gestures we observe.

Another way of pursuing this question is to ask how gesturing affects the way listeners process verbal information. Do gestures help engage a listener's attention? Do they activate imagistic or motoric representations in the listener's mind? Do they become incorporated into representations that are invoked by the listener when the conversation is recalled? One hypothesis is that gestures facilitate the processes by which listeners construct mental models of the events and situations described in an arrative. Communication has been defined as the process by which representations that exist in one person's mind come to exist in another's [472].

On a second construal, the question "What do conversational hand gestures tell us?" concerns the intrapersonal functions of gesture—here, the role they play in speech production. It might be paraphrased "How does gesturing affect us when we speak? The "us" in this interpretation is the speaker, and the about what has to do with the ideas the speaker is trying to articulate in speech. Our response to this question is that gestures are an intrinsic part of the process that produces speech, and that they aid in the process of lexical access, especially when the words refer to concepts that are represented in spatial or motoric terms. They "tell us" about the concepts underlying our communicative intentions that we seek to express verbally. In this way, conversational gestures may indirectly serve the function conventionally attributed to them. That is, they may indeed enhance the communicativeness of speech, not by conveying information that is apprehended visually by the addressee, but by helping the speaker formulate speech that more adequately conveys the communicative intention.

Considering the functions of conversational gestures reminds us that although linguistic representations derive from propositional representations of experience, not all mental representation is propositional. Spatial knowledge and motoric knowledge may have their own representational formats, and some components of emotional experience seem to be represented somatically. These representations (perhaps along with others) will be accessed when we recall and think about these experiences. However, when we try to convey such experiences linguistically, we must create new representations of them, and there is some evidence that so doing can change how we think about them. For example, describing a face makes it more difficult to recognize that face subsequently [473], and this "verbal overshadowing" effect, as it has been termed, is not limited to representations of visual stimuli [474, 475, 476]. Linguistic representations may contain information that was not part of the original representations, or omit information that was. It is possible that gestures affect the internal representation and experience of the conceptual content of the speech they accompany, much as facial expressions are believed to affect the experience of emotion.

8.1.1 cooperative communication

Tourists manage to survive and interact effectively in many situations in foreign cultures, in which no one shares their conventional language, precisely by relying on such naturally meaningful forms of gestural communication.

The central claim in Tomasello's book "Origins of human communication" is that to understand how humans communicate with one another using a language and how this competence might have arisen in evolution, we must first understand how humans communicate with one another using natural gestures. The first uniquely human forms of communication were pointing and pantomiming. The social-cognitive and social-motivational infrastructure that enabled these new forms of communication then acted as a kind of psychological platform on which the various systems of conventional linguistic communication (all 6,000 of them) could be built. Pointing and pantomiming were thus the critical transition points in the evolution of human communication, already embodying most of the uniquely human forms of social cognition and motivation required for the later creation of conventional languages.

The problem is that, compared with conventional human languages (including conventionalized sign languages), natural gestures would seem to be very weak communicative devices, as they carry much less information "in" the communicative signal itself.

Suppose that you and I are walking to the library, and out of the blue I point for you in the direction of some bicycles leaning against the library wall. Your reaction will very likely be "Huh?" as you have no idea which aspect of the situation I am indicating or why I am doing so, since, by itself, pointing means nothing. But if some days earlier you broke up with your boyfriend in a

particularly nasty way, and we both know this mutually, and one of the bicycles is his, which we also both know mutually, then the exact same pointing gesture in the exact same physical situation might mean something very complex like "Your boyfriend's already at the library (so perhaps we should skip it)." On the other hand, if one of the bicycles is the one that we both know mutually was stolen from you recently, then the exact same pointing gesture will mean something completely different. Or perhaps we have been wondering together if the library is open at this late hour, and I am indicating the presence of many bicycles outside as a sign that it is.

And so our question is: how can something as simple as a protruding finger communicate in such complex ways, and do so in such different ways on different occasions? Any imaginable answer to this question will have to rely heavily upon cognitive skills of what is sometimes called mindreading, or intention-reading. Thus, to interpret a pointing gesture one must be able to determine: what is his intention in directing my attention in this way? But to make this determination with any confidence requires, in the prototypical instance, some kind of joint attention or shared experience between us (Wittgenstein's [1953] forms of life; Bruner's [1983] joint attentional formats; Clark's [1996] common conceptual ground).

For example, if I am your friend from out of town and there is no way I could be familiar with your ex-boyfriend's bicycle, then you will not assume that I am indicating it for you. This is true even if, by some miracle, I do indeed know that this is his bicycle, but you do not know that I know this. In general, for smooth communication it is not enough that you and I each know separately and privately that this is his bicycle (and even that the other knows this); rather, this fact must be mutually known common ground between us. And in the case in which it is common ground between us that this is his bicycle, but not that the two of you have just broken up (even if we each know this privately), then you will probably think that I am indicating your boyfriend's bicycle as a way of encouraging our entrance into the library, not discouraging it. The ability to create common conceptual ground—joint attention, shared experience, common cultural knowledge is an absolutely critical dimension of all human communication, including linguistic communication with all of its he's, she's, and it's.

The other remarkable aspect of this mundane example of human pointing, from an evolutionary perspective, is its prosocial motivation. I am informing you of your ex-boyfriend's likely presence or the location of your stolen bicycle simply because I think you would want to know these things. Communicating information helpfully in this way is extremely rare in the animal kingdom, even in our closest primate relatives (in chapter 2 we will deal with examples such as warning cries and food calls). Thus, when a whimpering chimpanzee child is searching for her mother, it is almost certain that all of the other chimpanzees in the immediate area know this. But if some nearby female knows where the mother is, she will not tell the searching child, even though she is perfectly capable of extending her arm in a kind of pointing gesture. She will not tell the child because her communicative motives simply do not include informing others of things helpfully.

In contrast, human communicative motives are so fundamentally cooperative that not only do we inform others of things helpfully, but one of the major ways we request things from others is simply to make our desire known in the expectation that they will volunteer help. Thus, I may request a drink of water by simply stating that I want one (informing you of my desire), knowing that, in most instances, your tendency to be helpful (and our mutual knowledge of this) turns this act of informing into what is effectively a full-blown request.

Human communication is thus a fundamentally cooperative enterprise, operating most naturally and smoothly within the context of (1) mutually assumed common conceptual ground, and (2) mutually assumed cooperative communicative motives. The fundamentally cooperative nature of human communication is, of course, the basic insight of Grice (1957, 1975) [477, 478].

It turns out that human cooperation is unique in the animal kingdom in many ways, both

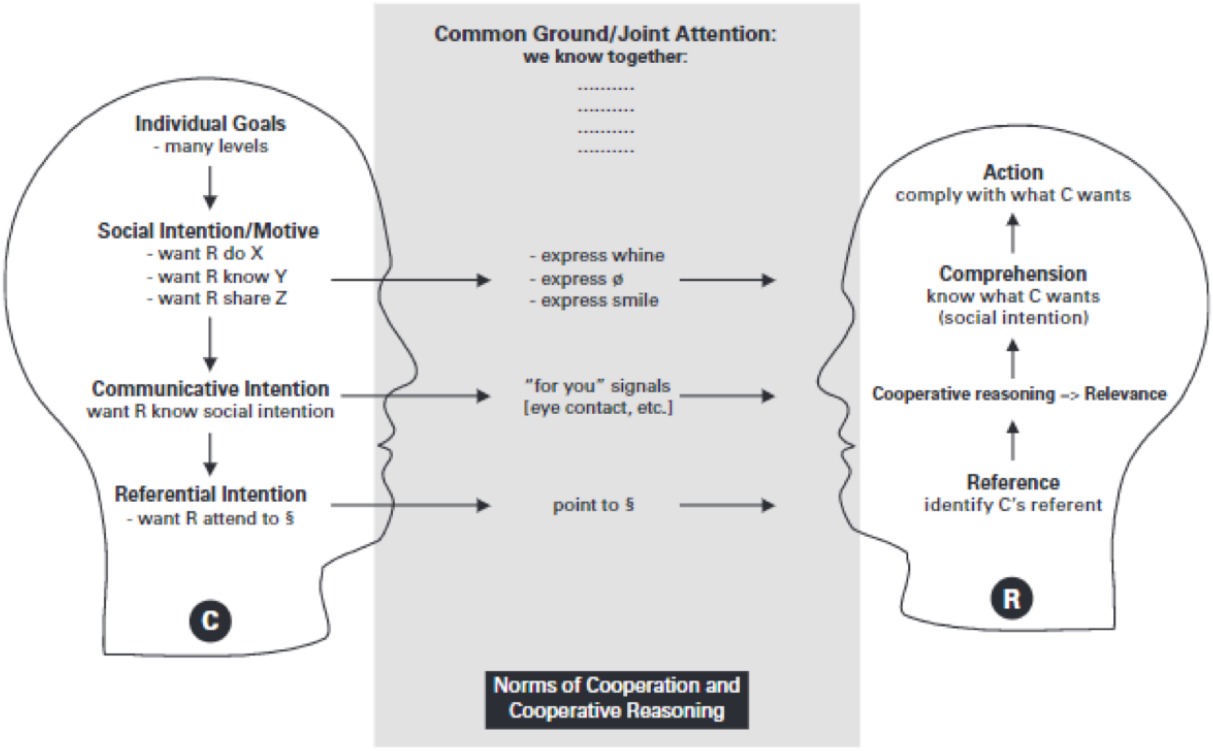


Figure 8.1: communication architecture proposed by Tomasselo [140]

structurally and motivationally

1) Human cooperative communication emerged first in evolution (and emerges first in ontogeny) in the natural, spontaneous gestures of pointing and pantomiming. 2) Human cooperative communication rests crucially on a psychological infrastructure of shared intentionality, which originated evolutionarily in support of collaborative activities, and which comprises most importantly: social-cognitive skills for creating with others.

Specifically, human cooperation is structured by what some modern philosophers of action call shared intentionality or "we" intentionality [479, 480, 481]. In general, shared intentionality is what is necessary for engaging in uniquely human forms of collaborative activity in which a plural subject "we" is involved: joint goals, joint intentions, mutual knowledge, shared beliefs—all in the context of various cooperative motives.

The proposal is thus that human cooperative communication—whether using "natural" gestures or "arbitrary" conventions—is one instance, albeit a special instance, of uniquely human cooperative activity relying on shared intentionality [482]. The skills and motivations of shared intentionality thus constitute what we may call the cooperative infrastructure of human communication.

Fig. 8.1 depicts all of the different components of the cooperation model of human communication, and something of their interrelations. Beginning in the top left and following the arrows, very sketchily: I as communicator have many goals and values that I pursue in my life: my individual goals. For whatever reason, I feel that you can help me on this occasion with one or more of them, by helping me or accepting my offer of information (which I want to make for my own reasons) or sharing attitudes with me: my social intention. The best way for me to get your help, or to help you, or to share with you in this situation is through communication, and so I decide to make mutually manifest to us (in our current joint attentional frame) a communicative act; this is my

communicative intention (perhaps indicated by "for you" signals such as eye contact or with some expression of motive). Given my signal of a communicative intention, I draw your attention to some referential situation in the external world—my referential intention—which is designed (along with some expression of motive) to lead you to infer my social intention via processes of cooperative reasoning, since you are naturally motivated to find out why I want to communicate with you (based on mutual assumptions or norms of cooperation). You thus first attempt to identify my referent, typically within the space of our common ground, and from there attempt to infer my underlying social intention, also typically by relating it to our common ground. Then, assuming you have comprehended my social intention, you decide whether or not to cooperate as expected.

8.2 Attention and Gaze

Attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information while ignoring other perceivable information, which has also been described as the allocation of limited cognitive processing resources. Human attention could be driven by internal goals or external stimulation. Goal-driven attention is referred to as top-down or endogenous attention, whereas stimulus-driven attention is referred to as bottom-up or exogenous attention [483]. Allocating attention over short time periods can be referred to as phasic orienting, while maintaining attention over longer time periods is referred to as sustained attention, or vigilance [484]. Endogenous and exogenous attention systems interact and compete over short time periods to guide behavior.

We don't indistinguishably perceive all the stimuli around us since our ability to attend to the things around us is limited in terms of both capacity and duration. Actually, human attention acts somewhat like a spotlight, highlighting the details that we need to focus on and casting irrelevant information to the sidelines of our perception. For example, in a party full of all kinds of noises, you find yourself still able to tune out the irrelevant sounds and focus on the amusing story that your dinner partner shares, which is a good example of the so-called selective attention. Selective attention is the process of focusing on a particular object in the environment for a certain period of time. Human need selective attention to tune out unimportant details and focus on what really matters to save our precious and limited cognitive resource.

In fact, we see much less than we think. Inattentional blindness is a perceptual phenomenon discovered by psychologists that can support selective attention. A best-known study, called Invisible Gorilla Test [485], demonstrates inattentional blindness vividly, which is shown in Fig. 8.2. In the experiment, subjects are asked to watch a video where a group of students wearing white or black T-shirts pass a basketball among themselves. Subjects are asked to count the number of times the players wearing white T-shirts pass the ball. And researchers find that subjects often fail to notice a person in a gorilla suit who appears in the center of the video scene. Another experiment also illustrates people's blindness to change, in which an experimenter held a map and asked a random pedestrian for direction. While the pedestrian was looking at the map, they replaced the experimenter with a different one. Nearly 50 percent of the pedestrians failed to notice that they were talking to a different person [486].

In the domain of computer vision, efforts have been made in modelling the mechanism of human attention, especially the bottom-up attentional mechanism. There are two kinds of models to mimic the bottom-up saliency mechanism. One way is based on the spatial contrast analysis and the other way is based on the frequency domain analysis. Social attention is one special form of attention that involves the allocation of limited processing resources in a social context. Previous studies on social attention often regard how attention is directed towards socially relevant stimuli such as

Figure 8.2: The Invisible Gorilla Test. © 2020 Simons, Daniel J. Reprinted, with permission, from Ref. [485].

faces and gaze directions of other individuals. Social attention operates at two polarizing states: attending-to-others and attending-to-self, which mark the two ends of an otherwise continuum spectrum of social attention. For a given behavioral context, the mechanisms underlying these two polarities might interact and compete with each other in order to determine a saliency map of social attention that guides human behaviors. An imbalanced competition between these two behavioral and cognitive processes will cause cognitive disorders and neurological symptoms such as autism spectrum disorders and Williams syndrome.

Gaze is an important cue for human attention in social interaction. Gaze direction provides a number of potential social cues which may be utilized by an individual to learn about the external (other individuals, objects, events, *etc.*) or internal (emotional and intentional) states. But gaze is not the only cue that is used to determine the focus of another individual's direction of attention. The whole head, in particular the orientation in which it is directed is a sufficient indicator of attention direction. In some cases, the eyes are not visible and the only cue available for processing is the head direction. If the head is occluded or in shadow, the orientation of the body provides a sufficient cue for communication, as shown in Fig. 8.3. If all cues are available for processing, a hierarchy of importance exists whereby the eyes provide more important cues than the head, and the head is more important cue than the body. Determining the direction of another individual's attention is easier to establish from larger visual cues, such as the head. However, the eyes present a more precise indicator of where another is looking, though they are much smaller than the head. An evidence is that humans have a larger extent of white sclera either side of the dark central iris compared to other non-human primates. This ratio may be one of the factors which has allowed humans to use the orientation of other individual's eyes for learning about objects in the environment in referential active communication [487].

As people speak, their gaze periodically fluctuates toward and away from their conversational partner. Some investigators have interpreted gaze directed at a conversational partner as an expression of intimacy or closeness [488, 489, 490, 491]. However, Butterworth [492] argues that gaze direction is affected by two complex tasks speakers must manage concurrently: planning speech, and monitoring the listener for visible indications of comprehension, confusion, agreement, interest, *etc.* [493, 494]. When the cognitive demands of speech planning are great, Butterworth argues, speakers avert gaze to reduce visual information input, and, when those demands moderate, they redirect their gaze toward the listener, especially at places where feedback would be useful. Such communicative effects could involve two rather different mechanisms. In the first place, many nonverbal behaviors are to some extent under the individual's control, and can be produced voluntarily.

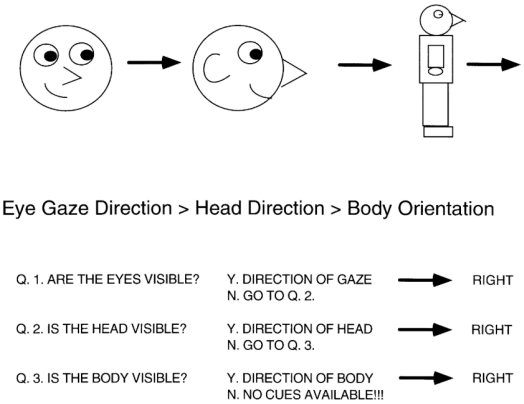


Figure 8.3: Different indicators for attention direction. © 2020 N.J. Emery. Reprinted, with permission, from Ref. [487]

For example, although a smile may be a normal accompaniment of an affectively positive internal state, it can at least to some degree be produced at will. Social norms, called "display rules," dictate that one exhibit at least a moderately pleased expression on certain social occasions. In the second place, nonverbal behaviors that serve noncommunicative functions can provide information about the noncommunicative functions they serve. For example, an excessive amount of gaze aversion may lead a listener to infer that the speaker is having difficulty formulating the message.

8.3 Gaze Communication

Gaze communication is the most primitive form of human communication, whose underlying social-cognitive and social-motivational infrastructure acted as a psychological platform on which various linguistic systems could be built [140]. Although verbal communication has become the primary form in social interaction, gaze communication still plays an important role in conveying hidden mental state and augmenting verbal communication [495, 470]. Evidence from psychology suggests that eyes are a cognitively special stimulus, with unique "hard-wired" pathways in the brain dedicated to their interpretation and humans have the unique ability to infer others' intentions from eye gazes [487]. To understand human-human communication in the social scene better, we not only need natural language processing (NLP), but also require a systematical study of human gaze communication mechanisms.

Over the past decades, lots of research [496, 497, 498, 499] on the types and effects of social gazes have been done in cognitive psychology and neuroscience communities. Typical eye gaze types include *Mutual gaze*, *Gaze aversions*, *Referential gaze*, *Gaze following*, *Joint attention* [495, 487].

Mutual gaze occurs when two agents have eye contact or look into eyes of each other. It is the strongest mode of establishing a communicative link between human agents. Mutual gaze can capture attention, initialize a conversation, maintain engagement and expresses feelings of trust and extroversion.

Gaze aversion happens when gaze of one agent is shifted away from another in order to avoid mutual gaze. Averted gaze expresses distrust, introversion, fear, and can also modulate intimacy, communicate thoughtfulness or signal cognitive effort such as looking away before responding to a question.

Referential gaze happens when one agent trys to use gaze to induce another agent's attention to a certain stimuli. Referential gaze shows intents to share or request something.

Gaze following occurs when one agent perceives gaze from another and follows to contact with the stimuli the other is attending to. In the visual domain, this involves perceiving the social partner's gaze, translating between his/her reference frame and our own by replicating or simulating the other's viewpoint, and extending our attention to include the other's putative visual focus. The ability to follow gaze is prerequisite to the ability to jointly attend and infer another's intentions and goals from his/her bodily behavior [500].

Eye gaze cues influence human attention within a tenth of a second. Attention is allocated in the direction of gaze despite the fact that the gaze cues had no predictive value and were thus irrelevant. Actually, human follow gaze even when explicitly informed that cues are counterpredictive of target location. Gaze following, which requires only those cognitive resources that arise within a few hundred milliseconds of stimulus onset, might be a fully modular behavior: fast, simple, reflexive, and once triggered, unalternable. To accurately follow gaze, we interpret another's bodily orientation in spatial relation to our own. This rich understanding of others' bodies and of three-dimensional visual space permits us to take their perspective, following their gaze geometrically to objects outside our immediate visual field [500].

Most animals appear to understand others' gaze as a vector within a rich, three-dimensional environment. Like humans, many animals appear to have expectations about what they should see when they follow another's gaze: if they find nothing, they do a double check. And both human and nonhuman animals modulate their gaze following behavior based on context. Gaze following can synchronize attentional shifts toward a common stimulus. Thus, gaze following plays pivotal role in achieving joint attention [500].

Joint Attention appears when two agents have the same intention to share attention on a common stimuli and both know that they are sharing attention with each other. Joint attention consists of several phases, including mutual gaze to establish communication channel, referential gaze to draw attention to a stimuli, following gaze to check the referred stimuli, and mutual gaze again to confirm the joint attention. Joint attention is located at the intersection of a complex set of capacities that serve our cognitive, emotional, and action-oriented relations with others. It involves social cognition, our ability to understand others, what they intend, and what their actions mean. Thus, joint attention is a significant first step for individuals to represent others' minds and develop Theory of Mind (ToM). Only when we have joint attention can we start to consider and understand others' perspective, attention and even intention. Also, without joint attention, different individuals cannot truly communicate and update their common mind. Infants need to learn to form joint attention with others to enter a more deeply social world of interconnecting attitudes and experiences.

What is joint attention? Are we in joint attention when I see you looking at something and I follow your gaze and turn to look at it too? When we go to see a film, are we in joint attention with all the other audience members? Although joint attention is crucially important in human social life, as a field we have not yet come to a full agreement on what exactly joint attention is.

Some researchers also call it shared attention in some literatures. Some researchers assume that shared attention and joint attention are different. For example, Fig. 8.4 shows a kind of definition of joint attention and a categorization of different types of gaze interaction patterns. Many situations that typically have been thought of as joint attention situations can actually involve individual, parallel attention rather than truly joint, shared attention. Or we could say that these joint attention situations have different degree of sharedness. Nevertheless, there are two important criteria for true joint attention: 1) the motivation to share attention in the first place and 2) that the participants know together that they are sharing attention [500].

If two individuals happen to look at the same object simultaneously, they are not in joint attention but just parallel individual attention. If an individual follow the other's gaze to look at

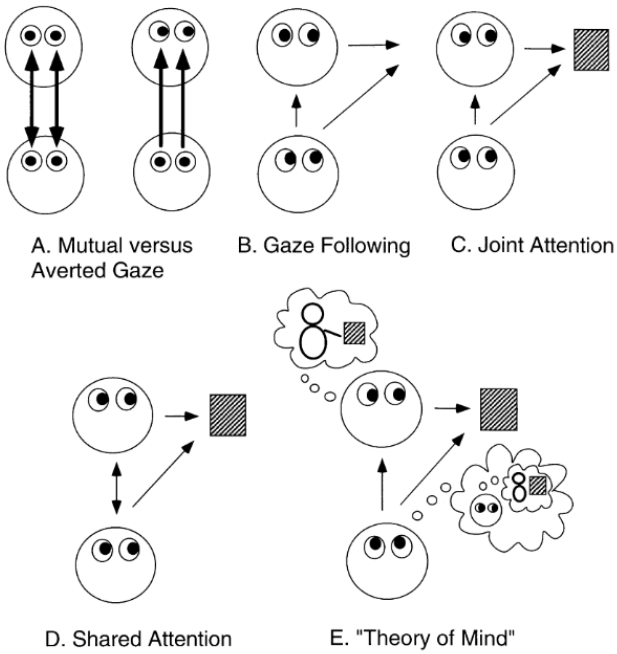


Figure 8.4: One kind of categorization of different gaze interaction patterns. © 2000 N.J. Emery. Reprinted, with permission, from Ref. [487].

the object, it's still not joint attention since they don't have the mutual knowledge of each other's attention. Only when they share their attention through some communication channel, such as a mutual gaze or voice, can we say that they have true joint attention. Fig. 8.5 (a) indicates the classic joint attention triangle. When two individuals get into joint attention, they not only know the object individually, but also know the other individual's attention on the same object, and even more, they know that the other individual knows that they know, and so on... The two minds are just like two mirrors reflecting each other infinitely. See Fig. 8.5 (b). But this recursive mind reading approach to model joint attention seems to have problem because 1) the processing demands are just too high and humans don't achieve mutual knowledge in this infinitely recursive way and 2) the two individuals are not truly joint together. It is basically two individual, solitary, parallel perspectives that never meet in the middle. There has to be some concept like "we-attention" to model true joint attention. And the sharing of attention in true joint attention involves communication, which may be something as simple as a meaningful look [500].

There are roughly two kinds of joint attention, "top-down" and "bottom-up", as shown in Fig. 8.6 [500]. In the top-down situation, the person who wishes to initiate joint attention actively directs the other person's attention to something. In this situation, three types of communicative looks are usually involved. The first type of look is an initiation look by the initiator to the recipient, which serves to get the recipient's attention. This look is an "invitation to interact" and opens the channel of communication between the two partners. It signals the initiator's communicative intention. The second type of communicative look is a reference look toward the object or event that the initiator wants to call attention to. It signals the initiator's referential intention and is usually accompanied by a gesture like a pointing or nodding towards the object. These two looks thus serve to open the joint attentional interaction and establish the topic or referent of it. The third type of look in the top-down joint attention situation is the sharing look. If the initiation look serves to open the joint attentional triangle, the sharing look serves to close it. Whereas the initiation look is

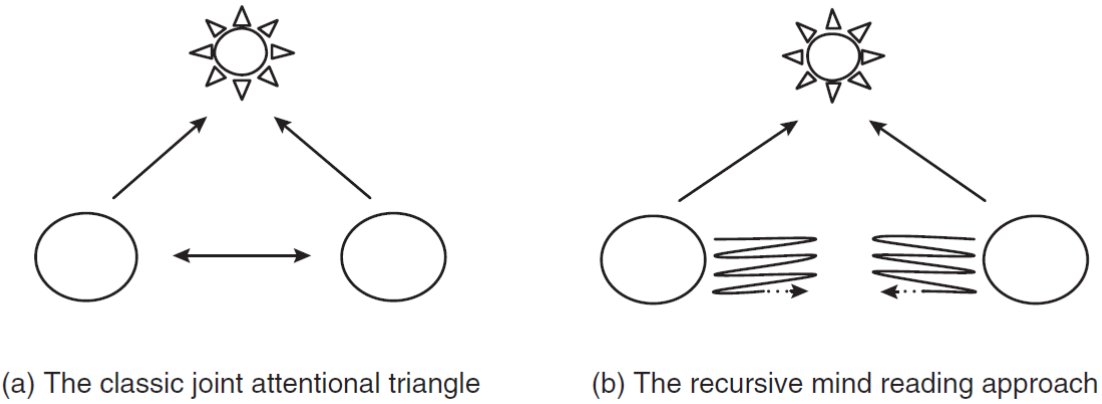


Figure 8.5: Depictions of different approaches to joint attention. © 1999 Seemann, Axel (ed.). Reprinted, with permission, from Ref. [500].

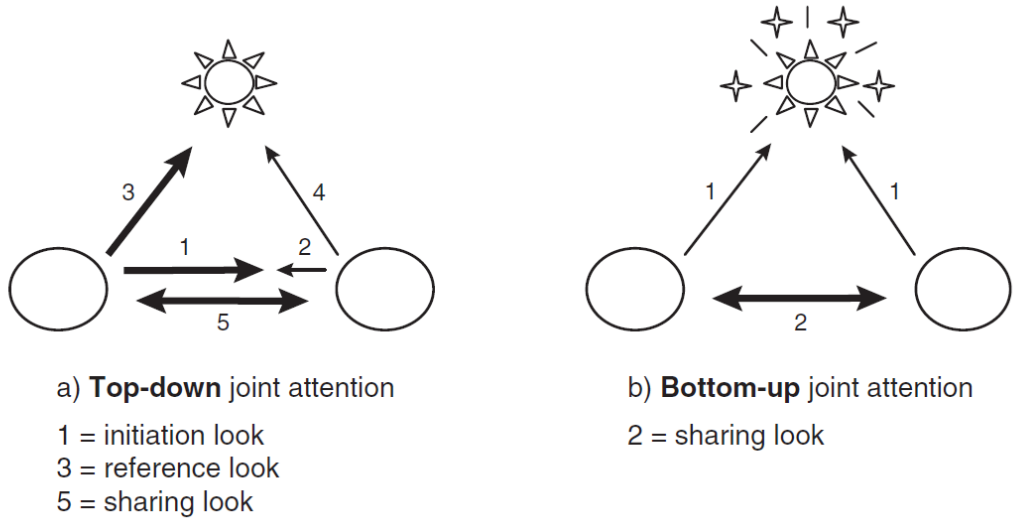


Figure 8.6: Two kinds of joint attention, "top-down" and "bottom-up." © 1999 Seemann, Axel (ed.). Reprinted, with permission, from Ref. [500].

relatively one-sided, the sharing look is bidirectional, with both partners participating equally. This look is what turns parallel or recursive or not-yet-shared attention into truly joint, shared attention. In the bottom-up joint attention situation, in contrast, the referent draws attention to itself because of its saliency. In this situation, the referent is given by the context so no referential look (or gesture) is needed. Typically, only one communicative look is needed in the bottom-up situation: the sharing look to the partner, although the sharing look might be slightly more complicated in this situation than in the top-down situation because some initial communicative intent must also be contained in this look, since the channel of communication is being opened at the same moment as the "triangle" is being closed with the sharing look.

Communication makes things public and thus creates commitments and obligations. Once two individuals get into joint attention, they can not deny it anymore.

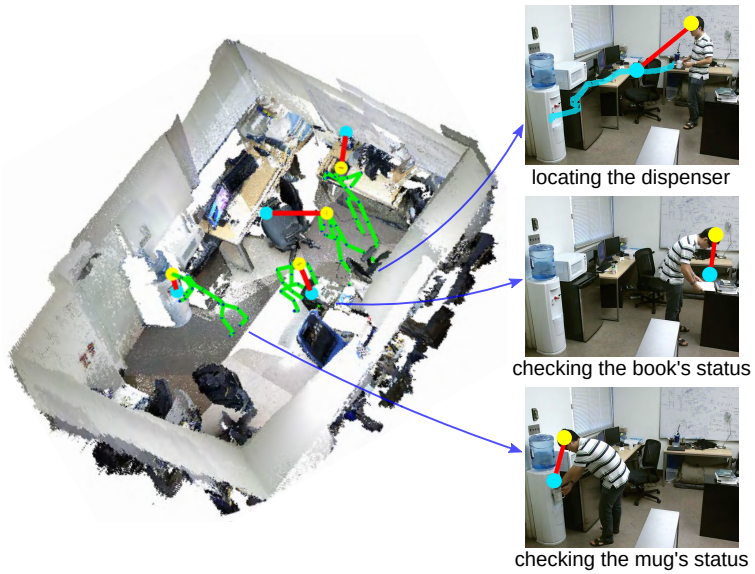


Figure 8.7: Human attention and intentions in a 3D scene. © 2017 Ping Wei *et al*. Reprinted, with permission, from Ref. [501]

8.4 Inferring Human Attention by Learning Latent Intentions

Inferring 3D human attention at scene scale is a challenging problem. First, in 3D space, human attention has weak observable features but huge degrees of freedom. As Fig. 8.7 shows, at the scale of daily-activity scenes, it is hard to obtain effective features of eyes or faces that are directly related to the human attention. Moreover, the human activity sequence data captured by RGB-D sensors are noisy. Different human activities present various poses, motions, and views, which makes it hard to precisely estimating the attention across different activities.

Human attention is related to human intentions [24]. The attention driven by different intentions presents different observation features and motion patterns. Land *et al*. [24] divided the roles of human fixations into four categories: *locating objects*, *directing hands*, *guiding an object to approach another*, and *checking an object's status*. As Fig. 8.7 shows, when the person's intention is to *locate the dispenser*, his attention sweeps from the table to the dispenser; while fetching water from the dispenser, his intention is to *check* if the mug is full and his attention steadily focuses on the mug.

The driving rules of intentions acting on attention can be independent of activity categories. For example, in Fig. 8.7, the attention driven by the intention *checking status* always presents as *steadily focusing*, even in different activities. This phenomenon makes it possible to infer the attention with the same rules across different activities. However, these driving rules are hidden and should be learned from data.

Ping Wei *et al*. proposes a probabilistic method to infer 3D human attention by jointly modeling attention, intentions, and their interactions. The attention and intention are represented with features extracted from human skeletons and scene voxels. Human intentions are taken as latent variables which guide the motions and forms of human attention. Conversely, the human attention reveals the intention features. Attention inference is modeled as a joint optimization with latent human intentions.

They adopt an EM-based approach to learn the model parameters and mine the latent intentions. Given an RGB-D video with human skeletons captured by the Kinect camera, a joint-state dynamic programming algorithm is utilized to jointly infer the latent intention, the 3D attention direction, and the attention voxel in each video frame.

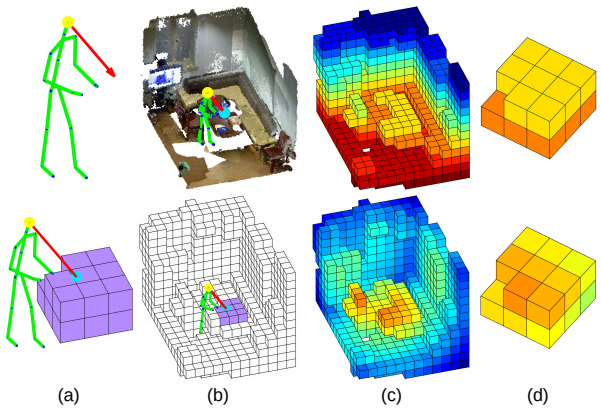


Figure 8.8: Attention and intention representation. (a) The attention direction and voxel. (b) Scene point clouds and voxels. (c) The voxel height and distance. (d) The voxel features. © 2017 Ping Wei *et al.* Reprinted, with permission, from Ref. [501].

8.4.1 Attention and Intention Representation

Each video frame includes RGB-D data and a 3D human skeleton, which are recorded by a Kinect camera. The 3D human skeleton is a collection of 3D location coordinates of the body joints, as shown in Fig. 8.8 (a). The scene point clouds defined by the scene depth data are converted into voxels, as shown in Fig. 8.8 (b). A voxel is a cube in 3D point clouds and it is like a pixel in 2D images. They define attention and intention features based on the 3D human skeletons and the scene voxels.

Attention

In 3D space, human attention includes two attributes: the direction and the voxel, as shown in Fig. 8.8 (a). The attention direction is a 3D vector with unit length which describes the sight line direction from the human head to what is looked at. In the attention direction, the voxel at which the sight line intersects with the scene point clouds is the attention voxel.

In daily activities, the directions of human body parts imply the attention directions. For example, when a human is manipulating an object with the hands, the directions from the head to the hands strongly signal the attention direction. They define the observation features of attention directions with eight directions extracted from human skeletons, such as the normal vector of the head and shoulder plane, the directions from the head to the hands, *etc.*

To normalize the data, all human skeletons are aligned to a reference skeleton with similarity transformation. The eight observation directions are defined on the aligned skeletons. The encapsulation of the eight normalized direction vectors is the observation feature of the attention.

Intention

In their work, intentions are discrete latent variables and describe the human attention motivation. The observation feature of an intention is the encapsulation of the attention feature and the voxel feature. The attention feature is defined in previous section. It characterizes the attention direction patterns in intentions.

The voxel feature is defined with the attention voxel and its neighbouring voxels, as shown in Fig. 8.8 (c) and Fig. 8.8 (d). The voxel feature is composed of the height part and the distance part. Around the attention voxel, they define a $Nx \times Ny \times Nz$ cubic grid of voxels, where Nx,

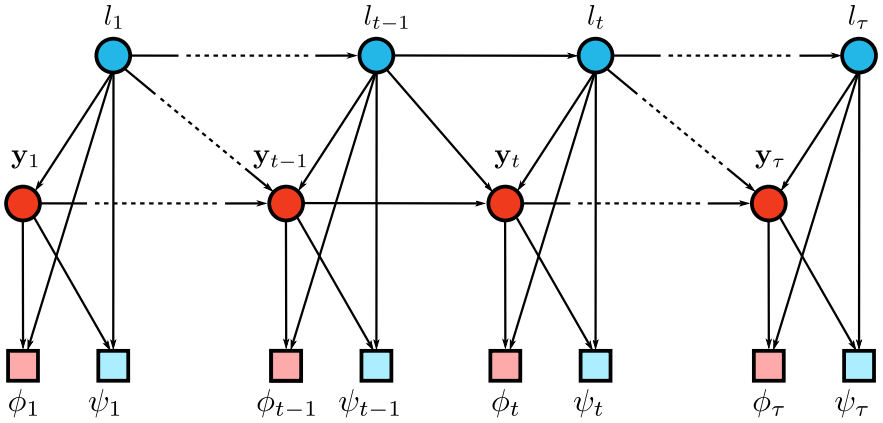


Figure 8.9: Joint probabilistic model of human attention and latent intentions. © 2017 Ping Wei *et al.* Reprinted, with permission, from Ref. [501].

Ny, and Nz are voxel numbers along the axis X, Y, and Z, respectively. The height feature is a $Nx \times Ny \times Nz$-dimensional vector whose entries correspond to the $Nx \times Ny \times Nz$ voxels in the cubic grid. The value of each entry is the height of the corresponding voxel relative to the floor. The distance feature is defined in a similar way but the vector entry value is the distance from the voxel to the human head.

The height feature reflects the spatial configuration of the attention voxels. The distance feature characterizes the human-scene interaction.

8.4.2 Model

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_\tau)$ be a video sequence of length τ. Each video frame $\mathbf{x}_t$ includes a 3D human skeleton and the scene voxels. Given $\mathbf{X}$, the goal is to infer the attention direction and the attention voxel in each video frame. Let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_\tau)$ be the attention direction sequence, where $\mathbf{y}_t$ denotes the attention direction in frame $\mathbf{x}_t$.

In each frame, they introduce a latent variable l_t to represent the latent intention. $\mathbf{l} = (l_1, ..., l_\tau)$ denotes the intention sequence of all the frames in $\mathbf{X}$.

They use a probabilistic model to jointly represent $\mathbf{X}$, $\mathbf{l}$, $\mathbf{Y}$, and their relations in time and 3D space, as shown in Fig. 8.9. The joint probability is

$$\begin{aligned} p(\mathbf{X}, \mathbf{l}, \mathbf{Y}|\boldsymbol{\theta}) = p(l_1) \prod_{t=1}^{\tau} p(\psi(\mathbf{x}_t)|l_t, \mathbf{y}_t) \prod_{t=2}^{\tau} p(l_t|l_{t-1}) \\ \cdot p(\mathbf{y}_1|l_1) \prod_{t=1}^{\tau} p(\phi(\mathbf{x}_t)|\mathbf{y}_t, l_t) \prod_{t=2}^{\tau} p(\mathbf{y}_t|\mathbf{y}_{t-1}, l_t, l_{t-1}). \end{aligned} \tag{8.1}$$

$\boldsymbol{\theta}$ is the set of model parameters. $\psi(\mathbf{x}_t)$ and $\phi(\mathbf{x}_t)$ are the intention and attention features, respectively, extracted from frame $\mathbf{x}_t$ as defined in Section 8.4.1. They are abbreviated as ψ_t and ϕ_t below. $p(\psi_t|l_t, \mathbf{y}_t)$ represents the intention identification and $p(\phi_t|\mathbf{y}_t, l_t)$ is the attention observation probability.

$p(l_t|l_{t-1})$ and $p(\mathbf{y}_t|\mathbf{y}_{t-1}, l_t, l_{t-1})$ describe transition relations of intentions and attention in two successive frames, respectively. $p(l_1)$ and $p(\mathbf{y}_1|l_1)$ characterize the initial states of the intention and the attention, respectively.

As Fig. 8.9 shows, the model is a joint representation of the intention and the attention. The intentions guide not only the attention observations but also the attention transitions. Conversely, the intention observation features depend on the voxels observed by the human.

The model is similar but different from the switching dynamic models [502, 503]. In the model, the latent variables of attention and intentions have different observation features.

Attention Model

They model human attention under the framework of the linear dynamic system (LDS) [504]. Different from the conventional LDS, they introduce an additional layer of latent variables to control the observation and motion patterns.

Initial attention $\mathbf{y}_1$ is modeled as:

$$\begin{aligned} &\mathbf{y}_1 = \boldsymbol{\mu}_{l_1} + u, \\ &u \sim \mathcal{N}(0, \mathbf{V}_{l_1}), \end{aligned} \tag{8.2}$$

where $\boldsymbol{\mu}_{l_1}$ is the prior value of $\mathbf{y}_1$ conditioned on intention l_1. u is the noise which follows Gaussian distribution with mean 0 and covariance $\mathbf{V}_{l_1}$. The initial attention probability is

$$p(\mathbf{y}_1|l_1) = \mathcal{N}(\mathbf{y}_1|\boldsymbol{\mu}_{l_1}, \mathbf{V}_{l_1}). \tag{8.3}$$

Attention observation describes the generation relation of the attention and the observation, which is formulated as:

$$\begin{aligned} &\phi_t = \mathbf{C}_{l_t}\mathbf{y}_t + \mathbf{v}, \\ &\mathbf{v} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{l_t}), \end{aligned} \tag{8.4}$$

where $\mathbf{v}$ is the noise which follows Gaussian distribution with mean 0 and covariance $\boldsymbol{\Sigma}_{l_t}$. The generation matrix $\mathbf{C}_{l_t}$ is governed by the intention l_t, which reflects the intention constraints on the attention observations. The attention observation probability is

$$p(\phi_t|\mathbf{y}_t, l_t) = \mathcal{N}(\phi_t|\mathbf{C}_{l_t}\mathbf{y}_t, \boldsymbol{\Sigma}_{l_t}). \tag{8.5}$$

Attention transition describes the temporal relations between attention in successive frames, which is formulated as

$$\begin{aligned} &\mathbf{y}_t = \mathbf{A}_{l_{t-1},l_t}\mathbf{y}_{t-1} + \mathbf{w}, \\ &\mathbf{w} \sim \mathcal{N}(0, \boldsymbol{\Gamma}_{l_{t-1},l_t}), \end{aligned} \tag{8.6}$$

where $\mathbf{w}$ is the noise which follows Gaussian distribution with mean 0 and covariance $\boldsymbol{\Gamma}_{l_{t-1},l_t}$. The transition matrix $\mathbf{A}_{l_{t-1},l_t}$ is related to the intentions in successive frames, which reflects the intention constraints on the attention motions. The transition probability model is

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, l_t, l_{t-1}) = \mathcal{N}(\mathbf{y}_t|\mathbf{A}_{l_{t-1},l_t}\mathbf{y}_{t-1}, \boldsymbol{\Gamma}_{l_{t-1},l_t}). \tag{8.7}$$

Intention Model

Intention model is composed of three parts: initial intention, intention identification, and intention transition.

Initial intention describes the prior knowledge about the intention in the first frame. It is formulated as:

$$p(l_1 = i) = \boldsymbol{\lambda}^i, \tag{8.8}$$

where $\boldsymbol{\lambda}$ is a discrete probability vector, and its ith entry $\boldsymbol{\lambda}^i$ represents the probability of the ith intention category.

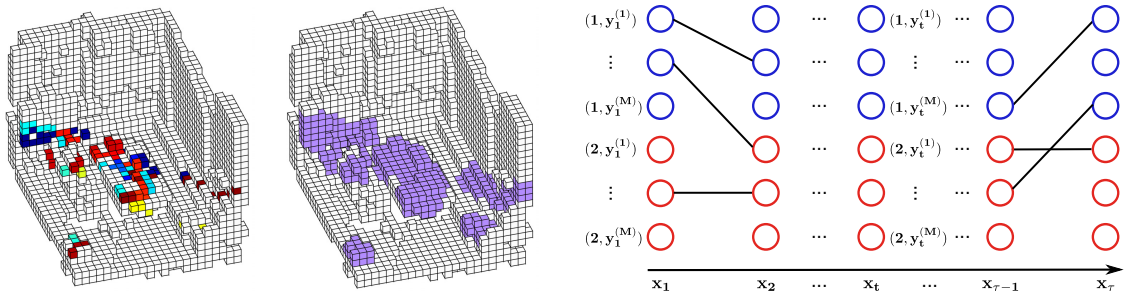


Figure 8.10: Joint-state dynamic programming. (a) Seed voxels in a video. The warmer colors indicate more recent time. (b) Candidate voxels. (c) Inference on a sequence, where two intention states are used to illustrate the algorithm. © 2017 Ping Wei *et al*. Reprinted, with permission, from Ref. [501].

Intention identification is formulated as

$$p(\psi_t|l_t, \mathbf{y}_t) \propto p(l_t|\psi_t, \mathbf{y}_t, \boldsymbol{\omega}). \tag{8.9}$$

$p(l_t|\psi_t, \mathbf{y}_t, \boldsymbol{\omega})$ is the posterior probability and $\boldsymbol{\omega}$ is the parameter of a linear classifier. The classifier is trained with Support Vector Machine and the scores output by the classifier are converted to probabilities [505].

The intention observation is dependent on the attention voxels related to the attention direction $\mathbf{y}_t$, which reflects the joint relations between the intentions and the attention.

Intention transition describes the relations of intentions in two successive frames, which is represented as

$$p(l_t = j|l_{t-1} = i) = \boldsymbol{\Lambda}^{ij}, \tag{8.10}$$

where $\boldsymbol{\Lambda}$ is the transition matrix. The entry $\boldsymbol{\Lambda}^{ij}$ in the ith row and jth column is the probability of the transition from the ith intention category to the j intention category.

8.4.3 Inference

Given an RGB-D video $\mathbf{X}$, they aim to infer the 3D human attention in each video frame, which is formulated as

$$\mathbf{Y}^* = \arg\max \ p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}), \tag{8.11}$$

where

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \sum\nolimits_{\mathbf{l}} p(\mathbf{l}, \mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}). \tag{8.12}$$

Dynamic programming is one of the most widely-used algorithms to interpret temporal sequences [506]. However, the attention and intentions are correlated, which means the conventional dynamic programming is inapplicable in the task.

They adopted a joint-state dynamic programming method to solve Eq. (8.11). The general procedures of the algorithm include: 1) in each video frame, a seed voxel is proposed, as shown in Fig. 8.10 (a); 2) the seed voxel generates candidate attention voxels in a cube around the seed, as shown in Fig. 8.10 (b); 3) the candidate voxels and all intentions are combined to form joint states; a joint state includes an attention voxel (direction) and an intention; 4) the dynamic programming [506] is performed on these joint states to produce the attention voxels (directions) and the latent intentions, as shown in Fig. 8.10 (c).

In each frame, they use attention features extracted from human skeletons to propose possible attention directions, which intersect with the scene to produce the seed voxels. Around the

seed voxel, they define a cube containing M neighbouring voxels as candidate attention voxels. Connecting the human head and these candidate voxels generates a set of candidate directions $\mathcal{Y}_t = \{\mathbf{y}_t^{(1)}, ..., \mathbf{y}_t^{(M)}\}$. In each frame, the joint state space is formed with $\mathcal{Y}_t$ and all possible intentions.

8.4.4 Learning

Let $\boldsymbol{\theta} = \{\boldsymbol{\mu}_i, \mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\Sigma}_i, \mathbf{A}_{ij}, \boldsymbol{\Gamma}_{ij}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}\}$ be all the parameters of the model. The subscripts i, j indicate parameters of different intentions. Given N videos and their attention sequences $\{(\mathbf{X}^1, \mathbf{Y}^1), ..., (\mathbf{X}^N, \mathbf{Y}^N)\}$, the goal is to learn $\boldsymbol{\theta}$ from the N samples by maximizing the likelihood function,

$$\boldsymbol{\theta}^* = \arg\max \sum_{n=1}^{N} \ln p(\mathbf{X}^n, \mathbf{Y}^n | \boldsymbol{\theta}), \tag{8.13}$$

where

$$p(\mathbf{X}^n, \mathbf{Y}^n | \boldsymbol{\theta}) = \sum_{\mathbf{l}^n} p(\mathbf{X}^n, \mathbf{l}^n, \mathbf{Y}^n | \boldsymbol{\theta}). \tag{8.14}$$

$\mathbf{l}^n$ is the latent intention sequence of the nth video sample.

Inspired by the general EM algorithm, they optimize Eq. (8.13) with the following steps.

1) Initialize $\mathbf{l}^n$ for each training sequence ($n = 1, ..., N$) and compute corresponding $\boldsymbol{\theta}^{\text{old}}$ with Eq. (8.16).
2) Compute the optimal latent intention sequence $\mathbf{l}^{n*}$ for each training sequence ($n = 1, ..., N$),

$$\mathbf{l}^{n*} = \arg\max p(\mathbf{l}^n | \mathbf{X}^n, \mathbf{Y}^n, \boldsymbol{\theta}^{\text{old}}). \tag{8.15}$$

3) Compute new parameter $\boldsymbol{\theta}^{\text{new}}$ by optimizing

$$\boldsymbol{\theta}^{\text{new}} = \arg\max \sum\nolimits_{n=1}^{N} \ln p(\mathbf{X}^n, \mathbf{l}^{n*}, \mathbf{Y}^n | \boldsymbol{\theta}) \tag{8.16}$$

4) If the convergence condition is met, stop and output the results; else set $\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{\text{new}}$ and return to step 2).

In step 1), they use k-means to cluster the intention features and produce the initial intention labels. In step 2), they compute the optimal latent intention sequence $\mathbf{l}^{n*}$ with the dynamic programming. In step 3), Eq. (8.16) is optimized by computing derivatives of the log likelihood function with respect to the parameters.

8.5 Jointly Inferring Human Attention and Intentions in Complex Tasks

Given an RGB-D video where a human performs a task, they want to answer three questions simultaneously: 1) where the human is looking-attention prediction; 2) why the human is looking there-intention prediction; and 3) what task the human is performing-task recognition. Wei *et al.* proposed a hierarchical model of human-attention-object (HAO) to represent tasks, intentions, and attention under a unified framework and to jointly infer human attention, intentions, and tasks from videos, as shown in Fig. 8.11.

A task is a complex goal-driven human activity and performing a task is a process of eye-hand coordination [508], as the task *mop floor* shown in Fig. 8.12. Human attention describes where a human is looking. It includes the attributes of 3D location, 3D direction, and 2D location, as

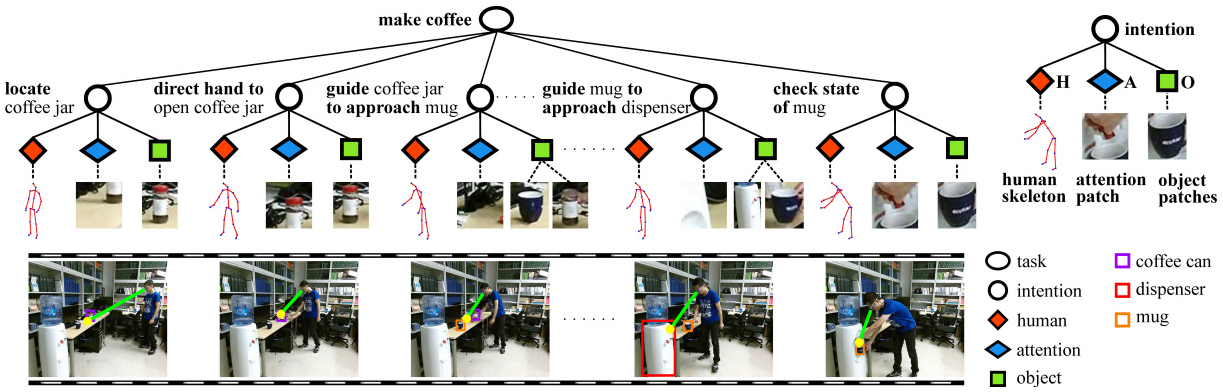


Figure 8.11: Human-attention-object (HAO) graph. The image patch under the attention node is the attention area where the human looks. © 2018 Ping Wei *et al*. Reprinted, with permission, from Ref. [507].

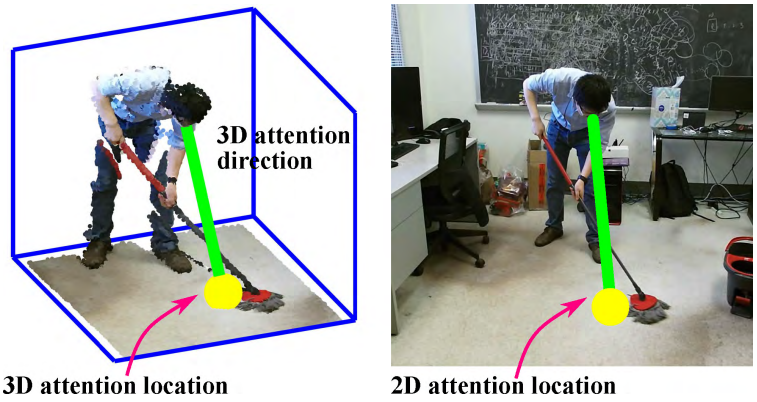


Figure 8.12: Human attention and intention in the task *mop floor*. While mopping the floor, the person is looking at the floor and his intention is checking if the floor has been cleaned or not. © 2018 Ping Wei *et al*. Reprinted, with permission, from Ref. [507].

shown in Fig. 8.12. A task is represented as sequential intentions which transition to each other. An intention is composed of the human pose, attention, and objects.

As the saying goes, "*eyes are the windows to the soul*." Human attention and intentions are closely related to each other in a task. By perceiving where a human is looking, they can infer the human's intentions. For example, in the task *make coffee* shown in Fig. 8.13, while fetching water from the dispenser, the person's attention focuses on the mug and his intention is to check the mug's state (full or not). On the other hand, human intentions drive human attention, which makes attention present different characteristics in different intentions. For example, in Fig. 8.13, when the person's intention is to check the mug's state, his attention focuses on the mug; when the person's intention is to locate the mug, his attention rapidly moves on the desk.

Eye or face features are often used to estimate human gazes. However, in large-scale daily-activity scenes, it is hard to obtain usable eye or face features due to low resolution. In this case, human body feature is an alternative to infer gazes.

8.5.1 Model

Fig. 8.11 illustrates the proposed HAO model. The graph contains four layers which correspond to the task, intentions, attention-bridged human body and objects, and the video, respectively.

A task is divided into several intentions in time domain. As shown in Fig. 8.11, the task *make*

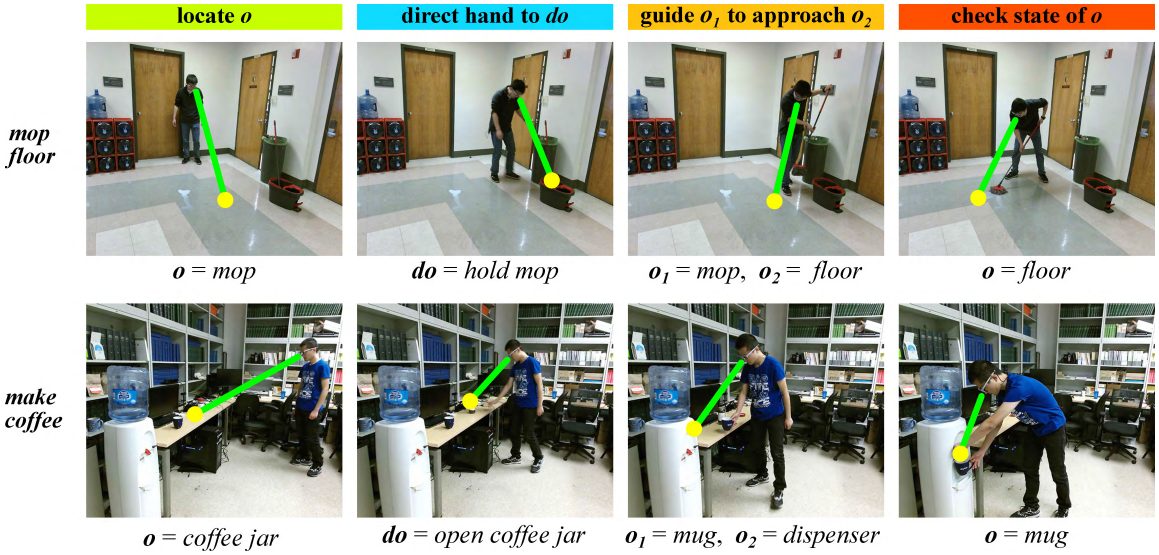


Figure 8.13: Four basic types of intentions when humans perform tasks. © 2018 Ping Wei *et al*. Reprinted, with permission, from Ref. [507].

coffee is composed of eight sequential intentions, such as *locate coffee jar*, *guide mug to approach dispenser*, *check state of mug*, *etc*. These intentions can transition to each other.

Intentions are revealed by cues of human bodies, human attention, and objects. Therefore, an intention is decomposed into the human pose, the human attention, and the intention-related objects, as shown in Fig. 8.11. The human attention bridges the human body and the objects.

Representation and Formulation

They use RGB-D videos recorded by motion capture technology like Kinect as inputs. Each frame includes an RGB image, a depth image, and a 3D human skeleton composed of 3D joint locations.

Let $\mathbf{I} = \{I_t | t = 1, ..., \tau\}$ be an input RGB-D video with length τ. I_t is the RGB-D frame at time t.

$\mathbf{H} = \{(\mathbf{h}_t, \mathbf{x}_t) | t = 1, ..., \tau\}$ is the human pose feature sequence. $\mathbf{h}_t$ and $\mathbf{x}_t$ are the appearance and geometric features extracted from the 3D skeleton at time t, respectively.

S is the task label of the input video. $\mathbf{L} = \{l_t | t = 1, ..., \tau\}$ is the human intention sequence of the video, where l_t is the intention label of the frame at time t.

$\mathbf{Y} = \{y_t | t = 1, ..., \tau\}$ is the human attention sequence. $\mathbf{y}_t$ is the *3D attention direction* in the t-th frame. It is defined as a unit 3D vector starting from the human head. The intersection point of the 3D attention direction and the scene point cloud is the *3D attention location*. With depth data, the 3D attention point is projected onto the image to form the *2D attention location*.

In the t-th RGB frame, they define a square image patch centered at the 2D attention point to extract the attention appearance feature $\mathbf{a}_t$. This image patch is like a central area where the human is looking, as shown in Fig. 8.11.

In the t-th frame, suppose $\mathbf{o}_t = (o_t^1, ..., o_t^m)$ is a bounding box collection of m intention-related objects, such as *mug* and *coffee jar* in the intention *guide coffee jar to approach mug*. These bounding boxes are proposed by the Faster R-CNN object detectors. With depth values of the RGB-D data, the 2D centers of object bounding boxes are projected onto the 3D space to form the objects' 3D locations $\mathbf{z}_t = (z_t^1, ..., z_t^m)$.

The energy that the input video is labeled with the task S, the intention $\mathbf{L}$, and the attention

$\mathbf{Y}$ is defined as

$$\begin{aligned}\mathcal{E}(\mathbf{Y},\mathbf{L},S|\mathbf{I},\mathbf{H}) &= \underbrace{\sum_{t=1}^{\tau}\Phi(\mathbf{h}_t,\mathbf{a}_t,\mathbf{o}_t,l_t)}_{\text{feature matching}} \\ &+ \underbrace{\sum_{t=1}^{\tau}\Psi(\mathbf{x}_t,\mathbf{y}_t,\mathbf{z}_t,l_t)}_{\text{HAO geometric relation}} + \underbrace{\sum_{t=2}^{\tau}\Gamma(\mathbf{y}_{t-1},\mathbf{y}_t,l_{t-1},l_t)}_{\text{attention and intention transition}} .\end{aligned} \tag{8.17}$$

$\Phi(\cdot)$ is the feature matching energy; $\Psi(\cdot)$ describes the relations among the human body, attention, and objects; $\Gamma(\cdot)$ represents the temporal transitions of attention and intention. Since the relation between a task and its intentions is a hard constraint, they omit S in the right side of Eq. (8.17).

Feature Matching of HAO

The feature matching term is written as

$$\Phi(\mathbf{h}_t,\mathbf{a}_t,\mathbf{o}_t,l_t) = \phi_1(\mathbf{h}_t,l_t) + \phi_2(\mathbf{a}_t,l_t) + \phi_3(\mathbf{o}_t,l_t) \tag{8.18}$$

Human pose matching $\phi_1(\mathbf{h}_t,l_t)$ describes the compatibility of the pose feature $\mathbf{h}_t$ and the intention l_t. With the 3D skeleton, they compute the differences between each joint and other joints, and concatenate the difference vector of each joint to form $\mathbf{h}_t$. Using pose features of all intention classes, they train a classifier with logistic regression for pose classification. The probability output by the classifier is used as $p(l_t|\mathbf{h}_t)$. The energy is

$$\phi_1(\mathbf{h}_t,l_t) = -\log p(l_t|\mathbf{h}_t). \tag{8.19}$$

Attention feature matching $\phi_2(\mathbf{a}_t,l_t)$ describes the compatibility between the attention feature $\mathbf{a}_t$ and the intention l_t. They train a CNN classifier with the VGG16 model on the square attention patch samples. The score output from the network is used as the attention patch labeling probability $p(l_t|\mathbf{a}_t)$. Fig. 8.14 shows two examples of the probability maps. The attention matching energy is

$$\phi_2(\mathbf{a}_t,l_t) = -\log p(l_t|\mathbf{a}_t). \tag{8.20}$$

Object matching represents the compatibility between the object features in the video frame and the object classes related to the intention. $(o_t^1,...,o_t^m)$ is the object bounding boxes related to the intention l_t. They fine-tune Faster R-CNN models on the training data to detect objects in each frame. The score output from the Faster R-CNN detector is used as an object's probability $p(o_t^i)$. The energy of all related objects in the frame is

$$\phi_3(\mathbf{o}_t,l_t) = -\frac{1}{m}\sum\nolimits_{i=1}^{m}\log p(o_t^i). \tag{8.21}$$

Geometric Relations of HAO

The human attention bridges the human body and the objects. The geometric relation term $\Psi(\mathbf{x}_t,\mathbf{y}_t,\mathbf{z}_t,l_t)$ describes the location and direction constraint of the human pose, attention, and objects. It is written as

$$\Psi(\mathbf{x}_t,\mathbf{y}_t,\mathbf{z}_t,l_t) = \psi_1(\mathbf{x}_t,\mathbf{y}_t,l_t) + \psi_2(\mathbf{z}_t,\mathbf{y}_t,l_t) \tag{8.22}$$

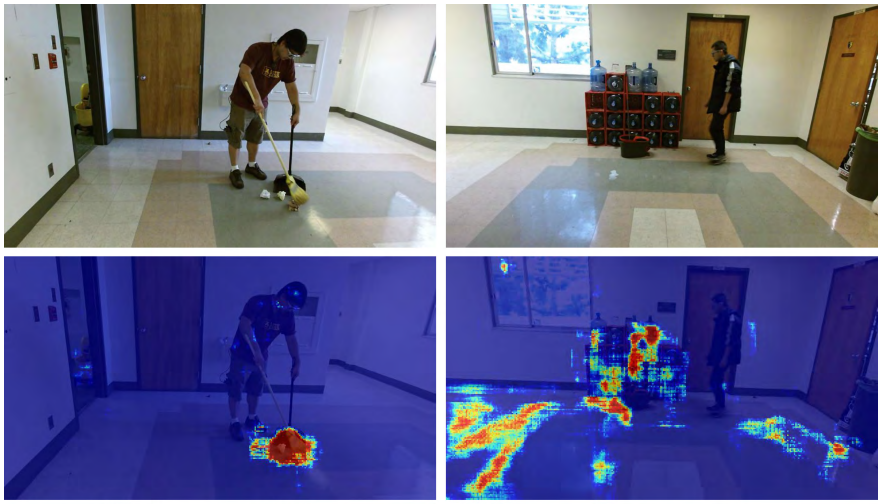


Figure 8.14: Attention map. Each map pixel value is the probability that the human looks at the pixel with the intention shown below. © 2018 Ping Wei *et al*. Reprinted, with permission, from Ref. [507].

Human pose and attention relation $\psi_1(\mathbf{x}_t, \mathbf{y}_t, l_t)$ describes the constraint between the 3D attention direction and the human pose. In daily-activity scenes, the body part directions imply the attention directions. For example, when a human manipulates objects with hands, the direction from the head to the hands implies the attention direction.

They adopt a similar method to the work [501] to model the pose and attention relations. Eleven 3D vectors are extracted from the 3D human skeleton, such as the normal vector of the head and shoulder plane, the direction from the head to the hands, *etc*. These 3D vectors are concatenated as the attention direction feature $\mathbf{x}_t$.

They train a regression model from the attention direction feature to the 3D attention direction with a 3-layer fully-connected neural network f. For an attention feature $\mathbf{x}_t$, the network f estimates a hypothesized 3D attention direction $f(\mathbf{x}_t)$.

The relation between the human attention direction $\mathbf{y}_t$ and $f(\mathbf{x}_t)$ is defined as

$$\begin{aligned} \mathbf{y}_t &= f(\mathbf{x}_t) + \mathbf{w}_{l_t}, \\ \mathbf{w}_{l_t} &\sim \mathcal{N}(\boldsymbol{\mu}_{l_t}, \boldsymbol{\Sigma}_{l_t}), \end{aligned} \tag{8.23}$$

where $\mathbf{w}_{l_t}$ is a noise variable following Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{l_t}, \boldsymbol{\Sigma}_{l_t})$. The geometric energy is written as

$$\psi_1(\mathbf{x}_t, \mathbf{y}_t, l_t) = -\log \mathcal{N}(\mathbf{y}_t | f(\mathbf{x}_t) + \boldsymbol{\mu}_{l_t}, \boldsymbol{\Sigma}_{l_t}). \tag{8.24}$$

The intention l_t in $\boldsymbol{\mu}_{l_t}$ and $\boldsymbol{\Sigma}_{l_t}$ suggests different geometric relations in different intentions, which reflects the constraints of intentions on attention.

Attention and object relation $\psi_2(\mathbf{z}_t, \mathbf{y}_t, l_t)$ describes the constraint between the human attention location and the object locations in 3D space. The attention location is closely related to the object location, but not necessarily the same. For example, in the intention *locate mug*, the attention location shifts from the nearby areas to the mug.

Suppose $\tilde{\mathbf{y}}_t$ is the 3D attention location. It is the intersection point of the 3D attention direction $\mathbf{y}_t$ and the scene point cloud. The relation between the attention location $\tilde{\mathbf{y}}_t$ and the object bounding

box o_t^i is formulated as

$$\begin{aligned} \mathbf{z}_t^i &= \tilde{\mathbf{y}}_t + \mathbf{v}_{l_t,\tilde{o}_t^i}, \\ \mathbf{v}_{l_t,\tilde{o}_t^i} &\sim \mathcal{N}(\lambda_{l_t,\tilde{o}_t^i}, \mathbf{\Lambda}_{l_t,\tilde{o}_t^i}), \end{aligned} \tag{8.25}$$

where $\tilde{o}_t^i$ is the object class label of the box o_t^i. $\mathbf{z}_t^i$ is the object's 3D location. $\mathbf{v}_{l_t,\tilde{o}_t^i}$ is a noise variable following Gaussian distribution $\mathcal{N}(\lambda_{l_t,\tilde{o}_t^i}, \mathbf{\Lambda}_{l_t,\tilde{o}_t^i})$. The subscripts $l_t, \tilde{o}_t^i$ in $\lambda_{l_t,\tilde{o}_t^i}$ and $\mathbf{\Lambda}_{l_t,\tilde{o}_t^i}$ suggests that the attention-object relations are different for different intentions and object classes.

The relation energy of multiple objects in the frame is

$$\psi_2(\mathbf{z}_t, \mathbf{y}_t, l_t) = -\frac{1}{m}\sum_{i=1}^{m} \log \mathcal{N}(\mathbf{z}_t | \tilde{\mathbf{y}}_t + \lambda_{l_t,\tilde{o}_t^i}, \mathbf{\Lambda}_{l_t,\tilde{o}_t^i}). \tag{8.26}$$

Temporal Transition of Attention and Intention

$\Gamma(\mathbf{y}_{t-1}, \mathbf{y}_t, l_{t-1}, l_t)$ represents the transitions of attention and intention in time domain. It is written as

$$\Gamma(\mathbf{y}_{t-1}, \mathbf{y}_t, l_{t-1}, l_t) = \gamma_1(\mathbf{y}_{t-1}, \mathbf{y}_t) + \gamma_2(l_{t-1}, l_t). \tag{8.27}$$

Attention transition $\gamma_1(\mathbf{y}_{t-1}, \mathbf{y}_t)$ describes the temporal relations between attention directions in two successive frames. It is formulated as a linear dynamic system [509, 501]:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Q}_{l_{t-1},l_t}\mathbf{y}_{t-1} + \mathbf{u}_{l_{t-1},l_t}, \\ \mathbf{u}_{l_{t-1},l_t} &\sim \mathcal{N}(0, \mathbf{\Upsilon}_{l_{t-1},l_t}), \end{aligned} \tag{8.28}$$

where $\mathbf{Q}_{l_{t-1},l_t}$ is the transition matrix. $\mathbf{u}_{l_{t-1},l_t}$ is a noise variable following Gaussian distribution $\mathcal{N}(0, \mathbf{\Upsilon}_{l_{t-1},l_t})$. The attention transition energy is

$$\gamma_1(\mathbf{y}_{t-1}, \mathbf{y}_t) = -\log \mathcal{N}(\mathbf{y}_t | \mathbf{Q}_{l_{t-1},l_t}\mathbf{y}_{t-1}, \mathbf{\Upsilon}_{l_{t-1},l_t}). \tag{8.29}$$

$\mathbf{Q}_{l_{t-1},l_t}$ and $\mathbf{\Upsilon}_{l_{t-1},l_t}$ are both related to the intentions l_{t-1} and l_t, which reflects the fact that the motion patterns of human attention are constrained by human intentions.

Intention transition $\gamma_2(l_{t-1}, l_t)$ represents the transition relations between different intentions. They model the transition as a Markov process. $p(l_t = j | l_{t-1} = i) = d_{ij}$ is the transition probability between two intentions in successive frames. The transition energy is defined as

$$\gamma_2(l_{t-1} = i, l_t = j) = -\log p(l_t = j | l_{t-1} = i). \tag{8.30}$$

8.5.2 Inference

Given an input RGB-D video $\mathbf{I}$ with 3D human skeletons $\mathbf{H}$, they aim to jointly output: 1) the human intention in each frame; 2) the 3D attention direction in each frame; and 3) the task label of the video. This problem is formulated as

$$(\mathbf{Y}, \mathbf{L}, S)^* = \arg\min \ \mathcal{E}(\mathbf{Y}, \mathbf{L}, S | \mathbf{I}, \mathbf{H}). \tag{8.31}$$

They use an algorithm similar to beam search [510] to solve Eq. (8.31), as shown in Fig. 8.15. It includes three procedures.

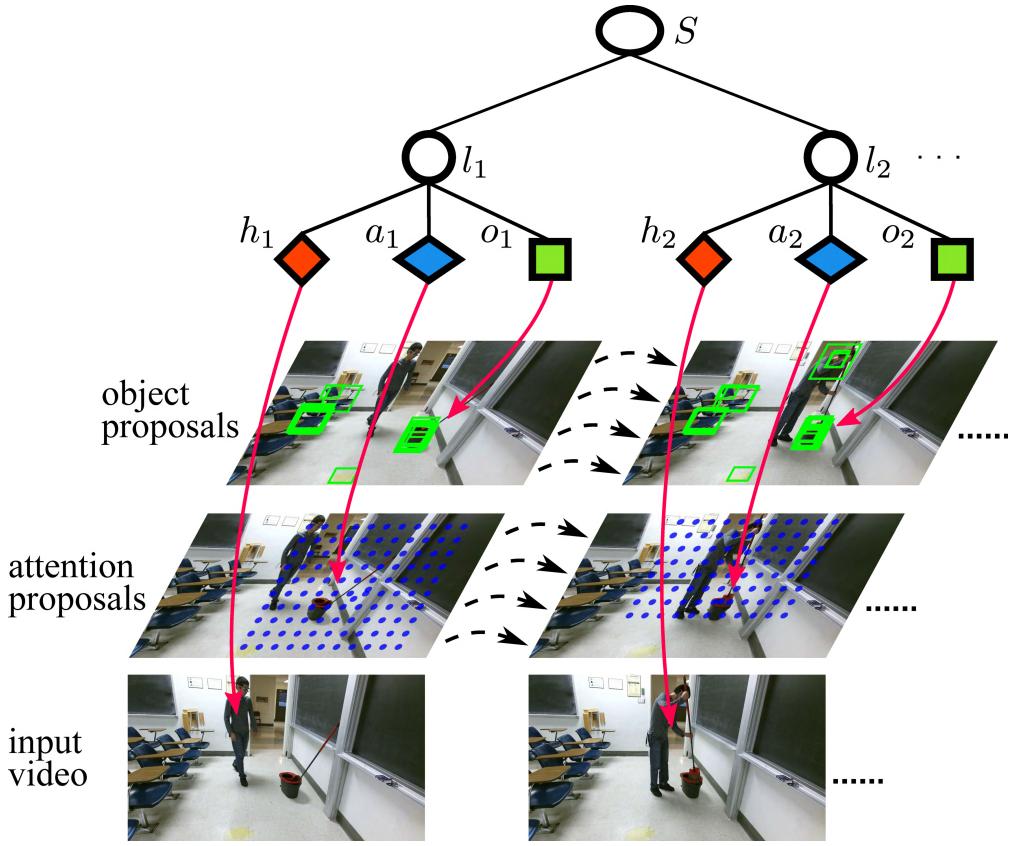


Figure 8.15: Inference algorithm. For clarity, only parts of the proposed object boxes and attention points are visualized. © 2018 Ping Wei *et al*. Reprinted, with permission, from Ref. [507].

1) Proposing hypothesized attention points. The possible attention points on RGB images are proposed according to human poses. As introduced in Section 8.5.1, with the pose feature $\mathbf{x}_t$, a hypothesized 3D attention direction $f(\mathbf{x}_t)$ is computed with the network f. A 3D attention point derived from $f(\mathbf{x}_t)$ is projected onto the image plane to form a 2D location. Around this location, they propose a group of possible 2D attention points, as shown in Fig. 8.15. The point range and step are empirically defined. Each 2D point is attached a probability vector of all possible intentions computing with the attention matching model in Eq. (8.20).

2) Proposing hypothesized objects. They use Faster R-CNN to detect all possible objects related to all the tasks and intentions in each frame, as shown in Fig. 8.15. Each detected box has the probabilities of all object classes.

3) Graph-guided optimization. With the hypothesized attention points and objects, the goal is to select optimal attention points, objects, intentions, and the task label in each video frame to minimize $\mathcal{E}(\mathbf{Y}, \mathbf{L}, S|\mathbf{I}, \mathbf{H})$.

From training samples, they construct HAO graphs for each task category. These graphs specify the intentions, related objects, the geometric and temporal relations. Let $\mathbf{I}_t$ be the video clip from time 1 to t. The graph-guided optimization is summarized as follows:

i) In frame I_t, all possible combinations of attention points, object bounding boxes, and intention labels for each task category are generated according to the HAO graph structure. Each of such combination is taken as one hypothesized joint label of frame I_t.

ii) The union of one joint label of I_t and one joint label sequence of the past video $\mathbf{I}_{t-1}$ forms a hypothesized joint label sequence of the video $\mathbf{I}_t$. The energy of the hypothesized joint label

Figure 8.16: **Shared attention is everywhere in our daily life.** Shared attention is a crucial first step towards social interaction, the primary basis of social intelligence and a precursor of Theory of Mind [512] © 2018 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [513]. .

sequence is computed with Eq. (8.17). At time t, all hypothesized joint label sequences are sorted according to their energies. The J joint label sequences with lowest energies are kept and others are pruned.

iii) The step i) and step ii) are iterated frame by frame until the video ends. The joint label sequence with the lowest energy is the output result, which includes the task label, human attention and intentions for each frame.

8.6 Shared Attention

Shared attention is defined as the attention focus shared by two or more individuals on one object or human [487]. Shared attention differs from joint attention in a subtle way and in the literature the two terms are used interchangeably [487]. Shared attention is everywhere in daily life and they can observe it every now and then in almost all social interactions. Imagine in a party, usually humans can easily recognize a group of people with shared attention and what exactly is the shared attention in the group at present. They can join the group and form shared attention with them naturally and instantly. However, patients with autism may feel it difficult to interact with people around them since they lack the ability to build shared attention with others [511]. Fig. 8.16 shows some examples of shared attention in social scenes and how shared attention shifts temporally as well as who are currently involved in the shared attention.

In order to be clarified with the concept of shared attention, they formulate the problem as follows: shared attention is the gaze focus shared by two or more individuals on one object or human; given a video clip, the task is to detect which frames contain shared attention and where is the shared attention in those frames. To tackle this problem,they collect a new dataset VideoCoAtt and build a deep spatial-temporal neural network with four modules: gaze estimation module, region proposal module, spatial detection module and temporal optimization module. The intuitions for building such a deep neural network architecture are as follows:

1) Firstly, gaze direction, which can be utilized to learn external environment state and inter-

nal mental state, is a key feature for shared attention detection. The strongest and most direct indication of human gaze direction is the closeup image patch of human head. They need to detect human heads in videos and predict gaze directions for each detected head.

2) Secondly, gaze direction is of course important, but still not the whole story. Shared attention is more than gaze intersection. According to their definition, there must be an object or human body part as the carrier of shared attention, which means the shared attention detection task is object-driven. Thus, bounding box proposals of object or human body parts, such as laptop, human face, *etc.* is another key feature for the task. They didn't use saliency models (like [514, 515]) because shared attention is more influenced by social group interaction instead of visual importance, and people engaged in shared attention are not free-viewing and may not look at the most salient object in the environment. They use a generic object proposal generation method to generate all potential bounding boxes independent of the categories.

3) Shared attention may last for a while before termination. Temporal information is a good constraint to make the detection results more accurate and robust. The input to the model is just a video clip without any other additional annotation, and the output is a shared attention heatmap for each video frame and the final shared attention prediction results can also be inferred based on the shared attention heatmap.

8.6.1 Model Architecture

Shared attention usually locates at the objects or human body parts gazed by two or more people simultaneously. Obviously, human gaze and target objects in the context environment are essential for inferring shared attention in social scene videos. Thus the shared attention detection model comprises of four modules:

1) the gaze estimation module that extracts individual gaze directions to generate a gaze heatmap for the whole scene;

2) the region proposal module that extracts region proposals from the context environment;

3) the spatial detection module that combines the gaze heatmap and the region proposal map to detect shared attention in spatial space; and

4) the temporal optimization module that utilizes inter-frame correlation to optimize the predicted shared attention heatmap in temporal space. An illustration of the whole model architecture is presented in Fig. 8.17.

Gaze and Region Proposal Modules

Gaze Estimation Module. Suppose for an input frame I_t in a video sequence $\{I_t\}_{t=1,\dots,T}$, the head detector outputs a set of head locations $q_{t,i} = (x_{t,i}^{min}, y_{t,i}^{min}, x_{t,i}^{max}, y_{t,i}^{max})$, $i = 1, 2, \dots, n$, where n could be zero when no head is detected in frame I_t (see the red rectangles in Fig. 8.18 (a) and (c)). The corresponding closeup image patch for head location $q_{t,i}$ is cropped out from I_t and denoted as $w_{t,i}^h, i = 1, 2, \dots, n$. They then use a batch of neural network layers $\Psi(\cdot)$ to regress a gaze direction $d_{t,i} \in [-1, 1]^2$ (yellow arrows in Fig. 8.18 (a) and (c)) for the input image patch $w_{t,i}^h$:

$$d_{t,i} \triangleq (d_{t,i}^x, d_{t,i}^y) = \Psi(w_{t,i}^h). \tag{8.32}$$

They use a Gaussian distribution to model the variation of a gaze ray with respect to the predicted primary gaze direction $d_{t,i}$, and the probability distribution is

$$P(\theta_{t,i}|d_{t,i}) \propto \frac{1}{\sigma} \exp\{-\frac{\theta_{t,i}^2}{2\sigma^2}\}, \tag{8.33}$$

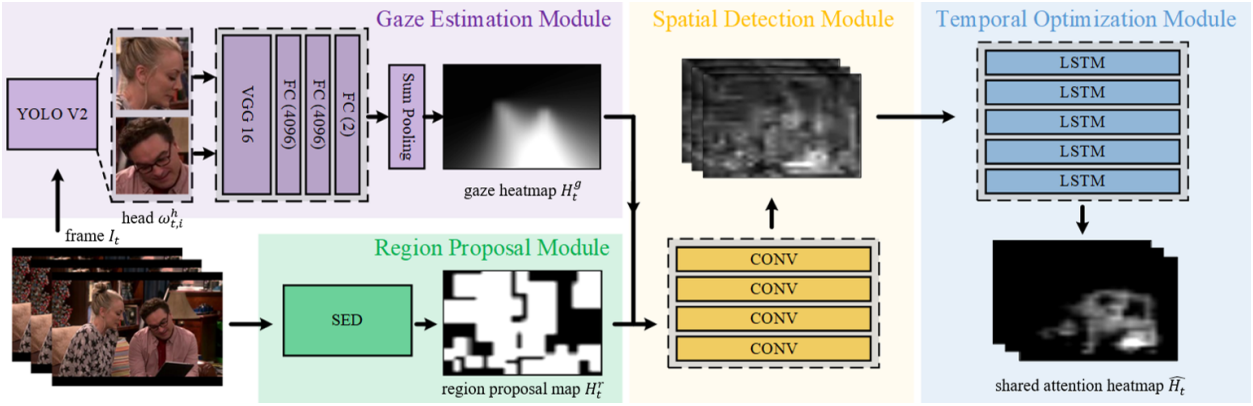


Figure 8.17: **Illustration of the model architecture.** The gaze estimation module and the region proposal module extract two key features of individuals and the scene context from raw input videos. The subsequent spatial detection module integrates the outputs from the two base modules to perform shared attention detection on a single frame. The temporal optimization module utilizes temporal constraints to optimize the predicted shared attention heatmap. © 2018 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [513].

(a) (b) (c) (d) (e)

Figure 8.18: **Illustration of gaze heatmap H_t^g generation procedure.** With detected head position $q_{t,i}$ (red rectangles in (a)(c)) and corresponding predicted gaze direction $d_{t,i}$ (yellow arrows in (a)(c)), they first generate individual gaze heatmap $H_{t,i}^g$ in (b) and (d), and then get the final gaze heatmap H_t^g in (e) via sum-pooling all the gaze heatmaps in (d). © 2018 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [513].

where $\theta_{t,i}$ is the angle between a gaze ray and the predicted primary gaze direction $d_{t,i}$. With detected head position $q_{t,i}$ and corresponding predicted gaze direction $d_{t,i}$, they compute $\theta_{t,i}$ for each grid in the image and then use Eq. (8.33) to get the probability for this grid to be gazed at by head $q_{t,i}$. After a gaze heatmap $H_{t,i}^g$ (see Fig. 8.18 (b) and (d)) for each head position $q_{t,i}$ is prepared, they generate the final gaze heatmap H_t^g (Fig. 8.18 (e)) of size $M \times N$ via Sum-Pooling $\{H_{t,i}^g\}_i$:

$$H_t^g = \sum_{i=1}^n H_{t,i}^g = \sum_{i=1}^n \phi(\Psi(w_{t,i}^h), q_{t,i}), \tag{8.34}$$

where $\phi(\cdot)$ indicates the gaze heatmap generator based on Eq. (8.33). More illustrations about the gaze heatmap generation procedure are shown in Fig. 8.18.

Region Proposal Module. To exploit context information, they use a region proposal module $Z(\cdot)$ to generate a binary region proposal map H_t^r of size $M \times N$ for input image I_t:

$$H_t^r = Z(I_t) \tag{8.35}$$

This module is implemented by Structured Edge Detector (SED) [516] to get region bounding boxes $\{b_{t,i}, i = 1, 2, \ldots, m\}$ for each frame I_t and then setting all the pixel values within the bbx proposals to 1 and all other pixel values outside to 0.

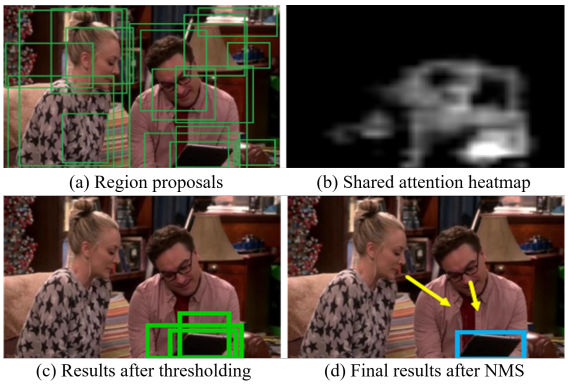


Figure 8.19: **Illustration of inference process.** Given (a) proposal bounding boxes and (b) shared attention heatmap, they first compute the score for each bounding box by accumulating all the confidence values inside the bounding box. (c) Then they select the bounding boxes with score higher than a certain threshold. (d) NMS is applied for generating final shared attention prediction. © 2018 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [513].

Spatio-temporal Shared Attention Network

The output feature maps of the gaze estimation module and the region proposal module are then fed to the subsequent spatial detection module and temporal optimization module for shared attention detection.

Spatial Detection Module. Shared attention detection is firstly conducted in a frame-by-frame style. They apply a spatial detection module $F(\cdot)$ that consists of several convolutional layers to combine the gaze heatmap H_t^g and region proposal map H_t^r for intra-frame shared attention detection:

$$\tilde{H}_t = F(H_t^g, H_t^r), \tag{8.36}$$

where $\tilde{H}_t$ indicates the intermediate shared attention heatmap output from the spatial detection module.

Temporal Optimization Module. To further exploit the temporal inter-frame constraints in videos, they add a temporal optimization module $LSTM(\cdot)$ that consists of several convolutional Long Short-Term Memory (convLSTM) network [517] layers to optimize the output shared attention heatmap $\tilde{H}_t$:

$$\{\hat{H}_t\}_t = LSTM(\{\tilde{H}_t\}_t), \tag{8.37}$$

where $\hat{H}_t$ denotes the eventual shared attention heatmap.

Learning and Inference

For the loss function, they apply the Mean Squared Error (MSE) between the predicted shared attention heatmap $\hat{H}_t$ and the ground truth shared attention binary map H_t:

$$L(\hat{H}_t, H_t) = \frac{1}{M \cdot N} \parallel \hat{H}_t - H_t \parallel^2, \tag{8.38}$$

where both $\hat{H}_t$ and H_t are of size $M \times N$.

The inference is possible given the predicted shared attention heatmap $\hat{H}_t$, based on which they can compute the cumulative score for each region proposal bounding box $b_{t,i}$. They only keep those proposal bounding boxes with a score higher than a threshold. Then they conduct a

Non-Maximum Suppression (NMS) [518] and treat the remaining bounding boxes as final shared attention prediction for frame I_t. See Fig. 8.19 for more detailed illustration.

Since there may be no shared attention or more than one shared attention in a scene, the model is designed to support multimodal predictions instead of regressing a single shared attention location.

8.6.2 Result Visualization and Analysis

Fig. 8.20 exhibits an internal visualization of shared attention detection results by the full model on some example frames. The *Gaze Heatmap* roughly features the attention of each individual in the social scene and is not enough to accurately feature shared attention. The *Region Proposal Map* gives some potential shared attention proposals and provides the important spatial constraints. *Single-frame Detection* combines the *Gaze Heatmap* and the *Region Proposal Map* to generate a preliminary shared attention heatmap, which still has too much noises. After the *Temporal Optimization* by convLSTM, the shared attention heatmap is much clearer and can provide more accurate shared attention distribution information. The final column in Fig. 8.20 compares the eventual shared attention prediction results (depicted in red rectangles) with the ground truth shared attention annotations (depicted in green rectangles). As shown, there are good predictions that can exactly locate the shared attention in the social scenes, like the prediction in the first example. However, there are also some false alarms existing. For example, The scene in the last row actually has only one shared attention, but the model gives two predictions located near the two human faces. This is an interesting failure example since whether the third person on the right side is looking at the person on the left side or the person in the middle is somehow ambiguous for the model to distinguish. That's why the shared attention heatmap gets two peaks for this example. But similar situation in the fifth scene is successfully solved by the model. Although they get some reasonable results in the experiments, they are still far from completely solving this problem.

8.7 Understanding Human Gaze Communication

Eye gaze is closely tied to what people are thinking and doing [519]. Gaze communication, as a major form of non-verbal communication, allows people to communicate with one another at the most basic level regardless of their familiarity with the prevailing verbal language system. Such social eye gaze functions thus transcend cultural differences, forming a kind of universal language [520]. During conversations, eye gaze can be used to convey information, regulate social intimacy, manage turn-taking, and convey social or emotional states. People also utilize eye gaze as approaches to determine objects around them, *i.e.*, human look at objects before naming or manipulating them [521]. People are also good at identifying the target of their partner's referential gaze and use this information to predict what their partner is going to say [522]. In a nutshell, gaze communication is omnipresent and multifunctional [520].

Fan *et al*. studied human gaze communication understanding by spatio-temporal graph reasoning [523]. With previous efforts and established terminologies, they distinguish daily social gaze communications of the atomic-level into six classes: *Single*, *Mutual*, *Avert*, *Refer*, *Follow*, *Share*, as shown in the left part of Fig. 8.21. The above atomic-level gazes capture the most general, core and fine-grained gaze communication patterns in human social interactions. They further study the long-term, coarse-grained gaze communications events with temporal compositions of the above six atomic-level gaze communication patterns and generalize into totally five gaze communication events, *i.e.*, *Non-communicative*, *Mutual Gaze*, *Gaze Aversion*, *Gaze Following* and *Joint Attention*, as illustrated in the right part of Fig. 8.21. Recognizing and understanding atomic-level gaze

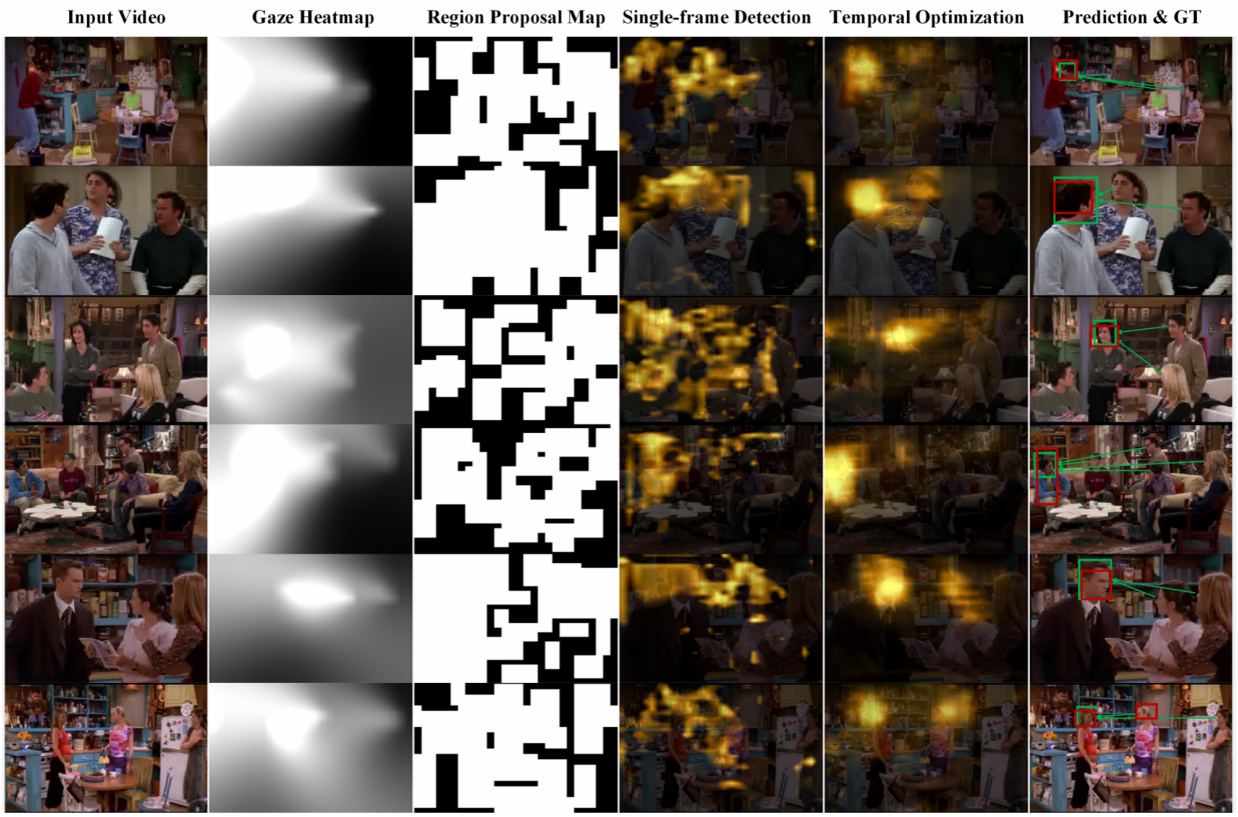


Figure 8.20: **Shared attention detection results on example frames.** With the input video frames, they show the outputs of the gaze estimation module and the region proposal module in the second and third columns. The *Single-frame Detection* column shows the shared attention heatmap $\tilde{H}_t$ trained on a single frame. The *Temporal Optimization* column shows the eventually optimized shared attention heatmap $\hat{H}_t$. The final prediction results (red rectangles) and the ground truth annotations (green rectangles) are presented in the last column. © 2018 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [513].

communication patterns are necessary and significant first-step for comprehensively understanding human gaze behaviors. To facilitate the research of gaze communication understanding in computer vision community, Fan *et al*. propose a large-scale social video dataset named *VACATION* (Video gAze CommunicATION). It contains 300 videos with 77,891 frames and complete annotations of human face and object bounding boxes, human attention, gaze communication type in atomic- and event-level. The video sequences in the dataset are elaborately collected to cover rich realistic social scenes, different cultures, and diverse appearances of actors/actresses, providing a solid foundation for human gaze behavior study. See Fig. 8.22.

8.7.1 Model Architecture

With the well established dataset, the task is defined as follows. Given a third-person social video sequence with the bounding boxes of human faces and objects, they aim to infer the social gazes for all the persons, build spatio-temporal attention graph and predict gaze communication relations for this video sequence in both atomic-level and event-level.

With this structured task that requires a comprehensive modeling of human-human and human-scene interactions in both spatial and temporal space, Fan *et al*. propose a novel spatio-temporal reasoning graph network for atomic-level gaze communication detection as well as an event network

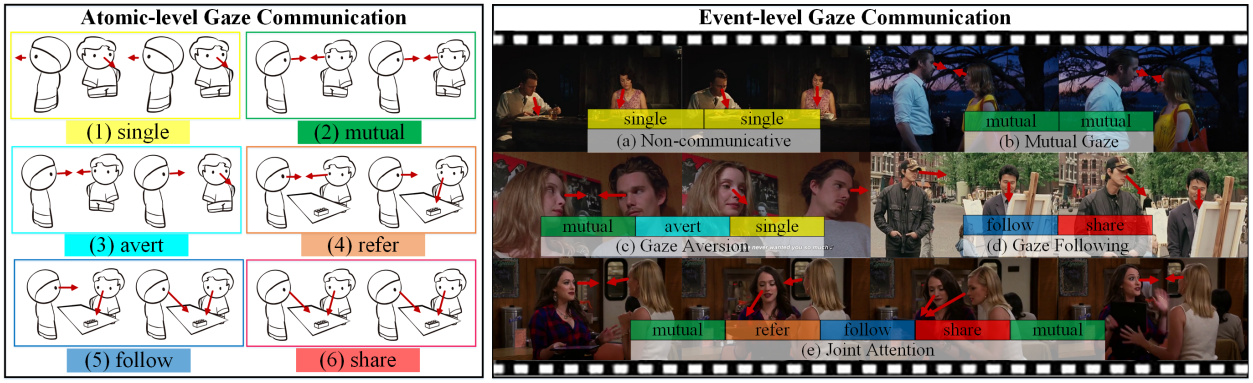


Figure 8.21: Fan *et al*. [523] study human gaze communication dynamics in two hierarchical levels: atomic-level and event-level. Atomic-level gaze communication describes the fine-grained structures in human gaze interactions, *i.e*., single, mutual, avert, refer, follow and share (as shown in left part). Event-level gaze communication refers to high-level, complex social communication events, including Non-communicative, Mutual Gaze, Gaze Aversion, Gaze Following and Joint Attention. Each gaze communication event is a temporal composition of some atomic-level gaze communications (as shown in right part). © 2019 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [523].

Figure 8.22: Example annotations of the *VACATION* dataset, showing that the dataset covers rich gaze communication behaviors, diverse general social scenes, different cultures, *etc*. It also provides rich annotations, *i.e*., human faces, gaze communication structures and labels. Human faces and related objects are marked by boxes with the same color of corresponding communication labels. White lines link entities with gaze relations in a temporal sequence and white arrows indicate gaze directions in the current frame. There may exist various number of agents, many different gaze communication types and complex communication relations in one frame, resulting in a highly-challenging and structured task. © 2019 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [523].

with encoder-decoder structure for event-level gaze communication understanding. The reasoning model learns the relations among social entities and iteratively propagates information over a social graph. The event network utilizes the encoder-decoder structure to eliminate the noise in social

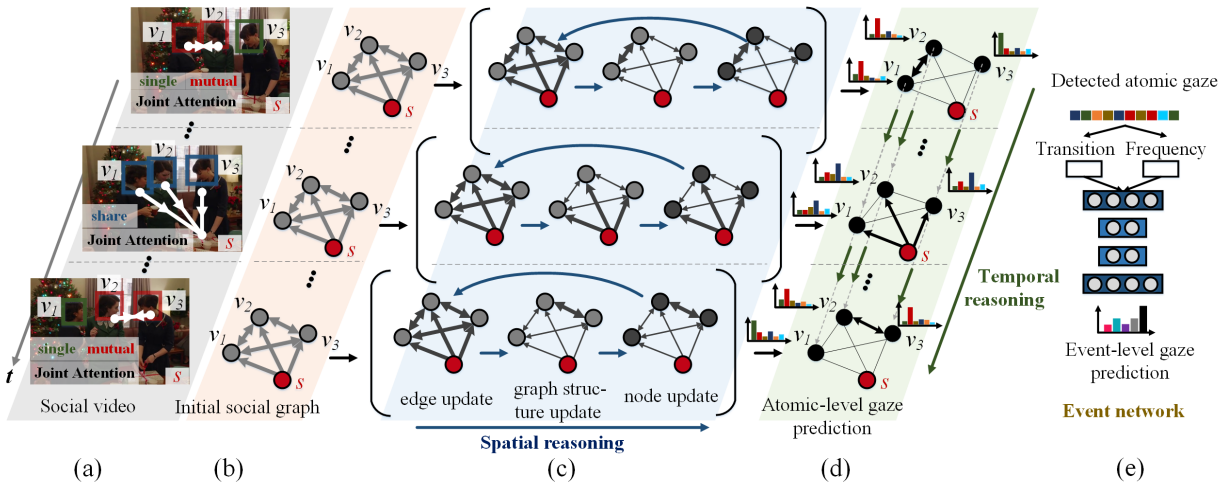


Figure 8.23: **Detailed architecture of the proposed spatio-temporal reasoning model** for gaze communication understanding. © 2019 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [523].

gaze communications and learns the temporal coherent for each event to classify event-level gaze communication.

Social Graph. They first define a social graph as a *complete graph* $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where nodes $v\in\mathcal{V}$ take unique values from $\{1,\cdots,|\mathcal{V}|\}$, representing the entities (*i.e.*, scene, human) in social scenes, and edges $e=(v,w)\in\mathcal{E}$, representing all the possible human-human gaze interactions or human-scene relations. (v,w) indicates a directed edge $v\rightarrow w$. There is a special node $s\in\mathcal{V}$ represents the social scene, and the other nodes $\mathcal{V}\backslash s$ are humans.

For node v, its *node representation/embedding* is denoted by a V-dimension vector: $\mathbf{x}_v\in\mathbb{R}^V$. Similarly, the *edge representation* (or *edge embedding*) for edge $e=(v,w)$ is denoted by an E-dimension vector: $\mathbf{x}_{v,w}\in\mathbb{R}^E$. Each human node $v\in\mathcal{V}\backslash s$ has an output state $l_v\in\mathcal{L}$ that takes a value from a set of atomic social gaze labels $\mathcal{L}=\{$*single*, *mutual*, *avert*, *refer*, *follow*, *share*$\}$. They further define an adjacency matrix $\mathbf{A}\in[0,1]^{|\mathcal{V}|\times|\mathcal{V}|}$ to represent the communication structure over the complete social graph $\mathcal{G}$, where each element $a_{v,w}$ represents the connectivity from node v to w.

Different from most previous graph neural networks that only focus on inferring graph- or node-level labels, the model aims to learn the graph structure $\mathbf{A}$ and the visual labels $\{l_v\}_{v\in\mathcal{V}\backslash s}$ of all the human nodes $\mathcal{V}\backslash s$ simultaneously.

To this end, the spatio-temporal reasoning model is designed to have two steps. First, in spatial domain, there is a message passing step that iteratively learns gaze communication structures $\mathbf{A}$ and propagates information over $\mathbf{A}$ to update node representations. Second, an LSTM is incorporated into the model for more robust node representation learning by considering temporal dynamics. A more detailed model architecture is schematically depicted in Fig. 8.23. In the following, the above two steps of the model will be described in detail.

Message Passing based Spatial Reasoning. Inspired by previous graph neural networks [524, 525, 526], the message passing step is designed to have three phases, an *edge update* phase, a *graph structure update* phase, and a *node update* phase.

The whole message passing process runs for N iterations for iteratively propagating information.

In n-th iteration step, the model first perform the edge update phase that updates edge representations $\mathbf{y}_{v,w}^n$ by collecting information from connected nodes:

$$\mathbf{y}_{v,w}^{(n)}=f_E(\langle\mathbf{y}_v^{(n-1)},\mathbf{y}_w^{(n-1)},\mathbf{x}_{v,w}\rangle), \tag{8.39}$$

where $\mathbf{y}_v^{(n-1)}$ indicates the node representation of v in $(n-1)$-th step, and $\langle \cdot, \cdot \rangle$ denotes concatenation of vectors. f_E represents an *edge update function* $f_E : \mathbb{R}^{2V+E} \rightarrow \mathbb{R}^E$, which is implemented by a neural network.

After that, the graph structure update phase updates the adjacency matrix $\mathbf{A}$ to infer the current social graph structure, according to the updated edge representations $\mathbf{y}_{v,w}^{(n)}$:

$$a_{v,w}^{(n)} = \sigma(f_A(\mathbf{y}_{v,w}^{(n)})), \tag{8.40}$$

where the connectivity matrix $\mathbf{A}^{(n)} = [a_{v,w}^{(n)}]_{v,w}$ encodes current visual communication structures, f_A is a *connectivity readout network* $f_A : \mathbb{R}^E \rightarrow \mathbb{R}$ that maps an edge representation into the connectivity weight, and σ denotes nonlinear activation function.

Finally, the node update phase update node representations $\mathbf{y}_v^{(n)}$ via considering all the incoming edge information weighted by the corresponding connectivity:

$$\mathbf{y}_v^{(n)} = f_V(\langle \sum\nolimits_w a_{v,w}^{(n)} \mathbf{y}_{v,w}^{(n)}, \mathbf{x}_v \rangle), \tag{8.41}$$

where f_V represents a *node update network* $f_V : \mathbb{R}^{V+E} \rightarrow \mathbb{R}^V$.

The above functions $f(\cdot)$ are all learned differentiable functions. In above message passing process, social communication structures are inferred in the graph structure update phase (Eq. (8.40)), where the relations between each social entities are learned through updated edge representations (Eq. (8.39)). Then, the information is propagated through the learned social graph structure and the hidden state of each node is updated based on its history and incoming messages from its neighborhoods (Eq. (8.41)). If we know whether there exist interactions between nodes (human, object), *i.e.*, given the groundtruth of $\mathbf{A}$, we can learn $\mathbf{A}$ in an *explicit* manner, which is similar to the graph parsing network [525]. Otherwise, the adjacent matrix $\mathbf{A}$ can be viewed as an attention or gating mechanism that automatically weights the messages and can be learned in an *implicit* manner; this shares a similar spirit to graph attention network [527].

Recurrent Network based Temporal Reasoning. Since the task is defined on a spatio-temporal domain, temporal dynamics should be considered for more comprehensive reasoning. With the updated human node representations $\{\mathbf{y}_v \in \mathbb{R}^V\}_{v \in \mathcal{V} \backslash s}$ from the message passing based spatial reasoning model, LSTM is further applied to each node for temporal reasoning. More specifically, the temporal reasoning step has two phases: a *temporal message passing* phase and a *readout* phase. They denote by $\mathbf{y}_v^t$ the feature of a human node $v \in \mathcal{V} \backslash s$ at time t, which is obtained after N-iteration spatial message passing. In the temporal message passing phrase, the information is propagated over the temporal axis using LSTM:

$$\mathbf{h}_v^t = f_{\text{LSTM}}(\mathbf{y}_v^t | \mathbf{h}_v^{t-1}), \tag{8.42}$$

where $f_{\text{LSTM}} : \mathbb{R}^V \rightarrow \mathbb{R}^V$ is an LSTM based temporal reasoning function that updates the node representation using temporal information. $\mathbf{y}_v^t$ is used as the input of the LSTM at time t, and $\mathbf{h}_v^t$ indicates the corresponding hidden-state output via considering previous information $\mathbf{h}_v^{t-1}$.

Then, in the readout phase, for each human node v, a corresponding gaze label $\hat{l}_v^t \in \mathcal{L}$ is predicted from the final node representation $\mathbf{h}_v^t$:

$$\hat{l}_v^t = f_R(\mathbf{h}_v^t), \tag{8.43}$$

where $f_R : \mathbb{R}^V \rightarrow L$ maps the node feature into the label space $\mathcal{L}$, which is implemented by a classifier network.

Event Network. The event network is designed with an encoder-decoder structure to learn the correlation of the atomic gazes and classify the event-level gaze communication for each video sequence. To reduce the large variance of video length, they pre-process the input atomic gaze

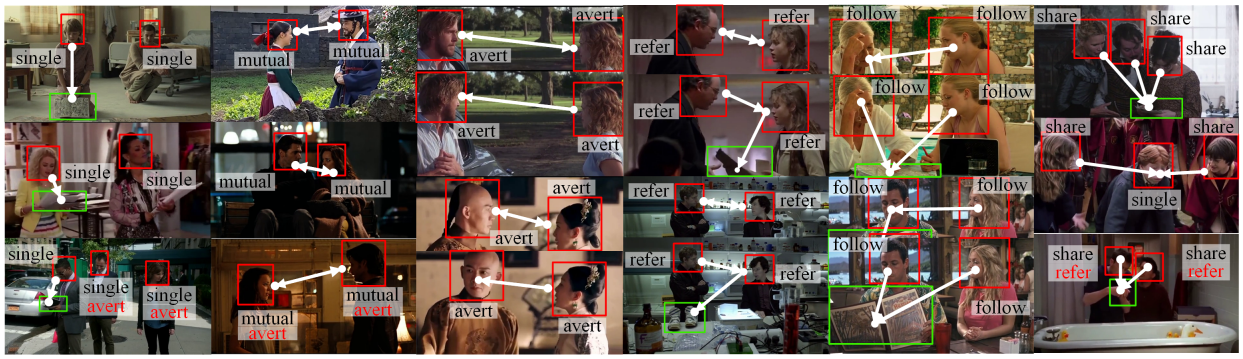


Figure 8.24: Qualitative results of atomic-level gaze communication prediction. Correctly inferred labels are shown in black while error examples are shown in red. © 2019 Lifeng Fan *et al*. Reprinted, with permission, from Ref. [523].

sequence into two vectors: i) the transition vector that records each transition from one category of atomic gaze to another, and ii) the frequency vector that computes the frequency of each atomic type. The encoder individually encodes the transition vector and frequency vector into two embedded vectors. The decoder decodes the concatenation of these two embedded vectors and make final event label prediction. Since the atomic gaze communications are noisy within communicative activities, the encoder-decoder structure will try to eliminate the noise and improve the prediction performance. The encoder and decoder are both implemented by fully-connected layers.

Here is a short summary of the whole spatio-temporal reasoning process. As shown in Fig. 8.23, with an input social video (a), for each frame, an initial complete graph $\mathcal{G}$ (b) is built to represent the gaze communication entities (*i.e.*, humans, and social scene) by nodes and their relations by edges. During the spatial reasoning step (c), edge representations are firstly updated using Eq. (8.39) (note the changed edge color compared to (b)). Then, in the graph structure update phase, the graph structure is inferred through updating the connectivities between each node pairs using Eq. (8.40) (note the changed edge thickness compared to (b)). In the node update phase, node embeddings are updated using Eq. (8.41) (note the changed node color compared to (b)). Iterating above processes leads to efficient message propagation in spatial domain. After several spatial message passing iterations, the enhanced node feature is fed into a LSTM based temporal reasoning module, to capture the temporal dynamics (Eq. (8.42)) and predict final atomic gaze communication labels (Eq. (8.43)). Then the event network is applied to reason about event-level labels based on previous inferred atomic-level label compositions for a long sequence in a larger time scale.

8.7.2 Result Visualization and Analysis

Fig. 8.24 shows some test result example visualizations by the full model for the atomic-level gaze communication classification task. The predicted communication structures are shown with bounding boxes and arrows. The full model can correctly recognize different atomic-level gaze communication type (shown in black) with effective spatial-temporal graph reasoning. Some error examples are also exhibited (shown in red).

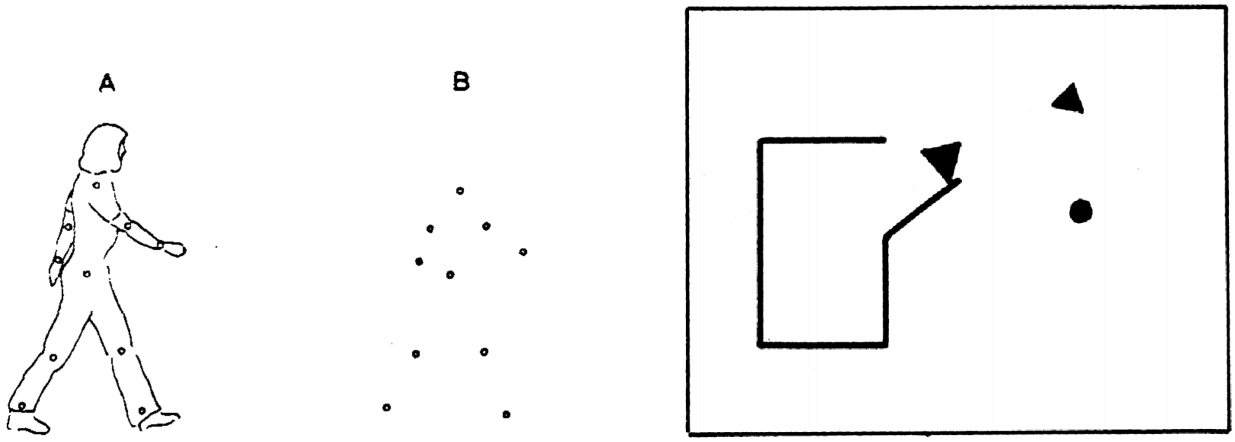


Figure 9.1: Illustration of the cognitive science experiment discussed: (Left) Observers immediately report seeing a walking pattern even without the outline contour of a person; (Right) Observers recognize the goal of each shape in a Heider-Simmel display.

Chapter 9

Intentionality

9.1 Introduction

Understanding human activities is one of the most fundamental problems in artificial intelligence and computer vision. As the most readily available learning source, there has been great effort put into video analysis in the computer vision field. However, there are more profound reasons why we have to look to the origin of human activity understanding. With studies and analyses of human motion perception rooted in the field of neuroscience [528]; Johansson's seminal work on visual perception of biological motion [529] first paved the way for the mathematical modeling of human action and automatic recognition. Notably, using a dot-representation of human motions instead of using pixel-based input, Johansson adopted a method to produce proximal patterns (*i.e.*, the moving light display experiment), which demonstrated that human perception of activities does not tightly couple with pixel-based features; human subjects can still perceive the semantics of activities from sparse representations of motions. Similarly, in the classic Heider-Simmel display [530], human subjects can directly and irresistibly perceive a story-like description of the observed motions just upon viewing simple shapes roaming around a space. These experiments set up a cornerstone for studying the underlying intents, rather than the superficial behaviour, that matters when we observe motions [531].

In fact, cognitive studies [123] also have shown that humans have a strong inclination to interpret

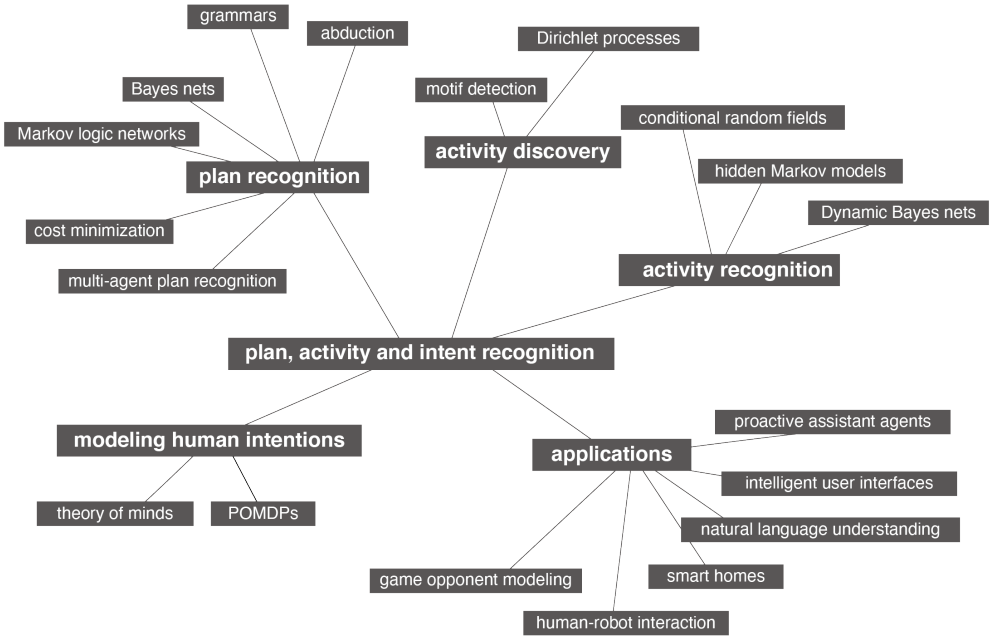


Figure 9.2: A map of research directions and applications related to understanding and reasoning with intents considered. Courtesy of Sukthankar [535]

events as a series of goals driven by the intentions of agents. As discovered by Gergely *et al*. [532], infants as young as 12 months old have the ability to inference about actions using a rationality principle: the assumption that intentional actions bring about the most efficient and economical path in order to achieve a goal. They further proposed the teleological stance theory [533], which states that infants are endowed with an interpretational system that allows teleogical reasoning and inference over actions depending on goals. Such a teleological stance has inspired various models for intent estimation as an inverse planning problem [124, 534].

In addition to perceptible goal achievements, intents often include other hidden status of agents (humans and animals), such as drinking water because of being "thirsty," "hungry" or "tired." Such transient status are similar to, but more complex than, the fluents of objects, and come with the following characteristics: (i) They are hierarchically organized in a sequence of goal status and are the main factors driving actions and events in a scene. (ii) They are oftentimes "dark," that is, not represented by pixels. (iii) Unlike the instant change of fluents in response to actions, intents are often formed across long spatial-temporal ranges. With the hope of truly understanding human behaviors and the pressing need for sophisticated and efficient autonomous agents, these characteristics add new challenges to research and have already boomed a large field covering topics in computer vision as well as artificial intelligence, see in Fig. 9.2.

In this chapter, we discuss on the topic of teleological reasoning and intent in both modeling and related applications. We provide a formulation for representing this process using Spatial-Temporal-Causal And-Or Graphs (STC-AOG) in Section 9.2. Next we use goal inference and action prediction as two typical teleological reasoning tasks to illustrate how we integrate properties about intents into real-world scenes in Section 9.3 and Section 9.4. We conclude the chapter with a discussion on possible future directions on this field.

9.2 Formulating Intents with STC-AoG

Understanding an event typically requires four types of knowledge need to be captured by a knowledge representation system: (i) spatial knowledge that expresses the physical configuration of the environment when performing the task; (ii) temporal knowledge which reveals the series of human actions in the process of the task, (iii) causal knowledge that conveys the status change of an object in each dynamic human action and (iv) theory of mind representations for modeling others' beliefs under social scenarios. Spatial knowledge empower the ability of spatial reasoning for getting better estimation about one's current status, while temporal and causal knowledge work together in deriving how things will behave and why things behaved like they did. Here we put less effort on the multi-agent social scenarios to make the problem clearer. To capture the first three types of knowledge required in understanding event and inferring latent intent, we here represent knowledge of the physical environment, consisting of objects, scenes, actions by a joint stochastic spatial, temporal, and causal And-Or Graph (STC-AoG).

The And-Or Graph (AoG) is defined as a 3-tuple $\mathcal{G} = (V, R, P)$, where $V = V^{\text{AND}} \cup V^{\text{OR}} \cup V^{\text{T}}$ consists of a disjoint set of And-nodes, Or-nodes, and Terminal nodes respectively. And-nodes represents decomposition (conjunction) of an entity into its constituent parts. Or-nodes represents alternative ways of decomposition and Terminal nodes represents grounding basic entities which serve as a basis for describing a scene or event. R represents a set of relations between Or-nodes or sub-graphs, each of which represents a generating process from a parent node to its children nodes. Given this definition, a parse graph is an instance of $\mathcal{G}$ where each Or-node decides one of its children.

On top of the basic AoG structure, we can define three different types of AoG for describing events. First, we represent spatial concepts through a stochastic Spatial And-Or Graph (S-AoG), where nodes in the S-AoG represent visual information of varying levels of abstraction over basic object entities. An And-node in this case signifies physical compositionality whereas an Or-node describes structural variation. Next, the hierarchical nature of actions leads us to represent actions by a stochastic Temporal And-Or Graph (T-AoG), where And-nodes correspond to a sequence of actions with the semantic of a macro action, Or-nodes correspond to alternative actions for completing the same macro action, Terminal-nodes are the basic action primitives we are concerned with. Finally, a Causal And-Or Graph (C-AoG) for encapsulating causality, where each cauasl node is a fluent change operator, transforming an input fluent to an output fluent by using an action from the T-AoG. In this way, we can easily represent an event like "make a hot meal" as a joint spatial, temporal, causal parse graph as shown in Fig. 9.3.

With the above definition given a clear specification on how each aspect of an event is formulated, we here clarify on how such a STC-AoG representation is used for representing intents and goals. As we can notice from the formulation, each parse graph of this joint AoG corresponds to a complete event. At some point before the event is finished, we can decompose a STC-pg into three different areas as shown in Fig. 9.4. We observe a sub-parse graph oftentimes when observing events that haven't ended yet. Given the knowledge about the full event, we will have three different type of sub-parse graphs over the full one. First, we have the current situation defined in a partial parse graph $pg_{0:t}$ describing the spatial-temporal changes of the event during time interval $[0, t]$. This part of the sub-parse graph show explanations on how events evolved during time $[0, t]$, which also correspond to the prior observations (or history) at current observation time point t. Next, we have the following partial parse graph to the current action period in $[t, t + D]$, which correspond to nodes in orange and triangle in red. This sub parse graph $pg_{t:t+D}$ represents what the event is currently evolving as and is oftentimes related to people's attention when working on a specific task. Finally, intents and plans are covered in the $pg_{t:t+T}$, which gives possible future actions and

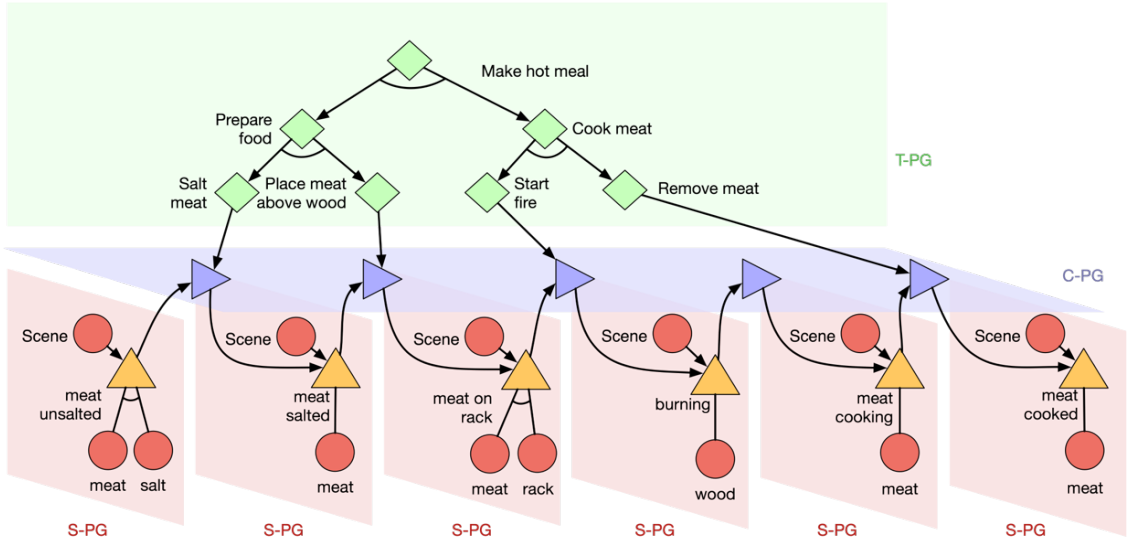


Figure 9.3: An STC-pg for representing the event of "make hot meal." From bottom to top, temporal nodes describe human action patterns at different level of resolution where macro actions are decomposed into action primitives. Spatial nodes describe the status of meat (*e.g.*, being salted or not) and causal nodes describe the transition of spatial entity status after applying actions. These three types of nodes jointly describe scene changes over time caused by human action effects and planning which serve as a in-depth representation of the event.

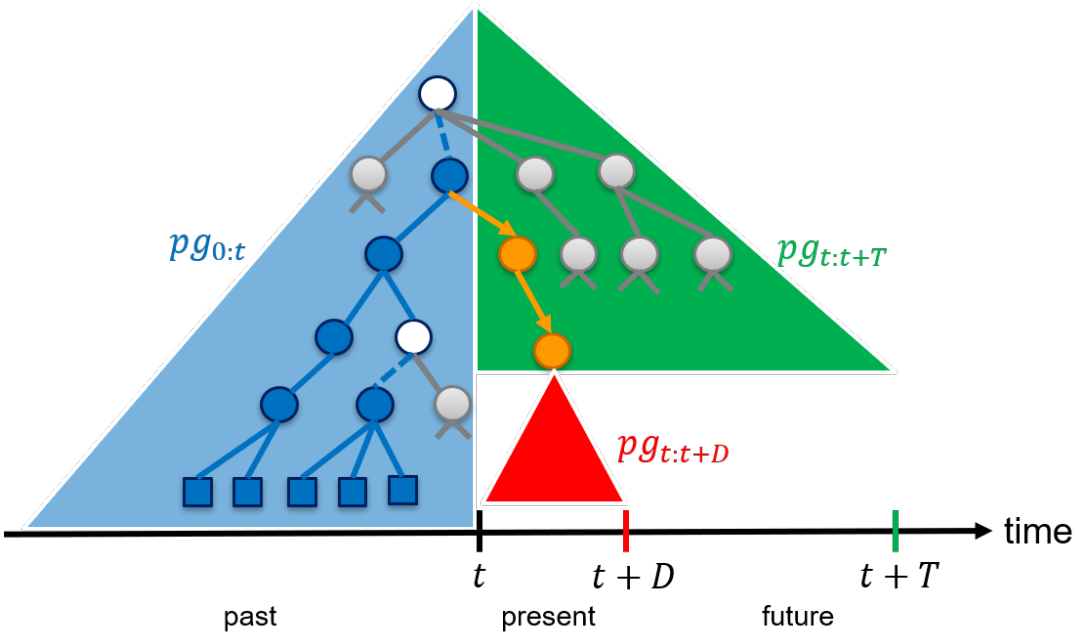


Figure 9.4: In this figure And-nodes are represented with center-colored nodes and Or-nodes are represented by center-white nodes. Connections between And-node to other nodes are in solid line and dotted line represent connection between Or-node to its predecessors. Parse graphs representing past are colored in blue, the partial parse that currently is being processed on is colored in orange and future parses are shaded in green.

goal states with the constraints provided by the full STC-AoG structure.

A perhaps more intuitive illustration of the decomposition of an STC-pg comes from taking

different perspectives. An human-performed event, when standing in the shoes of the performer, is actually represented by a sequence of desired states and corresponding actions for achieving this states. With the hierarchical nature in planning and perceptual organization, the corresponding STC-pg actually describes the detailed attribute values of scene, action during each step of the planning for this egocentric planner. Given this into consideration, when taking from an allocentric perspective and interpreting others' intents and goals, what we try to obtain is actually this ground-truth STC-pg that can summarize the planner's goals and actions. In contrast to egocentric agents who possess the ground-truth STC-pg, we need to estimate and select the most probable STC-pg from a set of possible parse graphs we are knowledgeable of that best explains the current observation. To this end, intent recognition and goal/action prediction is transformed into finding the best partial parse given the current observation over the STC-AoG that captures all knowledge of possible events. We skip the technical details about optimal parsing or partial parsing here as we have described algorithms and details in the second volume of the book. Next, we get into more details on two common applications of recognizing intents and goals, goal prediction and action prediction.

9.3 Inferring the Intentionality and Goals of Agents

As we mentioned in Section 9.2, intents and goals could be properly represented by a STC-pg. However, there are several critical issues with obtaining this structure, especially from an allocentric view when inferring others' intents and goals. This reflects to the second property we discussed about intents in Section 9.1. Intents are "dark," in the sense that it sometimes can not be represented by pixel. Take the example shown in Fig. 9.5, when a person heads to a food truck in the courtyard, there are extra efforts needed in obtaining the STC-pg which covers internal status of the person being hungry. Such a gap between planner's representation of events and observer's causes trouble for obtaining the full STC-AoG, as well as retrieving the most probable STC-pgs when inferring others' intents.

This problem is not entirely unsolvable and leads to the definition of functional objects. To a certain degree, much of human understanding depends on the ability to comprehend causality, this induces the concept of functional objects that bind the actions and potential effects to the objects themselves, *e.g.*, a chair is bind with sitting and relaxing. Other examples as shown in Fig. 9.6 include food truck (solving hunger), trashcan (throwing trash), or vending machine (solving thirst), *etc*. With the prior knowledge of this functional objects, we can solve the "dark" problem by correlating the goals and intents to the actual actions or states that agents want to achieve to interactions with functional objects. In this sense, we will only need the knowledge of food truck sell foods that solves hunger to interpret the event of one person approaching to it without observing the fact that he/she is hungry at the moment.

Additionally, we can make simple and proper inference with only functional objects and trajectories of agents even without the detailed STC-AoG representation. As studied by Xie *et al*. [92], people in pulic spaces are expected to intentionally take shortest paths (subject to obstacles) toward certain functional objects (*e.g.*, vending machine, picnic table, trash-can, *etc.*) where they can satisfy certain needs (*e.g.*, quench thirst). Colored trajectories in Fig. 9.6 show several trajectories of people. If you would notice, trajectories are naturally attracted to functional objects like food truck or vending machine, while being repelled from obstacles like foul odor or grassland. Without a detailed description of the plan of the agents (in STC-AoG form), we can already use such attracting/repelling properties for a rough estimation of goal recognition under the rationality principle.

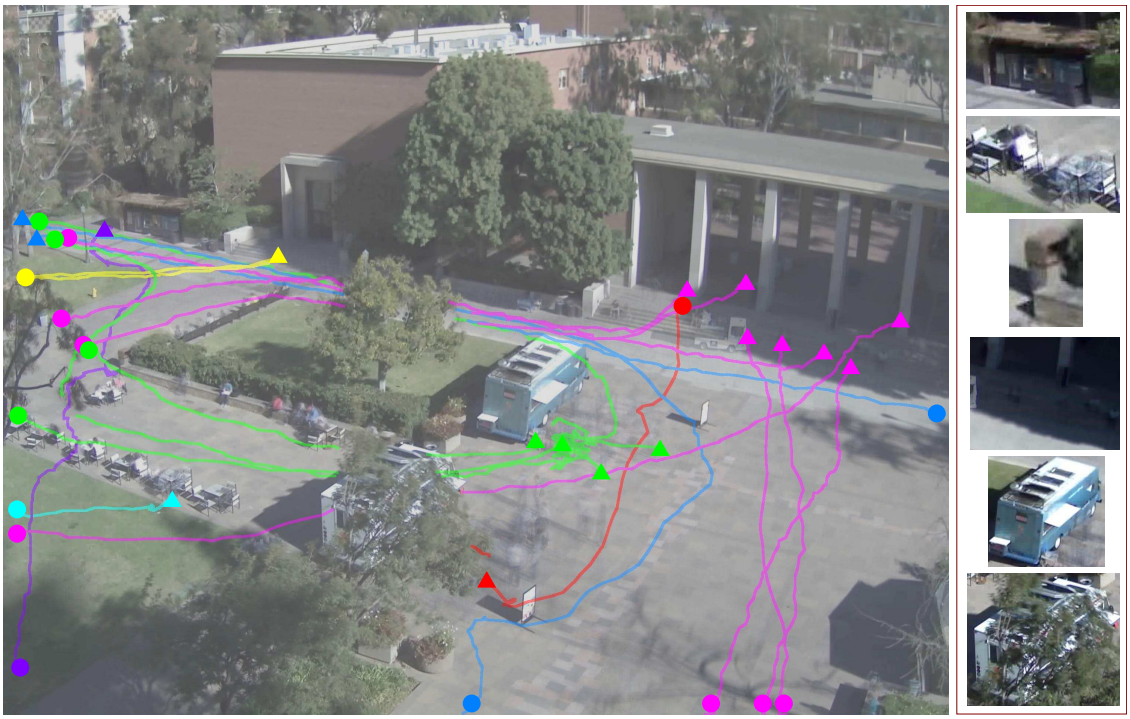


Figure 9.5: Example of human trajectories taken at UCLA courtyard. Courtesy of (Xie, 2018) [92].

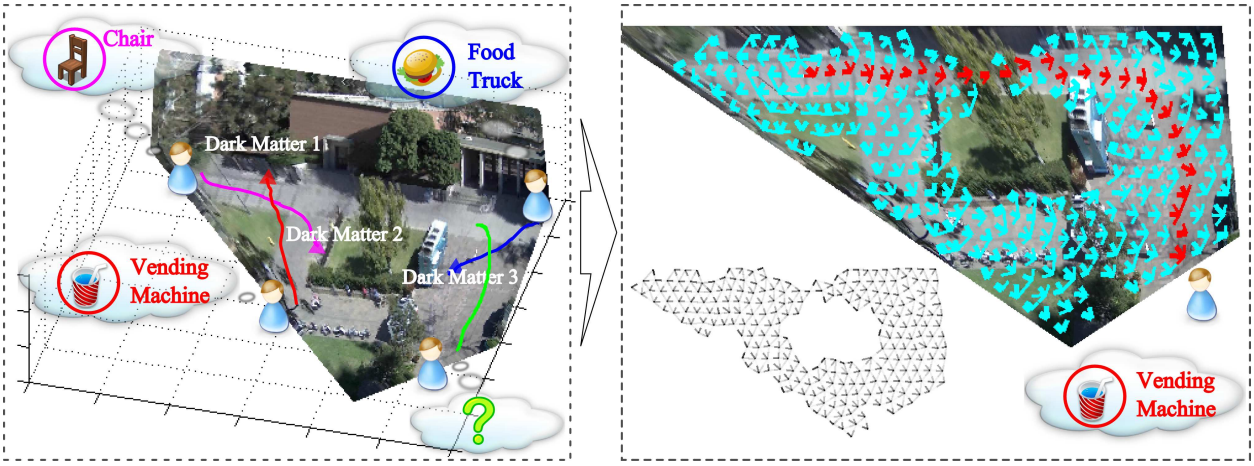


Figure 9.6: An example video where people driven by latent needs move toward functional objects. (Right) A top down visualization of prediction results of: (a) Inferring and localizing the person's goal destination; (b) Predicting a person's full trajectory (red); (c) Estimating the force field affecting the person (the blue arrows, where their thickness indicates the force magnitude; the black arrows represent another visualization of the same field.); and (d) Estimating the constraint map of non-walkable areas and obstacles in the scene (the "holes" in the field of blue arrows and the field of black arrows). Courtesy of (Xie, 2018) [92].

Such an idea was formulated by Xie *et al*. as a scene representation consists of layers of attraction propulsion fields based on objects' functionality and influence on human activities. They referred to these objects as "dark matter" because they are distinguishable from other objects primarily by the functionality to attract or repel people, not by their appearance. This definition comes by analogy to cosmology, where existence and properties of dark matter are hypothesized and inferred from its gravitational effects on visible matter. Table 9.1 lists examples functional objects they concerned

with that exert attraction and repulsion forces on people's trajectories. Under this scenario, intent and goal recognition are reflected by models' capability of inferring goal destinations, estimating the force fields affecting the person and predicting the person's full trajectory.

Examples of "dark matter"	Human need
Vending machine / Food truck / Table	Hunger
Water fountain / Vending machine	Thirst
ATM / Bank	Money
Chair / Table / Bench / Grass	Rest
News stand / Ad billboard	Information
Trash can	Hygiene
Bush / Tree	Shade from the sun

Table 9.1: Examples of human needs and objects that can satisfy these needs in the context of a public space. These functional objects appear as "dark matter" attracting people to approach them, or repelling people to stay away from them.

Agent-based Lagrangian Mechanics

We briefly discuss about the modeling of this "dark matter" formulation using force fields and least action principle in the Langragian Mechanics framework. At the scale of large scenes such as courtyard, people can be considered as "particles" whose shapes and dimensions are neglected, and their motion dynamics can be modeled within the framework of Lagrangian mechanics (LM) [536]. LM studies the motion of a particle with mass, m, at positions $\mathbf{x}(t) = (x(t), y(t))$ and velocity, $\dot{\mathbf{x}}(t)$, in time t, in a force field $\vec{F}(\mathbf{x}(t))$ affecting the motion of the particle. Particle motion in generalized coordinates system is determined by the Lagrangian function, $L(\mathbf{x}, \dot{\mathbf{x}}, t)$, defined as the kinetic energy of the entire physical system, $\frac{1}{2}m\dot{\mathbf{x}}(t)^2$, minus its potential energy, $-\int_{\mathbf{x}} \vec{F}(\mathbf{x}(t))d\vec{\mathbf{x}}(t)$,

$$L(\mathbf{x}, \dot{\mathbf{x}}, t) = \frac{1}{2}m\dot{\mathbf{x}}(t)^2 + \int_{\mathbf{x}} \vec{F}(\mathbf{x}(t))d\vec{\mathbf{x}}(t). \tag{9.1}$$

Action in such a physical system is defined as the time integral of the Lagrangian of trajectory $\mathbf{x}$ from t_1 to t_2: $\int_{t_1}^{t_2} L(\mathbf{x}, \dot{\mathbf{x}}, t)dt$. LM postulates that a particle's trajectory, $\Gamma(t_1, t_2) = [\mathbf{x}(t_1), ..., \mathbf{x}(t_2)]$, is governed by the principle of Least Action in a generalized coordinate system:

$$\Gamma(t_1, t_2) = \arg\min_{\mathbf{x}} \int_{t_1}^{t_2} L(\mathbf{x}, \dot{\mathbf{x}}, t)dt. \tag{9.2}$$

Since the classical LM is not directly applicable to the trajectory analysis domain, we can simplt extend it into Agent-based Lagrangian mechanics (ALM) by letting the physical system consists of a set of force sources. The first extension enables the particles to become agents with free will to select a particular force source from the set which can drive their motion. The second extension endows the agents with knowledge about the layout map of the physical system. Consequently, by the principle of Least Action, the agents can globally optimize their shortest paths toward the selected force source, subject to the known layout of obstacles. These two extensions can be formalized as follows.

Let i-th agent choose j-th source from the set of sources. Then, i's action, *i.e.*, the trajectory

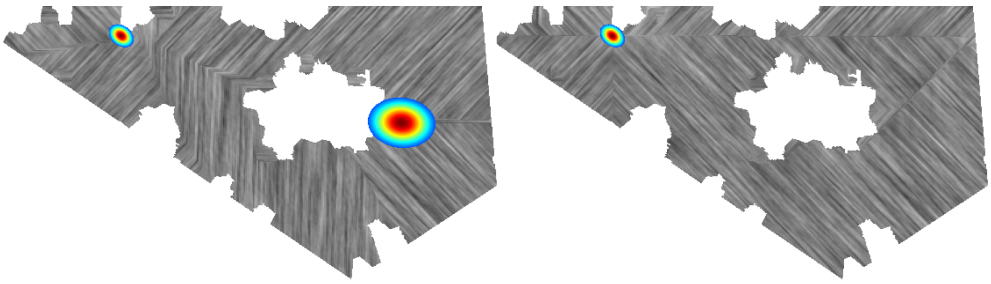


Figure 9.7: Visualizations of the force field for the scene from Fig. 9.6. (*left*) In LM, particles are driven by a sum of all forces; the figure shows the resulting fields generated by only two sources. (*right*) In ALM, each agent selects a single force $\vec{F}_j(\mathbf{x})$ to drive its motion; the figure shows that forces at all locations in the scene point toward the top left of the scene where the source is located. The white regions represent our estimates of obstacles. Repulsion forces are short ranged, with magnitudes too small to show here. Courtesy of Xie [92]

could be obtained by

$$\begin{aligned}&\Gamma_{ij}(t_1,t_2)\\&=\arg\min_{\mathbf{x}}\int_{t_1}^{t_2}\Big[\frac{1}{2}m\dot{\mathbf{x}}(t)^2+\int_{\mathbf{x}}\vec{F}_{ij}(\mathbf{x}(t))d\vec{\mathbf{x}}(t)\Big]dt,\\&\quad\text{s.t.}\quad\mathbf{x}(t_1)=\mathbf{x}_i,\ \mathbf{x}(t_2)=\mathbf{x}_j.\end{aligned}\tag{9.3}$$

Here we use the notation $\vec{F}_{ij}(x)$ to denote the net force coming from an attracting/repelling sources. In order to solve this optimization problem, certain approximation need to be applied. In the domain of public spaces, the agents cannot increase their speed without limit. Hence, every agent's speed is upper bounded by some maximum speed. Also, it is reasonable to expect that accelerations or decelerations of people along their trajectories in a public space span negligibly short time intervals. Consequently, the first term in Eq. (9.3) is assumed to depend on a constant velocity of the agent, and thus does not affect estimation of $\Gamma_{ij}(t_1,t_2)$. For simplicity, we can also assume that agents make only discrete displacements over a lattice of scene locations Λ (e.g., representing centers of superpixels occupied by the ground surface in the scene), *i.e.*, $d\vec{\mathbf{x}}(t)=\vec{\Delta\mathbf{x}}$. The second assumption is that the agent is reasonable and always moves along the direction of $\vec{F}_{ij}(\mathbf{x})$ at every location. We can therefore transfrom Eq. (9.3) into:

$$\begin{aligned}&\Gamma_{ij}(t_1,t_2)=\arg\min_{\Gamma\subset\Lambda}\sum_{\mathbf{x}\in\Gamma}|\vec{F}_{ij}(\mathbf{x})\cdot\vec{\Delta\mathbf{x}}|,\\&\quad\text{s.t.}\quad\mathbf{x}(t_1)=\mathbf{x}_i,\ \mathbf{x}(t_2)=\mathbf{x}_j.\end{aligned}\tag{9.4}$$

A globally optimal solution of (Eq. (9.4)) can be found with the Dijkstra algorithm. The end location of the predicted $\Gamma_{ij}(t_1,t_2)$ corresponds to the location of source j. It follows that estimating human trajectories can readily be used for estimating the functional map of the scene. To better model agents' trajectories we can consider three types of agents' behaviors as described in [534] in addition to force fields. "single" which indicates agents' intents on reaching one specific goal, "sequential" which indicates agents' intents to achieve several goals along the trajectory and "change of intent" when an agent may give up on the initial goal before reaching it, and switch to another goal.

Under the assumption that we have access to noisy trajectories of agents, observed over a given time interval in the video, $\Gamma'=\Gamma'(0,t_0)=\{\Gamma'_i(0,t_0):i=1,...,M\}$. Given these observations, we define latent trajectories of agents for any time interval, (t_1,t_2), including those in the future (*i.e.* unobserved intervals), $\Gamma=\Gamma(t_1,t_2)=\{\Gamma_i(t_1,t_2):i=1,...,M\}$. Each trajectory Γ_i is specified by

accounting for one of the three possible behaviors of the agent. Following the principle of Least Action, as specified in Section 9.3, an optimal trajectory $\Gamma_{ij}(t_1, t_2) = [\mathbf{x}(t_1) = \mathbf{x}_i, \ldots, \mathbf{x}(t_2) = \mathbf{x}_j]$ of a_i at location $\mathbf{x}_i$ moving toward $\mathbf{s}_j$ at location $\mathbf{x}_j$ minimizes the energy $\sum_{\mathbf{x}\in\Gamma_{ij}} |\vec{F}_{ij}(\mathbf{x}) \cdot \vec{\Delta\mathbf{x}}|$. The agent's behavior can thus be formulated as

$$\Gamma_i = \sum_j \Gamma_{ij} = \arg\min_{\Gamma\subset\Lambda} \sum_j \sum_{\mathbf{x}\in\Gamma} |\vec{F}_{ij}(\mathbf{x}) \cdot \vec{\Delta\mathbf{x}}|, \tag{9.5}$$

where the summation over j uses: (i) only one source for "single" intent (*i.e.*, $\Gamma_i = \Gamma_{ij}$ when $r_{ij} = 1$), (ii) two sources for "change of intent," and (iii) maximally n sources for "sequential" behavior. Note that for the "sequential" behavior the minimization in (Eq. (9.5)) is constrained such that the trajectory must *sequentially* pass through locations $\mathbf{x}_j$ of all sources $\mathbf{s}_j$ pursued by the agent.

Probabilistic Inference with MCMC

Using these definitions, we can infer the force field of function map M as well as agents' behaviour types Z by probabilistic inference over the joint posterior distribution over $W = \{M, Z\}$:

$$P(W|\Gamma, I) \propto P(W|I)P(\Gamma|W) = P(W|I) \prod_{i=1}^{M} P(\Gamma_i|W), \tag{9.6}$$

where the terms are decomposed into prior estimation of the force fields and the likelihood of trajectories given the function map. We skip the probabilistic formulation of the probability $P(W|I)$ which is discussed detailedly in [92]. The likelihood of trajectory Γ_i comes in a natural form from energy-based models where the energy that a_i must spend moving along the trajectory and probability can be formulated as:

$$P(\Gamma_i|W) \propto e^{-\lambda \sum_{j,\mathbf{x}\in\Gamma_{ij}} |\vec{F}_{ij}(\mathbf{x}) \cdot \vec{\Delta\mathbf{x}}|}, \tag{9.7}$$

where $\lambda > 0$. The likelihood in (Eq. (9.7)) models that when agent i is far away from a potential source, the total energy needed to cover that trajectory is bound to be large, and consequently uncertainty about agent i's trajectory is large. Conversely, as agent i gets closer to a force field source, uncertainty about the trajectory reduces.

Under this probabilistic formulation, given observations $\{I, \Gamma\}$, we can estimate the overall status over interval $(0, t_0)$ through the data-driven MCMC [537, 538] approach. In essence, each step of the MCMC proposes a new solution W_{new}. The decision to discard the current solution, W, and accept W_{new} is made based on the acceptance rate,

$$\alpha = \min(1, \frac{Q(W \to W_{\text{new}})}{Q(W_{\text{new}} \to W)} \frac{P(W_{\text{new}}|\Gamma, I)}{P(W|\Gamma, I)}). \tag{9.8}$$

If α is larger than a threshold uniformly sampled from $[0, 1]$, the jump to w_{new} is accepted. The posterior distribution of $P(W|\Gamma, I)$ is specified as above. Initial states of the MCMC algorithm is designed to be states that will not lose generality in the sampling process. New proposals W_{new} are generated based on initial states W, please refer to [92] for more details. We show an example of step-wise result for models' estimation during the MCMC sampling process in Fig. 9.8. We also shown an comparison of hyperparameter selection for λ which governs the uncertainty in likelihood modeling in Fig. 9.9.

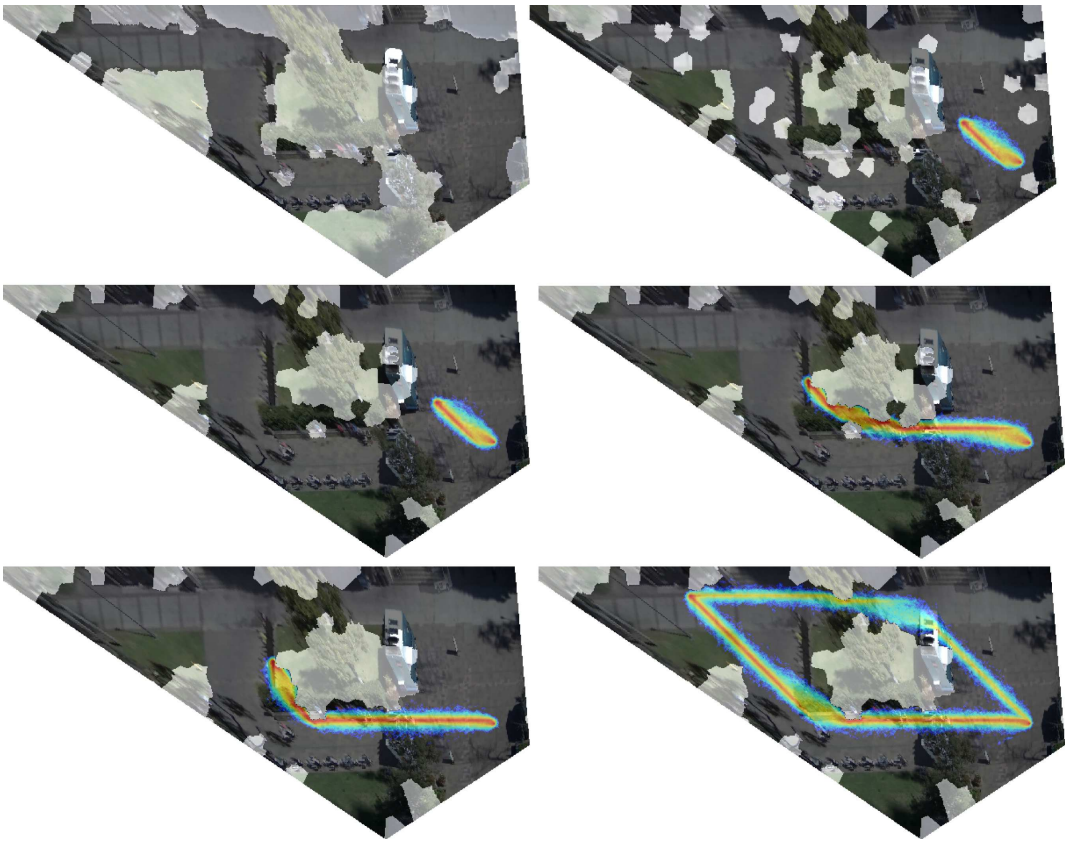


Figure 9.8: Top view of the scene from Fig. 9.6 with the overlaid illustration of the MCMC inference. The rows show the progression of proposals of the function map in raster scan (the white regions indicate obstacles), and trajectory estimates of agent i with goal on the right which gradually shifts to the location at the top-left of the scene during MCMC process, and finds two equally likely trajectories for this goal. Courtsey of Xie [92].

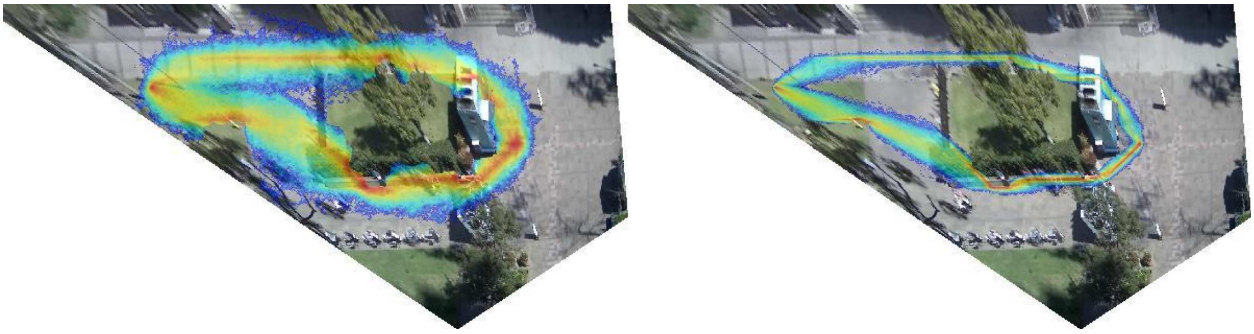
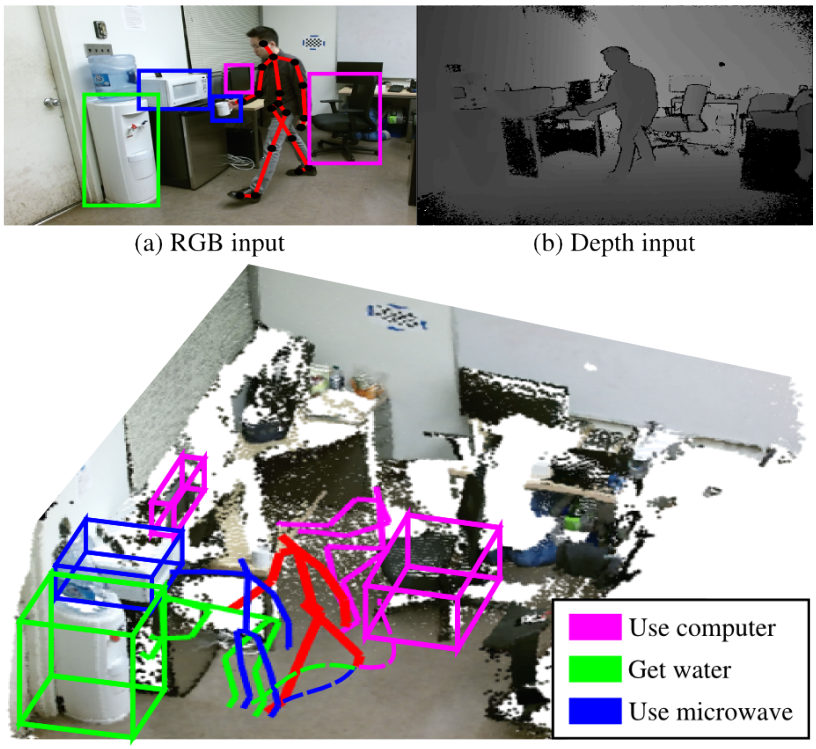


Figure 9.9: Top view of the scene from Fig. 9.6 with the overlaid trajectory predictions of a person who starts at the top-left of the scene, and wants to reach the dark matter in the middle-right of the scene (the food truck). A magnitude of difference in parameters $\lambda = 0.2$ (*on the left*) and $\lambda = 1$ (*on the right*) used to compute likelihood $P(\Gamma_{ij}|W)$ gives similar trajectory predictions. The predictions are getting more certain as the person comes closer to the goal. Warmer colors represent higher likelihood. Courtsey of Xie [92].

9.4 Predicting Human Intents in Daily activities

Compared to agents' trajectories, understanding daily active humans naturally requires more delicate reasoning due to its non-Markovian property and rich contexts between human and environ-



(a) RGB input (b) Depth input

(c) Activity prediction

Figure 9.10: What is he going to do? (a)(b) Input RGB-D video frames. (c) Activity prediction: human action with interacting objects, and object affordances (how the agent will perform the task). The red skeleton is the current observation. The magenta, green and blue skeletons and interacting objects are possible future states.

ments. As we previously discussed in Section 9.2, an event is decompositional spanning spatial, temporal and causal aspects. In this section, we dig deeper into how such a STC-AoG could be formulated and provide examples on how inference could be done over such representations.

The most proper correspondence to STC-AoG in algorithmic structures is Stochastic Context-Free Grammars (SCFG). Here we leave the detailed definition of SCFG to Book II as it is not our main focus. Such a grammar model need to capture features including human actions, objects, and their affordances. Down to earth, the problem of learning activities and inferring goals become the problem of learning grammar models from demonstration sequences and finding the best partial parse graphs. Common tasks related to intent recognition are transformed in a similar way. Recall in Fig. 9.4, attention prediction becomes the problem of finding the objects that the current sub-parse graph is at, future action prediction becomes finding the most probable next-action token given the current parsing results. Depending on the fineness of study, one can add on to these rough topics by making finer level inferences like predicting gaze direction, foot step location as well as hand-object interaction patterns. To provide an illustrative example, we use the task of future action prediction in this case to walk through the whole pipeline.

Consider the image from a video shown in Fig. 9.10 (a). The task of future action prediction is to predict what the possible future states are to some extent. For example, three possible future states are likely as in Fig. 9.10 (c). After applying basic computer vision tools, with the knowledge of the person grabing a cup, we should consequently make the prediction that the person is going to get water. We here show an example STC-AoG representation from Qi *et al.* [45] for representing events using spatial object/human features as well as actions in Fig. 9.11.

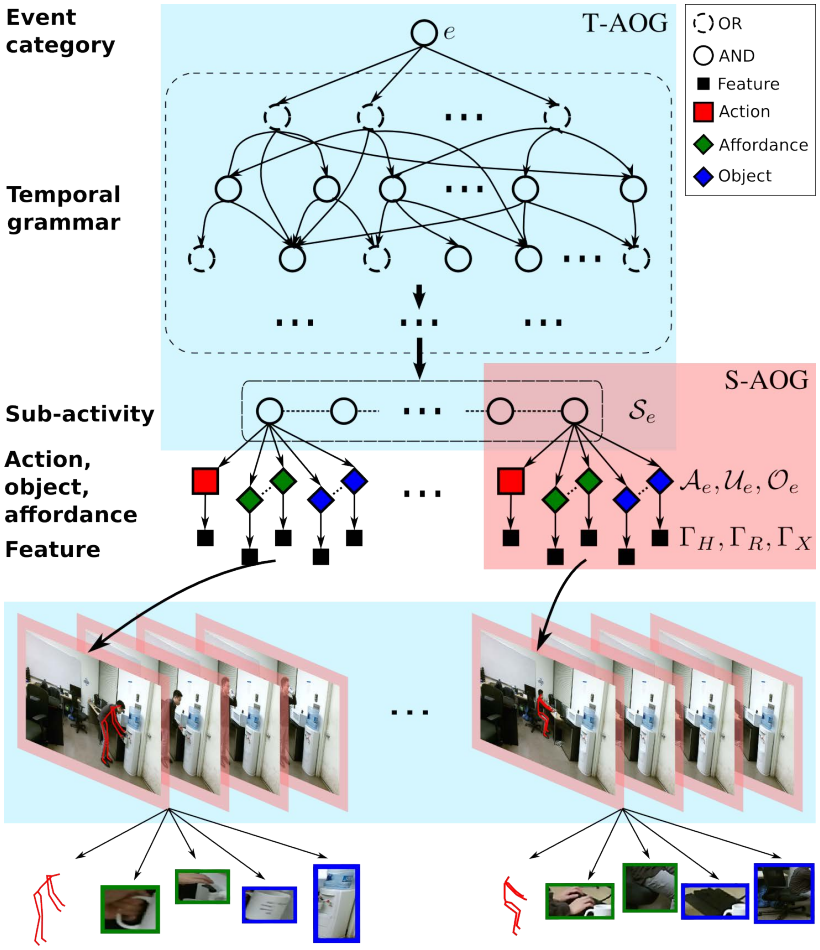


Figure 9.11: Illustration of the STC-AoG. The sky-blue area indicates the T-AoG, and the coral area indicates the S-AoG and the C-AoG is latent and therefore omitted in this graph. The T-AoG is a temporal grammar in which the root node is the activity and the terminal nodes are sub-activities. The S-AoG represents the state of a scene, including the human action, the interacting objects and their affordances.

The problem of inferring intentions and making predictions of future actions require the basis of learning events and organizing them into a STC-AoG structure. This learning process can be decomposed into two main parts: (i) learning the symbolic grammar structure for each event/task, and (ii) learning the parameters Θ of the underlying grammar model. Using grounded action sequences, one can simply learn a grammar model from all observed sequences by grammar induction tools. Such an algorithm learns the And-node and Or-nodes by generating significant patterns and equivalent classes which are basic approaches for general grammar induction problems.

Parameter learning over the learned grammar structure is obtained from maximum likelihood estimation (MLE). The optimal branching probabilities of Or-Nodes is simply given frequency of each alternative choice [154]. This is also the standard protocol of other grammar induction problems. We show an example grammar learned from these methods in Fig. 9.12.

Given this learned grammar model $\mathcal{G}$, we can conduct probabilistic inference for the most probable task that one agent is trying to work on. With spatial features and relationships aggregated

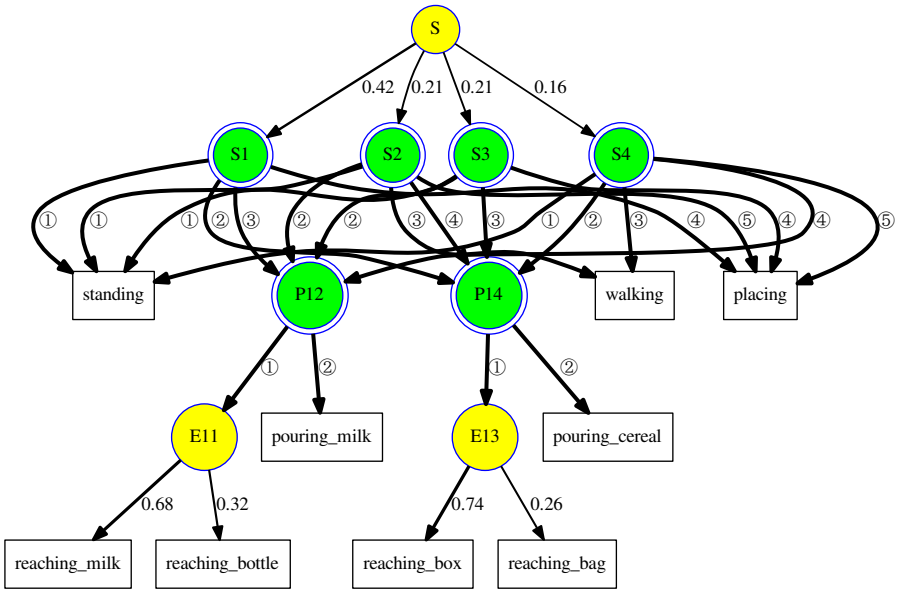


Figure 9.12: An example of a temporal grammar. The green and yellow nodes are And-nodes and Or-nodes respectively. The numbers on branching edges of Or-nodes represent the branching probability. The circled numbers on edges of And-nodes indicates the temporal order of expansion.

into Γ, the essential problem becomes:

$$\begin{aligned} PG^* &= \arg\max_{PG} p(PG|\Gamma, \mathcal{G}) \\ &\propto \arg\max_{PG} p(\Gamma|PG)p(PG|\mathcal{G}) \end{aligned} \tag{9.9}$$

which is an maximum a posteri inference for the most probable parse graph PG. The grammar prior of parse graph $p(PG|\mathcal{G})$ follows from the parsing probability of parse graph PG given grammar $\mathcal{G}$. Note that there are practical computational issues for enumerating possible PGs and also computing the probability, several methods was proposed with special focus on the computational efficiency of models, *e.g.*, Generalized Earley Parser(GEP) [539]. We skip the details of such methods as it is not our primary focus and please refer to Book II where we discussed the computational concerns in detail. An example of the step-wise parsing result is shown in Fig. 9.13.

The remaining question lies in making proper predictions about future states. Given the current parsing result PG_t of the observed video sequence, the STC-AoG could be used to predict the next sub-activity, action, which object the subject is going to interact with, and how the subject will interact with the object. Similarly, the problem of making the next action becomes

$$\begin{aligned} a^* &= \arg\max_{a \in A} p(a|G, PG_t) \\ &\propto \arg\max_{a \in A} \sum_{\substack{PG_{t+1} \\ l_{t+1}=[l_t,a]}} p(PG_{t+1}|G, PG_t) \end{aligned} \tag{9.10}$$

where l_t denotes the derived sentence of parse graph PG_t. This is done by trying all available future parse graphs PG_{t+1} that have a as the last token in the derived sentence l_{t+1} which concatenate on

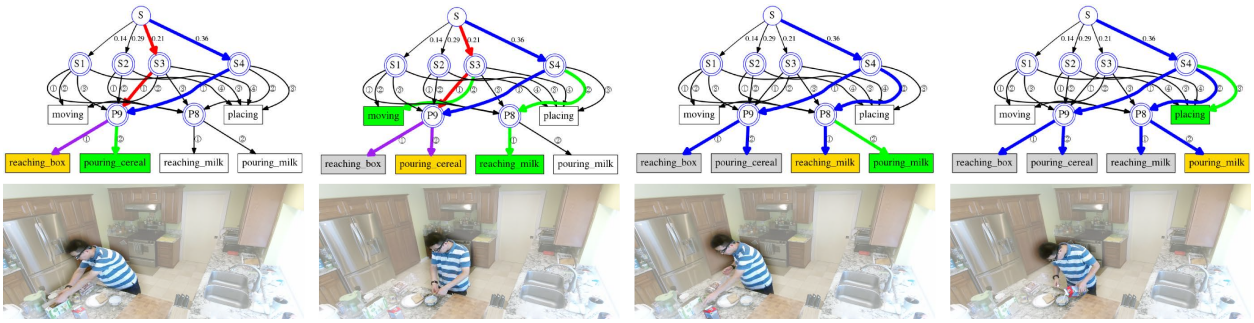


Figure 9.13: A simplified example illustrating the parsing and symbolic prediction process. In the first two figures, the red edges and blue edges indicates two different parse graphs for the past observations. The purple edges indicate the overlap of the two possible explanations. The red parse graph is eliminated from the third figure. For the terminal nodes, yellow indicates the current observation and green indicates the next possible state(s).

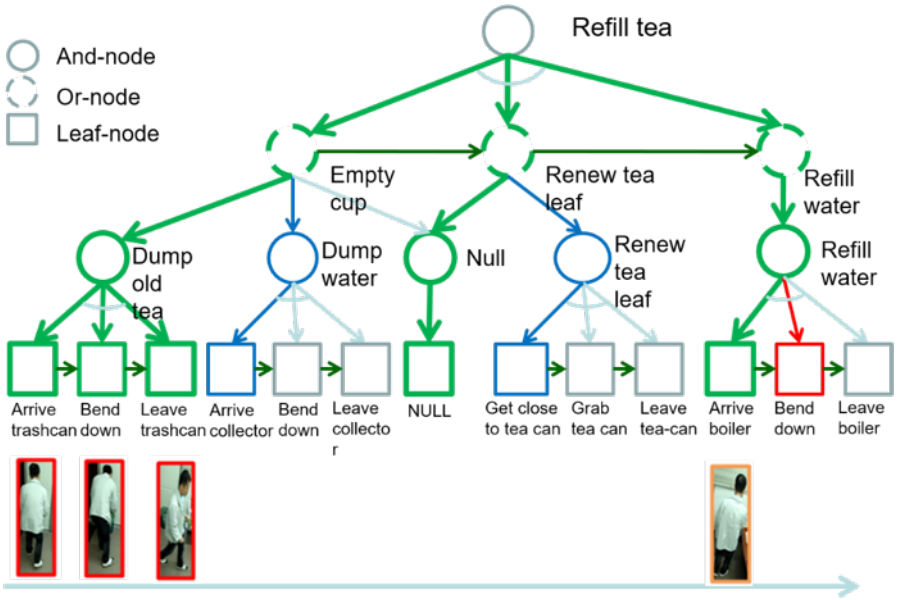


Figure 9.14: An example scenario of "refilling tea" in the office. Nodes in the figure corresponds to grammar terminals/non-terminals. Lines and nodes highlighted in green indicate the current parse graph PG_t. Lines and nodes colored in blue refers to possible alternatives parses for PG_t. Lines and nodes in color red show the grammatical correct next action, which is the only proper next action in this case, "blend down."

l_t with action a. With this approximation, the problem is simplified into calculating the grammar transition probability from partial parse graph PG_t to partial parse graph PG_{t+1}. As the grammar transition probability will be 0 for sequences that is not derivable by the grammar, only grammatical correct next actions will be proposed as shown in Fig. 9.14. Please also refer to Fig. 9.13 for a real example obtained from current method's computation.

9.5 Discussion

In Section 9.4, we briefly went through an example which takes temporal AoG as the primary entry point for learning event representations and making inferences. As we mentioned in Section 9.2, knowledge of events covers both spatial, temporal, and also causal aspects for description purposes.

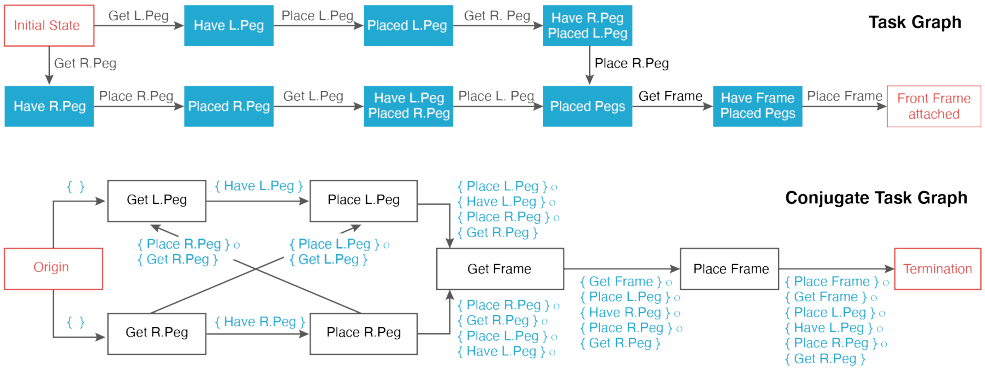


Figure 9.15: The task graph and conjugate task graph representation for the same task of assembling furniture: (Left) the task graph of assembling furniture where furniture status is modeled with nodes with actions modeled as edges between different nodes. (Right) the conjugate task graph of assembling furniture where actions are treated as nodes and states are modeled as edge transitions.

The missing causal flavor in the previous example is largely due to conjugation of states and actions. With temporal AoG focusing on the description of actions applied to object states, causal AoG focus more on the understanding of object fluent change chains. Similar to the task-graph and conjugate task graph representation, see in Fig. 9.15 commonly studied in robotics literature [540], temporal and causal AoGs share a similar conjugation relationship.

The critical debate between the two types of representations lie in the competition between efficiency and generalizability. As actions are already abstractions of object fluent changes, using the sequence of action to describe an event is definitely more efficient when compared to using object fluent change sequences. However, the limitation of following action patterns often produce the gap for models' performance between seen tasks and unseen tasks. With less knowledge of how fluents changed and the details of the goal states, following action patterns or routines naturally harm models' capability to perform similar inference in new environments. In contrast, using object fluents change sequence allow us to understand more about the details of goal statuses and values of each world state, which further helps generalization in new tasks in the same environment. Therefore, to balance the trade-off between efficiency and generalizability, we emphasize that both temporal and causal ingredients should be captured for a good and efficient model.

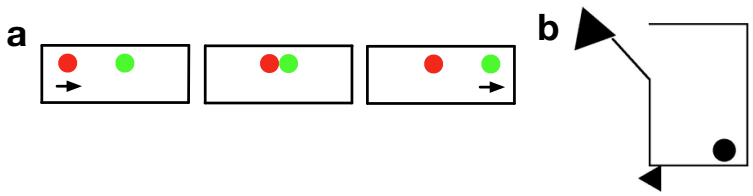


Figure 10.1: (a) A classic launching stimulus [99]. (b) A screenshot of the original Heider-Simmel movie [530].

Chapter 10

Animacy: Physical vs. Social Perception

In this chapter, we introduce how we can build computational model to study human perception of animacy in a unified framework.

10.1 Introduction

10.1.1 Background

Imagine you are playing a multi-player video game with open or free-roaming worlds. You will encounter many physical events, such as blocks collapsing onto the ground, as well as social events, such as avatars constructing buildings or fighting each other. All these physical and social events are depicted by movements of simple geometric shapes, which suffice to generate a vivid perception of rich behavioral, including interactions between physical entities, interpersonal activities between avatars engaged in social interactions, or actions involving both humans and objects.

This type of rich perception elicited by movements within simple visual displays has been extensively studied in psychology. On the one hand, classic work such as Michotte [99] has famously shown that people can perceive physical causality from a simple animation depicting a moving ball colliding with a stationary ball, which then appears to launch and move off (Fig. 10.1a). On the other hand, the motion of similar geometric shapes may generate an impression of agency. This phenomenon is termed as the perception of animacy [301]. For instance, the seminal work of Heider and Simmel [530] demonstrated that people also have a spontaneous perception of animacy when viewing simple geometric shapes moving around, where they described the shapes as characters with minds, personalities and have different relationships with one another (Fig. 10.1b).

Despite the clear evidence of people's strong abilities to perceive physical causality and social behaviors from limited and abstract inputs, it is, however, still unclear how these two types of

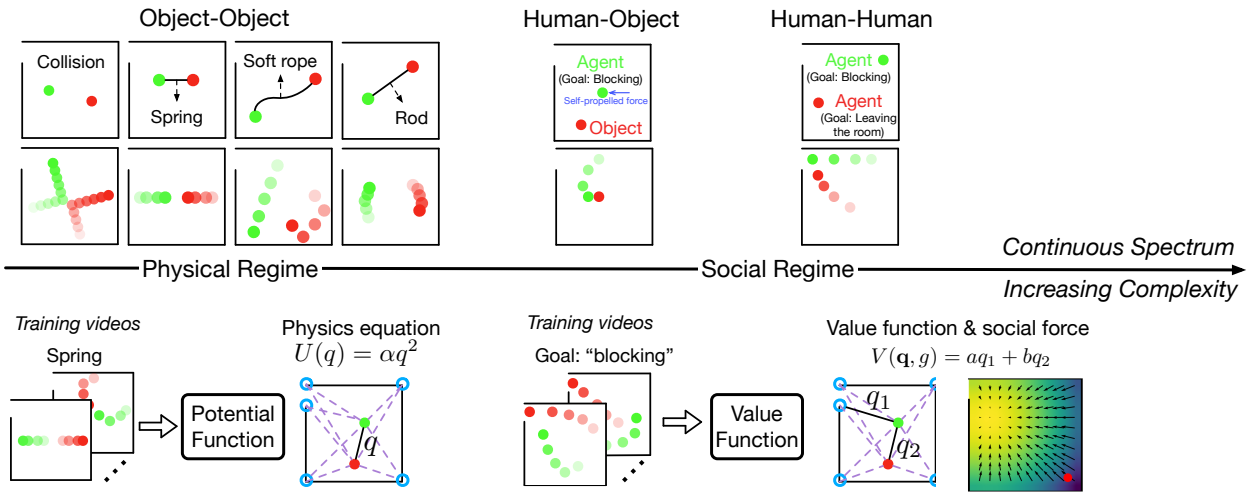


Figure 10.2: A continuous spectrum connecting physical systems and social behaviors. In this chapter, we focus on three types of interactions, human-human (HH), human-object (HO) and object-object (OO). A few examples are included by showing trajectories of the two entities. The dot intensities change from low to high to denote elapsed time. Accordingly, we will learn an increasingly complex model that include potential functions representing physical laws for inanimate objects as well as value functions representing social behaviors of human agents.

perception are connected. For a long time, researchers have been approaching these two domains separately. In the case of physical events, research has been focused on the perception and interpretation of physical objects and their dynamics, aiming to determine whether humans use heuristics or mental simulation to reason about intuitive physics (see a recent review by [190]). For social perception, some research has aimed to identify critical cues based on motion trajectories that determine the perception of animacy and social interactions [541, 301, 542, 543, 132]. There has also been work focusing on inferences about agents' intentions [534, 544, 545].

10.1.2 A Continuous Spectrum from Physics to Social Behaviors

In this chapter, we will present a unified view of these two domains. As illustrated in Fig. 10.2, the physical systems and social behaviors lie on a continuous spectrum ranging from the physical regime to the social regime. Given this unified view of physical and social perception, we introduce a unified computational framework for modeling both physical events and social events based on the movements of simple shapes. In particular, this framework that unifies the physical and social modeling in three ways.

First, a unified physical and social simulation for generating Heider-Simmel animations in which simple moving shapes vary in degrees of physical violation and the involvement of intention.

Second, a unified physical and social concept learning paradigm by formulating the concept learning process as the pursuit of generalized coordinates and the corresponding parsimonious potential energy functions.

Third, a unified psychological space that may reveal the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents.

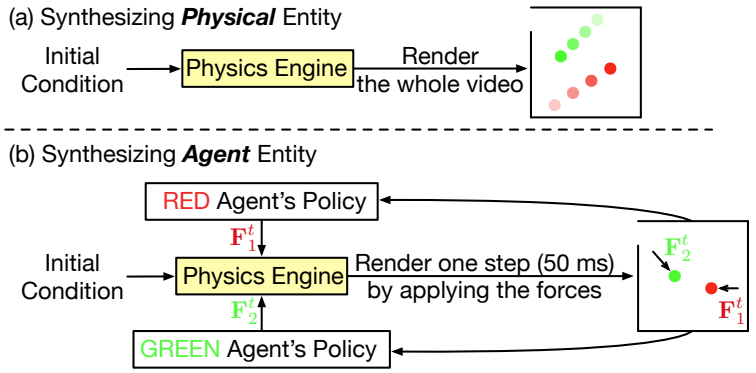


Figure 10.3: Overview of the joint physical-social simulation engine. For a dot instantiating a physical object, we randomly assign its initial position and velocity and then use physics engine to simulate its movements. For a dot instantiating a human agent, we use policies learned by deep reinforcement learning to guide the forces provided to the physics engine.

10.2 Heider-Simmel-type Animations in the Continuous Spectrum

Can we have a unified view of physical events with inanimate objects and social events with human agents? Can we create a continuous transition from objects to agents, and from agents back to objects? In other words, can we bridge physics and social behaviors? We believe that the first step towards addressing these questions should be building a simulation engine that can generate both physical interactions and social interactions in a principled manner, so that the two types of interactions can emerge in the same world.

Fig. 10.3 gives an overview of a joint physical-social simulation engine. Each video included two dots (red and green) and a box with a small gap indicating a room with a door. The movements of the two dots were rendered by a 2D physics engine (pybox2d [1]). If a dot represents an object, we randomly assigned the initial position and velocity, and then used the physics engine to synthesize its motion. Note that our simulation incorporated the environmental constraints (*e.g.*, a dot can bounce off the wall, the edge of the box), but did not include friction. If a dot represents an agent, it was assigned with a clearly-defined goal (*e.g.*, leaving room) and pursued its goal by exerting self-propelled forces (*e.g.*, pushing itself towards the door).

10.2.1 Interaction Types

As summarized in Fig. 10.2, there are three types of interactions, including human-human (HH), human-object (HO) and object-object (OO) interactions. When synthesizing the agents' motion, we set two types of goals for the agents, *i.e.*, "leave the room" (g_1) and "block the other entity" (g_2).

In addition to the three general types of interactions,there are also sub-categories of interactions to capture a variety of physical and social events. For OO animations, theere are four events – collision, connections with rod, spring and soft rope. For HH animations, we varied the "animacy degree" (AD) of the agents by controlling how often they exerted self-propelled forces in the animation. In general, a higher degree of animacy associates with more frequent observations about violation of physics, thus revealing self-controlled behaviors guided by the intention of an agent. The animacy manipulation introduced five sub-categories of HH stimuli with five degrees of animacy—7%, 10%, 20%, 50%, and 100%.

[1] https://github.com/pybox2d/pybox2d

10.2.2 Unified Physical and Social Concept Learning via Potential and Value Functions

As illustrated in Fig. 10.2, we intend to acquire physical and social concept by progressively increasing the complexity of a unified model. Specifically, in the social regime, we will learn potential functions representing physical laws. However, they would not be able to adequately represent agent behaviors exhibited in the social regime, for which we will learn value functions that capture the key social concepts that can interpret the intentional movements (or plans) of social agents.

Potential Functions for Physical Systems

To introduce the basic idea of using potential functions as a type of representation for physical systems, let us first look at the comparison between Lagrangian mechanics (based on potential energy) and Newtonian mechanics (direct force analysis).

Consider a system of N particles with the same mass (*i.e.*, $m_i = m$, $\forall i = 1, \cdots, N$) where their positions are $(\mathbf{x}_1(t), \mathbf{x}_2(t), \cdots, \mathbf{x}_N(t))$ in Cartesian coordinates at time t. The surrounding environment (context) is denoted as c. The Lagrangian of this system is defined as

$$L = L(\mathbf{x}_1, \cdots, \mathbf{x}_N, \dot{\mathbf{x}}_1, \cdots, \dot{\mathbf{x}}_N, t) = T - U, \tag{10.1}$$

where $T = T(\dot{\mathbf{x}}_1, \cdots, \dot{\mathbf{x}}_N, t) = \sum_{i=1}^{N} \frac{1}{2} m \dot{\mathbf{x}}_i(t)^2$ is the kinetic energy of all entities and U is the potential energy. When there are only conservative forces in the system, the potential energy solely depends on the coordinates of the entities, *i.e.*, $U = U(\mathbf{x}_1, \cdots, \mathbf{x}_N, t)$. For convenience, we may drop the notation t sometimes.

From the Euler-Lagrange equation, we may derive the motion of equations for each entity:

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\mathbf{x}}_i} - \frac{\partial L}{\partial \dot{\mathbf{x}}_i} = 0, \quad \forall i = 1, \cdots, N. \tag{10.2}$$

By plugging in T and U, this in fact gives us Newton's second law:

$$m\dot{\mathbf{x}}_i = \mathbf{F}_i = -\frac{\partial U(\mathbf{x}_1, \cdots, \mathbf{x}_N)}{\partial \mathbf{x}_i}. \tag{10.3}$$

This implies that as an alternative approach to conducting explicit force analysis which is often extremely difficult in complex systems, we can instead derive forces from a few scalar functions, *i.e.*, potential energy functions. This advantage becomes more significant when we adopt suitable generalized coordinates which constitutes potential energy functions in simple forms.

Formally, we may convert the Cartesian coordinates of the N entities into a generalized coordinate system $\mathbf{q} = (q_j)_{j=1}^{D}$, where D is usually the number of degrees of freedom in the system. Each dimension is derived from a transformation function $q_j = \phi_j(\mathbf{x}_1, \cdots, \mathbf{x}_N, c)$, where c is the context (*e.g.*, surrounding environment) of the current system. These coordinates' first-order derivatives $\dot{\mathbf{q}} = (\dot{q}_j)_{j=1}^{D}$ become generalized velocities accordingly. Here, ϕ_j could be understood as a type of state representation extracted from the raw observations. Based on the generalized coordinates, we can redefine the Lagrangian:

$$L = L(\mathbf{q}, \dot{\mathbf{q}}) = T - U, \tag{10.4}$$

where $T = T(\mathbf{q}, \dot{\mathbf{q}}) = T(\dot{\mathbf{x}}_1, \cdots, \dot{\mathbf{x}}_N)$ is the kinetic energy, and V is the potential energy. Again, if we only consider conservative forces, we will have $U = U(\mathbf{q})$. The Euler-Lagrange equation still holds for the generalized coordinates:

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = 0, \quad \forall j = 1, \cdots, D. \tag{10.5}$$

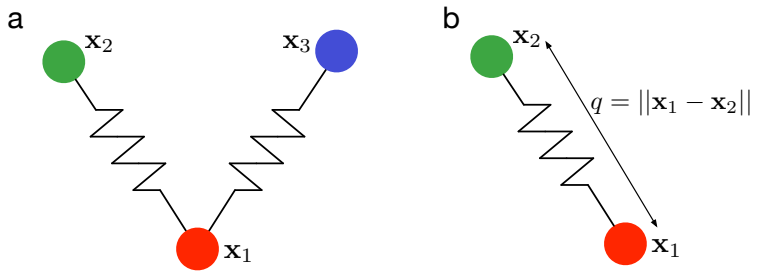


Figure 10.4: Systems with circles and springs. (a) Two entities (circles) connected by a massless spring. The Cartesian coordinates of the two entities are $\mathbf{x}_1$ and $\mathbf{x}_2$. The potential energy of this system can be defined by using just one variable, *i.e.*, the distance between the two entities. (b) Three entities connected by two massless springs.

The resulting equations of motion describe the dynamics of the system as a whole in terms of how generalized coordinates (*i.e.*, the physical quantities of interest) change over time. We can map the motion back to individual entity's Cartesian coordinates based on the transformation functions ϕ_j:

$$m\dot{\mathbf{x}}_i = \mathbf{F}_i = -\sum_{j=1}^{D} \frac{\partial U(\mathbf{q})}{\partial q_j} \frac{\partial \phi_j}{\partial \mathbf{x}_i} \quad \forall i = 1, \cdots, N. \tag{10.6}$$

The use of generalized coordinates allows us to greatly simplify the derivation of forces (or dynamics) for entities in a system, which ultimately results in a parsimonious model to describe the dynamics of a system. Therefore, by constructing the most suitable generalized coordinates, the key characteristics of a system may naturally emerge from raw observations. Consider the spring system shown in Fig. 10.4a as an example. Assume the equilibrium length of the spring is l and its constant is k, then potential energy of this system can be conveniently defined by only one variable – the distance between the two entities (or equivalently the length of the spring). Let $q = \phi(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||$, the potential is $U(q) = \frac{1}{2}k(q-l)^2$, which is a simple quadratic function of q. Based on Eq. (10.6), we can derive the forces applied to the two entities accordingly: $\mathbf{F}_1 = -k(q-l)(\mathbf{x}_1 - \mathbf{x}_2)/q$, $\mathbf{F}_2 = -k(q-l)(\mathbf{x}_2 - \mathbf{x}_1)/q$.

Multiple independent potential energy functions may coexist in a complex system, and the overall potential energy is simply the sum of all individual potential energy functions. This naturally leads to a modular design, where the potential energy of any system is a combination of atomic potential energy functions as bases. For instance, in Fig. 10.4b, by defining generalized coordinates $q_1 = ||\mathbf{x}_1 - \mathbf{x}_2||$ and $q_2 = ||\mathbf{x}_1 - \mathbf{x}_3||$, the overall potential energy can be decomposed into two functions associated with the two springs: $U(\mathbf{q}) = U_1(q_1) + U_2(q_2)$. If the two springs have the same property, then the potential energy can be further simplified by reusing the same atomic function: $U(\mathbf{q}) = U(q_1) + U(q_2)$.

To enforce sparsity, we assume a polynomial form for each potential function. Specifically, we consider a potential function such as $U_j(q_j) = \mathbf{w}_j^\top [q_j, q_j^2]$, where $\mathbf{w}_j$ are parameters of the polynomial function.

When we have multiple atomic potential energy functions in a system, it is often important to identify when each function will be present or effective in terms of yielding forces to the entities. Some potential energy functions like the ones in Fig. 10.4 are always effective. But there are also functions with limited effective spatial ranges. For instance, to approximate the force an entity receives when bouncing off a wall (here we assume perfectly elastic collision) as shown in Fig. 10.5, we can imagine that when the entity is expected to violate the non-overlapping constraint (the distance between the entity and the wall can not be smaller than a threshold) in a very short

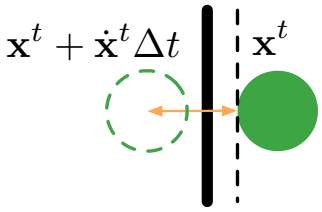


Figure 10.5: A circle bouncing off a wall. The generalized coordinate in this case can be derived as the expected violation after a short period of time Δt based on the entity's current position $\mathbf{x}^t$ and velocity $\dot{\mathbf{x}}^t$.

period of time (Δt) based on its current position and velocity, there will be an effective potential energy function applied to the entity. In fact, this potential can be approximated by a spring (with a very large constant $k \gg 1$ and a equilibrium length of distance threshold) connecting the contact point and the entity. This type of approximation has been previously introduced in robotics literature as well [546].

If we denote $\delta_j(q_j)$ to be the triggering condition function, then we may define the complete potential energy as

$$U(\mathbf{q}) = \sum_{j=1}^{D} \delta_j(q_j) U_j(q_j). \tag{10.7}$$

Value Functions for Social Behaviors

In the joint simulation engine, everything is generated in a physics engine. It is natural to derive the generalized coordinates and the corresponding potentials regardless of whether an entity is an object or an agent. Consequently, similar ideas discussed for modeling physical systems may also be applied to modeling the goals and relations in social behaviors as illustrated in Fig. 10.6a. Suppose an agent with free will can exert self-propelled forces to purse its goal. Then its plan or policy w.r.t. a certain goal can be represented as the force exerted by itself given its current state and the context. By assuming rationality of the agent's plan, the force should be explained by certain potential function associated with its goal and its relations with the environment and other agents, which can be seen as a form of value function defined on semantically meaningful measurements (*i.e.*, generalized coordinates) such as the distance between its current position and its goal position, or the relative spatial displacement between itself and other agents. By seeking the simplest generalized coordinates and the corresponding sparse functions of potential energy, important concepts in social behaviors, such as goals and relations could naturally emerge as well.

With this analogy, the Cartesian coordinates $(\mathbf{x}_i)_{i=1}^N$ coupled with the context c are the states of the agents, and the generalized coordinates q_j are equivalent to the sufficient statistics in describing the observed social scenario. Let the agents' goals be $g_i \in \mathcal{G}$, where $\mathcal{G}$ is a set of all possible goals, then an agent's behavior is guided by a potential energy function defined in Cartesian coordinates, *i.e.*, $U_i(\mathbf{x}_1, \cdots, \mathbf{x}_N, G, c)$. We then use a potential energy function defined in generalized coordinates to equivalently represent the goal directed value function for agent i as follows

$$V_i(\mathbf{q}, g_i) = -U_i(\mathbf{x}_1, \cdots, \mathbf{x}_N, G, c). \tag{10.8}$$

Let $X = \{\mathbf{x}_i\}_{i=1}^N$. The plan of agent i can be derived in a step-by-step manner by Eq. (10.6), *i.e.*,

$$\mathbf{F}_i(\mathbf{x}_i | X_{-i}, g_i) = \sum_{j=1}^{D} \frac{\partial V(\mathbf{q}, g_i)}{\partial q_j} \frac{\partial \phi_j}{\partial \mathbf{x}_i}, \quad \forall i = 1, \cdots, N. \tag{10.9}$$

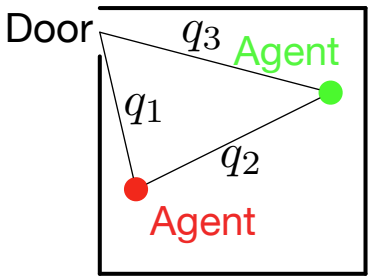


Figure 10.6: Illustration of social concepts as generalized coordinates. (a) An example of generalized coordinates in social systems. The (q_1, q_2, q_3) here are potentially the most critical variables in describing this social system. q_1 and q_3 here reveal the potential goal (*i.e.*, the door) for both agents, so an attraction potential term could explain the behavior of "leaving the room." q_2 can represent the relation between the agents. *e.g.*, the "chasing" behavior could be modeled by a potential term that only depends on q_2. (b) The generalization of (a) where the generalized coordinates and the potential energy function can be preserved; we only need to modify the transformation from raw observations to the generalized coordinates.

For instance, in Fig. 10.6a, if the red agent tries to leave the room, then its motion will be driven by potential $V(q_1)$. Similarly, if the green agent aims to catch the red agent, then it is driven by a potential $V(q_2)$.

Thus, learning sparse value functions through generalized coordinates takes a straightforward approach in explaining the rational behaviors demonstrated by the agents since it allows us to derive the optimal policy directly from the inferred value of states in addition to discovering the goals. This method may also help us discover sub-goals (*i.e.*, different value function terms) in the optimal plans. Finally, the explicit modeling of generalized coordinates can potentially improve the generalization of the learned optimal plans as well since we can simply remap any new environment to the same coordinate system by only changing $\phi_j(\cdot)$; the previously learned value functions and the corresponding optimal plans can be preserved. For instance, the generalized coordinates and value functions constructed based on the environment in Fig. 10.6a can be transferred to the new scenario in Fig. 10.6b where the new position of the door will only affect the coordinate transformation for q_1 and q_3.

We summarize the main advantages of constructing generalized coordinates and the corresponding potential energy functions as follows:

- **Generalized coordinates as effective representations of a system**. The change in $\mathbf{q}$ are the effective change of a system, *i.e.*, $\partial U(\mathbf{q})/\partial \mathbf{q}$. By pursing the coordinates that results in the simplest $U(\mathbf{q})$, we are essentially pursuing a sparse model for the system. For physical systems, such representations will reveal physical concepts, whereas in social systems, they may denote important concepts of goals and social relations.
- **"Compression" of optimal planning**. Optimal planning is complex and time consuming. However, given demonstrations (observed trajectories of agents), we may compress these optimal plans into a few value functions. Consequently, instead of searching for an optimal plan from scratch every time, we may derive forces from the value functions and roll out the whole plan step-by-step starting from the initial state. We may deploy this plan directly, or use it as a starting point and further refine it to compensate the errors in the learned value functions. Similarly, we can also take advantage of the derived forces to conduct inverse planning for Bayesian goal inference.
- **Knowledge transfer**. When the surrounding environment changes, the value function defined on generalized coordinates, $V(\mathbf{q})$, may be preserved. In order to derive forces for the entities in the new environment, we only need to change the coordinate transformations, *i.e.*, $q_j =$

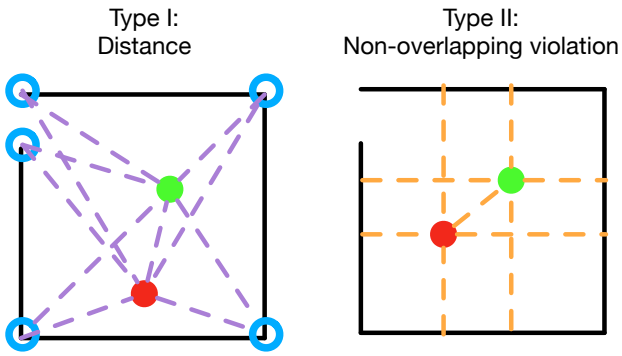


Figure 10.7: Two types of candidates of generalized coordinates shown as the purple and orange dashed lines respectively. The blue circles highlight the reference points used for extracting the first type of candidate coordinates.

$\phi_j(\mathbf{x}_1, \cdots, \mathbf{x}_N, c)$.

A Sketch of the Learning Algorithm

Problem setup. In an N-entity system, we may observe the context (environment) c, and the trajectories of all entities $\Gamma_i = \{(\mathbf{x}_i^t, \dot{\mathbf{x}}_i^t)\}_{t=1}^T$, where the length of each step is Δt, and the total length is $T\Delta t$. We assume that all entities have the same mass m and there are only conservative forces in the system. From the trajectories, we may also compute the ground-truth force each agent i receives at time step t, *i.e.*, $\mathbf{F}_i^t$. The goal is to learn a model (generalized coordinates and potential energy functions) which can predict the forces given the observations.

Proposals of generalized coordinates. From bottom-up proposals, we obtain a pool of candidates for generalized coordinates, $\mathbb{Q} = \{q_j\}_{j=1}^D$. Note that many of them may be redundant and will not be selected by the final model. In particular, these candidates can arise from two types of proposals:

i) Distance between two geometric shapes. As shown in Fig. 10.7, this can be the distance between two entities (*e.g.*, the one in Fig. 10.4) or the distance between an entity and a part of the context (*e.g.*, the one in Fig. 10.5). The corresponding potential energy functions are always triggered, *i.e.*, $\delta_j(q_j) = 1$.

ii) Expected constraint violation as illustrated in Fig. 10.5. When there is violation, q_j represents the expected overlapped length; otherwise $q_j = 0$. The triggering condition is consequently defined as $\delta_j(q_j) = \mathbb{1}(q_j > 0)$.

Note that for social behaviors, we do not consider the second type of the generalized coordinates.

Pursuing a set of atomic potential energy functions. The final potential energy function consists of a set of atomic potential energy functions, each of which is defined as $U_k(q_k)$, $k \in \mathbb{S} \subset \mathbb{Q}$, where $\mathbb{S}$ is a set of generalized coordinates selected from the candidate pool $\mathbb{Q}$. The final potential energy will be used for predicting the forces for each entity:

$$\hat{\mathbf{F}}_i^t = -\sum_{k \in \mathbb{S}} \delta_k(q_k^t) \frac{\partial U_k(q_k^t)}{\partial q_k^t} \frac{\partial \phi_k^t}{\partial \mathbf{x}_i^t}. \tag{10.10}$$

Finally, we define an MSE loss for the force prediction as the learning objective function:

$$L(\mathbb{S}, \Omega = (w_k)_{k=1}^K) = \mathbb{E}\left[\frac{1}{2}||\mathbf{F}_i^t - \hat{\mathbf{F}}_i^t||_2^2\right]. \tag{10.11}$$

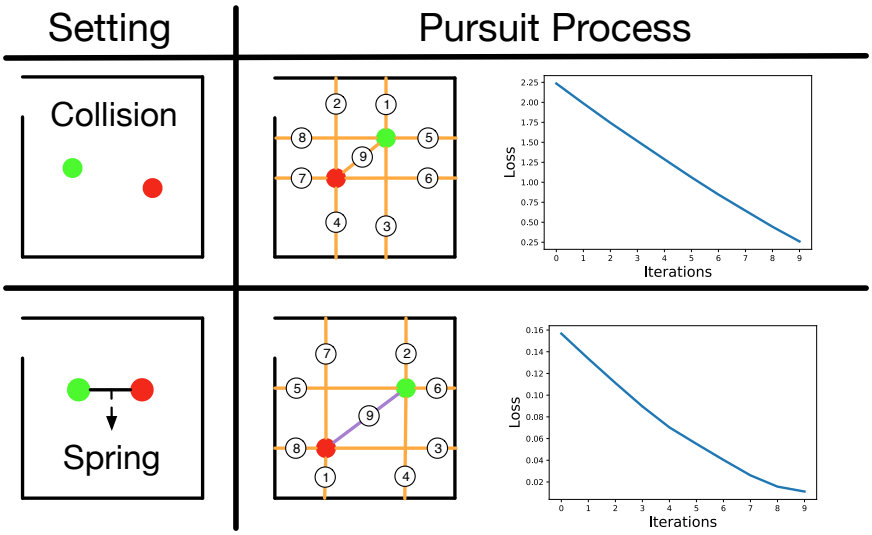


Figure 10.8: Learning process of two physical systems. The purple and orange lines are the selected generalized coordinates from the first and the second type of candidates respectively; each number indicates the iteration when the corresponding generalized coordinate was selected.

The pursuit of the final model is essentially the search of the optimal generalized coordinates $\mathcal{S}$ and the parameters Ω of the corresponding potential energy functions that minimize the above loss (along with some regularization for sparsity). For computational efficiency, we adopt a greedy pursuit, where we start from an empty set of generalized coordinates, then at each iteration, we augment the final model with the candidate generalized coordinate that has not yet been selected in previous iterations and yields a fitted potential energy function with the largest loss reduction. The iterative pursuit is repeated until there is no significant loss reduction anymore.

Learning Results

We generated collision and spring (with several different spring lengths) physical systems shown in Fig. 10.2, each had 50 videos as training examples. Fig. 10.8 shows the learning process of two systems.

We also used the same approach to pursue value functions for two goals depicted in Fig. 10.2 for HH videos. In practice, we used 50 videos of an agent fleeing the room successfully to learn the potential energy functions for the goal of "leaving the room," and used another 50 videos of an agent successfully blocking another agent or attempting to block it without success for the goal of "blocking." Fig. 10.9 shows generalized coordinates and the derived forces fields based on the learned model for both goals. We find that using Lasso can help discover more meaningful goal-directed potentials for social behaviors by enforcing sparsity for the potential energy function of each generalized coordinate.

Physics Inference

By giving the positions and velocities of the two entities at time t, *i.e.*, $\mathbf{x}_i^t$, $\dot{\mathbf{x}}_i^t$, $i = 1, 2$, we can predict the physical forces each entity receives at t and consequently their future velocities at $t+1$, $\hat{\dot{\mathbf{x}}}_i^{t+1}$, $i = 1, 2$. By comparing with the ground truth $\dot{\mathbf{x}}_i^{t+1}$, we can evaluate to what degree an entity's

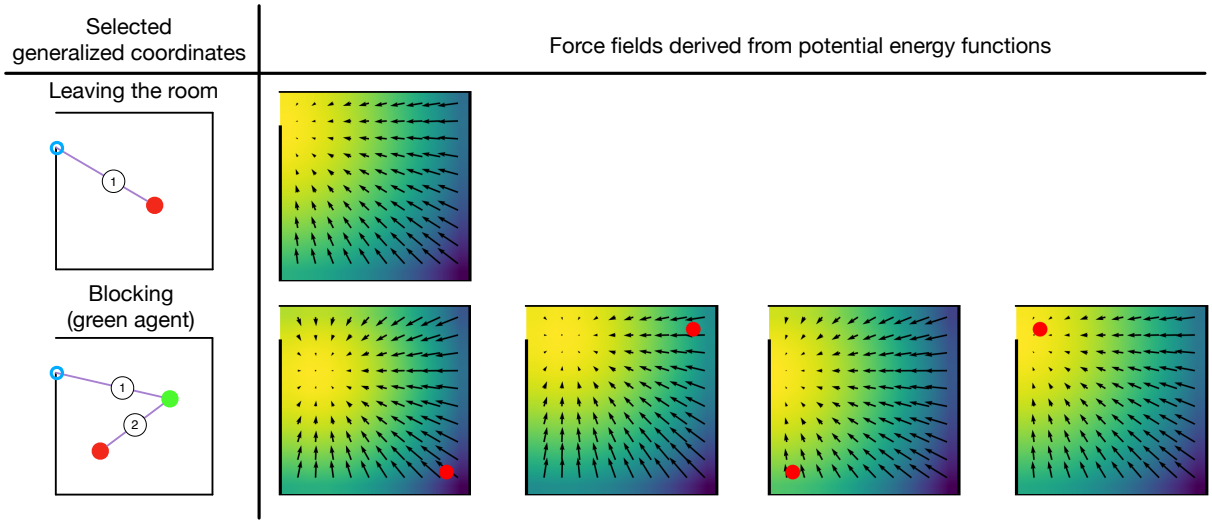


Figure 10.9: Learning results of two goals. Left: selected generalized coordinates; right: the learned value functions and force fields derived from the value functions, where the red circle represents the position of the other agent, and the color of the background indicate the value of a state (blue to yellow indicates low value to high value). An agent will move towards high value positions and move away from low value positions.

motion is inconsistent with physics predictions:

$$\mathcal{D}_i = \frac{1}{T}\sum_{t=1}^{T} ||\dot{\mathbf{x}}_i^t - \hat{\dot{\mathbf{x}}}_i^t||_2^2, \quad \forall i = 1, 2. \tag{10.12}$$

In practice, there are multiple physical systems, each of which will give different predictions. Since we do not know which system an observation belongs to, we can enumerate all learned physical systems and select the one that yields the lowest prediction error, which we may use as the physical violation measurement.

Intention Inference

The force fields illustrated in Fig. 10.9 give us the expected moving direction at each location given the goal of the agent and the position of the other agent. Inspired by the classic FRAME model [4, 547] which was originally used for modeling texture and natural images, we may treat a field derived from the learned model as filters of motion for a given goal at different locations. The basic idea is illustrated in Fig. 10.10. Specifically, the filter response at location $\mathbf{x}_i$ for agent i with goal g_i and the other agent being at $\mathbf{x}_j$ can be defined as

$$h(\dot{\mathbf{x}}_i|\mathbf{x}_i, \mathbf{x}_j, g_i) = cos(\theta) = \frac{\hat{\mathbf{F}}_i(\mathbf{x}_i|\mathbf{x}_j, g_i)^\top \dot{\mathbf{x}}_i}{||\hat{\mathbf{F}}_i(\mathbf{x}_i|\mathbf{x}_j, g_i)|| \cdot ||\dot{\mathbf{x}}_i||}, \tag{10.13}$$

where θ is the angle between the observed moving direction $\dot{\mathbf{x}}_i$ and the expected moving direction from the predicted force $\hat{\mathbf{F}}_i$ in Eq. (10.9). By dividing the whole space into R discrete regions ($R = 4$ in this work), where each region has a location set $\mathbb{X}_r$, we can define the likelihood of observing an agent with a goal having a certain trajectory Γ_i as

$$p(\Gamma_i|g_i, \Gamma_j) = \frac{1}{Z(\Lambda)} \exp\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R} \mathbb{1}(\mathbf{x}_i^t \in \mathbb{X}_r)\lambda_r h(\dot{\mathbf{x}}_i^t|\mathbf{x}_i^t, \mathbf{x}_j^t, g_i)\right\} q(\Gamma), \tag{10.14}$$

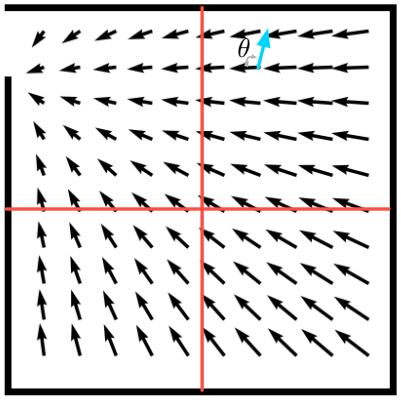


Figure 10.10: Illustration of the idea of motion filters. Suppose the blue arrow is the observed velocity of an agent at a given moment, then we may use the angle θ between to measure the fitness of the observed motion and the expected goal-directed motion (*i.e.*, using $cos(\theta)$ as the filter response). We divide the space into four regions to compute the likelihood of an agent is pursuing a specific goal.

where $q(\Gamma_i) = \prod_{t=1}^{T} q(\dot{\mathbf{x}}_i^t)$ is a background model for all moving directions without pursuing a specific goal (we assume a uniform distribution for $q(\dot{\mathbf{x}}_i^t)$), $\Lambda = (\lambda_1, \cdots, \lambda_R)$ is the parameter for the likelihood corresponding to the R regions, and $Z(\Lambda)$ is the normalization term. We may write $Z(\Lambda)$ as

$$Z(\Lambda) = E_{q(\Gamma)}\left[\exp\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R}\mathbb{1}(\mathbf{x}_i^t \in \mathbb{X}_r)\lambda_r h(\dot{\mathbf{x}}_i^t|\mathbf{x}_i^t, \mathbf{x}_j^t, g_i)\right\}\right]. \tag{10.15}$$

Since we assume a uniform distribution for the background velocity, it is easy to show that $Z(\Lambda) = 1$. Then parameter λ_r in the likelihood can be estimated as the every filter responses of trajectories in training examples in region r. Finally, we define the intention measurement as the log-likelihood ratio of a trajectory following the optimal plan for pursuing *any* goal over the background trajectory model:

$$\mathcal{L}_i = \max_{g\in\mathcal{G}} \log p(\Gamma_i|g, \Gamma_j) - \log q(\Gamma_i), \quad \forall i = 1, 2. \tag{10.16}$$

10.3 Human Experiment

To test how well the computational model can explain human perception of physical and social events, we conducted the following human experiment.

10.3.1 Participants

30 participants (mean age = 20.9; 19 female) were recruited from UCLA Psychology Department Subject Pool. All participants had normal or corrected-to-normal vision. Participants provided written consent via a preliminary online survey in accordance with the UCLA Institutional Review Board and were compensated with course credit.

10.3.2 Stimuli and Procedure

850 videos of Heider-Simmel animations were generated from the simulation engine, with 500 HH videos (100 videos for each AD level), 150 HO videos, and 200 OO videos (50 videos for each sub-category). Videos lasted from 1 s to 1.5 s with a frame rate of 20 fps. By setting appropriate initial

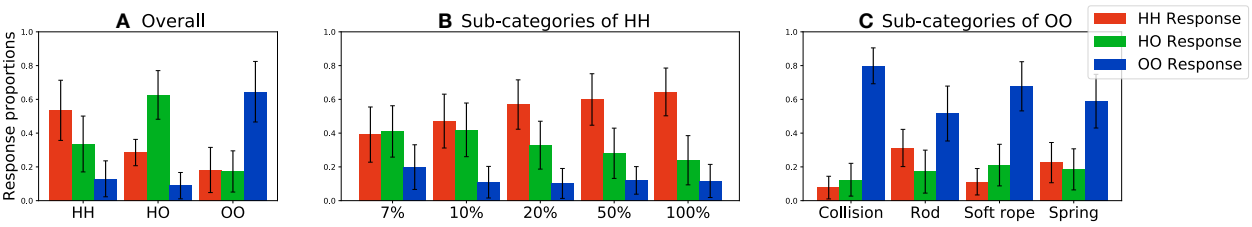


Figure 10.11: Human response proportions of interaction categories (a) and of the sub-categories (b,c) in the experiment. Error bars indicate the standard deviations across stimuli.

velocities, the average speeds of dots in OO videos were controlled to be the same as the average speeds of dots in HH with 100% ADs (44 pixel/s). The dataset was split into two equal sets; each contained 250 HH, 75 HO, and 100 OO videos. 15 participants were presented with set 1 and the other 15 participants were presented with set 2.

Stimuli were presented on a 1024×768 monitor with a 60 Hz refresh rate. Participants were given the following instructions: "In the current experiment, imagine that you are working for a security company. Videos were recorded by bird's-eye view surveillance cameras. In each video, you will see two dots moving around, one in red and one in green. Your task is to 'identify' these two dots based on their movement. There are three possible scenarios: human-human, human-object, or object-object." Videos were presented in random orders. After the display of each video, participants were asked to classify the video into one of the three categories.

10.3.3 Results

Human response proportions are summarized in Fig. 10.11. Response proportion of human-human interaction swas ignificantly greater than the chance level 0.33 ($t(499) = 25.713$, $p < .001$). For HO animations, response proportion of human-object interaction was significantly greater than the other two responses ($p < .001$). Similarly, response proportion of object-object was greater than the other two responses ($p < .001$) for OO animations. These results reveal that human participants identified the main characteristics of different interaction types based on dot movements.

Next, we examined human responses to the sub-categories within the HH and OO animations. We first used the animacy degree as a continuous variable and tested its effect on human responses in the HH animations. With increases in degree of animacy in HH, the response proportion of human-human interaction increased significantly as revealed by a positive correlation ($r = .42$, $p < .001$). This finding suggests that humans are sensitive to the animacy manipulation in terms of the frequency with which self-propelled forces occurred in the stimuli. For the OO animations, the response proportion for object-object interaction among the four sub-categories yielded significant differences ($F(3, 196) = 34.42$, $p < .001$ by an ANOVA), with the most object-object responses in the collision condition, and the least in the rod condition. Pairwise comparisons among the four-categories show significant difference between collision and everything else ($p < .001$), between soft rope and rope ($p < .001$), and also between soft rope and string ($p = .018$); there is a marginally significant difference between rod and string ($p = .079$).

We then combined human responses and the model-derived measures for each animation stimulus to depict the unified psychology space for the perception of physical and social events. Fig. 10.12 presents the distributions of 100 HH videos with 100% animacy degree, 150 HO videos, and 200 OO videos, all in this unified space. In this figure, an animation video is indicated by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. Specifically, the values of its RGB channels are determined by the aver-

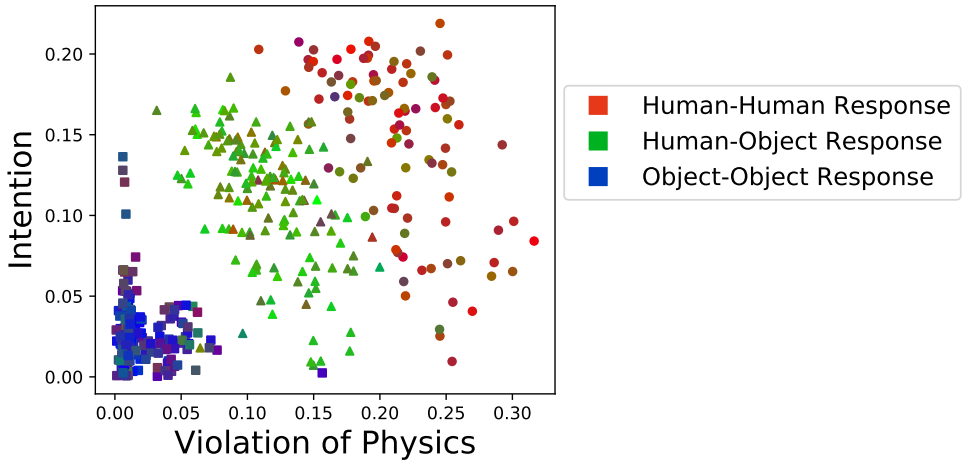


Figure 10.12: Constructed psychological space including HH animations with 100% animacy degree, HO animations, and OO animations. In this figure, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. The two coordinates of the space are the averaged measures between the two entities, as the measure of the degree of violation of physical laws (horizontal) and the measure of maximum log-likelihood ratio of goal-directed trajectory over the background model indicating the presence of intention. The mark shapes of data points correspond to the interaction types used in the simulation for generating the corresponding stimuli (circle: HH, triangle: HO, square: OO).

age human-human responses in red, human-object responses in green, and object-object responses in blue. The mark shapes of data points correspond to the interaction type used in the simulation for generating the synthesized animations. The coordinates of each data point were calculated as the model-derived measures averaged across the two entities in an animation, *i.e.*, Eq. (10.12) for physical violation and Eq. (10.16) for the log-likelihood ratio of the trajectory of an entity is driven by a goal. The resulting space showed clear separations between the animations that were judged as three different types of interactions. Animations with more human-human interaction responses (red marks) clustered at the top-right corner, corresponding to great values of intention and strong evidence signaling the violation of physics. Animations with high responses for object-object interactions (blue marks), located at the bottom left of the space, show low values of intention index and little evidence of violation of physics. Animations with high responses for human-object interactions (green marks) fell in the middle of the space.

To quantitatively evaluate how well the model-derived space accounts for human judgments, we trained a classifier using the coordinates derived in the space shown in Fig. 10.12 as input features ($\mathcal{D}$ and $\mathcal{L}$ for the indices of physical violation and intention respectively). For each ground-truth type of interactions $y \in \{\text{HH}, \text{HO}, \text{OO}\}$, we fit a 2D Gaussian distribution $p_y(\mathcal{D}, \mathcal{L})$, using half of the stimuli as training data. Then for a given animation with the coordinates of $(\mathcal{D}, \mathcal{L})$, the classifier predicts $p(y|\mathcal{D}, \mathcal{L}) = \frac{p_y(\mathcal{D},\mathcal{L})}{\sum_y p_y(\mathcal{D},\mathcal{L})}$ for animations in the remaining half of the stimuli. The correlation between the model predictions and average human responses was 0.815 ($p < .001$) based on 2-fold cross-validation. Using a split-half reliability method, human participants showed an inter-subject correlation of 0.728 ($p < .001$). Hence, the response correlation between model and humans closely matched inter-subject correlations, suggesting a good fit of the unified space as a generic account of human perception of physical and social events based on movements of simple shapes.

We examined the impact of different degrees of animacy on the perception of social events, and how different subcategories of physical events affect human judgments on interaction types. The

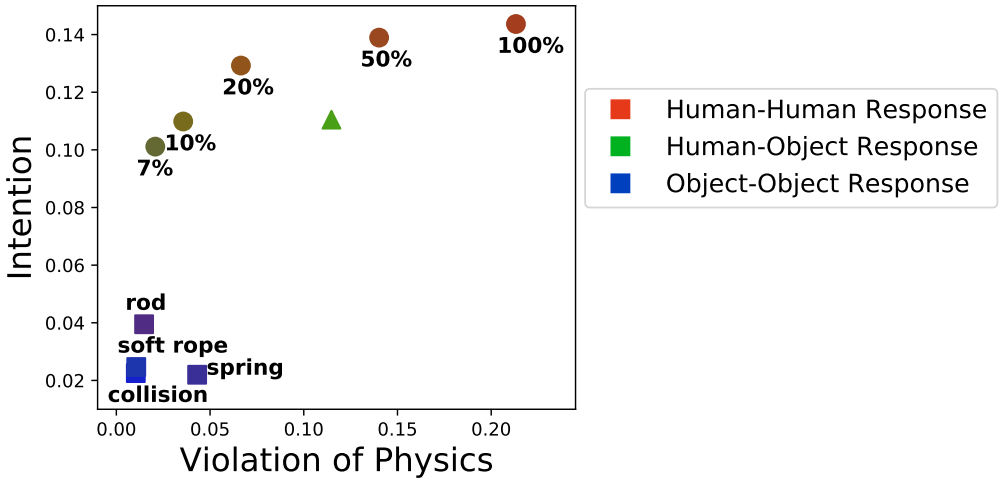


Figure 10.13: Centers of all types of stimuli.

unified space provides a platform to compare these fine-grained judgments. Fig. 10.13 shows the centers of the coordinates and the average responses for each of the sub-categories. We first found that, with a decreased degree of animacy, the intention index in HH animations was gradually reduced towards the level of HO animations. Meanwhile, human judgments of these stimuli varying from low to high degree of animacy transited gradually from human-object responses to human-human responses, consistent with the trend that the data points moved along the physics axis. Among all physical events, the rod and spring conditions showed the highest intention index and the strongest physical violation, respectively, resulting in a greater portion of human-human interaction responses than the other categories.

Chapter 11

Theory of Mind Representations

11.1 Introduction to Theory of Mind

Theory of mind (ToM) refers to the ability to understand one's own and others' mental states, which was firstly studied in psychology and cognitive science [548]. As has been shown, the ability to perform mental simulations of others increases rapidly since the young infant phase [549, 550, 551, 552]. The ability of ToM allows reasoning about others' mind, and is vital in a multi-agent environment because each agent's choice affects the payoff of other agents [553, 554].

Multi-agent systems, ranging from two-player games to the human society, have been studied across many domains. For individual agents to maximize their values in such environments, they must learn to interact with and against others, as well as understand the consequences of their actions.

Contemporary discussions of Theory of Mind have their roots in philosophical debate most broadly, from the time of Descartes' Second Meditation, which set the groundwork for considering the science of the mind. Theory of Mind (ToM) is defined by Premack and Woodruff [555] in the highly influential article "Does the Chimpanzee have a theory of mind?" as "an individual imputing mental states (like beliefs, desires and intentions) to himself and others, to make predictions, specifically about the behavior of other organisms." Further, they differentiated between ToM for motivation (*i.e.*, another organism's valuation, intention, purpose, goal) and ToM for knowledge (*i.e.*, another organism's belief states or learned schemas / scripts).

Since this initial empirical investigation of ToM in nonhuman primates, experimental approaches probing and characterizing ToM capacities have been introduced by psychological and behavioral economics research [556]. Tasks here will be discussed in two dimension following [556]: interactivity and uncertainty.

Observation under divergent knowledge and environmental uncertainty. One of the most prominent tasks in ToM research is the so-called false belief task [557]. After observing a social scene that comprises a change in the physical environment that the observed agent is unaware of (inducing a false belief), participants have to predict that observed agent's behavior. Following a similar general idea as false belief reasoning, Baker and colleagues [558] introduced a perspective taking scenario. In a grid world environment, an observed agent with unknown preferences is placed in an environment containing different choice options with varying subjective value to the agent. Based on the agent's trajectory and the environment layout like occlusion, participants will choose the option that is most valuable to the agent. There are also other tasks including social influence [559], learning about expertise [560], observational inverse learning [561] that require the participants to make some choices based on the observation of other agents.

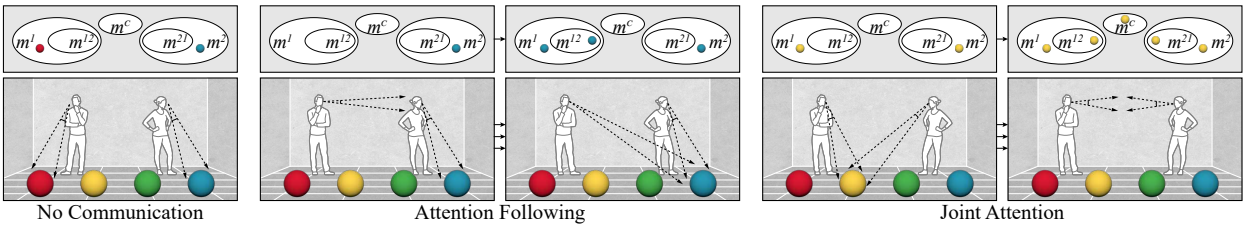


Figure 11.1: **Multi-agent belief dynamics in nonverbal communication.** Different communication events emerge from the social interactions and construct agents' beliefs. In this paper, the belief dynamics are modeled by "five minds" (top) and maintained by a hierarchical energy-based model that tracks each agent's mental state (m^1 and m^2), their estimated belief about other agent's mental state (m^{12} and m^{21}), and the shared common mind (m^c). The concept of the shared "common mind" avoids the infinite recursion issue in prior work.

Interactive tasks. Interactive tasks require the participant to interact with other agents instead of only passively observing others' action. In a fully observable environment, other agents' state and payoff can be directly known by the participants. Well-known settings comprise "prisoner's dilemma" and "stag hunt" [562, 563]. Very few experimental approaches to date have combined asymmetric distribution of information, environmental uncertainty and interactivity. One example is the multi-agent tiger task. In a scenario where two players have to learn which of two doors hides a pot of gold and which hides a dangerous tiger, after each action players receive half of the partner's outcome in the cooperative setting, while in the competitive scenario half of their partner's outcome is subtracted from their own outcome [564]. These tasks are also adapted into suitable tasks to test the theory of mind of non-human primates and individuals with Autism [565, 566].

Modeling. For fully observable tasks, experimental economics and behavioral game theory focus on finding the converging choices of all the agents called Nash equilibrium. In AI field, "Recursive Modeling Method" (RMM) [567] deploys mind recursion in agent planning for single-step game setting. Interactive POMDP (I-POMDP) [568] model is proposed to deal with partial observation and asymmetric information for sequential planning. Neural networks implicated in ToM were successfully identified using standard neuroimaging methods [569, 570]. The studies reported in [571] establish for the first time that a region in the human temporo-parietal junction (called the TPJ-M) is involved specifically in reasoning about the contents of another person's mind.

11.2 Spatiotemporal social event parsing and mental representation

In this section, we propose representations for spatiotemporal social event parsing, as well as new mental representation, called "five minds" that accounts for the triadic relation and "common mind;" this representation is embedded in a hierarchical graphical model with a six-level structure. Table 11.1 lists common notations and their definitions.

We adopt the representation proposed by [572, 573, 140] based on experimental psychology, infant study, and animal cognition, wherein the communication during social interactions heavily relies on the "common mind" after only one or two levels of recursive reasoning of mental state. Below, we use the term "mind" in human/animal studies and the term "mental state" in computational models interchangeably.

Formally, all agents' minds M_t at time t is represented as a set, forming a "five mind" representation:

$$M_t = \{m_t^1, m_t^2, m_t^{12}, m_t^{21}, m_t^c\}, \qquad t = 1, \ldots, T, \tag{11.1}$$

Table 11.1: Common notations for parsing social events

Notation	Description
$I = \{I_t\}_{t=1,\ldots,T}$	The input image sequence, where T is the total number of frames.
$h_t^i = (x_t^i, p_t^i, g_t^i)$	The detected human agent i at time t, where $x_t^i \in \mathbb{R}^3$ denotes the spatial position, $p_t^i \in \mathbb{R}^{3\times 26}$ the skeleton pose, and $g_t^i \in \mathbb{R}^3$ gaze direction.
$o_t^j = (x_t^j, c_t^j, d_t^j)$	The detected object j at time t, where $x_t^j \in \mathbb{R}^3$ denotes the spatial location, $c_t^j \in \mathbb{C}$ the object category, and $d_t^j \in \{1, \ldots, N_o\}$ the object ID; $\mathbb{C}$ is the object category set.
$H = \{h_t^i : i = 1, \ldots, N_h\}$	The detected human agents in the video, where N_h is the total number of agents. Without losing generality, we assume $N_h = 2$ in this paper.
$O = \{o_t^j : j = 1, \ldots, N_o\}$	The detected objects in the video, where N_o is the total number of objects.

where m_t^1 and m_t^2 denote two agents' mind, m_t^{12} and m_t^{21} denote the agent's belief about the other agent's mind, and m_t^c denotes their common mind. Each mind is defined as $m_t = \{(e_t^i, A(e_t^i)) : i = 1, \ldots, N_{e,t}\}$ with a set of entities e^i (*e.g.*, an object) and their attributes $A(e^i)$ (*e.g.*, 3D location).

From a top-down perspective, the change within this mental representation along time constructs the *belief dynamics* $\{\Delta\boldsymbol{\mathcal{M}}\}$ between two agents, derived from the spatiotemporal parsing of the video. The parsing is represented by a spatiotemporal parse graph [154] $pg = (pt, E)$, a hierarchical graphical model that combines a parse tree pt and the contextual relation E on terminal nodes; Fig. 11.2 gives an example. Here, a parse tree $pt = (V, R)$ includes the vertex set with a six-level hierarchical structure $V = V_r \cup V_b \cup V_e \cup V_s \cup V_f \cup V_t$ and the decomposing rule R, where V_r is the root set and contains only one element—the node that represents the entire video, V_b the set of belief dynamics forming "five minds," V_e the set of communication events, V_s the set of interactive segments, V_f the set of frame-based static scenes, and V_t the set of all the detected instances in a single scene. Specifically:

- The belief dynamics are conditioned on communication events, grouped by interactive segments. In this paper, we define four types of belief dynamics: *occur*, *disappear*, *update*, *null*.
- A communication event $e \in V_e$ is one of the three typical nonverbal communication events: *No Communication*, *Attention Following*, and *Joint Attention*, as shown in Fig. 11.1.
- An interactive segment $s \in V_s$ is the decomposition of a communication event and represented by the 4D spatiotemporal features $\Phi_s = (\Phi_s^1, \Phi_s^2)$ extracted from detected entities. These features describe social interactions, including both unary Φ_s^1 and pair-wise features Φ_s^2.
- The contextual relation E is represented by an attention graph $\mathcal{G}_s$ established based on 4D features, wherein the node represents an agent or an object in the scene, and an edge is connected between two nodes if there is directed attention detected among the two entities from the visual

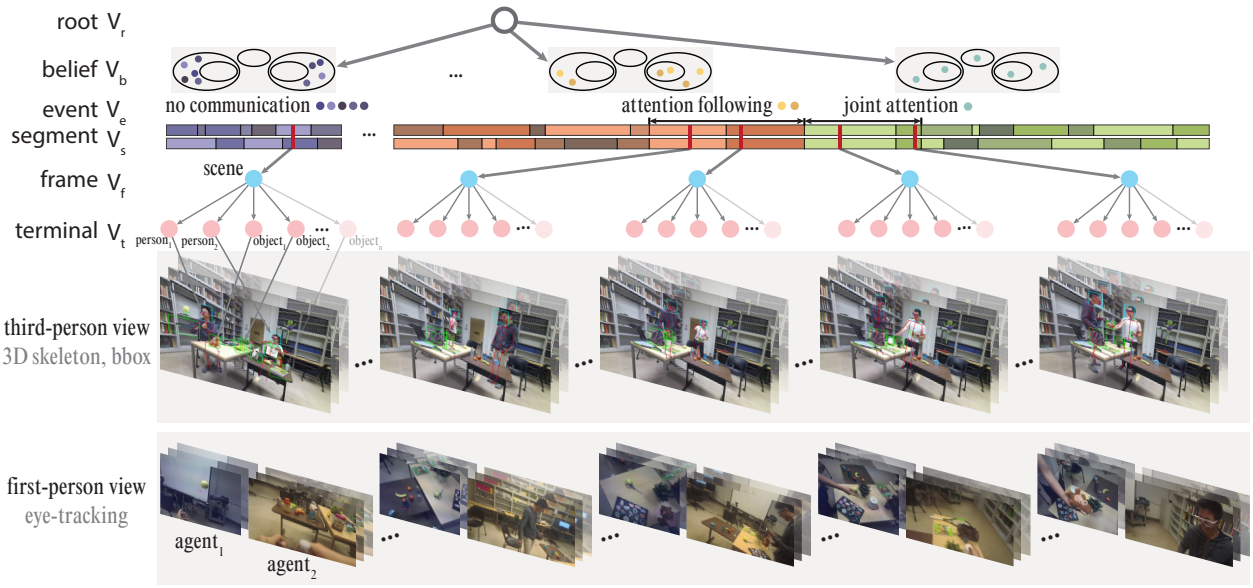


Figure 11.2: **A parse graph of a social event with a six-level hierarchical structure.** V denotes vertex sets in the hierarchy. The root node V_r corresponds to the entire video. The set of belief dynamics V_b emerges from the lower-level communication events (see also Fig. 11.1). Communication events in V_e decompose into lower-level interactive segments in V_s; these segments are social primitives learned unsupervisedly. Each frame of the scene in V_f further decomposes into several terminal nodes in V_t, grounded into entities detected from videos. The colored dots in the V_e layer represent belief changes triggered by communication events. Note that belief dynamics are accumulated over time; we only illustrate the most significant changes.

inputs.

11.3 Example of theory-of-mind in communication

Fig. 11.3 shows one example of Theory of Mind in Communication. The lecturer M is talking in the front of the classroom. The person on the left, denoted as H, noticed that time is up, and thus raised the iPad up, on which is a timer, so as to remind the lecturer M of the remaining time. However, the lecturer M is not looking into the direction of H, and didn't notice H's message. The person sitting in the right side, denoted as G, noticed that H is raising his iPad, and also noticed that M didn't see H's message. After a while, G decided to help, and he raised his arm to attract M's attention. M noticed G's arm and looked to G. Then, G puts down his arm and points to H to refer M's attention to H. M follows G's pointing and looks to the direction of H, and thus finally build a communication channel with H via mutual gaze. M successfully noticed H's message now, and nods his head to H as a signal of receiving the message. M looks back and rushes to finish his talk, while H puts down the iPad. G, sitting in the back of the classroom, watches the whole procedure.

This example is simple and common in our daily life; there are rich communication elements in this simple example, including gaze, waving hand, pointing, *etc*. The nonverbal communication signals are performing much more important role in the scenario of this example. Actually, although natural language is one main communication method in human social interaction, nonverbal communication signals are also irreplaceable in almost all face-to-face communications. As Tomasello pointed out in his book [140], to understand how humans communicate with one another using a language and how this competence might have arisen in evolution, we must first understand how hu-

Figure 11.3: An example of Theory of Mind in communication involving three agents.

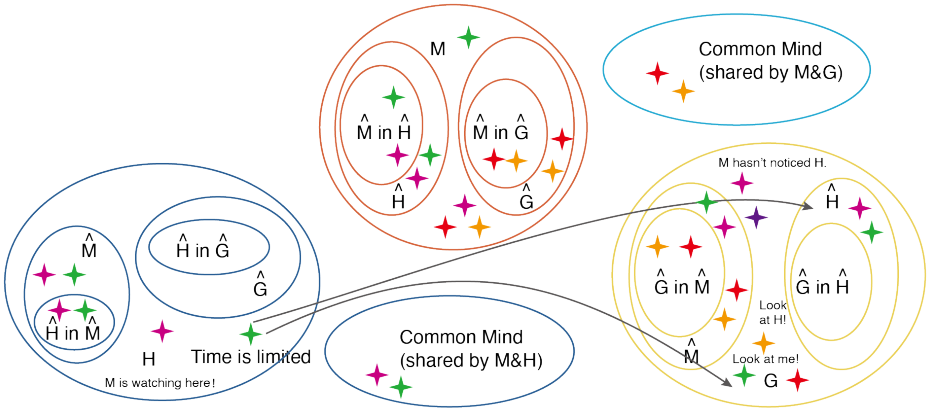


Figure 11.4: Different messages are transmitted in different minds.

mans communicate with one another using natural gestures. There are two basic types of great ape gesture, based on how they function communicatively: intention movements and attention getters. Attention getter serves to attract the attention of the recipient either with underlying social intention or referential intention. Attention getter is usually followed by mutual gaze for a verification that both the communicator and the recipient come to a common mode that they are going to open the communication channel. The following gestures, actions and gazes will be of great importance for understanding such a social interaction. For example, in the timer example, G's waving hand to attract M's attention is one attention getter. Pointing is one typical and signigicant communication gesture, and one of the first uniquely human forms of communication. A simple gesture of pointing could mean a lot, combined with different context and shared experience. The ability to create common conceptual ground–joint attention, shared experience, common cultural knowledge–is absolutely another critical dimension of all human communication. Shared intentionality is what is necessary for engaging in uniquely human forms of collaborative activity.

Fig. 11.4 shows how different messages are transported between different minds in the above example. We use different colors to distinguish these different messages with each other.

11.4 Inferring the Theory-of-Mind Dynamically

11.4.1 Probabilistic Formulation

Based on our proposed new representation, an energy-based probabilistic formulation is derived, capable of parsing the communication events that emerged from the raw pixel inputs. We illustrate the detailed description of the model learning and joint inference procedures in Algorithm 3.

To infer the optimal parse graph pg^* from raw video sequence I, we formulate the video parsing of social events as an MAP (maximum a posteriori) inference problem:

$$pg^* = \arg\max_{pg} P(pg|H,O)P(H,O|I) = \arg\max_{pg} P(H,O|pg)P(pg)P(H,O|I), \tag{11.2}$$

where $P(H,O|I)$ is the detection score of agents and objects in the video, $P(pg)$ is the prior model, and $P(H,O|pg)$ is the likelihood model. Below, we detail the prior and likelihood model one by one.

Prior The prior model $P(pg)$ measures the validness of parse graph; all the nodes in the parse graph should be reasonably parsed from the root node. We model the prior probability of pg as a Gibbs distribution: $P(pg) = \frac{1}{Z_1}\exp\{-\mathcal{E}(pg)\} = \frac{1}{Z_1}\exp\{-\mathcal{E}_{aggr} - \mathcal{E}_{evt} - \mathcal{E}_{be}\}$, where $\mathcal{E}_{aggr}$ is the aggregation prior, $\mathcal{E}_{evt}$ the communication event prior, and $\mathcal{E}_{be}$ the belief dynamics prior. Specifically:

- The aggregation prior encourages the algorithm to focus more on high-level communication pattern, instead of being trapped into trivial primitives that results in fragments; this design prevents the spatiotemporal parsing from being too brittle. Hence, the aggregation prior is defined to be proportional to the total number N_e of events composed of interactive segments: $\mathcal{E}_{aggr} = \lambda_1 \frac{N_e}{T}$.
- The communication event prior leverages the knowledge of transition and co-occurrence frequencies of communication events, defined as

$$\mathcal{E}_{evt} = -\frac{\lambda_2 \sum_{i,j,\mathbb{1}^{trans}(e_i,e_j)=1} \log p^{trans}(e_i,e_j)}{\sum_{i,j}(\mathbb{1}^{trans}(e_i,e_j)=1)} - \frac{\lambda_3 \sum_{i,j,\mathbb{1}^{occ}(e_i,e_j)=1} \log p^{occ}(e_i,e_j)}{\sum_{i,j}(\mathbb{1}^{occ}(e_i,e_j)=1)}, \tag{11.3}$$

  where $p^{trans}(e_i,e_j)$ and $p^{occ}(e_i,e_j)$ are based on frequencies from the dataset, and $\mathbb{1}^{trans}$ and $\mathbb{1}^{occ}$ are indicator functions that reflects the spatiotemporal relations among events.
- $\mathcal{E}_{be}$ models the prior of belief dynamics, which helps to prune some invalid configurations, such as two consecutive *occur*s or an *occur* after an *update*. The prior model is defined as $\mathcal{E}_{be} = -\lambda_4 \sum_{j=1}^{N_e} \log p^M(\Delta\boldsymbol{\mathcal{M}}_j|e_j)$, where

$$p^M(\Delta\boldsymbol{\mathcal{M}}_j|e_j) = \prod_t p(\Delta M_{t+1}|\Delta M_t, e_j)\, p(\Delta M_t|e_j), \tag{11.4}$$

  where $\Delta\boldsymbol{\mathcal{M}}_j$ is the set of belief dynamics occurred within the communication events e_j.

Spatiotemporal Likelihood The likelihood model measures the consistency between the parse graph and the ground-truth observed data. Since our model has a hierarchical structure, we split the likelihood into three energy terms, corresponding to the three crucial layers above the parsing of a single frame in the parse graph; the parsing of the single frame provides H and O as the input:

$$P(H,O\,|pg) = P(H,O|V_b,V_e,E) = \frac{1}{Z_2}\exp\left\{-\mathcal{E}^{comp}(H,O|V_e,E) - \mathcal{E}^{evt}(H,O|V_e,E) - \mathcal{E}^{be}(H,O,V_e|\{\Delta\boldsymbol{\mathcal{M}}\})\right\}. \tag{11.5}$$

- The first energy term $\mathcal{E}^{comp}$ constrains the communication event composed by the interactive segments, so that the features within one composition are similar enough, whereas the features between two consecutive compositions are significantly distinct:

$$\begin{aligned}\mathcal{E}^{comp}(H,O|V_e,E) = \mathcal{E}(\boldsymbol{\Phi}|V_s,E) = \frac{\lambda_5}{N_e}\sum_{j=1}^{N_e}\left(\frac{1}{T_j}\sum_t \mathcal{D}(\phi_{j,t},\phi_{j,t+1})\right) \\ - \frac{\lambda_6\sum_{i,j,\mathbb{1}^{trans}(e_i,e_j)=1}\mathcal{D}(\psi(\Phi_i),\psi(\Phi_j))}{\sum_{i,j}(\mathbb{1}^{trans}(e_i,e_j)=1)} - \frac{\lambda_7\sum_{i,j,\mathbb{1}^{occ}(e_i,e_j)=1}\mathcal{D}(\psi(\Phi_i),\psi(\Phi_j))}{\sum_{i,j}(\mathbb{1}^{occ}(e_i,e_j)=1)}\end{aligned} \tag{11.6}$$

  where $\Phi_i = \{\phi_{i,t}\}$ is the set of features within the interactive segment s_i, $\psi(\cdot)$ is the wavelet transform [574], and $\mathcal{D}(\cdot)$ is the distance measurements between two sets of extracted features.
- The second energy term $\mathcal{E}^{evt}$ is the negative communication event classification score with respect to the detected feature set $\boldsymbol{\Phi} = \{\Phi_j\}$ and the constructed attention graph set $\boldsymbol{\mathcal{G}} = \{\mathcal{G}_j\}$. This second term is defined as $\mathcal{E}^{evt}(H,O|V_e,E) = \mathcal{E}(\boldsymbol{\Phi},\boldsymbol{\mathcal{G}}|V_e)$ and encodes all the entities in the scene extracted from visual input, which can be solved by a traditional MLE:

$$\mathcal{E}(\boldsymbol{\Phi},\boldsymbol{\mathcal{G}}|V_e) = -\frac{1}{N_e}\sum_{j=1}^{N_e}\lambda_8\log p(\Phi_{\Lambda_j},\mathcal{G}_{\Lambda_j}|e_j) = -\frac{1}{N_e}\sum_{j=1}^{N_e}\lambda_8\log p(e_j|\Phi_{\Lambda_j},\mathcal{G}_{\Lambda_j}) - C, \tag{11.7}$$

  where Λ_j is the set of indexes of the interactive segments decomposed from e_j, and C is a constant.
- The third energy term $\mathcal{E}^{be}$ models the belief dynamics in all five minds:

$$\begin{aligned}\mathcal{E}^{be}(H,O,V_e|\{\Delta\boldsymbol{\mathcal{M}}_j\}) &= -\frac{1}{N_e}\sum_{j=1}^{N_e}\lambda_9\log p(\Delta\boldsymbol{\mathcal{M}}_j|H,O,V_e) \\ &= -\frac{1}{N_e}\sum_{j=1}^{N_e}\left(\frac{1}{T_j}\sum_t\lambda_9\log p\left(\Delta M_{j,t+1}|g_{j,t+1},e_j,\{\Delta M_{j,t'}|t'\in[t_j^s,t]\}\right)\right),\end{aligned} \tag{11.8}$$

  where t_j^s is the starting frame of the communication event e_j, and $g_{j,t+1}$ is the attention graph of frame $t+1$ under event e_j.

11.4.2 Learning Algorithm

The learning process follows a bottom-up procedure; the algorithm (i) parses each frame to extract the entities and relations, (ii) joint parses both interactive segments (proposals generated unsupervisedly) and communication events (with trained likelihood) by beam search (see the detailed algorithm in supplementary material), (iii) predicts the belief dynamics (with trained likelihood), and (iv) fine-tunes all the parameters to minimize the overall errors. Algorithm 3 details the overall procedure.

11.5 Emotional Quotient (EQ) Test

11.5.1 Introduction

With the rapid development of artificial intelligence, its applications are spread in all corners of society. Artificial intelligence not only provides convenience for social life in the form of functionalization (*e.g.*, face recognition, machine translation), but also lands in the field of service (*e.g.*, robot, assistant). Incorporating human emotional mechanisms allows AI to perform tasks that machines

Algorithm 3: Learning to parse social events

Input : Video $\{I_{train}\}$, ground truth V_e^* and V_b^*.
Output: Parameter sets Θ_1^* and Θ_2^*, and parse graph pg.
Init. : $H, O, \boldsymbol{\Phi}, \boldsymbol{\mathcal{G}}, \Theta_1^*, \Theta_2^* = \varnothing$; $L_1^*, L_2^* = +\infty$

for I_i *in* $\{I_{train}\}$ **do**
  H_i = humanDetectionWithReID(I_i), $H \leftarrow H \cup H_i$
  O_i = objectDetectionWithReID(I_i), $O \leftarrow O \cup O_i$
  $\boldsymbol{\Phi}_i$ = extractSTFeatures(H_i, O_i), $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \boldsymbol{\Phi}_i$
  $\boldsymbol{\mathcal{G}}_i$ = buildAttentionGraph(H_i, O_i, $\boldsymbol{\Phi}_i$), $\boldsymbol{\mathcal{G}} \leftarrow \boldsymbol{\mathcal{G}} \cup \boldsymbol{\mathcal{G}}_i$
end
$V_s \leftarrow$ Generate $\{s\}$ by unsupervised clustering.
`/* Train likelihood of` e_j `as in [523]` */
Train $p(e_j|\Phi_{\Lambda_j}, \mathcal{G}_{\Lambda_j})$ in Eq. (11.7) with ground-truth V_e^*.
`/* Finetune the parameter set` Θ_1^*. */
for $\Theta_1^{(i)} = (\lambda_1, \lambda_2, \lambda_3, \lambda_5, \lambda_6, \lambda_7, \lambda_8) \in \Omega_{\Theta_1}$ **do**
  Compute $\mathcal{E}^{comp}$ based on Eq. (11.6), given $\boldsymbol{\Phi}$ and $\Theta_1^{(i)}$.
  Compute $\mathcal{E}^{evt}$ based on Eq. (11.7), given $\boldsymbol{\Phi}$, $\boldsymbol{\mathcal{G}}$, $\Theta_1^{(i)}$.
  Infer V_e by dynamic programming beam search; see details in Algorithm 4.
  Calculate error L_1 between V_e and V_e^*.
  if $L_1 < L_1^*$ **then** $L_1^* \leftarrow L_1$. $\Theta_1^* \leftarrow \Theta_1^{(i)}$.
end
`/* Train belief dynamics likelihood` */
Train $p\left(\Delta M_{j,t+1} | g_{j,t+1}, e_j, \{\Delta M_{j,t'}\}\right)$ in Eq. (11.8) with V_b^*.
`/* Finetune the parameter set` Θ_2^*. */
for $\Theta_2^{(i)} = (\lambda_4, \lambda_9) \in \Omega_{\Theta_2}$ **do**
  for e_j *in* V_e **do**
    Compute the posterior probability of belief dynamics based on Eqs. (11.4) and (11.8).
    Predict the best $\hat{V}_b$ by MAP.
  end
  Calculate error L_2 between the best predicted belief dynamics $\hat{V}_b$ and the ground-truth V_b^*.
  if $L_2 < L_2^*$ **then** $L_2^* \leftarrow L_2$. $\Theta_2^* \leftarrow \Theta_2^{(i)}$.
end

cannot currently be programmed or trained to perform: machines can feel the effects of others and put themselves in the shoes of others by thinking about the benefits and drawbacks of their actions. Emotional factors such as curiosity, fear and surprise can regulate their behaviors. We may expect that intelligences will be able to express their inner states through communication with others and possibly influence decision-making.

With the emergence of robots, voice assistants, and virtual idols, AI is widely used in the services [575, 576]. At present, although AI has been applied to psychological counseling, legal counseling, elderly companionship, *etc.*, most of the existing task planning algorithms for EQ focus on how to model the strategies of other intelligence based on the "physical environment state",

Algorithm 4: Event inference via DP beam search

Input : $\boldsymbol{\Phi}$, $\boldsymbol{\mathcal{G}}$, V_s, $p(e_j|\Phi_{\Lambda_j}, \mathcal{G}_{\Lambda_j})$.
Output : V_e
Initialization: $V_e = \varnothing, \mathcal{B} = \{V_e, p = 0\}, m, n$.

```
while True do
    B' = ∅
    for {V_e, p} ∈ B do
        {e_i} = Next(V_s, V_e, m)
        if {e_i} ≠ ∅ then
        for each proposed e_i do
            p(V_e|Φ, G) = DP(V_e, p, e_i, Φ, G)
            V_e = V_e ∪ {e_i}; B' = B' ∪ {V_e, p}
        end
        else B' = B' ∪ {V_e, p}
    end
    if B' == B then return V_e = Best(B, 1)
    else D = Best(B', n); B = D
end
```

and there is a lack of research on EQ that combine emotion, personality, and behavioral habits in the academic community. There is a lack of research on EQ that combine emotions, personality, and behavioral habits. Research related to the emotional intelligence has also focused on expression recognition and generation: *i.e.*, judging the expression of a person through speech, pictures or videos of facial expressions [577]; some research has also used expression information to generate facial animations for use in speech, dialogue, *etc.* [578].

In recent years, AI has also gradually started to combine with some humanities and social sciences to expand the scenarios of its applications. In specific scenarios, AI can be involved in ethical judgment [579], legal discourse [580], moral code reasoning, and social value research [581]. In the context of social service and human-centered research, agents of high EQ will have direct benefits for human-computer interaction. We hope agents will be able to not only understand the overall environment, but also perceive model specific states, remember historical behaviors, and judge future actions for individuals, in order to achieve more humane and efficient communication and interaction.

This part of the book focuses on how to give intelligences an emotional quotient (EQ) that enables them to have a complete understanding and expression mechanism of human emotions, basic human value perceptions, and reasonable language and behavior guidelines in the service of social life. We summarize our research on the emotional intelligence of intelligent bodies into the following three aspects. First, we make artificial agent capable of understanding social relations, especially in a real-time and comprehensive manner. Second, agents are trained from multi-modalities (human language, movements, facial expressions) so that they can understand and express human emotions. robots may express emotions, have anthropomorphic images in service scenarios, and can have better emotional empathy with service recipients. Third, we want to give the intelligent body in line with the basic cognition of the universal values of human society, so that it can make a right and wrong judgment of its own behavior based on morality and common sense, so that we have a more stable, safe and trustworthy intelligent assistant.

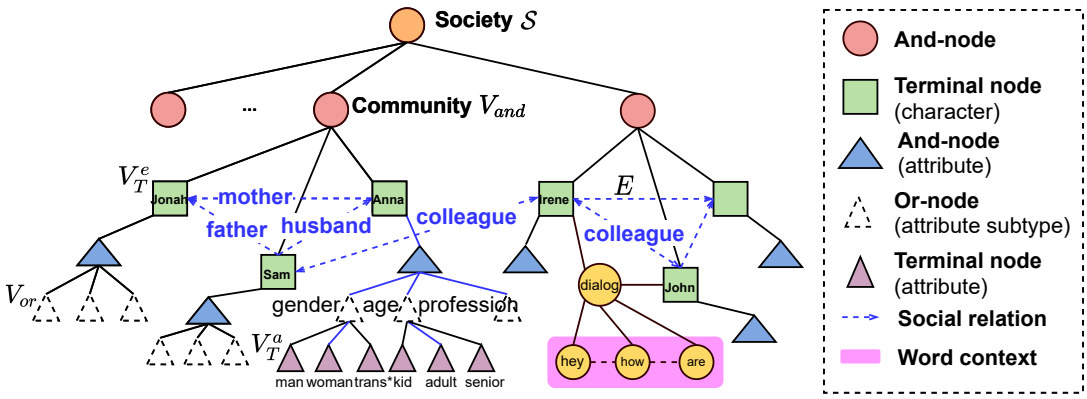


Figure 11.5: **SocAoG**: Attributed And-Or Graph representation of a social network. A parse graph determining each attribute and relation type is marked in blue lines. Dialogues are governed by the word context and associated human attributes and relations.

11.5.2 Incremental Graph Parsing for Social Relation Inference

Our goal is to construct a social network through utterances in dialogue. The network is a heterogeneous physical system [582] with particles representing entities and different types of edges representing social relations. Each entity is associated with multiple types of attributes, while each type of relation is governed by a potential function defined in human attribute and value space, acting as the social norm. The relations are often asymmetric, *e.g.*, A is B's father does not mean B is A's father. To model the network, we utilize an attributed And-Or Graph (A-AoG), a probabilistic grammar model with attributes on nodes. Such design takes advantage of the reconfigurability of its probabilistic context-free grammar to reflect the alternative attributes and relations, and the contextual relations defined on Markov Random Field to model the social norm constraints.

Graph-based Social Relation

The social network graph, named SocAoG, is diagrammatically shown in Fig. 11.5. Formally, SocAoG is defined as a 5-tuple:

$$\mathcal{G} =< S, V, E, X, P > \tag{11.9}$$

where S is the root node for representing the interested society. $V = V_{and} \cup V_{or} \cup V_T^e \cup V_T^a$ denotes all nodes' collection. Among them, And-nodes V_{and} represent the set of social communities, which can be decomposed to a set of entity terminal nodes, V_T^e, representing human members. Community detection is based on the social network analysis [583, 584], and can benefit the modeling of loosely connected social relations. Each human entity is associated with an And-node that breakdowns the attributes into subtypes such as gender, age, and profession. All the subtypes consist of an Or-node set, V_{or}, for representing branches to alternatives of attribute values. Meanwhile, all the attribute values are represented as a set of terminal nodes V_T^a. We denote E to be the edge set describing social relations, $X(v_i)$ to be the attributes associated with node v_i, and $X(\vec{e}_{ij})$ to be the social relation type of edge $\vec{e}_{ij} \in E$.

Relation Inference

Given P to be the probability model defined on SocAoG, a parse graph pg is an instantiation of SocAoG with determined attribute selections for every Or-node and relation types for every edge.

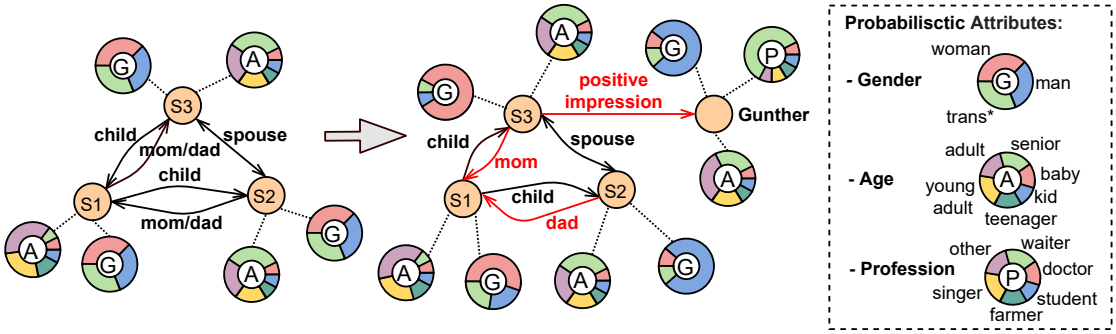


Figure 11.6: Our method iteratively updates the robot's belief of users' individual attributes and social relations, similar to human's reasoning process. The left and right graph show the established and updated belief, respectively.

For a dialogue session with T turns $D_T = \{D^{(1)}, D^{(2)}, ..., D^{(T)}\}$, where $D^{(t)}$ is the utterance at turn t, our method infers the attributes and social relations incrementally over turns:

$$\mathcal{G}_T = \{pg^{(1)}, pg^{(2)}, ..., pg^{(T)}\} \tag{11.10}$$

where $pg^{(t)}$ represents the belief of SocAoG at the dialogue turn t. We incrementally update the pg by maximizing the posterior probability:

$$pg^* = \arg\max_{pg} p(pg|D;\theta) \tag{11.11}$$

where pg^* is the optimum social relation belief, and θ is the set of model parameters.

For simplicity, we denote $X(v_i)$ as $\mathbf{v}_i$ and $X(\vec{e}_{ij})$ as $\mathbf{e}_{ij}$ in the rest of the paper. We introduce three processes, *i.e.*, α, β, and γ process, to infer any SocAoG belief pg^*. We start by rewriting the posterior probability as a Gibbs distribution:

$$\begin{aligned} p(pg|D;\theta) &\propto p(D|pg;\theta)p(pg;\theta) \\ &= \frac{1}{Z} exp\{-\mathcal{E}(D|pg;\theta) - \mathcal{E}(pg;\theta)\} \end{aligned} \tag{11.12}$$

where Z is the partition function. $\mathcal{E}(D|pg;\theta)$ and $\mathcal{E}(pg;\theta)$ are dialogue- and social norm-based energy potentials respectively, measuring the cost of assigning a graph instantiation.

Denoting a dialogue as a sequence of words: $D = \{w_1, w_2, ..., w_\mathcal{T}\}$, the dialogue likelihood energy term $\mathcal{E}(D|pg;\theta)$ can be expressed with a language model conditioned on the parse graph:

$$\begin{aligned} \mathcal{E}(D|pg;\theta) &= \sum_{t=1}^{\mathcal{T}} \mathcal{E}(w_t|\mathbf{c}_t, pg) \\ &= \sum_{t=1}^{\mathcal{T}} -\log(p(w_t|\mathbf{c}_t, pg)) \end{aligned} \tag{11.13}$$

where $\mathbf{c}_t = [w_1, ..., w_{t-1}]$ is the context vector. Intuitively, the word selection depends on the word context, the entities' attributes and their interpersonal relations.

We approximate the likelihood by finetuning a BERT-based transformer with a customized input format

$$\langle \texttt{[CLS]D[SEP]} v_{i_0}\mathbf{e}_{i_0j_0}v_{j_0}...v_{i_n}\mathbf{e}_{i_nj_n}v_{j_n}v_0\mathbf{v}_0...v_n\mathbf{v}_n \texttt{[SEP]} \rangle$$

which is a concatenation of the dialogue history D and a flattened parse graph string encoding the current belief. We call the estimation of pg from the dialogue likelihood $p(w_t|\mathbf{c}_t, pg)$ to be the α **process**. α process lacks the explicit constraints for social norms related to interpersonal relations and human attributes.

For the social norm-based potential, we design it to be composed of three potential terms:

$$\begin{aligned}\mathcal{E}(pg;\theta) = &-\beta \sum_{v_i,v_j\in V(pg)} \log(p(\mathbf{e}_{ij}|\mathbf{v}_i,\mathbf{v}_j)) \\ &-\gamma_l \sum_{\vec{e}_{ij}\in E(pg)} \log(p(\mathbf{v}_i|\mathbf{e}_{ij})) \\ &-\gamma_r \sum_{\vec{e}_{ij}\in E(pg)} \log(p(\mathbf{v}_j|\mathbf{e}_{ij}))\end{aligned} \tag{11.14}$$

where $V(pg)$ and $E(pg)$ are the set of terminal nodes and relations in the parse graph, respectively. We call the term $p(\mathbf{e}_{ij}|\mathbf{v}_i,\mathbf{v}_j)$ the β **process**, in which we bind the attributes of node v_i and v_j to update their relation edge $\mathbf{e}_{ij}$, in order to model the constraint on relations from human attributes. Reversely, we call the terms $p(\mathbf{v}_i|\mathbf{e}_{ij})$ and $p(\mathbf{v}_j|\mathbf{e}_{ij})$ the γ **process**, in which we use the social relation edge $\mathbf{e}_{ij}$ to update the attributes of node v_i and v_j. This models the impact of relation to the attributes of related entities. β, γ_l, and γ_r are weight factors balancing α, β and γ processes. Combining Eq. (11.12), Eq. (11.13), and Eq. (11.14), we get a posterior probability estimation $p(pg|D;\theta)$ of parse graph pg, with the guarantee of the attribute and social norm consistencies.

Here we also provide a reduced version of our model, $\text{SocAoG}_{\text{reduced}}$, which applies when characters' attributes annotation are not available for training[1]. With the same dialogue-based energy potential, We define the parse graph prior energy over a set of relation triangles:

$$\mathcal{E}(pg;\theta) = -\beta \sum_{\vec{e}_{ij},\vec{e}_{ik},\vec{e}_{jk}\in E(pg)} \log(p(\mathbf{e}_{ij}|\mathbf{e}_{ik},\mathbf{e}_{jk})). \tag{11.15}$$

Incrementally parsing the SocAoG is accomplished by repeatedly sampling a new parse graph $pg^{(t)}$ from the posterior probability $p(pg^{(t)}|D^{(t)};\theta)$. We utilize a Markov Chain Monte Carlo (MCMC) sampler to update our parse graph since the complexity of the problem caused by multiple energy terms.

11.5.3 Triangular Character Animation Sampling with Motion,Emotion and Relation

In this section, we first define the elements to make an animation. Then, we introduce the stochastic grammar to sample animations. In the sampling process, the norms of **valence**, **arousal**, **dominance** and **intimacy** run as constraints between the motion, emotion and social relation. Finally, we put them together to make our probabilistic model of ST-AOG.

Skeletal animation

Skeletal animation or rigging is a technique in computer animation in which a character/an agent is controlled by a hierarchical set of body joints. Let j denote one body joint that is characterized by its rotation (r_x, r_y, r_z) and position (x, y, z). A body pose p is defined as a set of joints $\{j_u\}_{u=1,2,...,n}$ controlling the whole body, where n is the total number of joints. Let f be a facial expression

[1]Both SocAoG and $\text{SocAoG}_{\text{reduced}}$ do not need attribute annotation during inference once trained.

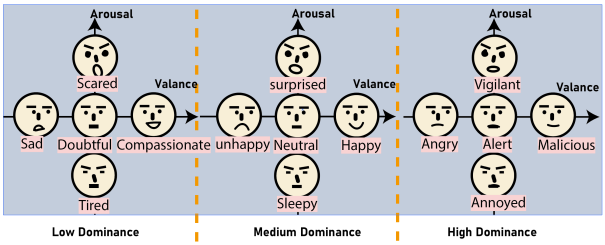


Figure 11.7: The norms of VAD and facial expressions

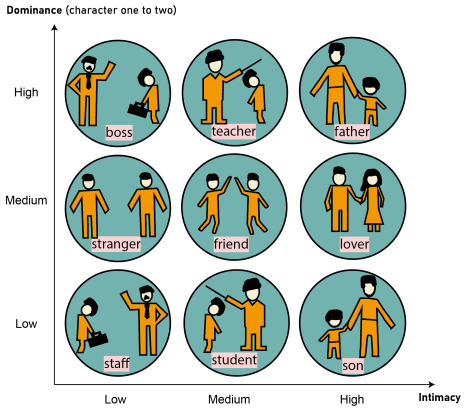


Figure 11.8: Examples of different relation types

characterized by the its valance, arousal and dominance (v, a, d), which we will discuss in details in the next part. Then we make animations by designing the body pose p and facial expression f for the k-th frame at time t_k. A **motion** m is defined as a sequence of body poses and an **emotion** e is defined a sequence of facial expressions corresponding to the key frames:

$$m = \{(p_u, t_u)\}_{u=1,2,...} \tag{11.16}$$

$$e = \{(f_u, t_u)\}_{u=1,2,...} \tag{11.17}$$

11.5.4 Norms of valance, arousal, dominance and intimacy

The norms valence, arousal and dominance (VAD) are fairly standardized to assess environmental perception, experience, and psychological responses [585]. Valance v describes the pleasantness of a stimulus, arousal a quantifies the intensity of emotion provoked by a stimulus, and dominance d evaluates the degree of control. Fig. 11.7 shows different facial expressions with different degrees of valence, arousal and dominance.

We bring another concept: intimacy [586] to the norms to make them support the definition of different types of the social relation. Specifically, intimacy i describes the closeness of the relationship. Define the relation r between two agents as their relative dominance and intimacy (d, i). Fig. 11.8 shows some possible social relations along with their dominance-intimacy scores.

Norms of valance, arousal, dominance and intimacy form the space to set constraints between motion, emotion and social relation, which we will discuss in details in the probabilistic model of ST-AOG.

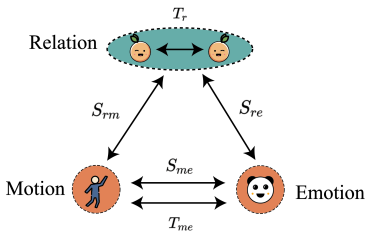


Figure 11.9: Spatial-Temporal relations between motion, emotion and social relation.

Representation of two-agent animations

Define a spatial-temporal And-Or Graph (ST-AOG),

$$\mathcal{G} = (R, V, C, P, S, T) \tag{11.18}$$

to represent the social-relational interaction between two characters, where R is the root node for representing the scene with two characters. V the node set, C the production rules, P the probability model. The spatial relation set S represents the contextual relations between terminal nodes and the temporal relation set T represents the time dependencies.

Node Set V can be decomposed into a finite set of nonterminal and terminal nodes: $V = V^{NT} \cup V^{T}$. The non-terminal nodes V^{NT} consists of two subsets V^{And} and V^{Or}. A set of **And-nodes** V^{And} is a node set in which each node represents a decomposition of a larger entity (*e.g.*, one body pose) into smaller components (*e.g.*, head pose and hand pose). A set of **Or-nodes** V^{Or} is a node set in which each node branches to alternative decompositions (*e.g.*, one relationship can be attributed to family or society). The selection rule of Or-nodes follows probability model P, which is defined as a multinomial distribution. The **terminal nodes** V^{T} represent entities which have different meanings according to context. In this paper, the terminal nodes under the *relation branch* identify the relationship between the two characters (*e.g.*, a father and a son in one family), the ones under *motion branch* determine body poses with the positions and rotations of body joints, and the ones under *emotion branch* depict facial expressions from eyebrows, eyes, mouth and *etc.*

Spatial Relations S among nodes are represented by the horizontal links in ST-AOG forming Markov Random Fields (MRFs) on the terminal nodes. We define different types of potential functions for different cliques to encode different semantics between body motion m, emotion e, and social relation r.

$$S = S_{me} \cup S_{re} \cup S_{rm} \tag{11.19}$$

S_{me} sets constrains on the motion and emotion to ensure that the body movement supports the right emotion. For example, crying (rubbing eyes) can hardly be compatible with a smile. S_{re} regulates the emotion when social relation is considered. For example, we are unlikely to laugh presumptuously in the front of our bosses. Similarly, S_{rm} manages to select the suitable body motion under social relation.

Temporal Relations T among nodes are also represented by the horizontal links in ST-AOG to address time dependencies in animation. The temporal relations

$$T = T_{me} \cup T_{r} \tag{11.20}$$

are divided into two subsets. T_{me} encodes the temporal relation between motion and emotion, to ensure that they match at the right time. Finally, T_r describes to what extent the two agents' animations match temporally. For example, the reaction of one's shaking hand proposal should not be too late.

A hierarchical parse tree pt is an instantiation of the ST-AOG by selecting a child node for the Or-nodes and determining the terminal nodes. A parse graph pg consists of a parse tree pt, a number of spatial relations S and a number of temporal relations T on the parse tree:

$$pg = (pt, S_{pt}, T_{pt}) \tag{11.21}$$

Probabilistic model of ST-AOG

A scene configuration is represented by a parse graph pg, including animations and social relations of the two characters. The prior probability of pg generated by an ST-AOG parameterized by θ is formulated as a Gibbs distribution:

$$\begin{aligned} p(pg \mid \Theta) &= \frac{1}{Z} \exp\{-\mathcal{E}(pg \mid \Theta)\} \\ &= \frac{1}{Z} \exp\left\{-\mathcal{E}(pt \mid \Theta) - \mathcal{E}\left(S_{pt} \mid \Theta\right) - \mathcal{E}\left(T_{pt} \mid \Theta\right)\right\} \end{aligned} \tag{11.22}$$

where $\mathcal{E}(pg \mid \Theta)$ is the energy function of a parse graph, and $\mathcal{E}(pt \mid \Theta)$ is the energy function of a parse tree. $\mathcal{E}(S_{pt} \mid \Theta)$ and $\mathcal{E}(T_{pt} \mid \Theta)$ are the energy terms of spatial and temporal relations.

$\mathcal{E}(pt \mid \Theta)$ can be further decomposed into the energy functions of different types nodes. Since the And-nodes are deterministically expanded, we do not need an energy term for the And-nodes here. The energy terms of Or-nodes and terminal nodes are defined as the log-likelihood from probability model P.

$$\mathcal{E}(pt \mid \Theta) = \underbrace{\sum_{v \in V} \mathcal{E}_{\Theta}^{Or}(v)}_{\text{non-terminal nodes}} + \underbrace{\sum_{v \in V_T^r} \mathcal{E}_{\Theta}^{T}(v)}_{\text{terminal nodes}} \tag{11.23}$$

Spatial potential $\mathcal{E}\left(S_{pt} \mid \Theta\right)$ combines the potentials of the three types of cliques formed in the terminal layer, integrating semantic contexts mentioned previously for motion, emotion and relation.

$$\begin{aligned} p\left(S_{pt} \mid \Theta\right) &= \frac{1}{Z} \exp\left\{-\mathcal{E}\left(S_{pt} \mid \Theta\right)\right\} \\ &= \prod_{c \in C_{me}} \phi_{me}(c) \prod_{c \in C_{re}} \phi_{re}(c) \prod_{c \in C_{rm}} \phi_{rm}(c) \end{aligned} \tag{11.24}$$

We apply the norms of valence, arousal, dominance and intimacy to quantify the triangular constraints between motion, emotion and social relation:

- By the definition of social relation r, we can directly get its dominance and intimacy (d_r, i_r).
- For emotion e, which is a sequence facial expression, we consider its valance, arousal and dominance scores as the different between the beginning facial expression f_0 and ending facial expression f_1:

$$(v_e, a_e, d_e) = f_1 - f_0 = (v_{f_1}, a_{f_1}, d_{f_1}) - (v_{f_0}, a_{f_0}, d_{f_0}) \tag{11.25}$$

- To get the scores of a motion m, we first label the name N_m of the motion, such as *talk, jump* and *cry*. Then we can the valance, arousal and dominance scores from NRC-VAD Lexicon [587], which includes a list of more than 20,000 English words and their valence, arousal, and dominance scores.

$$m \rightarrow N_m \rightarrow (v_m, a_m, d_m) \tag{11.26}$$

Therefore, the relation S_{me} and its potential ϕ_{me} on the clique $C_{me} = \{(m, e)\}$ containing all the motion-emotion pairs in the animation, we define

$$\phi_{me}(c) = \frac{1}{Z^s_{me}} \exp\{\lambda^s_{me} \cdot (v_m, a_m, d_m) \cdot (v_e, a_e, d_e)^\top\} \tag{11.27}$$

Calculating Potentials ϕ_{rm} on clique $C_{rm} = \{(m, r)\}$ and ϕ_{re} on $C_{re} = \{(e, r)\}$ needs another variable i_{me} suggesting the intimacy score. i_{me} is defined as the distance $dist$ between the two agents compared with a standard social distance $dist_0$:

$$i_{me} = \frac{dist_0 - dist}{dist_0} \tag{11.28}$$

Then we can define

$$\phi_{re}(c) = \frac{1}{Z^s_{re}} \exp\{\lambda^s_{re} \cdot (d_r, i_r) \cdot (d_e, i_{me})^\top\} \tag{11.29}$$

$$\phi_{rm}(c) = \frac{1}{Z^s_{rm}} \exp\{\lambda^s_{rm} \cdot (d_r, i_r) \cdot (d_m, i_{me})^\top\} \tag{11.30}$$

Temporal potential $\mathcal{E}\left(S_{pt} \mid \Theta\right)$ combines two potentials for time control.

$$\begin{aligned} p\left(T_{pt} \mid \Theta\right) &= \frac{1}{Z} \exp\left\{-\mathcal{E}\left(T_{pt} \mid \Theta\right)\right\} \\ &= \prod_{c \in C^T_{me}} \psi_{me}(c) \prod_{c \in C^T_r} \psi_r(c) \end{aligned} \tag{11.31}$$

Potential ψ_{me} is define on clique $C^T_{me} = \{(t_m, t_e)\}$ representing the time to start a motion and an emotion. We assume the time discrepancy between them follows a Gaussian distribution.

$$\psi_{me}(c) = \frac{1}{Z^t_{me}} \exp\left(\lambda^t_{re} \cdot (t_m - t_e)^2\right) \tag{11.32}$$

Notice that so far the training parameters $\lambda^s_{me}, \lambda^s_{rm}, \lambda^s_{rm}, \lambda^t_{me}$ and partition functions $Z^s_{me}, Z^s_{re}, Z^s_{rm}, Z^t_{me}$ should be doubled since we have two characters in the scene.

At last, to match the animation for both characters, we assume that the time differences between ending time of their motions $t_{1,m}, t_{2,m}$ and emotions $t_{1,e}, t_{2,e}$ follow the Gaussian distribution.[2]

$$\begin{aligned} \psi_r(c) = &\frac{1}{Z^t_m} \exp\left(\lambda^t_m \cdot (t_{1,m} - t_{2,m})^2\right) \\ &+ \frac{1}{Z^t_e} \exp\left(\lambda^t_e \cdot (t_{1,e} - t_{2,e})^2\right) \end{aligned} \tag{11.33}$$

Here we have two additional parameters λ^t_m, λ^t_e and two more partition functions Z^t_m, Z^t_e.

11.5.5 Towards Socially Intelligent Agents with Mental State Transition and Human Utility

We first briefly introduce the game environment LIGHT, followed by the mental state modeling and utility formulation.

[2]We do not make any constraints on starting time of their motions because every motion has a fixed duration.

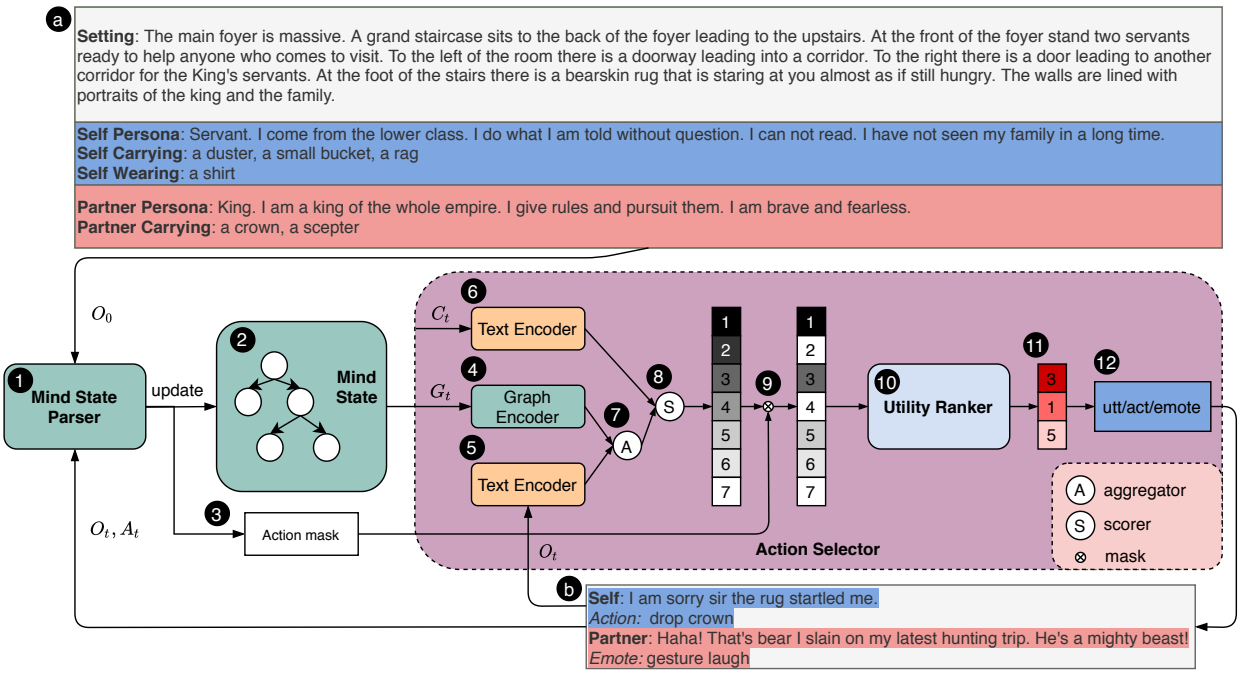


Figure 11.10: Socially Intelligent Agent Model Architecture with Mind State Parser and Utility Model.

LIGHT [588] is a large-scale crowdsourced fantasy text-adventure platform for studying grounded dialogues. Fig. 11.10 (a) shows a typical local environment setting, including location description, objects (and their affordances), characters and their personas. Agents can talk to other agents in free-form text, take actions defined by templates, or express certain emotions (Fig. 11.10 (b)). Agents could be role-played by either humans or machines. Our task is to build an agent to speak and act in LIGHT in a socially intelligent manner. To achieve this goal, we model the agent's mental state transition and incorporate human utility. The mind model is proposed to depict the agent's belief about the underlying states of the text world. Meanwhile, the utility model is designed to learn human preference in common social situations.

Mental State Modeling

Our goal is to construct and maintain the mental states among the theory of mind in dialogues. With the mental state grounding on the details of the local environment, the agent could simulate and reason the evolutionary status of the world and condition its speaking and actions. A graphical representation of the mental state is proposed, as illustrated in Fig. 11.11. All the agents, persona descriptions, objects and their descriptions, and setting descriptions are represented as nodes, which will change as the game location switches. The state of mind is described by the relational edges between these nodes. The mental state is updated with the observed dialogue history or actions, *e.g.*, *King gives the scepter to the servant* will result the servant is carrying the scepter. Such graphical representations are largely distributed among the theory of minds and they are updated in the following mental states:

- Level 0: Physical world
- Level 1: A's belief and desires; B's belief and desires
- Level 2: A's belief in A's mind, B's mind in B (self-conscious); B's belief in A's mind and A's belief in B's mind.

Note that our model only stays in the Level 1 due to the dataset limitation.

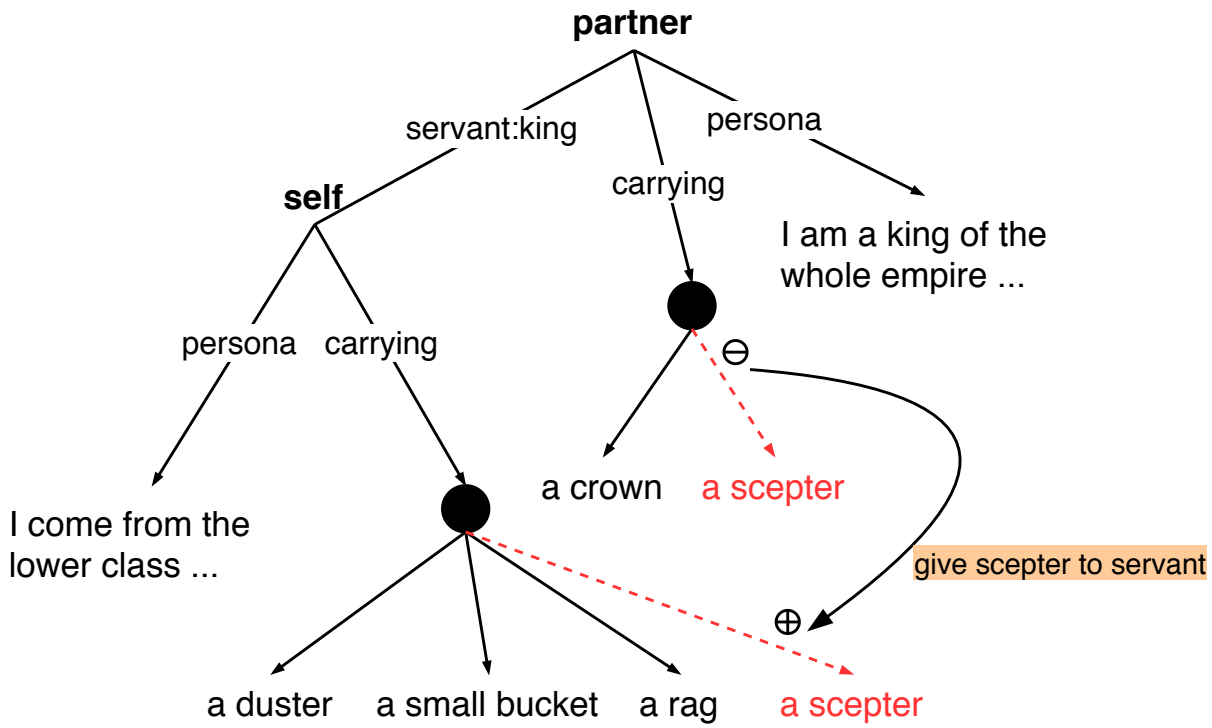


Figure 11.11: A graphical representation of the agent's mental state. Nodes are attributed with encoded natural language description of agents, objects and the environment. Agents' action trigger explicit topology changes of the graph.

Human Utility Modeling

We assume that the agent in the fantasy world would make near-optimal choices to maximize its utility. We denote the available alternatives to be a set of n exhaustive and exclusive utterances or actions $A = \{a_1, ..., a_i, ..., a_n\}$. The utility function $u(\cdot)$ describes the common preferences over the alternatives. For example, if a_i is more preferred than a_j, then $u(a_i) > u(a_j)$.

Our formation of human utilities takes the following two factors into consideration: (i) the task, speech, act or agent's emotion prediction, and (ii) the mental state constraints. As an example, since some actions could be impossible physically (one cannot drop an object if the agent is not carrying the object), the decision making process becomes a problem of maximizing the utility function that is subject to some constraints from the mental state, *i.e.*, $u(a|c)$, where c represents the context or constraints. Usually, we cannot find an analytical form of the utility function. However, what matters for preference ordering is which of the two options gives the higher expected utility, not the numerical values of those expected utilities.

The overall architecture of our proposed agent model is illustrated in Fig. 11.10. For each scenario, a setting description (Fig. 11.10 (a)) is provided by the LIGHT environment, which can include a description about the location, object affordances, agents' personas, and the objects that agents are carrying, wearing, or wielding. The free-form conversations, actions and emotions are logged during the communication as the observation history (Fig. 11.10 (b)). To begin with, a mental state parser will parse the setting descriptions into graph representation and initialize the agent's mental state (step 1 and 2). Besides the mental state updating, the parser also outputs an action mask that is aimed to rule out actions that are physically or causally impossible to take (step 3). A graph encoder (step 4) and a text encoder (step 5) will convert the mental state graph G_t

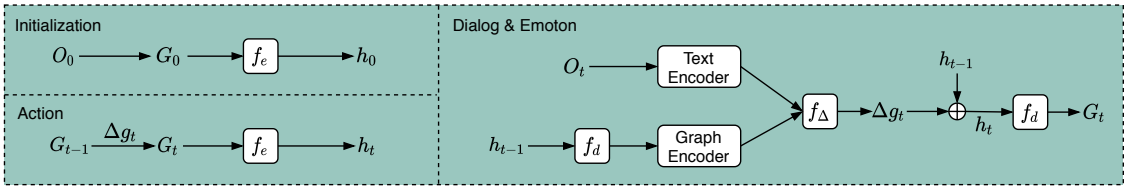


Figure 11.12: Overall Architecture of Hybrid Mind-State Parser

and the dialogue observation O_t into vector representations, respectively. The same text encoder will be used to encode the candidates C_t (step 6). In step 7, the context vectors are combined by a bi-directional attention aggregator, and each candidate is assigned a score with a Multi-layer Perceptron (MLP) (step 8). The action mask is then applied to get the feasible candidates under the current state-of-mind constraints (step 9). In step 10 and 11, the top three candidates from the last step will be fed into the utility model and re-ranked. Finally, the selected utterance/action/emotion is executed by the agent (step 12) and returned to the environment. Upon receiving the response from other agents in the environment, the new observation will be again parsed and used to update the agent's state of mind, and the cycle repeats. In the following, we will describe each component in more detail.

Mental State Modeling (Steps 1-2)

Fig. 11.12 describes the architecture of the mental state parser. The initial mental state graph G_0 is constructed by a ruled-based parser from the setting description O_0 and the graph is encoded by function f_e to a hidden state that is later used for graph update. At game step t, the mental state parser parses relevant information from observation O_t and update the agent's mental state from G_{t-1} to G_t. Considering that observations O_t typically convey incremental information from step $t-1$ to t, we generate the graph update Δg_t instead of the whole graph at each step

$$G_t = G_{t-1} \oplus \Delta g_t, \tag{11.34}$$

where $\oplus$ is the graph update operation. The graph update can be discrete and continuous, and there have been studies on the pros and cons of each updating method [589]. The discrete approach may suffer from an accumulation of errors but benefit from its interpretability. The continuous graph model needs to be trained from data, but it is more robust to possible errors. In this work, We propose a hybrid (discrete-continuous) method for updating the agent's state of mind by considering the characteristics of the LIGHT environment: since actions in LIGHT are template-based, it is more appropriate to adopt a discrete method for parsing; meanwhile, since utterances are challenging to be encoded into discrete representations, we apply a continuous update method instead.

11.6 Theory of Mind Inference in Games

Besides in real life scenarios, the significance of ToM can never be emphasized too much in games as well, which, most of the time, intrinsically involve multiple players and are partially observable to one or more of the them. Nonetheless, currently, most multiagent planning approaches focus on modeling other agents' policies based on only physical world states [590, 591, 592, 593]. To have more human-like agents that can interact with humans smoothly, we need to endow agents ToM, to be more specific, the ability to reason about other agents' mental states including their beliefs. This ability is very crucial in all kinds of multiagent games, such as cooperative games like Hanabi [594] and adversarial games like the policy-thief game we will later use as an example to explain the

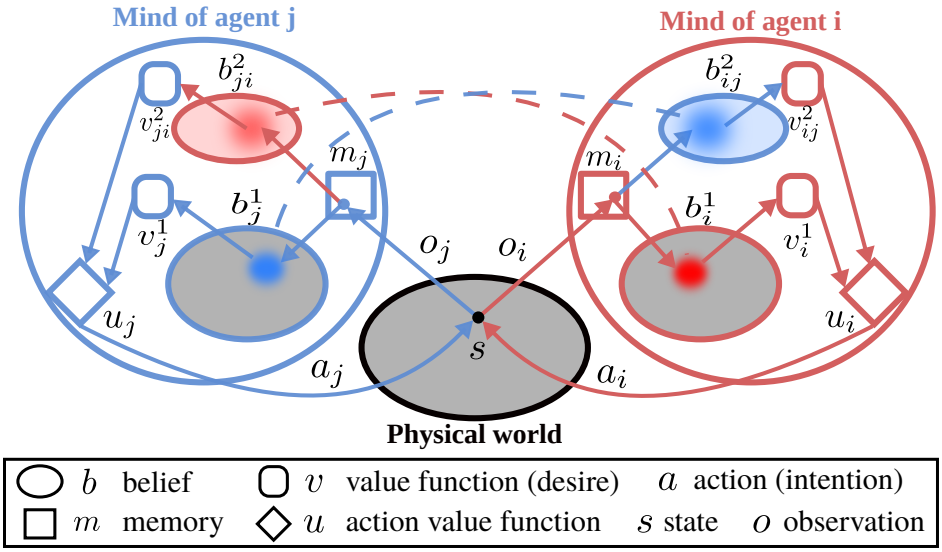


Figure 11.13: A ToM agent observes the world s and save the observation o into its memory m. The memory includes $o_{1:t}$ its past observations, and $a_{1:t-1}$ the performed actions. Based on the memory, it updates is belief b^1 about the world, belief b^2 about other agents' beliefs. Each level l of belief b^l has a corresponding value function $v^l(\cdot)$, and value functions are combined into an action value function u. Finally, the agent chooses an optimal action a based on u and changes the world state.

belief update modeling in ToM agents. In both types of the games, a player with ToM not only uses actions to interact with the world, but also intentionally manipulate other players' (teammates' or opponents') minds to achieve better performance. In general, the most common strategy is to reveal one's hidden information as much as possible to cooperative players and conceal, or even mislead, as much as possible to adversarial opponents [595].

We argue that there are two distinctions between ToM and non-ToM agents:

- ToM agents make predictions in a higher dimension (both physical and mental states).
- ToM agents intentionally change other agents' mental states, *i.e.*, beliefs about the world. In this way, a ToM agent changes others' future behaviors to maximize its own value [568].

Fig. 11.13 shows a ToM planning framework. In this framework, a ToM agent keeps beliefs (probability distributions) about the world state by Bayesian updates over time: at every time step, it updates the prior belief after performing an action and receiving an observation. It also keeps high-level beliefs about other agents' beliefs, which are represented by distributions of distributions. They are computed by a *nested Bayesian update*, which involves Bayesian updates of other agents' lower-level beliefs.

However, exact inference for nested belief updates is computationally very difficult for several reasons. These reasons include 1) other agents' model (*e.g.*, observation function) are required to compute the nested belief updates. 2) Belief updates for world states suffer from the curse of dimensionality, and nested belief updates aggravate this issue. In general, only approximate belief updates are computable.

To avoid the first problem, most methods assume that other agents' models are known [596, 597, 598, 599, 600, 601]. A more recent work [602] removes this constraint by adding a prior distribution to other agents' models.

To alleviate the second problem, various approximation methods for the belief update are proposed, including particle filter [598], state space pruning [597], and nested MDPs [603].

In this section, we take the policy-thief game as an example to illustrate a novel way of approxi-

mating the belief update, at the same time providing an alternative approach to learn other agents' models. By decomposing the belief update, we identify and approximate a "belief dynamics" term that is particularly computationally costly. The belief dynamics predicts how other agents' beliefs will change after one agent performs an action. In Section 11.6.1, we showed that this process can be modeled by a Markov probability transition. The transition kernel linearly transforms the current belief to a predicted belief, thus greatly reducing the computational complexity. We also showed that there always exists a kernel that transforms the true beliefs, hence theoretically this approximation can be exact. The kernel is also learnable, *e.g.*, by generator neural networks. Since this kernel characterizes another agent's belief update, agents essentially learn the other agents' models by learning their kernels.

In a police-thief game, the police agent needs to catch the thief while hiding its own identity from the thief. It can be shown that 1) ToM agents can accurately estimate the beliefs of other agents. 2) ToM agent can learn meaningful values over physical states and mental states. 3) Most importantly, they are able to intentionally change other agents' beliefs to achieve high values, outperforming other agents trained with state-of-the-art multiagent planning algorithms without modeling ToM.

11.6.1 Theory of Mind Belief Update

To act in uncertain environments due to noisy/partial observation, an agent tracks the physical world state s over time based on the actions it performed and its past observations. At each time step t, an agent i keeps a *belief* $b_{i,t}$, which is a probability distribution of the world state s_t given its memory $m_{i,t}$. The memory includes its past observations, $o_{i,1:t}$, and the performed actions, $a_{i,1:t-1}$.

In the rest of this section, we will use numbered superscripts to indicate the ToM level of the variables (*e.g.*, first level beliefs b^1). A first-order ToM agent tracks not only the physical world state, but also the mental states of other agents. Specifically, it tracks its state $s = (s^0, b^1)$ over time, where b^1 is all agents' first-level beliefs of the world state s^0. In other words, a first-order ToM agent i maintains two types of beliefs: first- and second-level beliefs. They are probability distributions of s^0 and b^1, respectively. The first-level belief is formulated as the probability of the world state given its past observation and actions:

$$b^1_{i,t}(s^0_t) = p(s^0_t | o_{i,1:t}, a_{i,1:t-1}) \tag{11.35}$$

The second-level belief $b^2_{ij,t}$ is defined as agent i's belief about agent j's first-level belief $b^1_{j,t}$:

$$b^2_{ij,t}(b^1_{j,t}) = p(b^1_{j,t} | o_{i,1:t}, a_{i,1:t-1}), \text{for } j \neq i. \tag{11.36}$$

At every time step t, an agent i updates its belief $b_{i,t-1}$ to $b_{i,t}$, according to the action $a_{i,t-1}$ performed at last time step and the observation $o_{i,t}$ received afterwards. This is called belief update, which we will discuss in details in the rest of this section. In general, the agent estimates the physical and mental states by Bayes filtering, which updates its belief each time an action is performed and a new observation arrives.

First-level Belief Update

The first-level belief of agent i about the world state at time t is $b^1_{i,t}(s^0_t) = p(s^0_t | o_{i,1:t}, a_{i,1:t-1})$. The Bayes filtering to update the belief at time t can be decomposed into a two-step process:

• *Prediction*. The agent updates its previous belief $b^1_{i,t-1}(s^0_{t-1})$ after taking an action $a_{i,t-1}$ by predicting how the state s^0_{t-1} will change:

$$\begin{aligned} p(s^0_t|o_{i,1:t-1}, a_{i,1:t-1}) &= \int_{s^0_{t-1}} p(s^0_t, s^0_{t-1}|o_{i,1:t-1}, a_{i,1:t-1}) ds^0_{t-1} \\ &= \int_{s^0_{t-1}} p(s^0_t|s^0_{t\text{-}1}, o_{i,t\text{-}1}, a_{i,t\text{-}1}) p(s^0_{t\text{-}1}|o_{i,1:t\text{-}1}, a_{i,1:t\text{-}2}) ds^0_{t\text{-}1} \\ &= \int_{s^0_{t-1}} \underbrace{p(s^0_t|s^0_{t-1}, o_{i,t-1}, a_{i,t-1})}_{\text{world dynamics}} \underbrace{b^1_{i,t-1}(s^0_{t-1})}_{\text{previous belief}} ds^0_{t-1} \end{aligned} \tag{11.37}$$

We can see that the agent updates its previous belief $b^1_{i,t-1}(s^0_{t-1})$ by applying a stochastic state transition function $p(s^0_t|s^0_{t-1}, o_{i,t-1}, a_{i,t-1})$. At this time, the agent has not received the new observation $o_{i,t}$. This prediction step computes how first-level beliefs will change after performing an action. Note that although we have uncertainty about the true world state, but this distribution transition process is deterministic for a certain time step.

• *Correction*. After receiving a new observation $o_{i,t}$, the agent corrects its prediction from the last step :

$$\begin{aligned} b^1_{i,t}(s^0_t) &= p(s^0_t|o_{i,t}, o_{i,1:t-1}, a_{i,1:t-1}) \\ &= \alpha\, p(s^0_t, o_{i,t}|o_{i,1:t-1}, a_{i,1:t-1}) \\ &= \alpha\, p(o_{i,t}|s^0_t) \underbrace{p(s^0_t|o_{i,1:t-1}, a_{i,1:t-1})}_{\text{first-level prediction}} \end{aligned} \tag{11.38}$$

where α is a normalizing constant as we apply the Bayes rule, $p(o_{i,t}|s^0_t)$ is the likelihood of observation $o_{i,t}$, and $p(s^0_t|o_{i,1:t-1}, a_{i,1:t-1})$ is the updated belief from the prediction step.

Second-level Belief Update

The second-level belief gets updated in a similar way. At each time step t, agent i maintains a second-level belief about agent: $b^2_{ij,t}(b^1_{j,t}) = p(b^1_{j,t}|o_{i,1:t}, a_{i,1:t-1})$, for $j \neq i$. The Bayes filtering is the same as the first-level belief update, except that the states are replaced by first-level beliefs:

• *Prediction*:

$$p(b^1_{j,t}|o_{i,1:t-1}, a_{i,1:t-1}) = \int_{b^1_{j,t-1}} \underbrace{p(b^1_{j,t}|b^1_{j,t-1}, o_{i,t-1}, a_{i,t-1})}_{\text{belief dynamics}} b^2_{ij,t-1}(b^1_{j,t-1}) db^1_{j,t-1} \tag{11.39}$$

Similar to the first-level belief update, the agent updates the previous belief $b^2_{ij,t-1}(b^1_{j,t-1})$ by applying a stochastic transition function $p(b^1_{j,t}|b^1_{j,t-1}, o_{i,t-1}, a_{i,t-1})$ defined on the first-level belief of agent j. We call this transition function the *belief dynamics*. It takes the observation as input since it contains information about the state, and the first-level belief changes differently under different states even if the same action is performed.

In principle, the belief dynamics should follow a similar Bayesian update as the first-level belief update of agent j itself. However, exact inference of the belief dynamics needs to marginalize out all possible j's observations and actions, and then perform a lower-level belief update. This is computationally very expensive or intractable. We discuss a proposed approximation method in Section 11.6.1.

• *Correction.* The belief is then updated by the observation:

$$\begin{aligned} b^2_{ij,t}(b^1_{j,t}) &= p(b^1_{j,t}|o_{i,t}, o_{i,1:t-1}, a_{i,1:t-1}) \\ &= \alpha\, p(o_{i,t}|b^1_{j,t}) \underbrace{p(b^1_{j,t}|o_{i,1:t-1}, a_{i,1:t-1})}_{\text{second-level prediction}} \end{aligned} \tag{11.40}$$

For simplicity, we still use α to represent the normalizing constant, with a value different from the first-level belief.

Belief Dynamics

The belief dynamics predicts how a belief of another agent, in the form of a probability distribution, will change stochastically when actions are performed over time. However, exact inference for nested belief update is computationally very expensive or intractable as discussed in Section 11.6.1. Approximated solutions have been proposed, such as particle filters [598] and bounded policy iteration [600]. Here we propose an alternative solution that is simple but effective.

The existing methods model the belief state transition in a general way similar to world state transitions (*e.g.*, particle filter). Let us denote the current belief as b_t, the predicted belief after performing an action as $\hat{b}_{t+1}$. The key observation of this method is that the Bayesian update process can always be described by a probability transition kernel. In other words, we can always find a probability transition kernel to transform b_t to $\hat{b}_{t+1}$. Formally, we have the following proposition.

Proposition 1. [3] *Let $\mathcal{S}$ be a measurable space and $s_t, s_{t+1} \in \mathcal{S}$. $\forall b_t(s_t) = p(s_t|o_{1:t}, a_{1:t-1})$ and $\hat{b}_{t+1}(s_{t+1}) = p(s_{t+1}|o_{1:t}, a_{1:t})$, there exists a Markov kernel $\kappa(s_t, s_{t+1}) \in \mathcal{S} \times \mathcal{S} \to [0, 1]$ such that (1) $\forall s_{t+1}$, $\hat{b}_{t+1}(s_{t+1}) = \int_{s_t} b_t(s_t)\kappa(s_t, s_{t+1})ds_t$, and (2) $\forall s_t$, $\int_{s_{t+1}} \kappa(s_t, s_{t+1})ds_{t+1} = 1$.*

Proof. Since s_t is independent of a_t, we have

$$\begin{aligned} \hat{b}_{t+1}(s_{t+1}) &= p(s_{t+1}|o_{1:t}, a_{1:t}) \\ &= \int_{s_t} p(s_{t+1}|s_t, o_{1:t}, a_{1:t}) p(s_t|o_{1:t}, a_{1:t}) ds_t \\ &= \int_{s_t} p(s_{t+1}|s_t, o_{1:t}, a_{1:t}) p(s_t|o_{1:t}, a_{1:t-1}) ds_t \\ &= \int_{s_t} p(s_{t+1}|s_t, o_{1:t}, a_{1:t}) b_t(s_t) ds_t \end{aligned}$$

Obviously $\forall s_t$, $\int_{s_{t+1}} p(s_{t+1}|s_t, o_{1:t}, a_{1:t}) = 1$.

Hence $\kappa(s_t, s_{t+1}) = p(s_{t+1}|s_t, o_{1:t}, a_{1:t})$ is a kernel that satisifies the conditions. □

This introduces a general linear form for the belief dynamics that can be extended to higher levels. Representing the beliefs by vectors, we can re-write the belief dynamics as:

$$\hat{b}^1_{j,t} = \kappa_{t-1} b^1_{j,t-1} \tag{11.41}$$

Parametrized by observations and actions, the kernel $\kappa(o, a)$ transforms the previous belief to the predicted belief. The kernel $\kappa(o, a)$ is to be learned. For example, it can be generated by a generator neural network. The kernel acts on the belief linearly, but the kernel itself is non-linear

[3] For simplicity, we omit the agent indices in the proposition.

in its parameters (*i.e.* the observations and actions). Due to its simple and general form, it is quite powerful and computationally favorable.

Since it is difficult to learn a perfect kernel, we can further add a noise term to this belief transition:

$$\hat{b}^1_{j,t} = \kappa_{t-1} b^1_{j,t-1} + \epsilon \tag{11.42}$$

Assuming the noise follows a multivariate Gaussian distribution as a convenient approximation, we have $\epsilon \sim \mathcal{N}(0, \Sigma_D)$ where Σ_D is the covariance matrix. Equivalently, we have $\hat{b}^1_{j,t} \sim \mathcal{N}(\kappa_{t-1} b^1_{j,t-1}, \Sigma_D)$ where $\kappa_{t-1} \triangleq \kappa(o_{i,t-1}, a_{i,t-1})$. After adding the noise, we normalize $b^1_{j,t}$ to ensure that it remains to be a distribution.

Given the above belief dynamics, we can efficiently compute the second-level belief update. If at the last time step we have $b^1_{j,t-1} \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$, then the prediction step (Eq. (11.39)) is a convolution of two Gaussian distributions. Hence the result will still be a Gaussian distribution. According to the second-level belief prediction (Eq. (11.39)) and correction (Eq. (11.40)), we have the following conclusions:

• The *prediction* step gives $b^1_{j,t} \sim \mathcal{N}(\mu_p, \Sigma_p)$, where

$$\mu_p = \kappa_{t-1} \mu_{t-1} \tag{11.43}$$

$$\Sigma_p = \Sigma_D + \kappa_{t-1} \Sigma_{t-1} \kappa_{t-1} \tag{11.44}$$

Hence we can obtain the μ_p and Σ_p by applying the transition kernel κ_{t-1}. To compute the correction step, we need a observation model. We can first learn an *belief estimation model* $\widetilde{b^1_{j,t}} = f(o_{i,t})$ that directly estimates the beliefs from observations up to a Gaussian noise: $\widetilde{b^1_{j,t}} \sim \mathcal{N}(b^1_{j,t}, \Sigma_o)$. Then we can rewrite the observation model $p(o_{i,t}|b^1_{j,t})$ as $p(\widetilde{b^1_{j,t}}|b^1_{j,t})$, which is a Gaussian distribution. Then the correction step is computing the predictive posterior given a Gaussian prior and a Gaussian likelihood. Since a multivariate Gaussian is the conjugate prior of itself, we have:

• The *correction* step gives $b^1_{j,t} \sim \mathcal{N}(\mu_t, \Sigma_t)$, where

$$\mu_t = (\Sigma_p^{-1} + \Sigma_o^{-1})^{-1} (\Sigma_p^{-1} \mu_p + \Sigma_o^{-1} \mu_o) \tag{11.45}$$

$$\Sigma_t = (\Sigma_p^{-1} + \Sigma_o^{-1})^{-1} \tag{11.46}$$

where μ_o is directly predicted by the belief estimation model. Σ_o is learned in the training phase of the belief estimation model by computing the covariance between predicted beliefs and ground truth beliefs.

The above equations give an efficient way to compute the second-level belief update. The above process agrees with the Kalman filter. When the state space is continuous, the belief is a continuous function instead of a discrete vector. Then infinite-dimensional Kalman filter can be adopted [604].

11.6.2 Theory-of-mind Planning

Based on the belief about the world state and other agents' mental states, a ToM agent chooses an optimal action based on a value function. This value function is defined on the belief space, and it is convex and piecewise-linear [568]. Usually this function is learned by value iteration. However, this is very hard due to the curse of dimensionality.

In the reinforcement learning literature, it is common to learn approximated value functions [605]. Here we approximate this value function by a linear function of the beliefs with intuitive semantic meanings. Specifically, we define the first-level value function as:

$$v^1_i(b^1_i) = \int_{s^0} b^1_i(s^0) v^0_i(s^0) ds^0 \tag{11.47}$$

where $v_i^0(s^0)$ is a value function defined on true world states. The second-level value function $v_{ij}^2(b_{ij}^2)$ measures agent i's value of its second-level belief b_{ij}^2 on agent j's belief. The second-level value function is similarly defined as:

$$\begin{aligned} v_{ij}^2(b_{ij}^2) &= \int_{s^0} \int_{b_j^1} b_{ij}^2(b_j^1) b_j^1(s^0) v_{ij}^0(s^0) db_j^1 ds^0 \\ &= \int_{s^0} \underbrace{\int_{b_j^1} b_{ij}^2(b_j^1) b_j^1(s^0) db_j^1}_{\text{Expectation of first-level belief}} v_{ij}^0(s^0) ds^0 \\ &= \int_{s^0} E_{b_{ij}^2}[b_j^1(s^0)]\, v_{ij}^0(s^0) ds^0 \end{aligned} \tag{11.48}$$

It is grounded to the actual mental states by the zero-level value function $v_{ij}^0(s^0)$, which is the value of agent i when agent j believes that the state is in s^0 with probability 1. Notice the difference between $v_i^0(s^0)$ and $v_{ij}^0(s^0)$: $v_i^0(s^0)$ is defined on the physical state while $v_{ij}^0(s^0)$ is defined on the mental state.

Then the first-order ToM agent i at time t chooses an optimal action $a_{i,t}^*$ that maximizes its future value, which combines the first- and second-level value functions:

$$\begin{aligned} a_{i,t}^* &= \arg\max_{a_{i,t}} u_i^1(b_{i,t}^1, a_{i,t}) + \sum_{j \neq i} u_i^2(b_{ij,t}^2, a_{i,t}) \\ &= \arg\max_{a_{i,t}} \underbrace{\int_{b_{i,t+1}^1} p(b_{i,t+1}^1 | b_{i,t}^1, a_{i,t}) v_i^1(b_{i,t+1}^1) db_{i,t+1}^1}_{\text{expected future physical state value}} \\ &\quad \underbrace{+ \sum_{j \neq i} \int_{b_{ij,t+1}^2} p(b_{ij,t+1}^2 | b_{ij,t}^2, a_{i,t}) v_{ij}^2(b_{ij,t+1}^2) db_{ij,t+1}^2}_{\text{expected future mental state value}} \end{aligned} \tag{11.49}$$

This way the agent considers the change of other agents' beliefs after taking an action. This is particularly important since it enables the first-order ToM agent to intentionally change other agents' beliefs. This changes others' future behaviors and thus maximizes the agent's own value.

11.6.3 Learning

Learning Belief Dynamics

To predict the future beliefs, the agent generates the transition kernel $\kappa(o_{i,t-1}, a_{i,t-1})$ given its observation $o_{i,t-1}$ and action $a_{i,t-1}$. Each column of this transition kernel κ sums to 1. This can be practically implemented as a generator neural network, which takes the observation and action as inputs and generates all the columns for κ. To ensure that each column sums to 1, a softmax activation function can be added before the final output of each column.

Learning Estimation Model

The belief estimation model $\widetilde{b_{j,t}^1} = f(o_{i,t})$ estimates the first-level belief of agent j given agent i's observation at time t. This can be learned by any classification model that outputs a probability for each state given the observation. The model can be trained by optimizing the difference (*e.g.* mean squared error, cross entropy) between the ground truth and estimated belief of agent j.

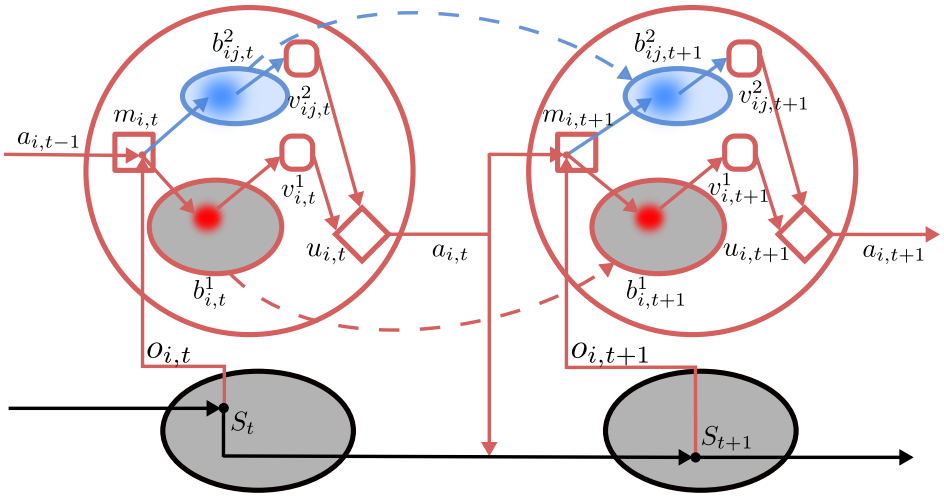


Figure 11.14: The decision-making process along time. A first-order ToM agent updates its first- and second-level beliefs at each time step after performing an action and receiving an observation. The agent then makes decision according to its updated belief and performs a new action.

Learning Zero-level Value Functions

Since the beliefs of agent i (*i.e.*, $b^1_{i,t}$ and $b^2_{ij,t+1}$) themselves are updated deterministically by Bayes filtering, the stochastic belief prediction $p(b^1_{i,t+1}|b^1_{i,t}, a_{i,t})$ in Eq. (11.49) is given by the Dirac delta function $\delta(b^1_{i,t+1} - \hat{b}^1_{i,t+1})$, where $\hat{b}^1_{i,t+1}$ is the belief after the performing the prediction step. Hence for first-level future value function we have:

$$\begin{aligned} u^1_i(b^1_{i,t}, a_{i,t}) &= \int_{b^1_{i,t+1}} p(b^1_{i,t+1}|b^1_{i,t}, a_{i,t}) v^1_i(b^1_{i,t+1}) db^1_{i,t+1} \\ &= \int_{b^1_{i,t+1}} \delta(b^1_{i,t+1} - \hat{b}^1_{i,t+1}) v^1_i(b^1_{i,t+1}) db^1_{i,t+1} \\ &= v^1_i(\hat{b}^1_{i,t+1}(a_{i,t})) = \int_{s^0} \hat{b}^1_{i,t+1}(s^0) v^0_i(s^0) ds^0 \end{aligned} \tag{11.50}$$

where $\hat{b}^1_{i,t+1}$ is the predicted future belief by performing the prediction step in the Bayes filtering after taking action $a_{i,t}$. Similarly, for the second-level future value function:

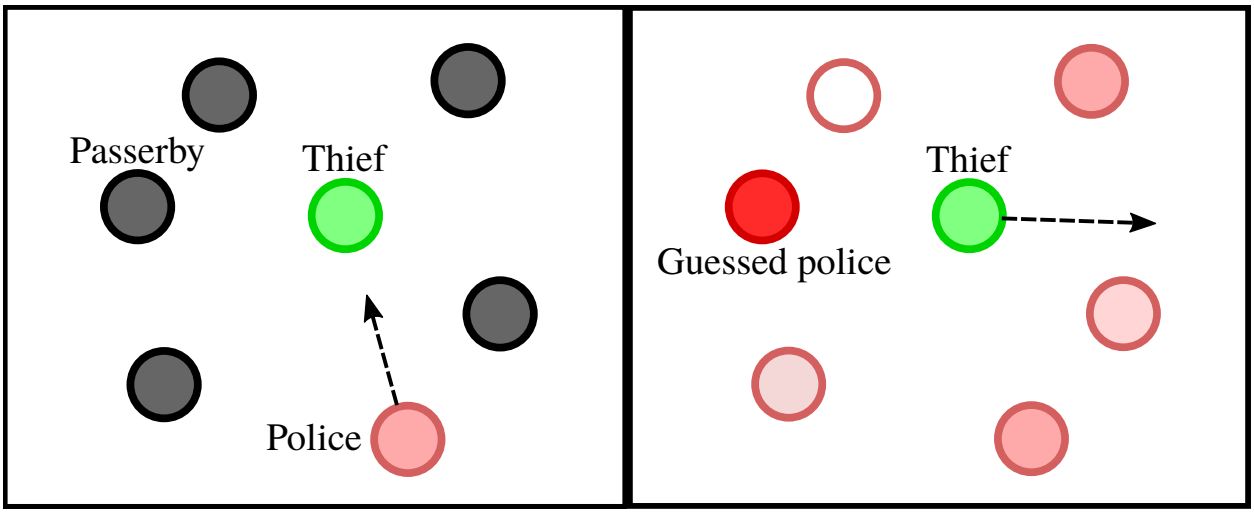
$$\begin{aligned} u^2_i(b^2_{ij,t}, a_{i,t}) &= v^2_{ij}(\hat{b}^2_{ij,t+1}(a_{i,t})) \\ &= \int_{s^0} E_{\hat{b}^2_{ij,t+1}}[b^1_{j,t+1}]\, v^0_{ij}(s^0) ds^0 \end{aligned} \tag{11.51}$$

Finally, the combined value function can be re-written as an inner product of beliefs and zero-level value functions:

$$q^1_i(b^1_{i,t}, a_{i,t}) + \sum_{j \neq i} q^2_i(b^2_{ij,t}, a_{i,t}) = \langle \hat{\boldsymbol{b}}, \boldsymbol{v^0} \rangle \tag{11.52}$$

where $\hat{\boldsymbol{b}}$ is the concatenation of $\hat{b}^1_i(s^0)$ and $E_{\hat{b}^2_{ij,t+1}}[b^1_{j,t+1}]$ for all $j \neq i$, and $\boldsymbol{v^0}$ is the concatenation of $v^0_i(s^0)$ and $v^0_{ij}(s^0)$ for all $j \neq i$.

From reinforcement learning's perspective, this formulation is a function approximation of the action-value functions. Specifically, $\hat{\boldsymbol{b}}$ can be interpreted as the feature vector extracted from the state, and $\boldsymbol{v^0}$ is the weight. Hence learning the zero-level value functions can be achieved by applying



(a) Ground truth (b) Thief's belief

Figure 11.15: The game setting. (Fig. 11.15a) shows the ground truth identities of all agents. (Fig. 11.15b) shows the thief's belief of each agent being the police (player). Darker colors indicates higher probabilities.

existing algorithms [605, 606, 607, 608] to learn the weights for a linearly approximated value function.

11.6.4 Example: Police-thief Game

We test the ToM agents in a police-thief game as shown in Fig. 11.15. There are three types of agents in this game: a police (the player) controlled by the tested algorithm, a thief (the game engine) controlled by a zero-order ToM policy, and several passersby that wander around randomly. The goal of the police is catching the thief by colliding with it, while the goal of the thief is to escape until the game exceeds a maximum time limit. In this game, the thief does not know which agent is the police, so it needs to maintain a belief of every agent being the police to escape. The police knows who is the thief, but it needs to hide its identity to prevent the thief from escaping. We benchmark the police's success rate on catching the thief to evaluate the multiagent planning methods.

Environment

We adopt the multiagent Particle Environment [609], which is a two-dimensional world with continuous space and discrete time. We use a closed-world setting: an agent will reappear on the opposite side of the world after crossing a boundary. The environment is physically simulated: an agent can accelerate, and agents can collide with each other. The observation for an agent is the positions of other agents, and the allowed actions are the accelerations in four directions. The police receives a 1.0 reward when it collides with the thief, otherwise a -0.1 reward at each time step.

Thief's Belief Update

As mentioned above, we use a zero-order ToM thief to recognize potential police and escape. The thief maintains a first-level belief about each agent's identity. In the beginning, the belief for each agent is uniform (0.5 probability being police for every agent). At each time step, the belief is updated by Bayes filtering given by Eq. (11.37) and Eq. (11.38). Since an agent's identity cannot

be changed by the thief's actions, the prediction step has no effect on the belief. For the correction step, we compute the observation likelihood as $p(o_{i,t}|s_t^0) = p(\theta|\mu, k)$. Here θ is the relative angle between the velocity of an agent and the line connecting the agent and the thief. $p(\theta|\mu, k) = \frac{e^{kcos(\theta-\mu)}}{2\pi I_0(k)}$ is the von Mises distribution. At every time step, the thief runs away from the agent that has the highest probability being the police.

We designed two settings in our experiments: slow-thief and fast-thief. In the slow thief setting, the thief will have a smaller acceleration than that of the police, and a directly chasing policy is sufficient to catch the thief. In the fast thief setting, the thief can easily escape from a direct chaser.

Comparative Methods

There has been a booming interest in the AI community to build algorithms that incorporate ToM into multiagent systems [134, 262]. For example, Bayesian Theory of Mind (BToM) [134, 558] predicts the mental states of humans. [610] proposed ToMnet, a neural network to predict the characteristic of an observed agent and its future behaviors. These are perception models that do not involve planning.

The most representative series of work on ToM planning is I-POMDP [596, 597, 598, 599, 600, 601], which extends the traditional POMDPs. It augments world states to interactive states to include beliefs of the intentional model of other agents (belief, reward function, observation function, *etc.*). However, solving I-POMDP can be extremely expensive and inefficient. The generalization to interactive states greatly increases the dimension of the state space, and this curse of dimensionality is exacerbated by the nested belief reasoning among agents.

Methods have been proposed to approximate I-POMDP. For example, I-PF [598] approximates the belief updates by particle filters. I-PBVI [597] constrains the interactive state space by computing a finite set of beliefs reachable from the initial belief over a certain horizon. I-POMDP Lite [603] uses a nested MDP to model other agents to approximate the exact I-POMDP policy.

The above approximation methods all assume that the agent knows exactly the models, except beliefs, of other agents. A more recent work [602] removes this constraint. It performs belief update using Bayesian inference and particle filtering by sampling other agents' models from prior distributions.

We compare the proposed method (full / ablated) with an existing ToM method and a state-of-the-art multiagent reinforcement learning algorithm. We use the following algorithms to learn the police's policy for benchmarking:

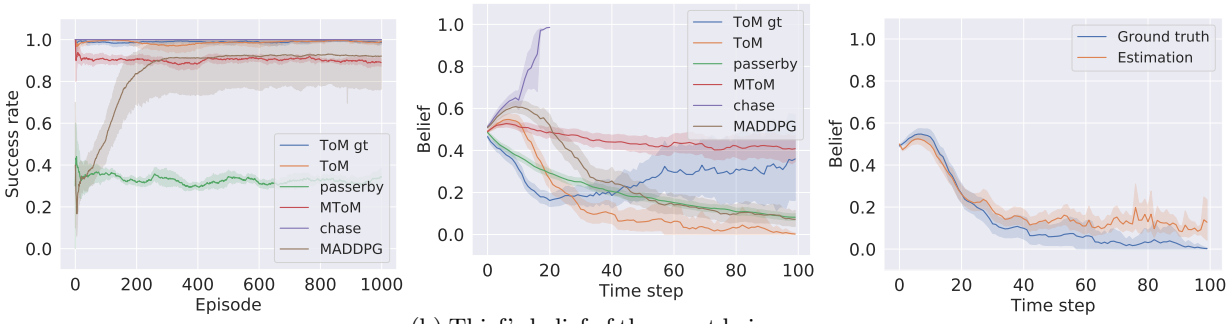
• ToM-gt. This is the ToM planning agent given the ground truth belief of the thief (computed by Bayesian filtering) at every time step. It does not perform belief updates, but it needs to learn the zero-level value functions. This serves as an ablative version of the full model.

• ToM. This is the full model. It learns the zero-level value functions, belief dynamics, and belief estimation models. It performs belief updates by Bayesian filtering. To learn the zero-level value functions, we use true online TD(λ) [608]. Therefore the value functions are estimated and updated at every time step during each episode. We use a 3-layer fully connected neural network for both the transition kernel generator and the belief estimation model. The size of hidden layers is 64 for the kernel generator, and 32 for belief estimation.

• MADDPG [591]. This is a state-of-the-art multiagent reinforcement learning algorithm that extends deep deterministic policy gradient (DDPG) [611] to a multiagent setting.

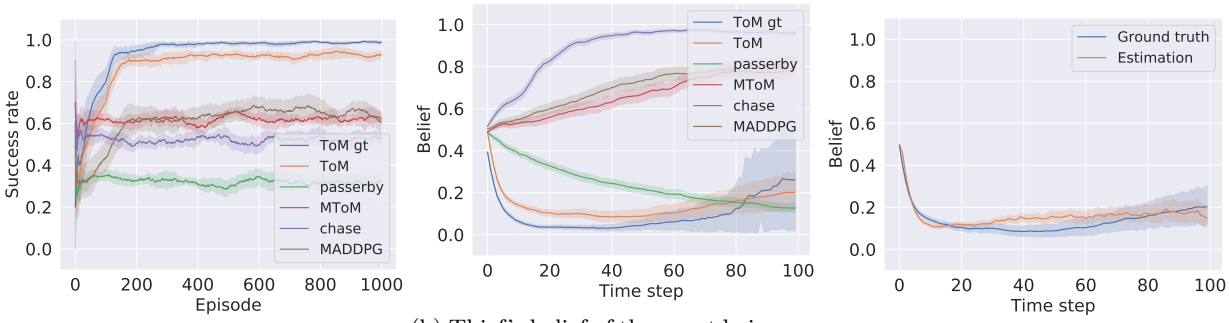
• MToM [612]. Another theory-of-mind planning approach. However, this approach only models the belief of policies rather than the belief of world states. We use the deep Q-Network (DQN) [321] for the initial policy estimation of MToM.

• Direct chasing. This agent always chooses the action that will minimize the distance between



(a) Average success rate (b) Thief's belief of the agent being police (c) Estimated belief

Figure 11.16: Experiment results in the **slow**-thief setting. Fig. 11.16a shows the average success rate of catching the thief in the training stage. Fig. 11.16b shows the probability within an episode that the thief predicts the agent is the police. The curves are averaged over 20 final episodes during training. From the beliefs we can see that the strategies of different methods vary, but they can achieve similar performance (as shown in Fig. 11.16b) in the slow-thief setting. Fig. 11.16c shows the estimated belief of the thief by the full ToM model within an episode, averaged over the final 20 episodes during training. The estimated beliefs are obtained by the two-level belief update, which employs a learned belief dynamics and belief estimation model. From the figure we can see that the ToM agent is able to estimate the thief's belief within a small error. All the plots use data generated from 10 independent training trials.



(a) Average success rate (b) Thief's belief of the agent being police (c) Estimated belief

Figure 11.17: Experiment results in the **fast**-thief setting. Fig. 11.17a shows that the ToM agent significantly outperforms other methods, achieving a success rate close to 1. Fig. 11.17b shows that the ToM agent intentionally lowers the thief's belief thus it achieves the high success rate. The other methods learn to chase the thief, hence they are recognized by the thief (the belief goes higher as time evolves). Fig. 11.17c shows that the ToM agent is able to estimate the thief's belief within a small error.

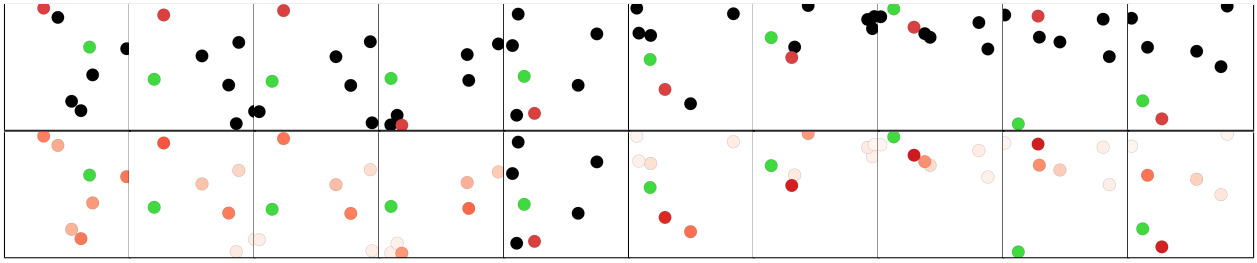
itself and the thief.

• Passerby. The same random walker as the other passersby.

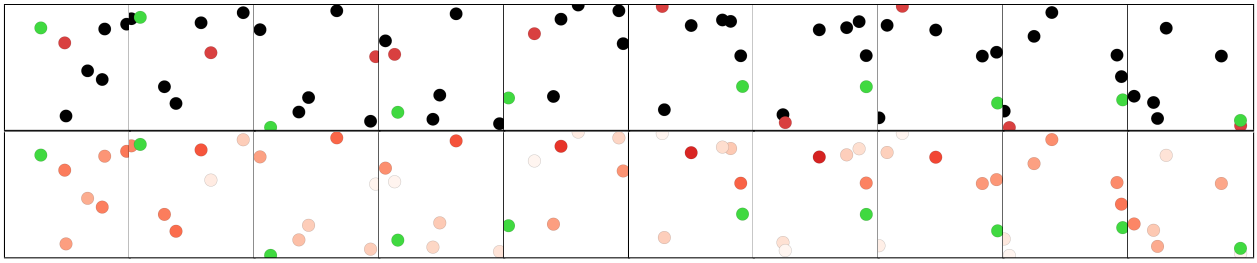
For simplicity, we only use the positions of the thief and the police itself as observations for every method. We also compute the distance between these two positions as a feature. The maximum episode length is 100, and each algorithm is trained for 1000 episodes. Qualitative results are shown in Fig. 11.18.

Empirical Results

We show some quantitative experiment results in Fig. 11.16 and Fig. 11.17. In both settings, we can see the proposed method achieves a high success rate and outperforms other methods.

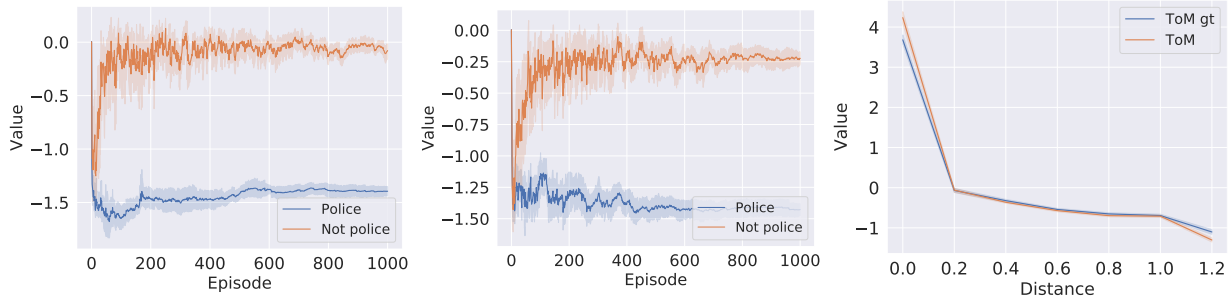


(a) An example episode played by MADDPG [591]



(b) An example episode played by ToM

Figure 11.18: Qualitative results in the **fast**-thief setting. Each group of two rows shows in the first row the ground truth state, and in the second row the beliefs of the thief along time (from left to right). The green one is the thief and the red one is the police. Darker colors indicate higher beliefs. Agents appear on the opposite side when they cross the boundary. Fig. 11.18a The thief recognizes the MADDPG agent as the police and escapes. Fig. 11.18b The ToM agent successfully hide its identity and catches the thief.



(a) Values of belief in ToM-gt training (b) Values of belief in ToM training (c) Value function of distance by ToM

Figure 11.19: Learned zero-level value functions in the **fast**-thief setting. Fig. 11.19a and Fig. 11.19b shows the learned values on the mental state learned by ToM-gt and ToM, *i.e.* the value of the ToM agent that the thief thinks it is/is not the police. The ToM agent successfully learns that being recognized as a police has a lower value. Fig. 11.19c shows the final learned value function of the distance between the ToM agent and the thief. The ToM agent learns that when the thief is very close, there is a high value to directly approach the thief. When the thief is far away, getting closer will not gain much value.

Slow-thief setting In the slow-thief setting, different methods show different belief curves (Fig. 11.16b) but they achieve similar performance (Fig. 11.16a). This is because the thief is easy to catch in this setting, so different methods converge to different types of policies. The value functions for ToM converges very quickly (within an episode) since we use true online TD(λ). It updates the value functions at every time step during every episode. In the beginning, the belief dynamics and estimation model are inaccurate, but it does not affect the performance. Even if the thief correctly recognizes the police, it can be caught by direct chasing. Typically, the thief is caught within 20 steps for all methods, and we can see a belief raise in that phase. In some episodes, the agent fails to catch the thief due to some random behavior. Therefore the curves go down after 20 steps similarly

to the curve of the passerby (a random walker).

Fast-thief setting In the fast-thief setting as a contrast, the police needs to hide its identity to catch the thief with a high success rate. From Fig. 11.17b we can see that the comparative non-ToM methods all converge to a greedy chasing behavior in this setting. The thief recognizes the police and escapes. The success rate is around 0.6 for non-ToM methods (including direct chasing), which is slightly higher than random walk that has a success rate of 0.4.

On the other hand, the ToM method achieves a success rate close to 1 by hiding its identity. From the learning curve (Fig. 11.17a) we can see that it takes some time for the ToM agent to learn the belief dynamics and estimation model to achieve the final performance. The ToM agent can intentionally lower the belief of the thief to achieve its goal. Comparing with the passerby (a random walker) we can see that the belief curve of the ToM agent has a sharp decline in the beginning. Hence this lowering is not a consequence of random movements. As a final result, the ToM agent significantly outperforms all other methods.

In both settings, the second-level belief update estimates the belief of the thief quite accurately as shown in Fig. 11.16c and Fig. 11.17c. Qualitative results are shown in Fig. 11.18 and Fig. 11.19 shows that the proposed algorithm learns meaningful zero-level value functions for both the physical state and the mental state.

11.6.5 Summary

In this section, we discussed the ToM integration in multiagent games and proposed a novel way to model the nested belief update, which alleviates the computation problem and provides an alternative way to learn other agents' models. It has been shown that the nested belief update can be modeled a a Markov probability transition, leading to a linear transformation. The transition kernel is learnable and provides an efficient nested belief update. In the illustrate the effectiveness of our approach in an adversarial multiagent game, the police-thief game and showed that the ToM agent can successfully learn other agents' belief updates and intentionally change other agents' beliefs to achieve its goal.

11.7 Theory of Mind in Practical Life

Theory of mind (ToM) has been studied by the field of cognitive science for several decades. However, most of previous work considers ToM in toy examples, such as simulation environment with several dots or geometric shapes. It is challenging to study the ToM in real scenes since the environment of the real scenes are too complex, the states cannot be defined accordingly, and the mental states of real agents are especially complicated to be modeled.

In this section we introduce a cognitive system which aims at applying the ToM in real scenarios to understand the false-belief behavior by joint inference of object states, robot knowledge, and human beliefs.

11.7.1 False Belief

Sally-Anne [613] is a seminal psychological test regarding human's social cognition in understanding *false-belief*—the ability to understand other's belief about the world may contrast with true reality. A cartoon version of the flagship implementation of Sally-Anne test is shown in the left of Fig. 11.20: Sally puts her marble in the box and left. While Sally is out, Anne moves the marble from the box to a basket. The test asks where Sally would look for her marble when she is back. In this

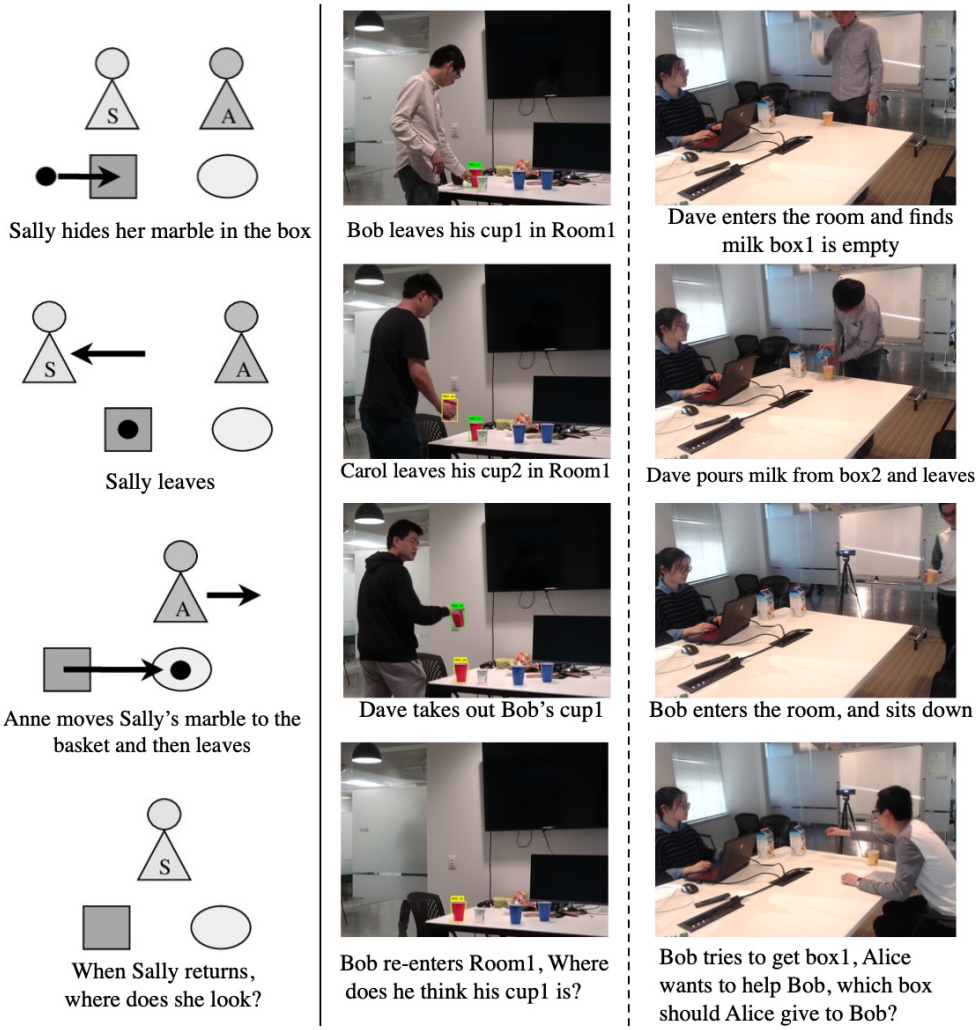


Figure 11.20: Left: The classic Sally-Anne test [613]. Middle and Right: Two false-belief scenarios in our dataset: belief test and helping test.

experiment, the marble would be inside the box according to Sally's false-belief, despite the fact that the marble is actually inside the basket. To answer this question correctly, a subject or an algorithm should understand and disentangle the object state (observation from the current frame), the (accumulated) knowledge, the belief of other agents, the ground truth/reality of the world, and most importantly, the concept of false-belief.

In order to endow the ability to understand false-belief to a robot system, the system should enable the following three capabilities with increasing depth in cognition:

- *Tracking small objects with occlusions across different views.* The objects in indoor environment (*e.g.*, cups) are usually small and have the similar appearance. Such objects are frequently occluded during human interactions. Additionally, each robot's camera view has few overlaps with others. The proposed method can address such challenging multi-view multi-object tracking problem.
- *Inferring human beliefs.* The state of an object normally does not change unless a human interacts with it. By identifying the interactions between human and objects, our system not only improves its tracking accuracy through a joint reasoning algorithm, but also supports the high-

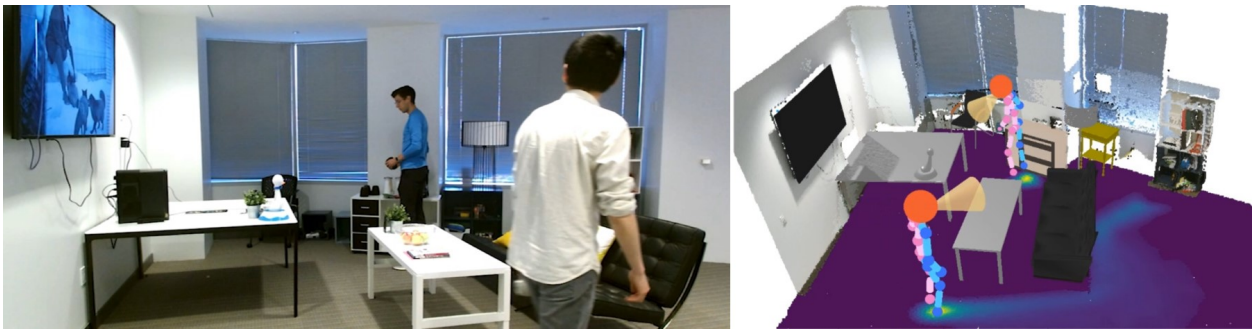


Figure 11.21: An illustration of modeling ToM in real life scenario

level cognitive capability; *e.g.*, knowing which object is interacted with which person, whether a person knows the state of the object has been changed.

- *Helping human by recognizing false-belief.* Giving the above tracking and reasoning results, the proposed algorithm can infer what a person's belief about the environment at a certain time, thereby capable of knowing whether and why the person has false-belief, so as to better assist the person given a specific context.

11.7.2 Cognitive Platform

To facilitate machines with ToM reasoning abilities from visual input, we first need to give machines deeper understanding about the scene. Initial scene understanding includes object detection and human pose estimation but we further figure out that all understanding needs to be in 3D environment to ease the ambiguity in 2D image plane, especially when distinguishing what and where a person is looking at or pointing to, which are fundamental aspects in belief updating and communication. See Fig. 11.21 as an example. Reasoning about Tom is built upon many modules discussed in previous chapters.

The first step includes what we have introduced in Chapter 1 about the 3d scene and human parsing. Then gaze estimation in the wild by incorporating information about head pose and pupil is essential for estimating where people's attention is in every time step. On the higher level what a person is doing is detected by action detector and further be parsed by image grammars as described in Chapter 4, which is followed by intention prediction and path planning.

Another problem worth to mention is that the robustness of mind reasoning relies heavily on the computer vision modules mentioned above, together with human and object tracking and ReID, especially in heavy occlusion cases.

With all the belief updates, 3D environment and intention prediction, now the machine can correct agents' false belief and even perform parental help.

Chapter 12

Explainable AI

12.1 Introduction

From low risk environments such as movie recommendation systems and chatbots to high risk environments such as self-driving cars, drones, military applications and medical-diagnosis and treatment, Artificial Intelligence (AI) systems are becoming increasingly ubiquitous [614, 615, 616, 617]. AI is finding its way into a wide array of applications in education, finance, healthcare, telecommunication, and law enforcement. In particular, AI systems built using black box machine learning (ML) models—such as deep neural networks and large ensembles [618, 619, 620, 621, 622, 623, 624, 625, 626]—perform remarkably well on a broad range of tasks and are gaining widespread adoption. However understanding the behavior of these systems remains a significant challenge as they cannot explain why they reached a specific recommendation or a decision. This is especially problematic in high risk environments such as banking, healthcare, and insurance, where AI decisions can have significant consequences. Therefore, much hope rests on explanation methods as tools to understand the decisions made by these AI systems.

Explainable AI (XAI) models, through explanations, make the underlying inference mechanism of AI systems transparent and interpretable to expert users (system developers) and non-expert users (end-users) [618, 619, 620, 627]. Explanations play a key role in integrating AI machines into our daily lives, *i.e.*, XAI is essential to increase social acceptance of AI machines. As the decision making is being shifted from humans to machines, **transparency** and **interpretability** achieved with reliable explanations is central to solving AI problems such as Safety (*e.g.*, *how to operate self-driving cars safely*), Bias & Fairness (*e.g.*, *how to detect and mitigate bias in ML models*), Justified Human Trust in ML models (*e.g.*, *how to trust the output of these AI systems to inform our decisions*), Model Debugging (*e.g.*, *how to improve my model by identifying points of model failure*), and Ethics (*e.g.*, *how to ensure that ML models reflect our values*) (Fig. 12.1).

In this chapter, we focus mainly on measuring and increasing **Justified Positive Trust** (JPT) and **Justified Negative Trust** (JNT) [628] in AI systems. We measure JPT and JNT by evaluating the human's understanding of the machine's (M) decision-making process. For example, let us consider an image classification task. Suppose if the machine M predicts images in the set C correctly and makes incorrect decisions on the images in the set W. Intuitively, JPT will be computed as the percentage of images in C that the human subject felt M would correctly predict. Similarly, JNT (also called as mistrust), will be computed as the percentage of images in W that the human subject felt M would fail to predict correctly. Note that this definition of justified positive and negative trust is domain generic and can be applied to any task. For example, in an AI-driven clinical world, our definitions of JPT and JNT can effectively measure how much doctors and patients understand

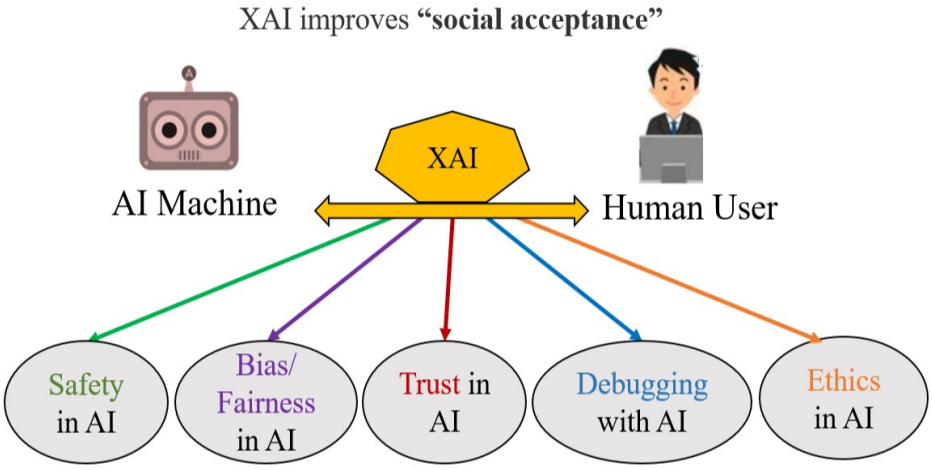


Figure 12.1: An AI machine that explains its predictions to human users will find more social acceptance. Therefore, explainable AI (XAI) models are the key in addressing the issues such as Safety in AI, Bias / Fairness in AI, Trust in AI, Model Debugging, and Ethics in AI.

the AI systems that assist in clinical decisions.

12.1.1 Introducing X-ToM: Explaining with Theory-of-Mind for Increasing JPT and JNT

Our work, in this chapter, is motivated by the following three key observations:

- **Attention is not a Good Explanation:** Previous studies have shown that trust is closely and positively correlated to the level of how much human users understand the AI system—*understandability*—and how accurately they can predict the system's performance on a given task—*predictability* [627, 618, 628, 620]. Therefore there has been a growing interest in developing explainable AI systems (XAI) aimed at increasing understandability and predictability by providing explanations about the system's predictions to human users [618, 619, 620, 621]. Current works on XAI generate explanations about their performance in terms of, *e.g.*, feature visualization and attention maps [622, 623, 624, 625, 626, 629]. However, solely generating explanations, regardless of their type (visualization or attention maps) and utility, *is not sufficient* for increasing understandability and predictability [630]. We verify this in our experiments (see Section 12.6).
- **Explanation is an Interactive Communication Process:** We believe that an effective explanation cannot be one shot and involves iterative process of communication between the human and the machine. The context of such interaction plays an important role in determining the utility of the follow-up explanations [631]. As humans can easily be overwhelmed with too many or too detailed explanations, interactive communication process helps in understanding the user and identify user-specific content for explanation. Moreover, cognitive studies [620] have shown an explanation can only be optimal if it is generated by taking user's perception and belief into account.
- **Defining a Collaborative Task for the Communication Process:** In our experiments, we found that it is difficult to evaluate the effectiveness of explanations without constraining the communication process. In our framework, we constrain the communication by explicitly defining a collaborative task for the human user to solve through the explanations. Based on how many tasks that are successfully solved by the user (and the number of explanations in the dialog), we

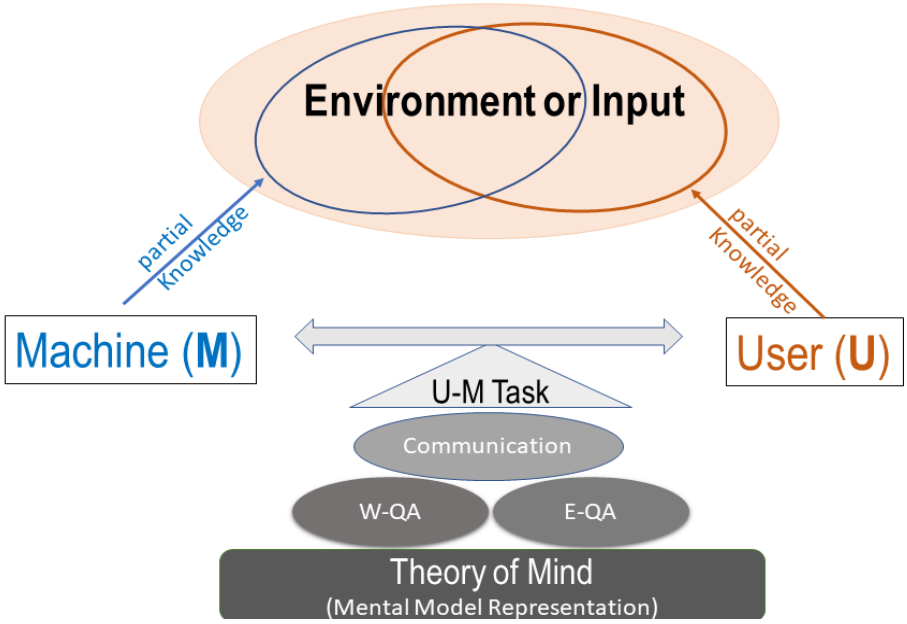


Figure 12.2: **XAI as Collaborative Task Solving**: Our interactive and collaborative XAI framework based on the Theory of Mind. The interaction is conducted through a dialog where the user poses questions about facts in the environment (W-QA) and explanation seeking questions (E-QA).

measure the effectiveness of the explanations.

Based on the above three key observations, we introduce an interactive explanation framework, **X-ToM**. In our framework, the machine generates sequence of explanations in a dialog which takes into account three important aspects at each dialog turn: (a) human's intention (or curiosity); (b) human's understanding of the machine; and (c) machine's understanding of the human user. To do this, we use Theory of Mind (ToM) which helps us in explicitly modeling human's intention, machine's mind as inferred by the human as well as human's mind as inferred by the machine. The ability to reason about other's perception and beliefs, in addition to one's own perception and beliefs, is often referred to as the Theory-of-Mind [632, 633, 555].

More specifically, in X-ToM, the machine and the user are positioned to solve a collaborative task, but the machine's mind (M) and the human user's mind (U) only have a partial knowledge of the environment (see Fig. 12.2). Hence, the machine and user need to communicate with each other, using their partial knowledge, otherwise they would not be able to optimally solve the collaborative task. The communication consists of two different types of question-answer (QA) exchanges—namely, a) Factoid question-answers about the environment (W-QA), where the user asks "WH"-questions that begin with `what`, `which`, `where`, and `how`; and b) Explanation seeking question-answers (E-QA), where the user asks questions that begin with `why` about the machine's inference. At each turn in the collaborative dialog, our X-ToM updates a model of human perception and beliefs, and uses this model for optimizing explanations in the next turn.

We argue that our interactive explanation framework based on ToM is practical and more natural for both expert and non-expert users to understand the internal workings of complex machine learning models. Furthermore, we also show that ToM facilitates in quantitatively measuring justified human trust in the machine by comparing all the three mental representations. To the best of our knowledge, this is the first work to derive explanations using ToM.

We applied our framework to three visual recognition tasks, namely, image classification, action recognition, and human body pose estimation. Using Amazon Mechanical Turk, we have collected

explanation dialogs by interacting with turkers through X-ToM framework. From there, X-ToM learned an optimal explanation policy that takes into account user perception and beliefs. Through our extensive human studies, we show that X-ToM allows the user to achieve a high success rate in visual recognition on blurred images, and does so very efficiently in a few dialog exchanges. We also found that the most popularly used attribution based explanations (viz. saliency maps) are not effective to improve human trust in AI system, whereas our Theory-of-Mind inspired approach significantly improves human trust in AI by providing effective explanations.

12.1.2 Contributions

The contributions of this work are threefold: (i) a new interactive XAI framework based on the Theory-of-Mind; (ii) a new collaborative task-solving game in the domain of visual recognition for learning collaborative explanation strategies; and (iii) a new objective measure of trust and quantitative evaluation of how humans gain increased trust in a given vision system.

12.2 Related Work

Generating explanations or justifications of predictions or decisions made by an AI system has been widely explored in AI. Most prior work has focused on generating explanations using feature visualization and attribution.

Feature visualization techniques typically identify qualitative interpretations of features used for making predictions or decisions. Recently, there has been an increased interest in developing feature visualizations for deep learning models, especially for Convolutional Neural Nets (CNNs) in computer vision applications, and Recurrent Neural Nets (RNNs) in NLP applications. For example, gradient ascent optimization is used in the image space to visualize the hidden feature layers of unsupervised deep architectures [634]. Also, convolutional layers are visualized by reconstructing the input of each layer from its output [624]. Recent visual explanation models seek to jointly classify the image and explain why the predicted class label is appropriate for the image [635]. Other related work includes a visualization-based explanation framework for Naive Bayes classifiers [636], an interpretable character-level language models for analyzing the predictions in RNNs [637], and an interactive visualization for facilitating analysis of RNN hidden states [638].

Attribution is a set of techniques that highlight pixels of the input image (saliency maps) that most caused the output classification. Gradient-based visualization methods [639, 640] have been proposed to extract image regions responsible for the network output. The LIME method proposed by [619] explains predictions of any classifier by approximating it locally with an interpretable model. Influence measures [641] have been used to identify the importance of features in affecting the classification outcome for individual data points.

More recently, apart from feature visualization and attribution techniques, other important lines of research in explainable AI explore dimensionality reduction techniques [642, 643] and focus on building models which are intrinsically interpretable [644, 645]. There are few recent works in the XAI literature that go beyond the pixel-level explanations. For example, the TCAV technique proposed by [646] aims to generate explanations based on high-level user defined concepts. Contrastive explanations are proposed by [647] to identify minimal and sufficient features to justify the classification result. [648] proposed counterfactual visual explanations that identify how the input could change such that the underlying vision system would make a different decision. More recently, few methods have been developed for building models which are intrinsically interpretable [644]. In addition, there are several works [620, 649, 650] on the goodness measures of explanation which aim to understand the underlying characteristics of explanations.

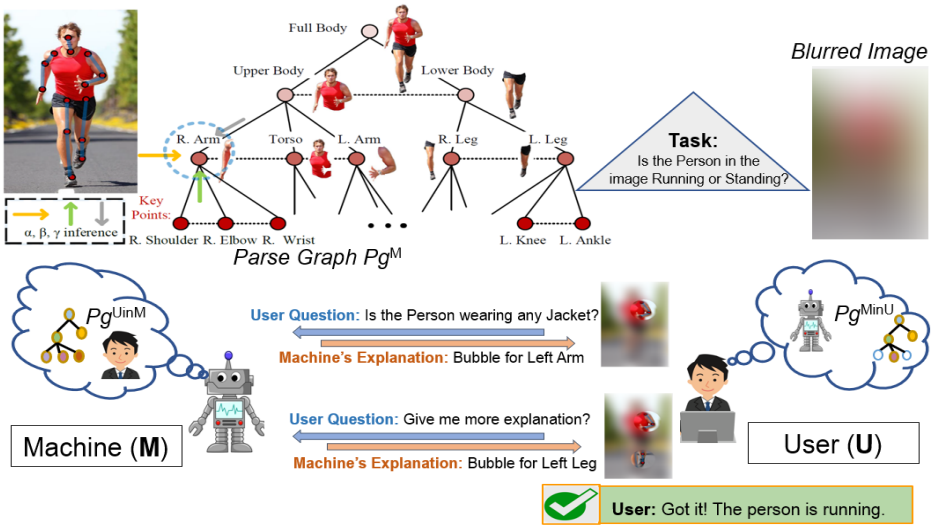


Figure 12.3: An example of the first phase of an X-ToM game aimed at estimating pg^{UinM}: The user is shown a blurred image and given a task to recognize if the person in the image is running or walking. X-ToM has access to the original (unblurred) image and pg^{M}. The user then asks questions regarding objects and parts in the image. Using the detections in pg^{M}, X-ToM provides visual explanations as "bubbles" that reveal the corresponding image parts in the blurred image. The generated explanations are used to update pg^{UinM}.

12.3 X-ToM Framework

Our X-ToM consists of three main components:

- A **Performer** that generates image interpretations (*i.e.*, machine's mind represented as pg^{M}) using a set of computer vision algorithms;
- An **Explainer** that generates maximum utility explanations in a dialog with the user by accounting for pg^{M} and pg^{UinM} using reinforcement learning;
- An **Evaluator** that quantitatively evaluates the effect of explanations on the human's understanding of the machine's behaviors (*i.e.*, pg^{MinU}) and measures human trust by comparing pg^{MinU} and pg^{M}.

12.3.1 X-ToM Game

An X-ToM game consists of two phases. The first phase is the collaborative task phase. The user is shown a blurred image and given a task to recognize what the image shows. X-ToM has access to the original (unblurred) image and the machine's (*i.e.*, **Performer's**) inference result pg^{M} (see Section 12.3.2). The user is allowed to ask questions regarding objects and parts in the image that the user finds relevant for his/her own recognition task. Using the detected objects and parts in pg^{M}, X-ToM **Explainer** provides visual explanations to the user, as shown in Fig. 12.3. This process allows the machine to infer what the user sees and iteratively update pg^{UinM}, and thus select an optimal explanation at every turn of the game (see Section 12.3.3). Optimal explanations generated by the **Explainer** are the key to maximize the human trust in the machine.

The second phase is specifically designed for evaluating whether the explanation provided in the first phase helps the user understand the system behaviors. The **Evaluator** shows a set of original (unblurred) images to the user that are similar to (but different from) the ones used in the first phase of the game (*i.e.*, the set of images shows the same class of objects or human activity). The user is then given a task to predict in each image the locations of objects and

parts that would be detected by the machine (*i.e.*, in pg^M) according to his/her understanding of the machine's behaviors. Based on the human predictions, the **Evaluator** estimates pg^{MinU} and quantifies human trust in the machine by comparing pg^{MinU} and pg^M (see Section 12.3.4).

12.3.2 X-ToM Performer (for Image Interpretation)

In this work, the visual tasks involve detecting and localizing human body parts, identifying their poses and attributes, and recognizing human actions from a given image. The AOG for this visual domain uses AND nodes to represent decompositions of human body parts into subparts, and OR nodes for alternative decompositions. Each node is characterized by attributes that pertain to the corresponding human body part, including the pose and action of the entire body. Also, edges in the AOG capture hierarchical and contextual relationships of the human body parts.

Our AOG-based performer uses three inference processes α, β and γ at each node. Fig. 12.3 shows an example part of the AOG relevant for human body pose estimation [651]. The α process detects nodes (*i.e.*, human body parts) of the AOG directly based on image features, without taking advantage of the surrounding context. The β process infers nodes of the AOG by binding the previously detected children nodes in a bottom-up fashion, where the children nodes have been detected by the α process (*e.g.*, detecting human's upper body from the detected right arm, torso, and left arm). Note that the β process is robust to partial object occlusions as it can infer an object from its detected parts. The γ process infers a node of the AOG top-down from its previously detected parent nodes, where the parents have been detected by the α process (*e.g.*, detecting human's right leg from the detected outline of the lower body). The parent node passes contextual information so that the performer can detect the presence of an object or part from its surround. Note that the γ process is robust to variations in scale at which objects appear in images.

12.3.3 X-ToM Explainer (for Explanation Generation)

The explainer, in the first phase of the game, makes the underlying α, β, and γ inference process of the performer more transparent to the human through a collaborative dialog. At one end, the explainer is provided access to an image and the performer's inference result pg^M on that image. At the other end, the human is presented a blurred version of the same image, and asked to recognize a body part, or pose, or human action depicted (*e.g.*, whether the person is running or walking). To solve the task, the human may ask the explainer various "what," "where" and "how" questions (*e.g.*, "Where is the left arm in the image.") We make the assumption that the human will always ask questions that are related to the task at hand so as to solve it efficiently. The explainer answers these questions using pg^M and justifies the answers by showing the corresponding visual explanations in the image (as illustrated in Fig. 12.4).

As visual explanations, we use "bubbles" [47], where each bubble reveals a circular part of the blurred image to the human. The bubbles coincide with relevant image parts for answering the question from the human, as inferred by the performer in pg^M. For example, a bubble may unblur the person's left leg in the blurred image, since that image part has been estimated in pg^M as relevant for recognizing the human action "running" occurring in the image.

Following the "principle of least collaborative effort" [652] and the aforementioned findings [620] that explanations should *not* overwhelm the human, our X-ToM explainer utilizes pg^M and pg^{UinM} (*i.e.*, the contextual and hierarchical relationships explicitly modeled in the AOG) for controlling the depth and breadth of explanations. To enable this control, each bubble is characterized by a number of parameters, including the amount of image reveal (*i.e.*, the unblurring level), size, and location in the image, to name a few. We use reinforcement learning to train the explainer to

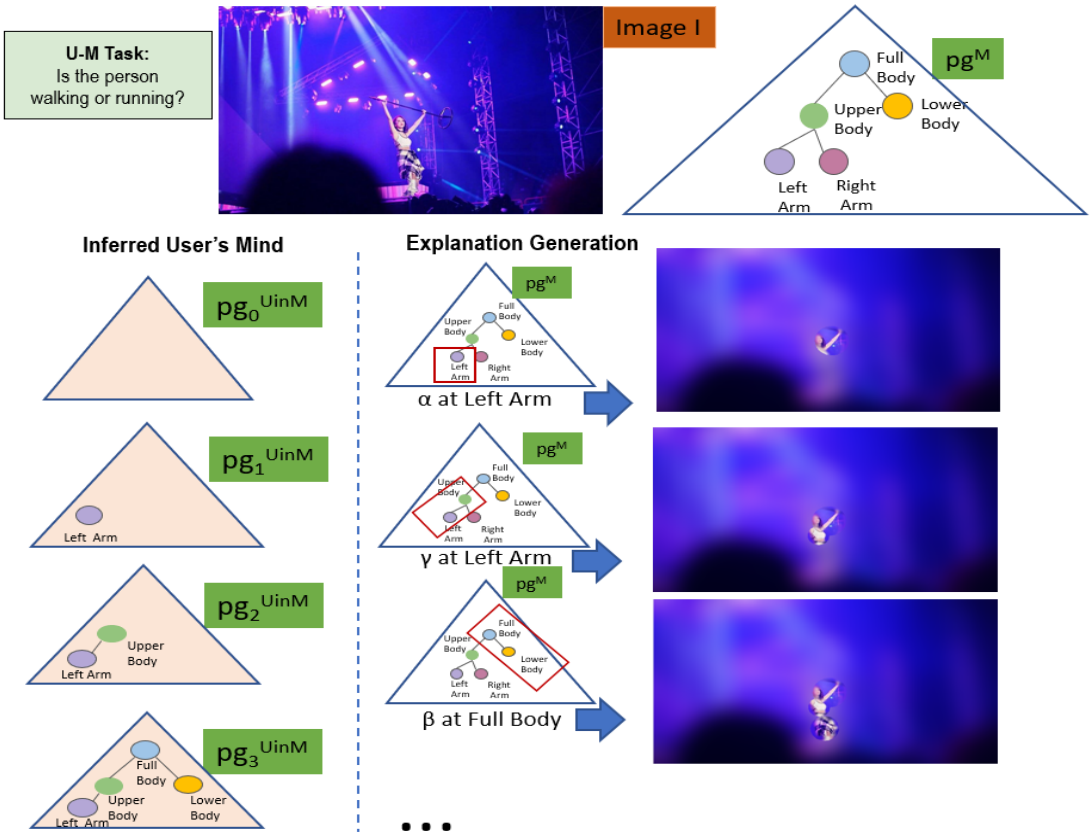


Figure 12.4: Illustration of the first phase in X-ToM game. The human is asked to solve the task "Is the person in the image walking or running?" The human may ask questions related to body parts and body poses. The machine reveals a bubble (of various sizes and scales) for each of those questions. The figure shows examples of explanations generated using α, β and γ processes and the updated inferred user's mind after each explanation.

optimize these parameters and thus provide optimal visual explanations.

12.3.4 X-ToM Evaluator (for Trust Estimation)

The second phase of the X-ToM game serves to assess the effect of the explainer on the human's understanding of the performer. This assessment is conducted by the evaluator. The human is presented with a set of (unblurred) images that are different from those used in the first phase. For every image, the evaluator asks the human to predict the performer's output. The evaluator poses multiple-choice questions and the user clicks on one or more answers (see Section 12.11.2 for more details on evaluator interface and questions). As shown in Fig. 12.5, we design these questions to capture different aspects of human's understanding of α, β and γ inference processes in the performer. Based on responses from the human, the evaluator estimates pg^{MinU}. By comparing pg^{MinU} with the actual machine's mind pg^{M} (generated by the performer), we have defined the following qualitative and quantitative metrics to quantitatively assess human trust [627, 653, 628, 654] in the performer:

- **Quantitative Metrics**:
  (1) *Justified Positive and Negative Trust:* It is possible for humans to feel positive trust with respect to certain tasks, while feeling negative trust (*i.e.*, mistrust) on some other tasks. The positive and negative trust can be a mixture of justified and unjustified trust [627, 628]. We

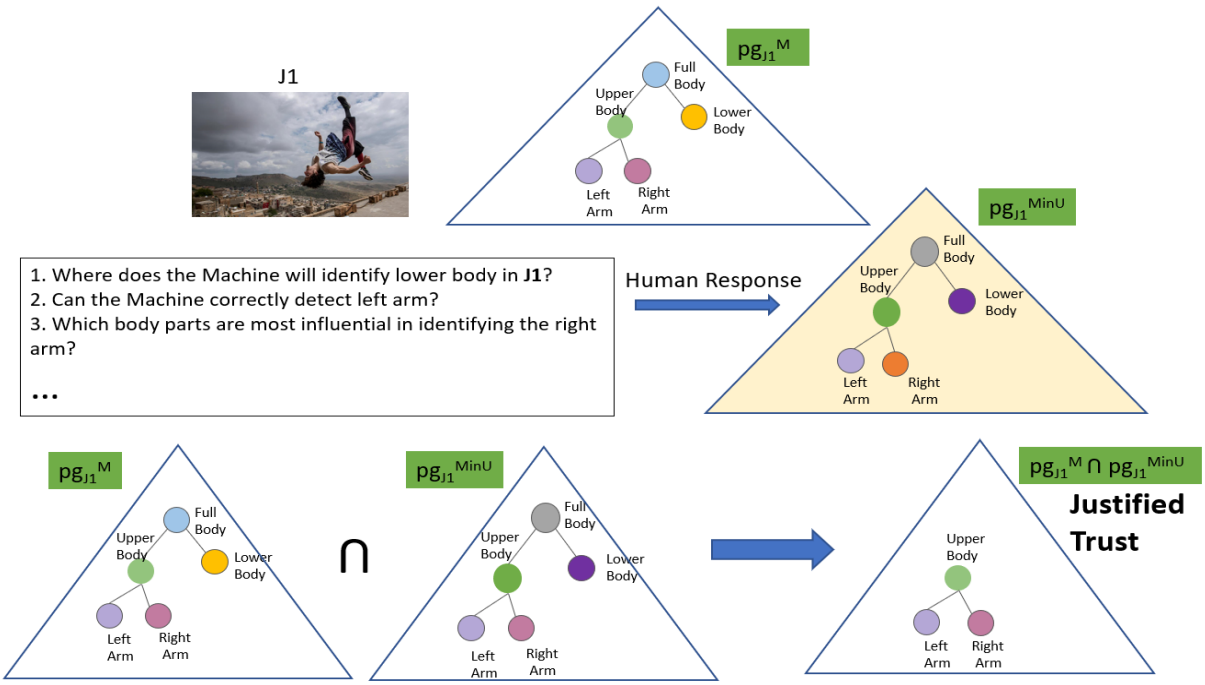


Figure 12.5: An example of second phase of X-ToM game where we estimate pg^{MinU} and also quantitatively compute justified trust.

compute justified positive trust (JPT) and negative trust (JNT) as follows:

$$\text{JPT} = \frac{1}{N}\sum_{i}\sum_{z=\alpha,\beta,\gamma} \Delta\text{JPT}(i,z),$$

$$\Delta\text{JPT}(i,z) = \frac{\|pg^{MinU}_{i,z,+} \cap pg^{M}_{i,+}\|}{\|pg^{M}_{i,+}\|},$$

$$\text{JNT} = \frac{1}{N}\sum_{i}\sum_{z=\alpha,\beta,\gamma} \Delta\text{JNT}(i,z),$$

$$\Delta\text{JNT}(i,z) = \frac{\|pg^{MinU}_{i,z,-} \cap pg^{M}_{i,-}\|}{\|pg^{M}_{i,-}\|},$$

where N is the total number of games played. z is the type of inference process. $\Delta\text{JPT}(i,z)$, $\Delta\text{JNT}(i,z)$ denote the justified positive and negative trust gained in the i-th turn of a game on the z inference process respectively. $pg^{MinU}_{i,z,+}$ denotes nodes in pg^{MinU}_{i} for which the user thinks the performer is able to accurately detect in the image using the z inference process. Similarly, $pg^{MinU}_{i,z,-}$ denotes nodes in pg^{MinU}_{i} for which the user thinks the performer would fail to detect in the image using the z inference process. $\|pg\|$ is the size of pg. Symbol $\cap$ denote the graph intersection of all nodes and edges from two pg's.

(2) *Reliance:* Reliance (Rc) captures the extent to which a human can accurately predict the performer's inference results without over- or under-estimation. In other words, Reliance is proportional to the sum of JPT and JNT.

$$\text{Rc} = \frac{1}{N}\sum_{i}\sum_{z=\alpha,\beta,\gamma}\Delta\text{Rc}(i,z),$$
$$\Delta\text{Rc}(i,z) = \frac{\|pg_{i,z}^{MinU} \cap pg_{i,z}^{M}\|}{\|pg_{i}^{M}\|}.$$

- **Qualitative Metrics**:
  (3) *Explanation Satisfaction (ES)*. We measure users' feeling of satisfaction at having achieved an understanding of the machine in terms of usefulness, sufficiency, appropriated detail, confidence, accuracy, and consistency. We ask them to rate each of these metrics on a Likert scale of 0 to 9.

12.4 Representation of Minds in X-ToM

The three minds pg^M, pg^{MinU}, and pg^{UinM} are sub-graphs of an And-Or Graph (AOG) defining all objects, parts, and their relationships and attributes of the visual domain considered. Our motivation to use AOGs for modeling the three mental states of the Theory of Mind stems from the following advantages. First, an AOG is a context-sensitive stochastic grammar [154] that can explicitly capture rich contextual and hierarchical relationships (spatial, temporal and causal). Second, AOG based representation and inference is a domain generic approach and the literature has abundantly demonstrated that AOG based systems, especially recent methods that combine deep learning and AOGs, are the top performers for a wide range of tasks in domains such as computer vision, natural language processing, and human-robot collaboration [644, 655, 656, 657, 651]. Third, since the result of visual recognition (*i.e.*, a parse graph) is a sub-graph of the AOG, image interpretations can be readily explained using the top-down, bottom-up, or contextual types of visual reasoning enabled by the AOG. Finally, and of great importance for XAI systems, the rich contextual and hierarchical nature of AOGs allows for formalizing and quantitatively evaluating human trust in the visual performer along both depth and breadth.

As AOG is interpretable, why not show Pg^M directly to the user as an explanation? It will be daunting to show the entire AOG since our AOG encodes hundreds of objects, parts, activities, attributes and other concepts as nodes. In addition, AOG has numerous edges. It might be possible to visualize a part of AOG, but it is not clear how to optimize which AOG subgraph would not overwhelm the user and maximize utility. The advantage of using our dialog based explanations is that, at each dialog turn, explainer can tailor the explanations based on the user's current perception and understanding (Miller *et al.*, 2017).

12.5 Learning X-ToM Explainer Policy

Given the following input: image I, task T assigned to the human, dialog history h_i of a sequence of generated bubbles, and question from the user q_i selected from a finite set of allowed questions $\mathcal{Q}(T)$ for task T, the explainer estimates an optimal explanation e_i at dialog turn i as

$$e_i = \arg\max_{e} U\left(e \mid pg^M, pg_i^{UinM}, q_i, h_i, T, I; \theta\right)$$

where U denotes the utility function parameterized by θ. The set of questions $\mathcal{Q}(T)$ is automatically generated from all concepts (objects, object parts, human activities, object attributes, *etc.*) that

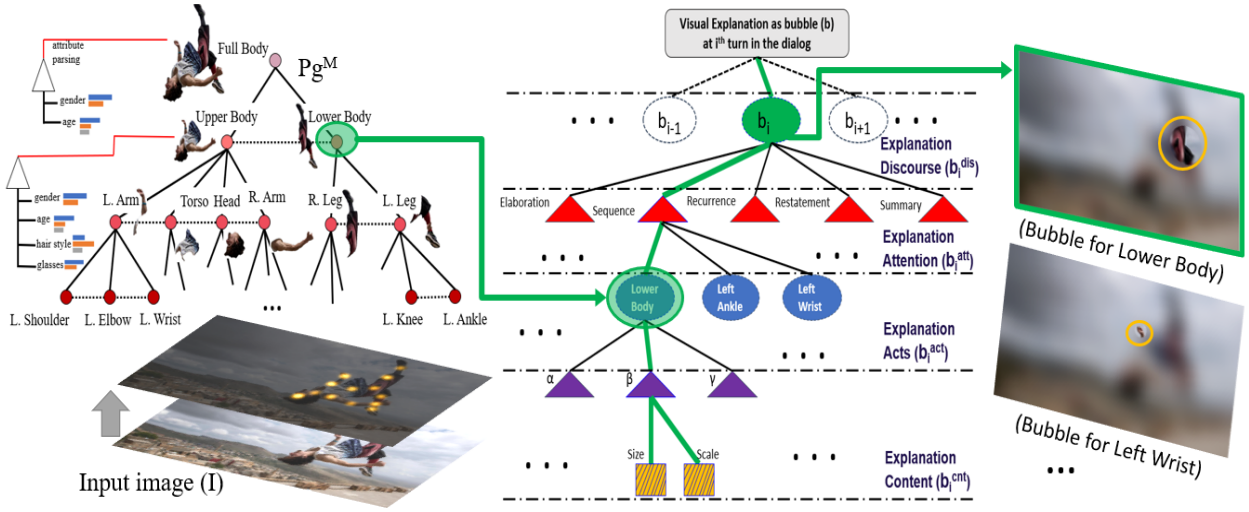


Figure 12.6: **Left**: The Machine interprets the image I as Pg^M; **Middle**: Hierarchical representation of the bubble using the four parameters: explanation content (b^{cnt}), explanation attention (b^{act}), explanation acts (b^{att}) and explanation discourse (b^{att}); **Right**: The Human receives visual explanations—bubbles—optimized by the X-ToM Explainer.

may appear in the image and are also modeled by the Performer. During interaction, the user is prompted to ask a question from this list[1].

As defined earlier, pg_i^{UinM} denotes the current estimate of human's mind, which is an empty graph without nodes and edges at the beginning of the X-ToM game. At every turn in the dialog, the explainer infers and updates pg_i^{UinM} by maximizing its posterior distribution based on h_i, T and q_i. Using a Bayesian approach, we define the posterior of pg_i^{UinM} as

$$p\left(pg_i^{UinM} \mid h_i, q_i, T\right) \propto$$
$$p\left(q_i \mid h_i, pg_i^{UinM}, T\right) p\left(h_i \mid pg_i^{UinM}, T\right) p\left(pg_i^{UinM}, T\right)$$

where $p\left(pg_i^{UinM}, T\right)$ is specified as a uniform prior. The likelihoods $p\left(q_i \mid h_i, pg_i^{UinM}, T\right)$ and $p\left(h_i \mid pg_i^{UinM}, T\right)$ are estimated based on the frequency of occurrence of the question $q = q_i$ and the dialog history $h = h_i$ over many X-ToM games played with human users. After updating pg_i^{UinM}, the selection of an optimal bubble, *i.e.*, explanation e_i, is cast as a sequential decision-making problem and formalized using reinforcement learning (RL). Below we specify the state, actions, reward, and policy of the RL framework.

RL State (s_i). The state of the explainer at dialog turn i consists of pg^M, pg_i^{UinM}, q_i, and h_i.

RL Action (a_i). The action space consists of all possible bubbles that can be generated from pg^M so that they reveal relevant image parts in the blurred image to the human. Each bubble b is characterized by the following four groups of parameters, as illustrated in Fig. 12.6:

(a) **Explanation Content**, b^{cnt}, is defined as the amount of visual information contained in the bubble. Our X-ToM uses the Gaussian scale-space [658] for measuring b^{cnt}. Specifically, we model "space" as a Gaussian with variance σ_1^2 governing the length of the radius (*i.e.*, spatial size) of the bubble. Also, we model "scale" as a Gaussian with variance σ_2^2 governing the amount of image unblur that the bubble reveals to the user. Given σ_1^2 and σ_2^2, we compute b_i^{cnt} as the differential

[1] A NLU component can be added to map users' free-form natural language questions to the list of interpretable questions.

entropy

$$b^{cnt} = 1 + \frac{1}{2}\log(4\pi^2\sigma_1^2\sigma_2^2)$$

Intuitively, a bubble with large "space" (*i.e.*, large size) and large "scale" (*i.e.*, high resolution) reveals a lot of information about the image. Conversely, a bubble with small "space" and "scale" reveals very little evidence. If the explainer always chose bubbles with small "space" and "scale," it would lead to inefficient dialogue for solving the task. On the other hand, if the explainer always chose bubbles with large space and scale, it would distract the human with unnecessary information and make it difficult for the human to understand the machine's internal representation and inference[2]. Thus, the explainer's goal is to find the bubble with an optimal b^{cnt}. In this work, we discretize "space" and "scale" of bubbles using $\sigma_1 \in \{1.15, 3.15, 4.5\}$, and $\sigma_2 \in \{1, 9, 15\}$.

(b) **Explanation Acts**, b^{act}, parametrizes the three types of visual explanations (*i.e.*, bubbles) that can be presented to the human, corresponding to the three inference processes in our AOG-based performer. Specifically, b^{act} can be: α, β, or γ explanation act. Note that using β and γ explanation acts (*i.e.*, bottom-up and top-down inference processes of the performer) allows for increasing depth of explanations.

(c) **Explanation Attention**, b^{att}, indexes a particular human body part from pg^M that is the current focus of the dialog with the human. In the work, the AOG explicitly models human body parts and their subparts, where pg^M infers only a subset of those appearing in the image.

(d) **Explanation Discourse**, b^{dis}, parametrizes discourse relations of the bubbles generated along the dialog with the human. In this work, we account for the dialog discourse for enforcing coherence among the explanations. In our experiments, we found the following five discourse relations [659, 652] to be sufficient and helpful:

- **Elaboration**. If bubble b_{i+1} provides additional details (*e.g.*, by increasing "scale" or "space") relative to the previous bubbles $h_i = b_{1...i}$, then b_{i+1} relates to the dialog history h_i with the *elaboration* relationship.
- **Sequence**. If the explanation attention b^{att} of bubble b_{i+1} is not part of the dialog history h_i, then b_{i+1} relates to h_i with the *sequence* relationship.
- **Recurrence**. If bubble b_{i+1} already exists in h_i, then the discourse relationship between b_{i+1} and h_i is called *recurrence*.
- **Restatement**. If the dialog history h_i already contains a bubble with the same explanation attention b^{att} as b_{i+1}, then b_{i+1} relates to h_i with the *restatement* relationship.
- **Summary** is a special case of the elaboration relationship. If an attention node of pg^M has been already explained in the dialog history h_i, and b_{i+1} has the same explanation attention but corresponds to a lower resolution and larger size bubble than the one in h_i, then b_{i+1} relates to h_i with the *summary* relationship.

RL Reward (r_i) Our reward function aims to maximize the success rate (ss), user confidence (cf), user satisfaction (sf) and minimize the cost (C_i) over the total number of bubbles. We estimate the cost of generating bubbles b_1,b_2,...,b_i as

$$C_i = \sum_{j=1}^{i} \frac{1}{b_j^{cnt}}$$

[2]For example, showing a very large bubble for revealing Left-Wrist will also reveal Left-Elbow to the human. This makes it harder for human to understand whether the machine is capable of detecting the exact location of Left-Wrist in the image. In addition, although larger bubbles can potentially minimize the number of turns, they transmit a large amount of information from machine to human. This effect may not be obvious in the current experimental set up, but will be significant in the situation where information to be transmitted is through text. Larger bubbles will correspond to longer textual descriptions.

RL Reward (r_i) is expressed in terms of a user feedback and cost associated with selecting the bubbles. At each dialog turn i, after choosing b_i, the explainer collects the following feedback from the user:

- *Success* (ss_i): The user is asked to solve the task based on $\{b_i, h_i\}$. The user's success indicates that the machine's dialog with the user had a high utility and the explanations made by the machine make sense and can help the user reach an understanding of the image. Therefore, if the user solves the task correctly, the explainer is rewarded with $\text{ss}_i = 1$; otherwise, ss_i = -1.
- *User confidence* (cf_i): It is possible that user might solve the task by chance without really understanding the task. We therefore additionally ask the user to report their confidence in solving the task on a scale of 1 to 5.
- *User satisfaction* (sf_i): We ask the user to rate the ordering of bubbles generated in the dialog, and their relevance for solving the task on a scale of 1 to 5.

To compute r_i, we also estimate the cost function C_i of generating bubbles b_1,b_2,...,b_i, defined as

$$C_i = \sum_{j=1}^{i} \frac{1}{b_j^{cnt}}, \tag{12.1}$$

where b^{cnt} is computed as follows:

$$b^{cnt} = 1 + \frac{1}{2}\log(4\pi^2\sigma_1^2\sigma_2^2). \tag{12.2}$$

Intuitively, a large C_i indicates that explanation content of the bubbles revealed is high.

Our reward function aims to maximize the success rate (ss), user confidence (cf), user satisfaction (sf) and minimize the cost (C_i) over the total number of bubbles. We estimate the cost of generating bubbles b_1,b_2,...,b_i as

$$r_i = \frac{1}{i}\exp(\frac{\text{ss}_i\ \text{cf}_i\ \text{sf}_i}{C_i}). \tag{12.3}$$

RL Policy and Training. The explainer operates under a stochastic policy, $\pi(a_i|s_i;\theta)$, which samples optimal bubbles conditioned on the state. This policy is learned by a standard recurrent neural network, called Long-Short Term Memory (LSTM) [660]. In this work we use a 2-layer LSTM parameterized by θ. Input to the LSTM is a feature vector representing the state s_i—specifically, a binary indicator vector of the AOG nodes and edges present in pg^M and pg_i^{UinM}, as well as indices of the question q_i and bubbles generated in h_i. The LSTM's output is the predicted quadruple $(b^{cnt}, b^{act}, b^{att}, b^{dis})$ of b_{i+1}. Thus, the goal of the policy learning is to estimate the LSTM parameters θ.

We use actor-critic with experience replay for policy optimization [661]. The training objective is to find $\pi(a_i|s_i;\theta)$ that maximizes the expected reward $J(\theta)$ over all possible bubble sequences given a starting state. The gradient of the objective function has the following form:

$$\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_i|s_i;\theta)\, A(s_i, a_i)] \tag{12.4}$$

where $A(s_i, a_i) = Q(s_i, a_i) - V(s_i)$ is the advantage function [662]. $Q(s_i, a_i)$ is the standard Q-function, and $V(s_i)$ is the baseline function aimed at reducing the variance of the estimated gradient. We use the same specifications of $Q(s_i, a_i)$ and $V(s_i)$ as in [662]. As in [662], we sample the dialog experiences randomly from the replay pool for training.

12.6 Experiments

We deployed the X-ToM game on the Amazon Mechanical Turk (AMT) and trained the X-ToM Explainer through the interactions with turkers. All the turkers have a bachelor's degree or higher. We used three visual recognition tasks in our experiments, namely, human body parts identification, pose estimation, and action identification. We used 1000 images randomly selected from Extended Leeds Sports (LSP) dataset [663]. Each image is used in all the three tasks. During training, each trial consists of one X-ToM game where a turker solves a given task on a given image. We restrict Turkers from solving a task on an image more than once. In total, about 2400 unique workers contributed in our experiments.

We performed off-policy updates after every 200 trials, using Adam optimizer [664] with a learning rate of 0.001 and gradients were clipped at [-5.0, 5.0] to avoid explosion. We used ϵ-greedy policy, which was annealed from 0.6 to 0.0. We stopped the training once the model converged. In our case, the X-ToM policy model converged after interacting with 3500 turkers. All our data and code will be made publicly available.

Elaboration	Sequence	Recurrence	Restatement	Summary
26%	48.7%	12.6%	5.1%	7.6%

Table 12.1: Distribution of observed discourse relations in the test trials

The trained X-ToM Explainer was applied to an additional 500 X-ToM games with AMT turkers for testing. Table 12.1 shows the percentage of discourse relations among bubbles found in the test interactions. As can be seen, the discourse relation `sequence` dominates other relations. This indicates that the X-ToM's most common explanation strategy is to prefer a bubble containing new evidence (that was not already shown to the user). Furthermore, the experiment has shown that 55.3% of the bubbles in the test trials were generated using α explanation act, 23.1% using β explanation act, and 21.6% using γ explanation act. The high percentage of β and γ explanation acts indicate that contextual evidence is not only helpful for the performer to detect but also for the explainer to explain.

12.6.1 AMT Evaluation of X-ToM Explainer

We conducted an ablation study to quantify the importance of taking the inferred human's mind into account for generating optimal explanations, *i.e.*, the ablated model does not explicitly represent and infer pg^{UinM}. Similar to X-ToM, the ablated model was also deployed and trained on AMT. The trained ablated model was again applied to an additional 500 X-ToM games with AMT turkers for testing. Table 12.2 compares X-ToM Explainer with the ablated model in terms of objective measures such as average success rate (ss), average number of bubbles, average rewards (r). X-ToM Explainer significantly outperforms the ablated model ($p < 0.01$) in terms of the overall reward. Although the success rates of both models are similar, the ablated model is found to use a significantly larger number of bubbles, which leads to lower overall reward.

Using an additional 100 X-ToM games on AMT, we further compare the explanations generated by our X-ToM Explainer with the explanations annotated by humans. We asked three graduate students (not the authors), to select the most appropriate bubbles for a given task. Bubbles that have been agreed upon by these three subjects were taken as the best explanations for the given task and image. In terms of maximizing the reward, we found that X-ToM Explainer performed significantly better than the human strategy of bubble selection ($p < 0.01$). However, we found

Model	#test trials	ss	#bubbles	r
X-ToM	500	**81.3**%	**10.5**	**0.91**
Ablated Model	500	77.1%	28	0.42
Human Strategy	100	78.9%	**6**	0.62

Table 12.2: Comparison of X-ToM with ablated and human baselines

that the average dialog length in the human explanations is 6, while the average dialogue length observed in the X-ToM explanations is 10.5, indicating that there is a possibility to further improve the quality of the X-ToM explanations. We leave this for future exploration.

12.6.2 Human Subject Evaluation on Justified Trust

Using X-ToM Evaluator, we conduct human subject experiments to assess the effectiveness of the X-ToM Explainer, that is trained on AMT, in increasing human trust through explanations. We recruited 120 human subjects from our institution's Psychology subject pool [3]. These subjects have no background on computer vision, deep learning and NLP (see Section 12.11.1 for more details). We applied between-subject design and randomly assigned each subject into one of the three groups. One group used X-ToM Explainer, and two groups used the following two baselines respectively:

- $\Omega_{\mathbf{QA}}$: we measure the gains in human trust only by revealing the answers for the tasks without providing any explanations to the human.
- $\Omega_{\mathbf{Salience}}$: in addition to the answers, we also provide saliency maps generated using attribution techniques to the human as explanations [639, 640].

Within each group, each subject will first go through an introduction phase where we introduce the tasks to the subjects. Next, they will go through familiarization phase where the subjects become familiar with the machine's underlying inference process (Performer), followed by a testing phase where we apply our trust metrics and assess their trust in the underlying Performer.

Fig. 12.7 compares the justified positive trust (JPT), justified negative trust (JPT), and Reliance (Rc) of X-ToM with the baselines.

As we can see, JPT, JNT and Rc values of X-ToM are significantly higher than Ω_{QA} and Ω_{Salience} ($p < 0.01$). *Also, it should be noted that attribution techniques ($\Omega_{Salience}$) did not perform any better than the Ω_{QA} baseline where no explanations are provided to the user.* This could be attributed to the fact that, though saliency maps help human subjects in localizing the region in the image based on which the performer made a decision, they do not necessarily reflect the underlying inference mechanism. In contrast, X-ToM Explainer makes the underlying inference processes (α, β, γ) more explicit and transparent and also provides explanations tailored for individual user's perception and understanding. Therefore X-ToM leads to the significantly higher values of JPT, JNT and Rc. This is one of the key results of our work, given the popularity of attribution techniques as the state-of-the-art explanations.

Fig. 12.8 shows the average explanation satisfaction rates obtained from each of the three groups. As we can see, subjects in X-ToM experiment group found that explanations were highly useful, sufficient and detailed compared to the baselines ($p < 0.01$). Interestingly, we did not find significant differences across the three groups in terms of other satisfaction measures: confidence, understandability, accuracy and consistency. We leave this observation for future exploration

[3]These experiments were reviewed and approved by our institution's IRB.

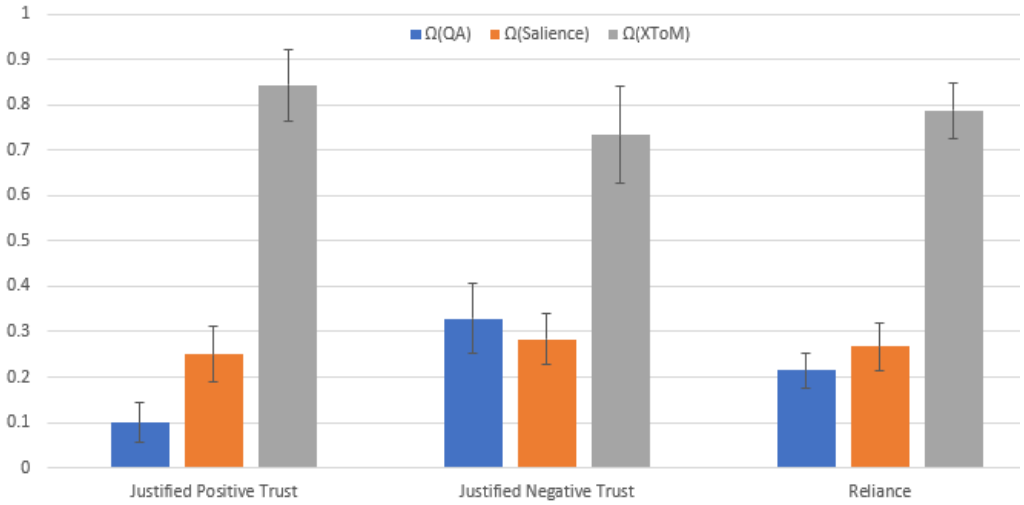


Figure 12.7: Gain in Justified Positive Trust, Justified Negative Trust and Reliance: X-ToM vs baselines (QA, Saliency Maps). Error bars denote standard errors of the means.

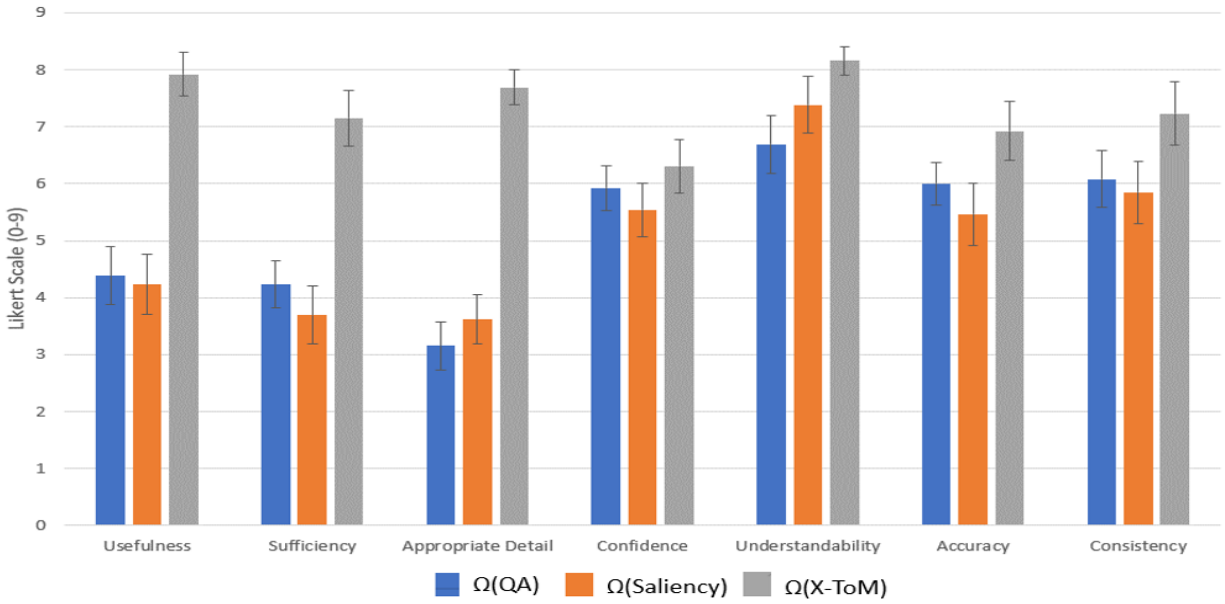


Figure 12.8: Explanation Satisfaction: X-ToM vs baselines (QA, Saliency Maps). Error bars denote standard errors of the means.

12.6.3 Gain in Reliance over time

We hypothesized that human trust and reliance in machine might improve over time. This is because, it can be harder for humans to fully understand the machine's underlying inference process in one single session. Therefore, we conduct an additional experiment with eight human subjects where the subjects' reliance is measured after every session. The results are shown in Fig. 12.9. As we expected, subjects' reliance increased over time. Specifically, reliance with respect to α inference process significantly improved only after 2.5 sessions. Reliance with respect to β and γ inference processes significantly improved after 4.5 sessions. It is clearly evident that, with more sessions, it is possible to further improve human reliance in AI system.

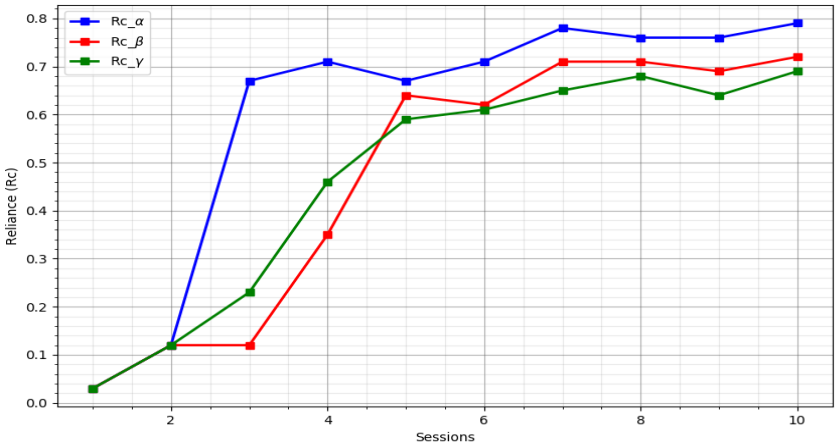


Figure 12.9: Gain in Reliance over sessions w.r.t. α, β and γ processes

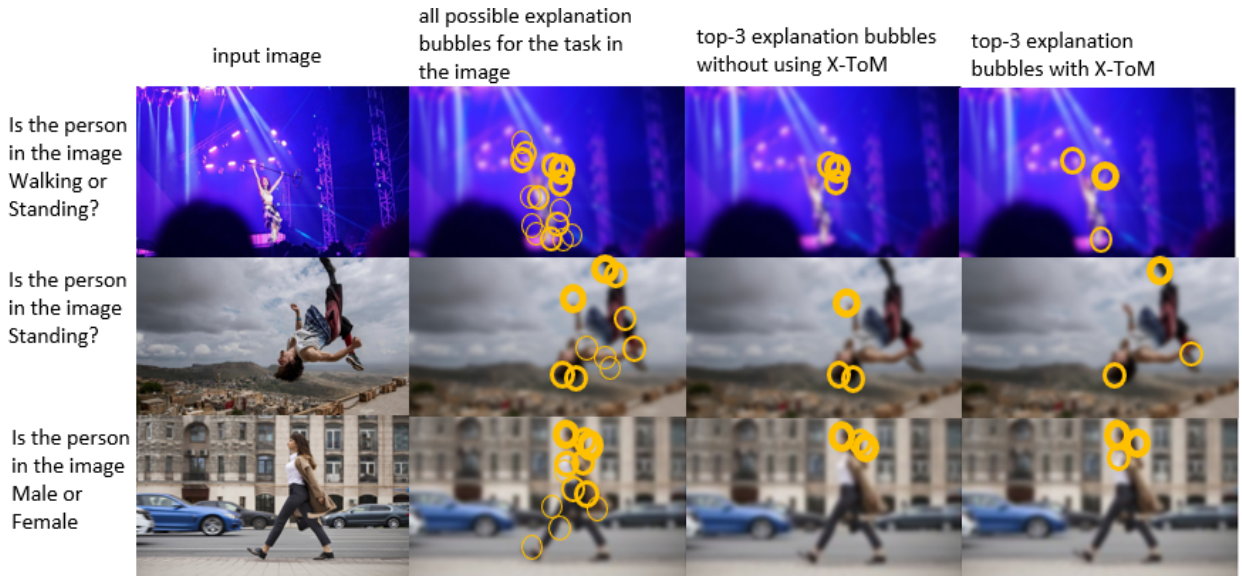
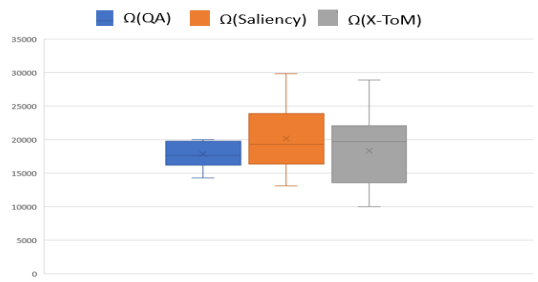


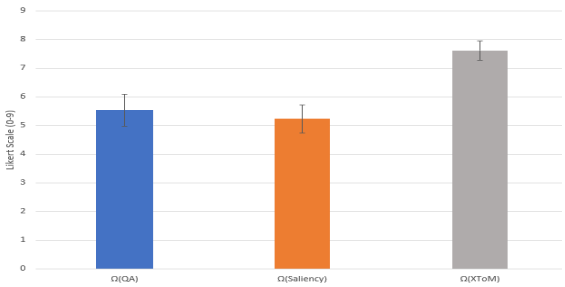
Figure 12.10: Top-3 best explanations generated with and without using X-ToM.

12.6.4 Case Study

Fig. 12.10 shows examples where the top-3 best explanations preferred by X-ToM are compared against the top-3 explanations generated by the attribution techniques. The first column shows the input image for the task. The second column shows all the evidence (*i.e.*, explanations in the form of bubbles, highlighted in yellow color) used in the machine's inference about the task. The thicker the bubble, the higher is its influence, for the machine, in interpreting the image. As we can see, attribution techniques chose the explanations only based on how influential they are for the machine in recognizing the image (third column). In contrast, since X-ToM maximizes the utility of explanations based on both influence values and user's model, explanations selected by the X-ToM (fourth column) are diverse and are more intuitive for humans to understand and solve the task efficiently. For example, for the first image, to aid the human user in solving the task "Is the person in the image walking," X-ToM generates the explanation bubbles based on left arm, right arm and lower body of the person, whereas attribution techniques generate the top-3 bubbles only based on right arm which clearly is not sufficient for the user to successfully solve the task.



(a) **Response Times** (in milliseconds per question). Error bars denote standard errors of the means.



(b) **Qualitative Reliance**. Error bars denote standard errors of the means.

In addition to the quantitative and qualitative metrics discussed in section 2.5, we also measure the following metrics for comparing our X-ToM framework with the baselines:

- **Response Time**: We record the time taken by the human subject in answering evaluator questions. Fig. 12.11a shows the average response times (in milliseconds per question) for each of the three groups (X-ToM, QA and Saliency Maps). We expected the participants in X-ToM group to take less time to respond compared to the baselines. However, we find no significant difference in the response times across the three groups.
- **Subjective Evaluation of Reliance**: We collect subjective Reliance values (on a Likert scale of 0 to 9) from the subjects in the three groups. The results are shown in Fig. 12.11b. These results are consistent with our quantitative reliance measures. It may be noted that subjects' qualitative reliance in Saliency Maps is lower compared to the QA baseline.

12.7 Case Study 2: Robot explanation

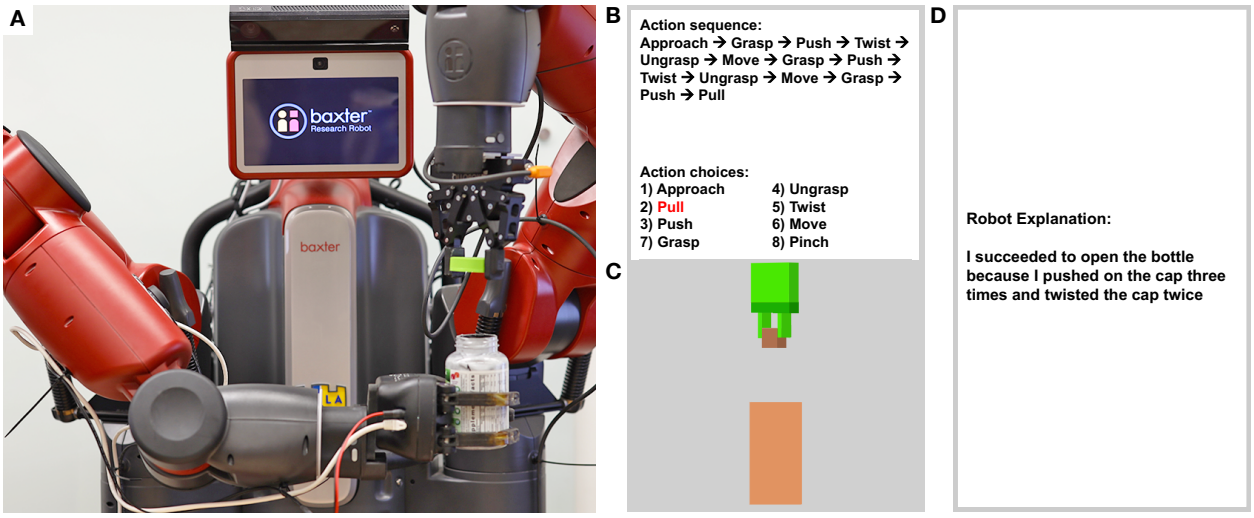
In this section, we present a case study that looks at how STC-AOGs can be used in combination with neural networks to explain robot behavior, which was presented originally in *Science Robotics* [665]. This case study aims to disentangle explainability from task performance, measuring each separately to gauge the advantages and limitations of two major families of representations—symbolic representations and data-driven representations—in both task performance and imparting trust to humans. The goals are to explore: (i) what constitutes a good performer for a complex robot manipulation task? (ii) How can one construct an effective explainer to explain robot behavior and impart trust to humans?

12.7.1 Experiment Domain

This case study develops an integrated framework consisting of a symbolic action planner using a stochastic grammar as the planner-based representation and a haptic prediction model based on neural networks to form the data-driven representation. The authors examine this integrated framework in a robot system using a contact-rich manipulation task of opening medicine bottles with various safety lock mechanisms. From the performer's perspective, this task is a challenging learning problem involving subtle manipulations, as it requires a robot to push or squeeze the bottle in various places to unlock the cap. At the same time, the task is also challenging for explanation, as visual information alone from a human demonstrator is insufficient to provide an effective explanation. Rather, the contact forces between the agent and the bottle provide the *hidden* "key" to unlock the bottle, and these forces cannot be observed directly from visual input.

To constitute a good performer, the robot system proposed here cooperatively combines multiple sources of information for high performance, enabling synergy between a high-level symbolic action planner and a low-level haptic prediction model based on sensory inputs. A stochastic grammar model is learned from human demonstrations and serves as a symbolic representation capturing the compositional nature and long-term constraints of a task [666]. A haptic prediction model is trained using sensory information provided by human demonstrations (*i.e.*, imposed forces and observed human poses) to acquire knowledge of the task. The symbolic planner and haptic model are combined in a principled manner using an improved Generalized Earley Parser (GEP) [539], which predicts the next robot action by integrating the high-level symbolic planner with the low-level haptic model. The learning from demonstration framework presented here shares a similar spirit of our previous work [667] but with a new haptic model and a more principled manner, namely the GEP, to integrate the haptic and grammar models. Computational experiments demonstrate a strong performance improvement over the symbolic planner or haptic model alone.

To construct an effective explainer, the proposed approach draws from major types of explanations in human learning and reasoning that may constitute representations to foster trust by promoting mutual understanding between agents. Previous studies suggest humans generate explanations from *functional* perspectives that describe the *effects* or *goals* of actions and from *mechanistic* perspectives that focus on behavior as a process [668]. The haptic prediction model is able to provide a functional explanation by visualizing the essential haptic signals (*i.e.*, *effects* of the previous action) to determine the next action. The symbolic action planner is capable of providing a mechanistic explanation by visualizing multiple planning steps (instead of just one) to describe the *process* of the task. The proposed robot system provides both functional and mechanistic explanations using the haptic model and symbolic planner, respectively.



E Summary of human subject groups and explanations presented

Group	RGB (A)	Symbolic (B)	Haptic (C)	Text (D)
Baseline	✓			
Symbolic	✓	✓		
Haptic	✓		✓	
GEP	✓	✓	✓	
Text	✓			✓

Figure 12.12: **Illustration of visual stimuli used in human experiment**. All five groups observed the RGB video recorded from robot executions, but differed by the access to various explanation panels. (**A**) RGB video recorded from robot executions. (**B**) Symbolic explanation panel. (**C**) Haptic explanation panel. (**D**) Text explanation panel. (**E**) A summary of which explanation panels were presented to each group.

12.7.2 Experiment Design

The human experiment aims to examine whether providing explanations generated from the robot's internal decisions foster human trust to machines and what forms of explanation are the most effective in enhancing human trust. We conducted a psychological study with 150 participants; each was randomly assigned to one of five groups. Our experimental setup consisted of two phases: familiarization and prediction. During familiarization, all groups viewed the RGB video, and some groups were also provided with an explanation panel. During the second phase of the prediction task, all groups only observed RGB videos.

The five groups consist of the baseline no-explanation group, symbolic explanation group, haptic explanation group, GEP explanation group, and text explanation group. For the baseline no-explanation group, participants only viewed RGB videos recorded from a robot attempting to open a medicine bottle, as shown in Fig. 12.12a. For the other four groups, participants viewed the same RGB video of robot executions and simultaneously were presented with different explanatory panels on the right side of the screen. Specifically, the symbolic group viewed the symbolic action planner illustrating the robot's real-time inner decision-making, as shown in Fig. 12.12b. The haptic group viewed the real-time haptic visualization panel, as shown in Fig. 12.12c. The GEP group viewed the combined explanatory panel, including the real-time robot's symbolic planning and an illustration of haptic signals from the robot's manipulator, namely both Fig. 12.12b–c. The text explanation group was provided a text description that summarizes why the robot succeeded or failed to open the medicine bottle *at the end* of the video, as shown in Fig. 12.12d. See a summary in Fig. 12.12e for the five experimental groups.

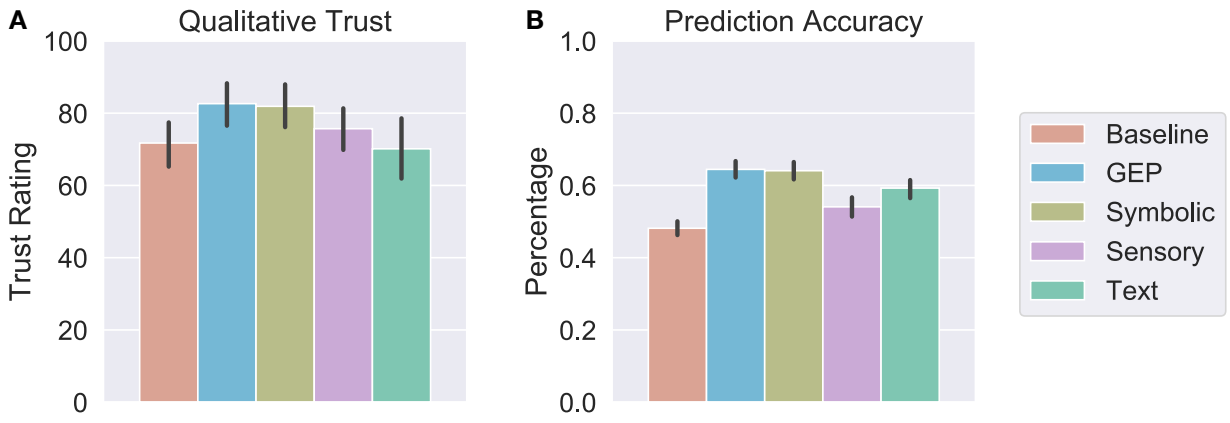


Figure 12.13: **Human results for trust ratings and prediction accuracy.** (**A**) Qualitative measures of trust: average trust ratings for the five groups. and (**B**) Average prediction accuracy for the five groups. The error bars indicate the 95% confidence interval. Across both measures, the GEP performs the best. For qualitative trust, the text group performs most similarly to the baseline group.

12.7.3 Results and Analysis

Fig. 12.13a shows human trust ratings from the five different groups. The analysis of variance (ANOVA) reveals a significant main effect of groups ($F(4, 145) = 2.848; p = 0.026$) with the significance level of 0.05. This result suggests that providing explanations about robot behavior in different forms impacts the degree of human trust to the robot system. Furthermore, we find that the GEP group with both symbolic and haptic explanation panels yields the highest trust rating, with a significantly better rating than the baseline group in which explanations are not provided (independent-samples t-test, $t(58) = 2.421; p = 0.019$). Interestingly, the GEP group shows greater trust rating than the text group in which a summary description is provided to explain the robot behavior ($t(58) = 2.352; p = 0.022$), indicating detailed explanations of robot's internal decisions over time is much more effective in fostering human trust than a summary text description to explain robot behavior. In addition, trust ratings in the symbolic group are also higher than ratings in the baseline group ($t(58) = 2.269; p = 0.027$) and higher than ratings in the text explanation group ($t(58) = 2.222; p = 0.030$), suggesting symbolic explanations play an important role in fostering human trust of the robot system. However, the trust ratings in the haptic explanation group are not significantly different from the baseline group, implying that explanations only based on haptic signals are not effective ways to gain human trust despite the explanations are also provided in real-time. No other significant group differences are observed between any other pairing of the groups.

The second trust measure based on prediction accuracy yields similar results. All groups provide action predictions above the chance-level performance of 0.125 (as there are 8 actions to choose from), showing that humans are able to predict the robot's behavior after only a couple of observations of a robot performing a task. The ANOVA analysis shows a significant main effect of groups ($F(4, 145) = 3.123; p = 0.017$), revealing the impact of provided explanations on the accuracy of predicting the robot's actions. As shown in Fig. 12.13b, participants in the GEP group yield significantly higher prediction accuracy than those in the baseline group ($t(58) = 3.285; p = 0.002$). Prediction accuracy of the symbolic group also yields better performance than the baseline group ($t(58) = 2.99; p = 0.004$). Interestingly, we find that the text group shows higher prediction accuracy than the baseline group ($t(58) = 2.144; p = 0.036$). This result is likely due to the summary text description providing a loose description of the robot's action plan; such a description decou-

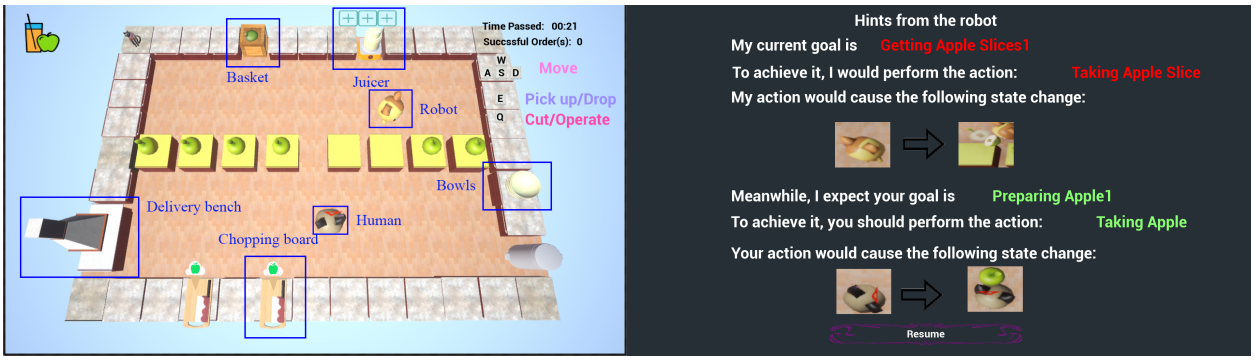


Figure 12.14: (a) A top-down view of our collaborative cooking game, where the user (the bottom character) collaborates with a robot (the top character) on some cooking tasks, *e.g.*, *making apple juice*. (b) The explanation interface exhibits the expected sub-tasks for both agents. Pre-conditions and post-effects of atomic actions are displayed as well.

ples the explanation from the temporal execution of the robot. The prediction accuracy data did not reveal any other significant group differences among other pairs of groups.

In general, humans appear to need real-time, symbolic explanations of the robot's internal decisions for performed action sequences in order to establish trust in machines when performing multi-step complex tasks. Summary text explanations and explanations only based on haptic signals are not effective ways to gain human trust, and the GEP and symbolic group foster similar degrees of human trust to the robot system according to both measures of trust.

12.8 Case Study 3: Explanation in human-machine workspace

We conducted a user study in a gaming environment to evaluate our algorithm, where participants can collaborate with agents on a virtual cooking task. The gaming environment and explanation interface are displayed in Fig. 12.14.

12.8.1 Experiment Domain

Our experiment domain is inspired by the video game **Overcooked**[4], where multiple agents are supposed to make use of various tools and take different roles to prepare, cook, and serve various dishes. Particularly, we use Unreal Engine 4 (UE4) to create a real-time cooking task, namely making *apple juice*. To finish the task, teammates need to take apples from the box and slice them with a knife near the chopping board. Three apple slices should be put into the juicer before producing and delivering apple juice. Fig. 12.14 shows a top-down view of the environment. The game interface is designed to be interactive (*e.g.*, object appearance will change after taking valid actions) so that people can easily play through.

To finish the task, each user needs to complete a sequence of 62 atomic actions, if acting optimally, and observe 5 different object fluent changes with a total state space around 10^9. An example task schedule is shown in Fig. 12.15.

12.8.2 Experiment Design

Hypotheses. The user study tests the following hypotheses with respect to our algorithm in the collaboration:

[4] `http://www.ghosttowngames.com/overcooked/`

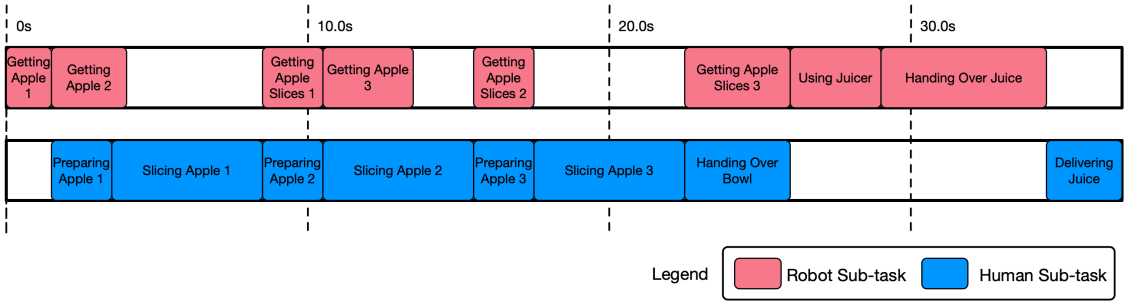


Figure 12.15: An example task schedule for *making apple juice*. The robot maintains the schedule to reflect its expectation on how the team should finish the task. Each color block represents a sub-task, performed by either robot or human. At a specific timing, we can assign tasks to both agents based on the schedule. *e.g.*, at 10.0s, the robot is *getting apple slices 1* while the user is supposed to be *preparing apple 2*. The schedule gets updated based on inferred human mental states.

- **H1: Task completion time.** Participants would collaborate with the robot more efficiently if the robot generates explanations based on the human mental state modeling, compared to the other conditions.
- **H2: Perception of the robot.** Participants would have higher perceived helpfulness and efficiency of the robot, as a result of receiving explanations based on the human mental state modeling, compared to the other conditions.

Manipulated Variables. We use a between-subject design for our experiment. In particular, users are randomly assigned to one of three groups and receive different explanations from the robot:

- **Control:** Users would not get any explanations from the robot. As a result, they can learn to finish the task by interacting with the environment.
- **Heuristics:** The robot gives explanations when there is no detected user action for a period of time. This serves as a simple heuristic for the robot to infer whether the user is having difficulties in finishing the task. The timing threshold is set to 9.3 seconds, based on the result of a pre-study in which users can actively ask for explanations when they get stuck.
- **Mind modeling:** The robot gives explanations when there is a disparity between robot and human mental states.

Study Protocol. Before starting the experiment, each participant signs an informed consent form. An introduction is given afterward, including rules and basic controls of the game. As a part of the introduction, participants are given three chances to work on a simple single-agent training task, to verify their understanding. Those who fail to complete the training task in one minute would not continue the study. This is a comprehension test to exclude people who do not understand game control.

Participants who finish training get to see further instructions before starting to collaborate with the robot. They are first educated about the goal of a collaboration task (*i.e.*, making *apple juice*) and what actions the team should perform to finish it. This is done to make sure every participant have sufficient knowledge to finish the task, so the impact of user-specific prior knowledge can be minimized. To prepare users to interact and communicate with the robot agent, we would also show them a top-down view of the level map (as shown in Fig. 12.14), the appearance of the robot agent as well as an example of an explanation. During the task, the team is required to make and serve two orders of dishes in the virtual kitchen. At the end of the study, each participant is asked to complete a post-experiment survey to provide background information and evaluate the robot teammate.

Measurement. In the background study, we have collected from users their basic demographic

information, education, as well as experience with video games.

Our objective measure is intended to evaluate the human-robot teaming performance and subjective measure is designed for evaluating users' perception of the robot. Our dependent measures are listed below:

- **Teaming performance.** We evaluate teaming performance by recording the time for the team to complete each order.
- **Perception of the robot.** We measure user's perception about the robot, in terms of its helpfulness and efficiency. Helpfulness is comprised of questions that measure users' opinion on the robot's ability to provide necessary help. Efficiency is comprised of questions that measure users' opinion on how efficiently and fluently the team is able to finish the task.

12.8.3 Results and Analysis

We recruited 29 subjects for our IRB-approved study from the university's subject pool. Most of the participants (69.3%) came from a non-STEM background. Their reported ages ranged from 17 to 36 (M=19.52, SD=2.89). All the participants have moderate experience with video games and have not played the video game **Overcooked**, which inspired our study design. Each participant got 1 course credit after completing the study. In addition, for ease of conducting the study, we discarded the data of 2 participants from the control group, as they got completely lost and failed to finish the designated task. As a result, there are 10 valid participants in the "mind modeling" and "heuristics" group, and 7 in the "control" group.

Generally, we use ANOVA to test the effects of different experimental conditions on teaming performance and subjective perception of the robot. Tukey HSD tests are conducted on all possible pairs of experimental conditions.

As shown in Fig. 12.16, we found marginally significant effects from "mind modeling" conditions on completion time of the first order ($F(2, 24) = 2.038, p = .152$). Post-hoc comparisons using the Tukey HSD tests revealed that teams could finish the first order significantly faster if users were under the "mind modeling" condition, compared to those under "control" ($p = .044$). The result is marginally significant compared to those in "heuristics" ($p = .120$), **confirming H1**. However, for the completion time of the second order, we did not find any significant effect ($F(2, 24) = 0.425, p = .658$). This is not surprising since users were asked to finish the same task twice. They could take advantage of their previous experience working with the robot for the second order. Intuitively, the quantitative result showed that our explanation generation algorithm helped non-expert users to finish the task efficiently on their first run, while those in the control group needed to complete the task once to be able to finish it with the same efficiency.

The factorial ANOVA also revealed a significant effect of the explanation system on the perceived helpfulness ($F(2, 24) = 4.663, p = .019$) and efficiency ($F(2, 24) = 4.136, p = .029$) of the robot (Fig. 12.17). **In support of H2**, post-hoc analysis with the Tukey HSD tests showed that the robot's perceived helpfulness was significantly higher under the "mind modeling" condition, compared to "control" ($p = .023$) and "heuristics" ($p < .01$). Users under the "mind modeling" were also more likely to believe the explanation system resulted in improved collaboration efficiency, compared to "heuristics" ($p = .026$) and "control" ($p < .01$).

12.9 Conclusions

This chapter presents X-ToM—a new framework for Explainable AI (XAI) and human trust evaluation based on the Theory-of-Mind (ToM). X-ToM generates explanations in a dialog by explicitly modeling, learning, and inferring three mental states based on And-Or Graphs—namely, machine's

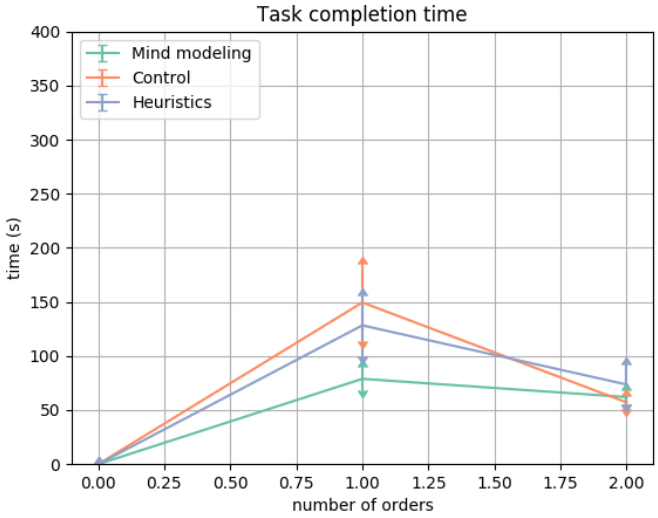


Figure 12.16: Time taken for the team to complete two orders under different testing conditions.

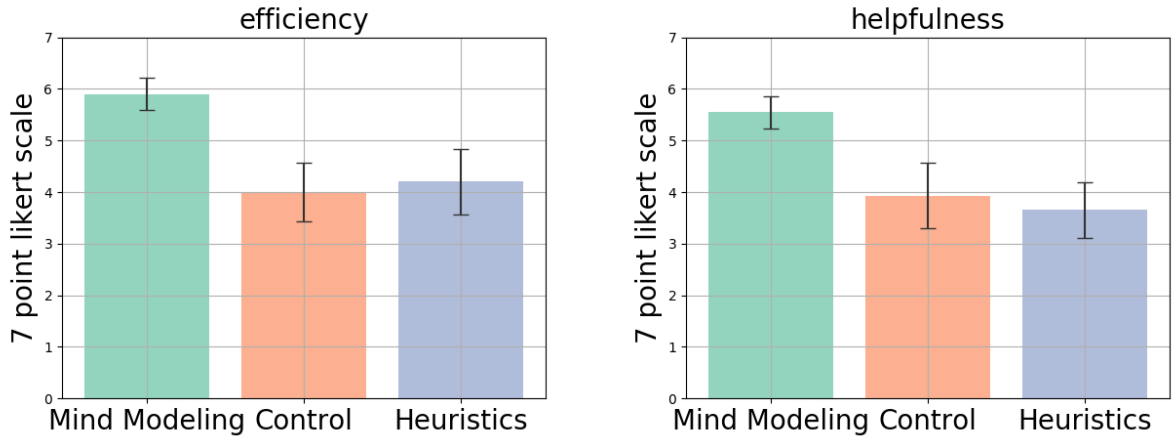


Figure 12.17: User's self-reported perception of the robot in terms of its efficiency and helpfulness.

mind, human's mind as inferred by the machine, and machine's mind as inferred by the human. This allows for a principled formulation of human trust in the machine. For the task of visual recognition, we proposed a novel, collaborative task-solving game that can be used for collecting training data and thus learning the three mental states, as well as a testbed for quantitative evaluation of explainable vision systems. We demonstrated the superiority of X-ToM in gaining human trust relative to baselines.

12.10 Acknowledgement

The work is supported by DARPA XAI N66001-17-2-4029.

Figure 12.18: Statistics (based on Age, First Language and Gender) of the 120 human subjects, from Psychology subject pool, participated in our study.

12.11 Appendix

12.11.1 Evaluation with Psychology Subject Pool

Fig. 12.18 shows the statistics (Age, First Language, Gender) of the 120 human subjects, recruited from our institution's Psychology subject pool.

12.11.2 X-ToM Evaluator Interface and Questions

Specifically, there are two main types of evaluator questions about the user's prediction: (1) whether the Performer would successfully or incorrectly detect objects, parts and other concepts encoded by AOG; and (2) which image parts are most influential for the Performer's successful or incorrect object detection. For example, the evaluator's questions include "which parts of the image are most important for the machine to recognize that the person is running," and "which small part of image contributes most to inferring the surrounding larger part of image." Figs. 12.19 to 12.21 show few sample screenshots (from our web interface) of the exact questions, on the detection of the body part "Left-Arm," that we pose to the subjects.

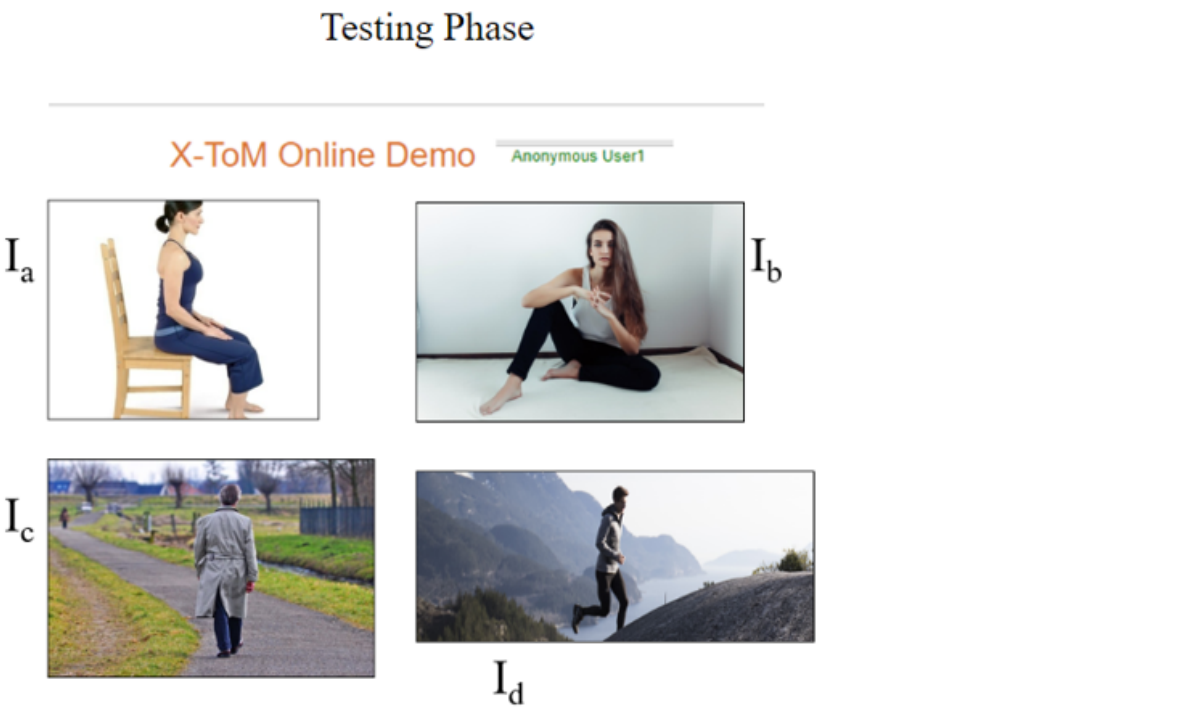


Figure 12.19: Sample evaluator questions

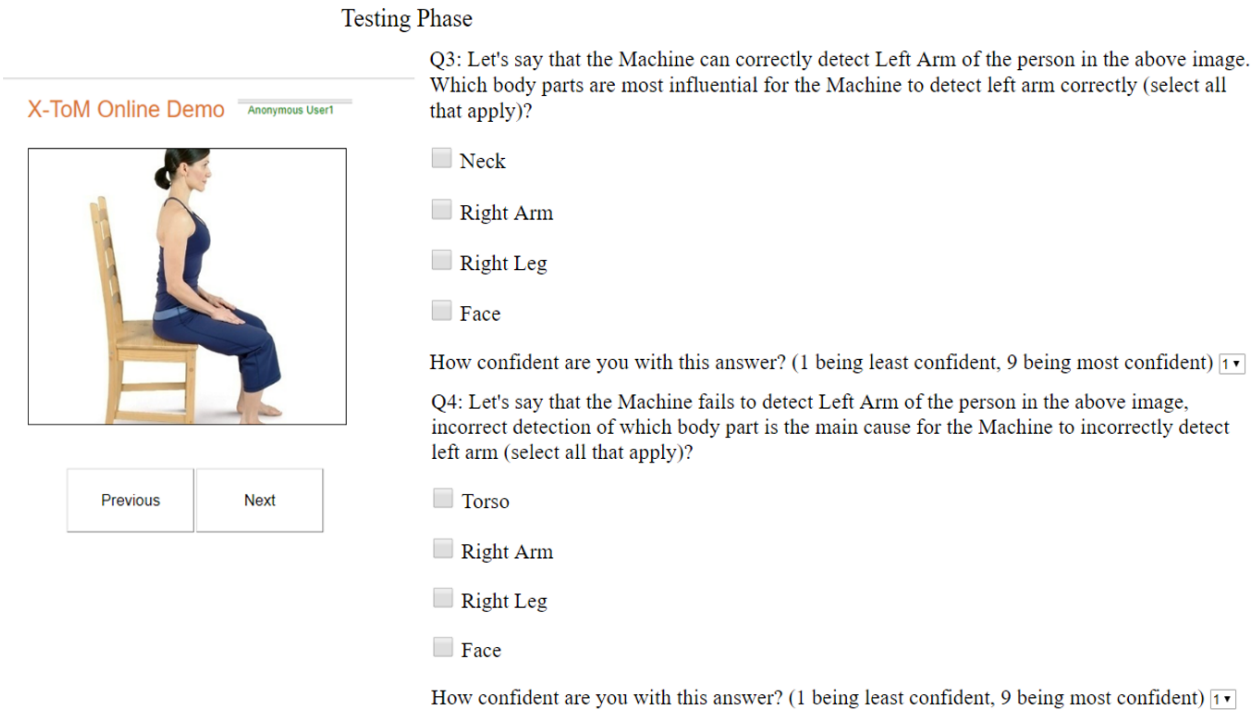


Figure 12.20: Sample evaluator questions

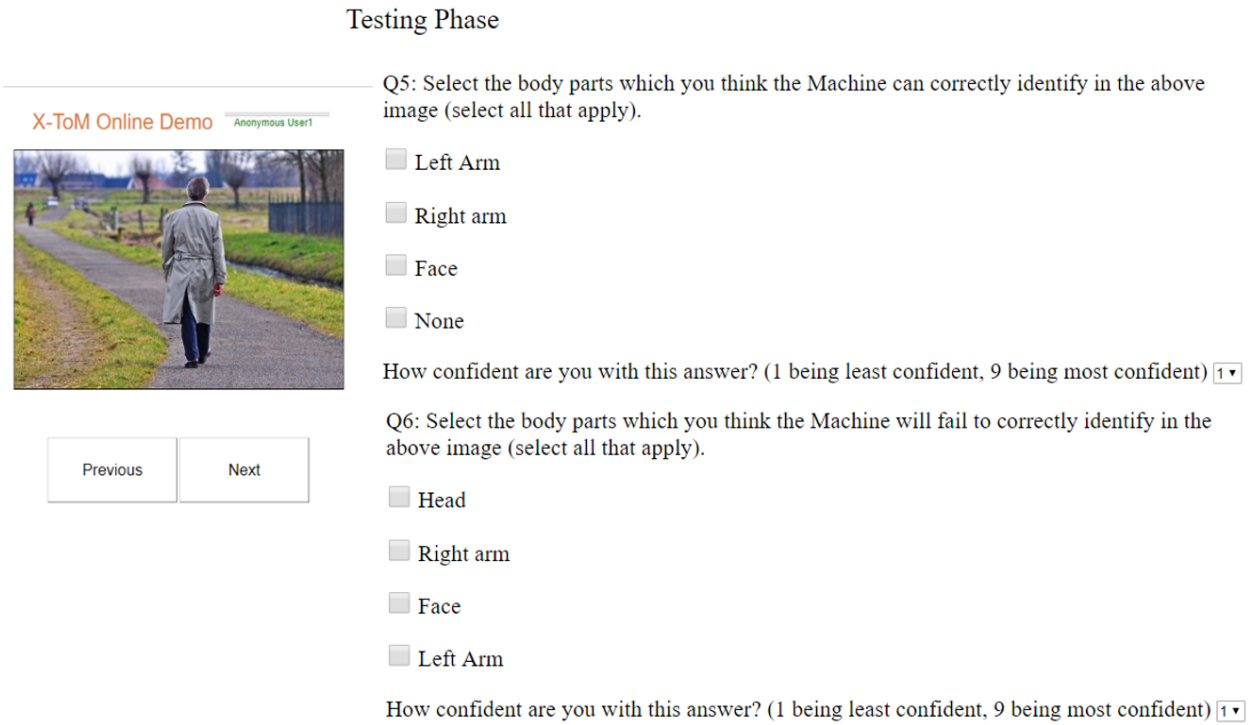


Figure 12.21: Sample evaluator questions

Chapter 13

Communicative Learning

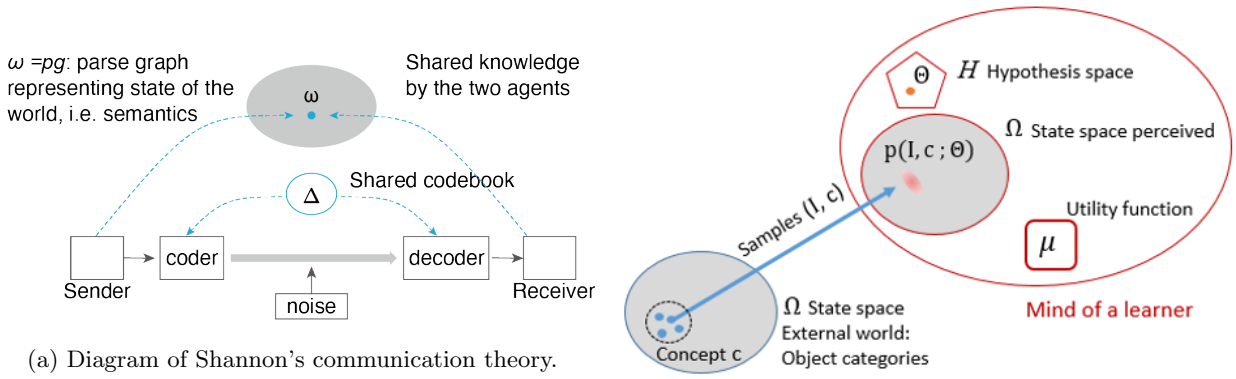
13.1 Introduction

As the advancement of information technology, the world is entering the era of **Big Data**. This deluge of data calls for automated methods of data analysis, which is the origin of machine learning, one of, if not the, most important stream of recent AI. Nonetheless, fitting models to explain patterns from extensive data, though well compatible to modern computers, is not the typical way of human learning. Human learning is a lifelong cognitive process of communicating with the physical and social world. In other words, rather than studying data by oneself, human learning happens through interaction with others. Its sophistication, effectiveness and complexity give rise to human intelligence—a phenomenon that AI is inspired to replicate. Decades of studies in cognitive psychology [139, 669] and anthropology & communications studies [140] have revealed that human communication and learning is built on many layers of cognitive infrastructures and protocols.

To fully grasp the essence of human learning, learning needs to be studied in a multi-agent system, where the agents communicate with each other in the process of learning and teaching. Each agent has a mind that consists of the agent's current knowledge of the environment, the agent's utility function and value, and the agent's goal and intent. These drive the learning and communication. In order for an agent to communicate with other agents effectively in the process of learning, the agent must know about other agents' knowledge, values and goals, *i.e.*, the agent must have a summary of the mind of any other agent. The ability for an agent to learn is determined by the agent's IQ, or capacity as a learner. The learning process may halt, at least temporarily, if certain conditions are met. For effective communication and learning, it is crucial to design a learning protocol that is optimal in terms of some criterion.

The current mathematical and statistical frameworks in communication, machine learning, and AI are still far from addressing the complexity of human learning and communications.

The mathematical theory of communication was concerned with sending messages through a noisy channel [670], as Fig. 13.1a illustrates. The sender and receiver share a codebook, and the message refers to some world state ω, *e.g.*, a parse graph. Shannon's theory has intentionally left out the "*semantics*" or "*meanings*" of messages. Sender and receiver are assumed to share common ground to make sense of the messages outside this framework. The limits of coding efficiency and channel capacity are based on a protocol that does not model mental states and motives of agents in sending messages. In our new communicative learning (CL) framework, messages are selected after *deliberations & reflection* using theory-of-mind representations, and carry *extra information* that is recoverable in a more effective communication protocol, *i.e.*, agents are capable



(a) Diagram of Shannon's communication theory.

(b) Diagram for statistical learning theory.

Figure 13.1: Diagrams of Shannon's communication theory and statistical learning theory.

of "*reading between lines*."

Statistical and machine learning theories went a step further. As show in Fig. 13.1b, a learner learns a concept c, which is defined as a set in a state space Ω [671] or a probability model θ in a hypothesis space $\theta \in \mathcal{H}$ [259], using random samples $\{(I_i, c_i), i = 1, ..., m\}$ drawn from an external world. Learning is driven by a pre-defined utility or loss function μ. In this setting, the PAC-learning theory [671] bounds the number of examples $n(\epsilon, \delta)$ needed to learn the concept with error $\leqslant \epsilon$ and confidence $> 1-\delta$. The bounds for PAC-learning and generalization are often defined on the capacity of the hypothesis space, and overly pessimistic [672, 673, 674, 675, 676]. In parallel to PAC learning, the minimax learning theory in the statistical learning literature defines the limit of the learned model as a lower bound,

$$LB(\theta, n) = \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{H}} \mathbb{E}_{\theta}[-\mu(\theta, \hat{\theta}(I_1, I_2, ..., I_n))] \tag{13.1}$$

These theories all make an inefficient assumption that examples in learning are *random samples*, while we argue that learning is a communication process where examples are *deliberate messages* by reflecting mental states of learner and teacher, and derive limits of learning in new CL protocol.

Deep learning and information bottleneck. Deep learning with convolutional neural networks (CNN) [677] maps input X to output Y (annotated classification labels) by learning k-layers of features $Z = (Z_1, ..., Z_k)$. Recently, [678, 679] tried to reveal the "blackbox" by an information bottleneck (IB) theory. That is, the representation Z should preserve the mutual information in X for Y: *i.e.*, $MI(Z, Y) - MI(X, Y) = 0$, and minimize its mutual information with the raw signal, *i.e.*, forgetting the features in X that are irrelevant to the task Y. By Lagrange multipliers, this becomes to learn in a hypothesis space of neural nets:

$$\theta^* = \arg\min MI(Z, X) - \beta MI(Z, Y) \tag{13.2}$$

The IB theory may give an explanation for how CNN works, but the layered features $Z = (Z_1, ..., Z_k)$ are features trained by big data for a specific task, *i.e.*, in a "big-data for small-task" paradigm, and remain uninterpretable. In contrast, to enable distributed intelligence and commonsense AI systems, agents must use *interpretable messages* and achieve common knowledge to communicate about world state and concepts: objects, scenes, actions, activities.

Complexity studies in CS and multi-agent systems in AI. In theoretical computer science, [680] studied a problem of communication complexity in distributed computation. *e.g.*, suppose the task is to compute a Boolean function $b(x, y)$, agent A knows augment x and agent B knows augment y, how much information exchange between two agents is necessary? This is extended in [681] using approximation theory. But this stream of work does not involve theory-of-mind representation either. The theory-of-mind representations are studied by multi-agent system [682], mostly focusing on toy examples (*e.g.*, [683, 684]). A recent work [598] extends POMDP to interactive POMDPs (I-POMDP), where the agent's belief is represented approximately by a set of samples, *i.e.*, particle filtering in sequential Monte Carlo. Then the *belief-of-belief* is represented by particles of particles and computation becomes infeasible. In this project, we will develop a parametrized representation for nested belief in the CL framework.

In this chapter, we propose a CL framework to investigate human-like learning. Such a framework encompasses and goes beyond existing machine learning paradigms. The framework is important for understanding real life learning and teaching, and is necessary for human-robot interactions in real life settings. To begin with, we introduce a formal definition of knowledge, from which we can define common knowledge and belief-of-belief.

13.2 Common Knowledge Representation

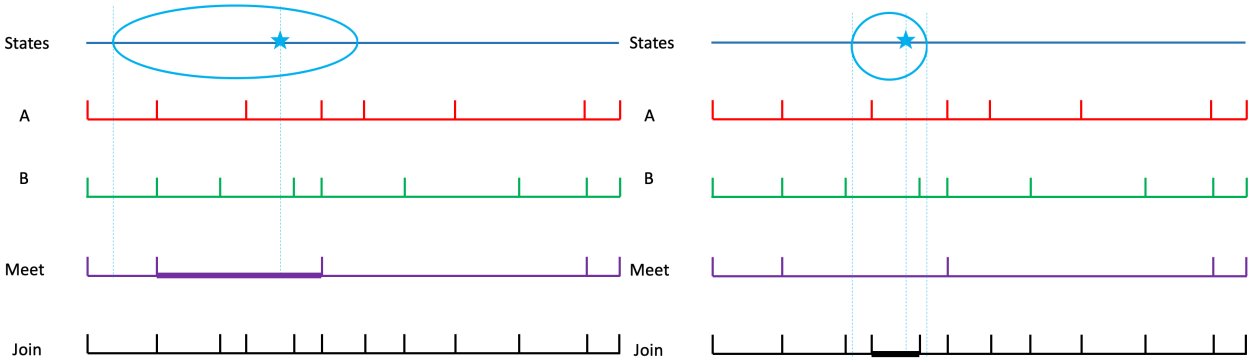
13.2.1 Overall Setting

There are two agents A and B in a physical world $\omega \in \Omega$. Each of them knows some facts about the world. They want to achieve certain tasks requiring some facts about the world to become their common knowledge. At the beginning of the process, agent A and B receives different sensory inputs from the common physical state ω. Then, they communicate with each other to infer some facts about the common physical state collaboratively. Each agent has a mind that includes the agent's model of the physical state, the agent's utility function or value, and the agent's intent and goal. For the agents to communicate with each other effectively, each agent must also model the minds of the other agents. The agents interact with each other and gradually reach common ground that consists of common knowledge shared by both agents as well as a common value function.

The key notion is that of "common knowledge." When we say that an event is "common knowledge," we mean more than just that both A and B know it; we require also that A knows that B knows it, A knows that B knows that A knows it, and so on. For example, if A and B are both present when the event happens and see each other there, then the event becomes common knowledge [685].

13.2.2 From Distributed Knowledge to Common Knowledge

Let $(\Omega, \mathcal{E}, P)$ denote a probability space where Ω is the set of states, $\mathcal{E}$ is the set of events, and P gives the probability assigned to each event. We make the assumption that both agents have access to the probability space, and all uncertainty is represented in the state. For agent $i \in \{A, B\}$, we define i's perception of the world as a partition $\mathbf{\Pi}_i$ of Ω. If the true state of the world is $\omega \in \Omega$, then i is informed of the element of $\mathbf{\Pi}_i$ that contains ω, which we denote as $\mathbf{\Pi}_i(\omega)$. An event $E \in \mathcal{E}$ is a subset of Ω. Agent i *knows* an event E at state ω if $\mathbf{\Pi}_i(\omega) \subseteq E$. In most cases, the set of events $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ are predefined. Both agents know the definition of every event. Namely, given an event, an agent always know in what states does that event happen. We can also regard $\mathcal{E}$ as a commonsense given a probability space. Here we use $\mathcal{P}(\Omega)$ to represent the power set of Ω and



(a) Common Knowledge (b) Distributed Knowledge

Figure 13.2: Aumann Knowledge Structure. The blue circle represents an event. The true state ω is indicated by a blue star. In Fig. 13.2a, the event is common knowledge as it contains the cell of the meet that includes ω. In Fig. 13.2b, the event is distributed knowledge (but not common knowledge) as it contains the relevant cell of the join. Everything that is common knowledge is also distributed knowledge.

the set of event is a subset of the power set, because not all combination of states are meaningful events that the agents care about. Given the same Ω, $\mathcal{E}$ can be task-dependent[1].

Name	Notation	Argument	Definition
Commonsense	$\mathcal{E}$	ω, E	Can tell whether $\omega \in E$
Perception	$\mathbf{\Pi}$	ω	Feel $\mathbf{\Pi}(\omega)$ when ω
Knowledge	$\mathbf{\Pi}(\omega) \subseteq E$	ω, E	Knows E at ω

Table 13.1: Notation Definitions

One could equivalently represent perception through random variables. Say that agents A and B observe the random variables O_A and O_B respectively. In the language of measure theory, these random variables are functions of the state. So if ω is the true state, then the agents observe $O_A(\omega)$ and $O_B(\omega)$ respectively. For each agent $i \in \{A, B\}$, construct the partition $\mathbf{\Pi}_i$ such that for each ω_1, $\mathbf{\Pi}_i(\omega_1) = \{\omega : O_i(\omega) = O_i(\omega_1)\}$. In other words, every cell $c \in \mathbf{\Pi}$ is a subset of Ω containing physical states generating the same observation.

$$\forall i \in \{A, B\}, \omega_1, \omega_2 \in \Omega, O_i(\omega_1) = O_i(\omega_2) \Leftrightarrow \mathbf{\Pi}_i(\omega_1) = \mathbf{\Pi}_i(\omega_2) \tag{13.3}$$

Suppose the observation functions O_A and O_B are known to both agents, or equivalently the partitions $\mathbf{\Pi}_A$ and $\mathbf{\Pi}_B$ are known to both agents. Under this assumption we will use the definition of common knowledge used by Aumann and Fagin [683]. If the true state is ω, then an event E is common knowledge at ω if and only if $\mathbf{\Pi}_C(\omega) \subseteq E$, where $\mathbf{\Pi}_C = \mathbf{\Pi}_A \wedge \mathbf{\Pi}_B$ (here, $\wedge$ indicates the meet of the two partitions). E is distributed knowledge at physical state ω if and only if $\mathbf{\Pi}_D(\omega) \subseteq E$, where $\mathbf{\Pi}_D = \mathbf{\Pi}_A \vee \mathbf{\Pi}_B$ (here, $\vee$ indicates the join of the two partitions). Intuitively, an event is distributed knowledge means that the agents verify the event if they combine their information, even if they cannot certify the event on their own. Common knowledge indicates that both agents can not only verify that the event has taken place, but can verify that the other agent can verify the event, and that the other agent can verify that they can verify the event, and so on. An illustration of common and distributed knowledge is shown in Fig. 13.2. We also give two simple examples using above knowledge representation.

[1]Notice that $\mathcal{E}$ is not a partition. Elements in $\mathcal{E}$ are not necessarily mutually exclusive and do not need to cover Ω

Example 13.2.1. *Consider the simple case where the state space consists of sets of integers. As a concrete example let* $\Omega = \{\{1,2,3,4\},\{3,4,5,6\},\{2,4,5,7\},\{2,3,5,8\},\{2,3,4,5\}\}$. *Label these elements as* $\omega_1,\ldots,\omega_5$ *respectively. Let agent A know the true state and agent B know nothing. In the language of observation functions,* $O_A(\omega)=\omega, O_B(\omega)=\varnothing$. *In the language of partitions, agent A's partition consists of singleton sets for each element, while agent B's partition contains only one cell. Here, the goal is for A to communicate its observed set to agent B. We will refer to the case that A knows the true state and B knows nothing as a* referential game *following [686].*

Example 13.2.2. *Consider situation of Example 13.2.1, but where the agents observe different features of the true state. Say that agent A only sees all the even numbers in the state, while B only sees all odd numbers in the state. This can be represented in the language of observation functions as* $O_A(\omega)=\{n|\forall n\in\omega, n\equiv 0\ (mod 2)\}, O_B(\omega)=\{n|\forall n\in\omega, n\equiv 1\ (mod 2)\}$. *In the language of partitions we have that the cells of* $\mathbf{\Pi}_A$ *are* $\{\omega_1,\omega_3,\omega_5\}$, $\{\omega_2\}$, *and* $\{\omega_4\}$ *and the cells of* $\mathbf{\Pi}_B$ *are* $\{\omega_1\}$, $\{\omega_2,\omega_4,\omega_5\}$, *and* $\{\omega_3\}$.

Knowledge Acquiring through Communication

With above definition of knowledge, we can model learning as a process of transferring information from distributed knowledge to common knowledge. Fig. 13.3 illustrates a toy example where an instance ω happens in a 1D Ω, *i.e.*, a time interval. The two agents A and B cannot observe ω, instead they observe some projections as input:

$$I_A = I_A(\omega) = (I_{A,1},...,I_{A,8}); I_B = I_B(\omega) = (I_{B,1},...,I_{B,9}) \tag{13.4}$$

Each input $I_{A,j}, I_{B,j} \in 0,1$ is binary: $=1$ if ω is on its right side and $=0$ if ω is on the left. We can define $\mathbf{\Pi}_{A/B}$ with $I_{A/B}$. If we define a probability measure p on Ω, then we can have $p(\omega)$ as the probability model on states and $p(E)=\sum_{\omega\in E} p(\omega)$ as the probability model on events. We can term an agent's uncertainty after receiving the observation as its imperceptibilty, defined by entropy of $p(\omega|\mathbf{\Pi}_{A/B})$, $H(p(\omega|\mathbf{\Pi}_{A/B})) = |\mathbf{\Pi}_{A/B}(\omega)|$, where $|\mathbf{\Pi}_{A/B}(\omega)|$ is the cardinality of the set.

As shown in Fig. 13.3, E_2 cannot be detected by A or B individually, but they can know E_2 after combining their knowledge. The goal is to pass minimum messages to transfer a distributed knowledge to common knowledge. In this example, the optimal messages for E_2 are:

$$m_1^{A\to B}(\omega) = I_{A,3}(\omega); m_1^{B\to A}(\omega) = I_{B,4}(\omega) \tag{13.5}$$

By only 1-round of messaging, E_2 becomes common knowledge. The perceived cell is compressed:

$$\mathbf{\Pi}_A(\omega, m_1^{B\to A}) = \{\omega' : I_A(\omega') = I_A(\omega) \wedge I_A(\omega') = m_1^{B\to A}\} \tag{13.6}$$

$$\mathbf{\Pi}_B(\omega, m_1^{A\to B}) = \{\omega' : I_B(\omega') = I_B(\omega) \wedge I_B(\omega') = m_1^{A\to B}\} \tag{13.7}$$

A/B's information gain from B/A's message is measured by the reduction of uncertainty/entropy,

$$\delta(m_1^{A/B\to B/A}) = \log \frac{|\mathbf{\Pi}_{A/B}(\omega)|}{\mathbf{\Pi}_{A/B}(\omega, m_1^{B/A\to A/B})} \tag{13.8}$$

Remark: When the agents gain information, this is represented through a refinement of their partitions. This is because new information can only narrow the set of possibly valid states, as the agents cannot loose previous information. This is equivalent to observing a new random variable, and computing the join of the corresponding partitions.

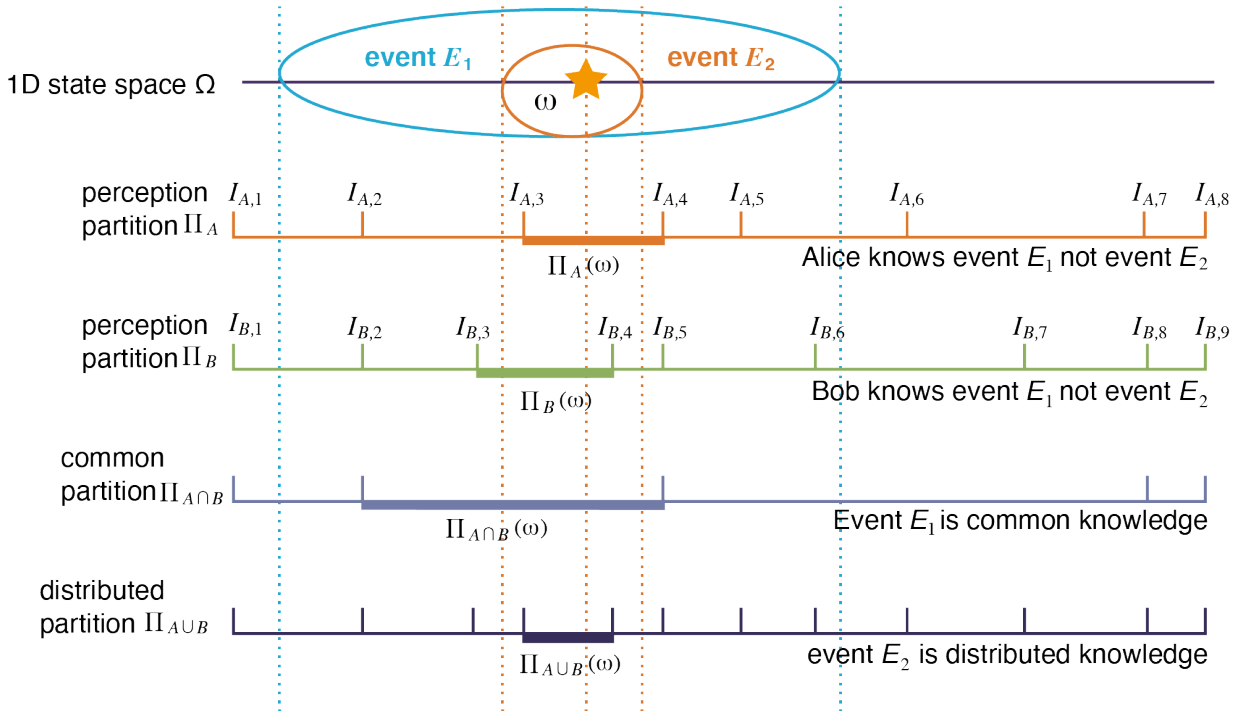


Figure 13.3: Common and distributed knowledge for inferring a state ω (star) in 1D space between A and B.

In CL, the agents are cooperative. Agent i's goal is to maximize its knowledge about the state (achieve the finest possible partition) but also for the other agent to achieve maximum information about the state. It does this through communicating its own information to its partner through messages, and through receiving messages from its partner and adding the content to its knowledge base. As the individual agents' partitions are refined, the common knowledge partition is refined. However the distributed knowledge partition remains fixed, as the agents cannot introduce knowledge that cannot be deduced from the combination of their observations through communication.

Goes beyond Shannon Limit with Theory of Mind Protocol

In the previous example, we show how learning can happen as transferring information from distributed knowledge to common knowledge. In fact, using different communication protocols, the efficiency of the communication can be also distinctive. Recall example 13.2.1. Suppose the true world state is $\omega_5 = \{2, 3, 4, 5\}$ and agent A wants to teach the true state to agent B by indicating a number that is included in the set. If the two agents are using the Shannon protocol, then agent A at least needs 4 messages to successfully identify ω_5. On the contrary, if agent A and B have ToM and communicate cooperatively, they can finish the teaching successfully with only 1 message. That is, $\omega_{1:4}$ all have unique identifier, namely, 1 for ω_1, 6 for ω_2, 7 for ω_3 and 8 for ω_4. Hence, suppose the intended state is one of the $\omega_{1:4}$, agent A will definitely use one of the $\{1, 6, 7, 8\}$ as the message. That is to say, as long as the message is not one of the $\{1, 6, 7, 8\}$, *e.g.* 2, then agent B can infer that the true state is ω_5. Fig. 13.4 compares the two protocols by visualizing the belief transition process.

Next, we generalize the 1D toy example in Fig. 13.3 by showing how to communicate distributed knowledge for high-dimensional spaces. Let state ω be an image and each agent has N_i "neurons"

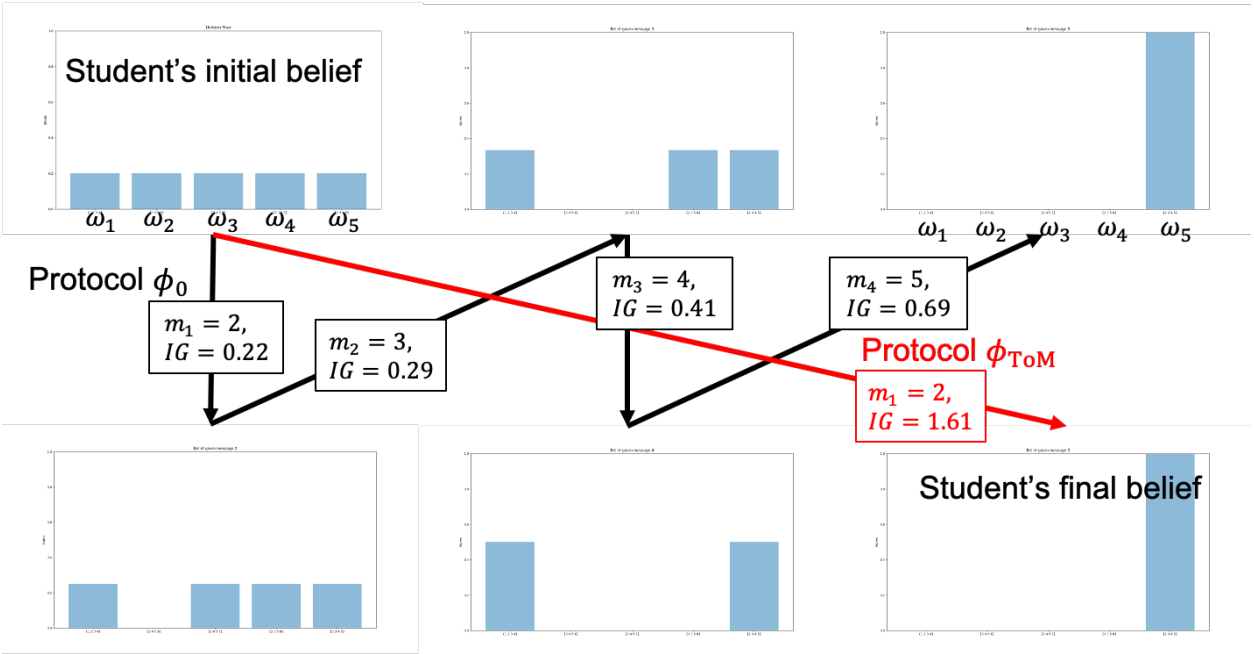


Figure 13.4: Comparison between the Shannon protocol and ToM protocol. the information gain is calculated as the decrease of the belief entropy after a message has been received. Here we assume that the initial belief of agent B is uniform over all possible world states, and agent B updates belief with Bayesian rule.

as its observations for $i \in A, B$:

$$I_i(\omega) = (I_{i,1}, ..., I_{i,N_i}), \text{ with } I_{i,j} = h_{i,j}(\omega), \forall i, j \tag{13.9}$$

Each neuron is an indicator or ReLu projection of the image, inside a layered network in Fig. 13.5 (d).

$$I_{i,j} = h_{i,j}(\omega) = \mathbf{1}(\langle \omega, \theta_{ij} \rangle \geqslant 0) \text{ or } = \max(0, \langle \omega, \theta_{ij} \rangle). \tag{13.10}$$

θ_{ij} is the weight of neuron h_{ij} and is a hyper-plane splitting the state space. Neuron in higher layers is a weighted poly-hyperplane and further splits cells. Thus we have the partitions $\mathbf{\Pi}_A$ and $\mathbf{\Pi}_B$ as cells shown in Fig. 13.5 (c). This is true for high-dimensional spaces and for multiple layer neural networks. Fig. 13.5 (d) zooms in two nested cells: $\mathbf{\Pi}_{A,a}$ is bounded by 8 neurons in red, and $\mathbf{\Pi}_{A,b}$ is bound by 4 neurons in green. When agent A knows an event $\omega \in \mathbf{\Pi}_{A,a}$ and tells B by

$$m^{A \to B}_{[1-8]}(\omega) = (h_{A,a_1}, ..., h_{A,a_8}). \tag{13.11}$$

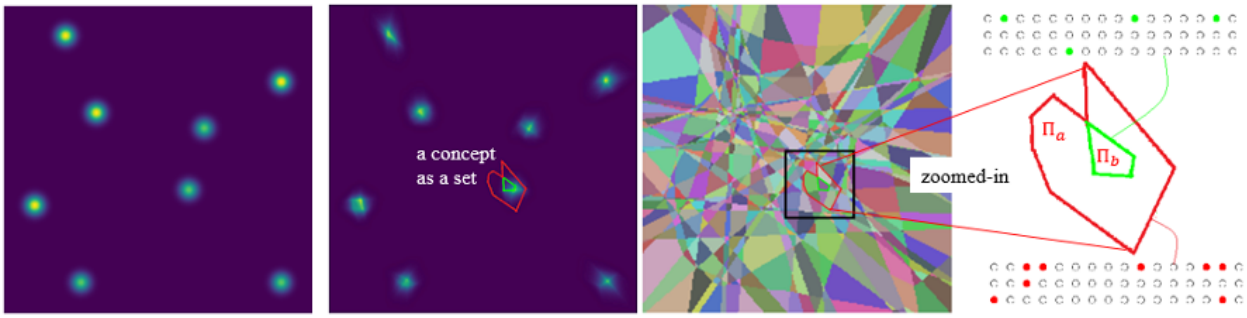
Then, agent B will refine its perception from $\mathbf{\Pi}_B(\omega)$ to $\mathbf{\Pi}_{A,a}$. an information gain by Shannon is:

$$\delta_{shannon}(m^{A \to B}_{[1-8]}(\omega)) = \log \frac{|\mathbf{\Pi}_B(\omega)|}{|\mathbf{\Pi}_{A,a}|} \tag{13.12}$$

. Similarly, if agent B has a ToM capability, he will read-between-lines: since Alice could but didn't send a *shorter message* using the 4 blue neurons, B infers that $\omega \notin \mathbf{\Pi}_{A,b}$. The new information that B gains from A's message is:

$$\delta_{ToM}(m^{A \to B}_{[1-8]}(\omega)) = \log \frac{|\mathbf{\Pi}_B(\omega)|}{\mathbf{\Pi}_{A,a}/\mathbf{\Pi}_{A,b}}. \tag{13.13}$$

$\mathbf{\Pi}_{A,a}/\mathbf{\Pi}_{A,b}$ means the set of elements that are in $\mathbf{\Pi}_{A,a}$ but not in $\mathbf{\Pi}_{A,b}$. It is obvious that $\delta_{ToM} > \delta_{Shannon}$.



(a) 2D state space and truth distribution. (b) Learned model $p(\omega;\Theta_{\mathbf{B}})$. (c) space partition Π and a concept (d) Concept with an internal cell

Figure 13.5: An example of partition in 2D space and inference of a state or concept using the ToM-protocol.

Remark: The ToM-protocol goes beyond Shannon's information by reflecting the minds of the other agent: agent A selects messages after deliberating what B knows, and B reasons why A sent this message not other plausible messages. This is a recursive mutual reasoning process. The level and complexity of the recusion will be discussed in later sections.

13.2.3 Belief over Belief

We can define a probability measure on $\mathcal{E}$ by assigning positive weights for each states. Namely, for agent $i \in \{A, B\}$ its belief for event E at state ω is

$$P(E|\mathbf{\Pi}_i, u_i) = \frac{\sum_{\omega\in\mathbf{\Pi}_i(\omega)\cap E} u_i(\omega)}{\sum_{\omega\in\mathbf{\Pi}_i(\omega)} u_i(\omega)} \tag{13.14}$$

where $u_i(\omega) \geqslant 0$ represents the weight assigned to the state ω by agent i.

Definition 6. *Suppose agent A's weight assignments are known to agent B. At state ω, agent B's belief about agent A's belief about event E can be defined as: $\forall \mu, \mu \in [0,1]$:*

$$\begin{aligned} P\big(P(E|\mathbf{\Pi}_A, u_A) = \mu|\mathbf{\Pi}_B, u_B, \omega\big) &= \sum_{c\in\mathbf{\Pi}_A} P(\mathbf{\Pi}_A(\omega) = c|\mathbf{\Pi}_B, u_B, \omega)\mathbf{1}\big(P(E|\mathbf{\Pi}_A = c, u_A) = \mu\big) \\ &= \sum_{c\in\mathbf{\Pi}_A} P(\mathbf{\Pi}_A(\omega) = c|\mathbf{\Pi}_B, u_B, \omega)\mathbf{1}\Big(\frac{\sum_{\omega\in c\cap E} u_A(\omega)}{\sum_{\omega\in c} u_A(\omega)} = \mu\Big) \end{aligned} \tag{13.15}$$

$$\forall c \in \mathbf{\Pi}_A, P(\mathbf{\Pi}_A(\omega) = c|\mathbf{\Pi}_B, u_B, \omega) = \frac{\sum_{\omega\in\mathbf{\Pi}_B(\omega)\cap c} u_B(\omega)}{\sum_{\omega\in\mathbf{\Pi}_B(\omega)} u_B(\omega)} \tag{13.16}$$

Remark: When there are finite number of cells in $\mathbf{\Pi}_A$, equation Eq. (13.15) is only non-zero for finite number of $\mu \in [0,1]$.

Cannot Agree to Disagree

Above definition of belief of belief will lead to an important fact, if the posterior of an event is a common knowledge between two agents, then their posterior must be the same.

Theorem 13.2.1. *For an event $A \in \mathcal{E}$, define q_A as the posterior probability equals to $p\big(A|\mathbf{\Pi}(\omega)\big)$ and $q_B = p\big(A|\mathbf{\Pi}(\omega)\big)$. If it is common knowledge that at state ω and $q_A = q_1, q_B = q_2$, then $q_1 = q_2$ [685].*

Proof. Let Π be the member of $\mathbf{\Pi}_A \wedge \mathbf{\Pi}_B$ that contains ω. Write $\Pi = \cup_j \pi_j$ where the π_j are disjoint members of $\mathbf{\Pi}_A$. Since $q_A = q_1$, throughout Π, we have $p(A \cap \pi_j)/p(\pi_j) = q_1$, for all j; hence $p(A \cap \pi_j) = q_1 p(\pi_j)$, and so by summing over j we get $p(A \cap \Pi) = q_1 p(\Pi)$. Similarly, $p(A \cap \Pi) = q_2 p(\Pi)$, and so $q_1 = q_2$. □

This theorem coincides with our goal in CL, whose goal is for two agents to share their individual information and agree on a common belief about a certain event. When a certain common belief is acquired, the learning process can be halted. In Section 13.7, we will discuss this topic in more detail. In next section, we give example usage of the ToM in some communication games, where two parties of the game have private information unknown to others and the success of the game requires the integration of both parties' information.

13.3 Applications: Referential Game

We discussed what is knowledge, how to model learning as a communication process transiting information from distributed knowledge to common knowledge and the advantage of the ToM protocol over Shannon protocols. In this section, we show how a ToM protocol can be acquired and illustrate the advantage of such a protocol over others. It is very common to study communication in multi-agent games [687, 688, 686]. In most communication games, there are two agents and each of them has some private information that the other doesn't know. The two have to work together to achieve a certain goal, whose completion requires the private information from both of the agents. In this section, we present a teacher-student scenario, in which only one of the agents has private information. This agent needs to teach the other agent its private information, playing the role of a teacher.

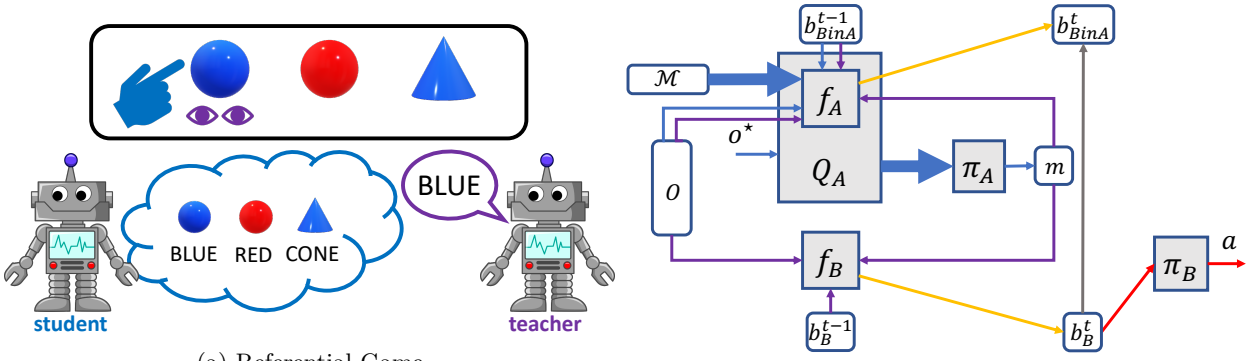
13.3.1 Referential Game

Game Definition

The referential game can be defined by a tuple $\langle A, B, \Omega, \mathcal{M}, \mathcal{A} \rangle$, where A and B stand for a teacher and a student. Ω is the instance space, where the distractors and targets are sampled from. $\mathcal{M}$ is the message space and $\mathcal{A}$ is the student's action space. In a specific game, a set of instances $O \subseteq \Omega$ is sampled from Ω as candidates, and one of the candidates $o^\star \in O$ is designated as the target, while the rest, $O/\{o^\star\}$, are distractors. The candidates O are available to both of the agents, while only the teacher knows the target, $o^\star$. Agents take turns in this game. In every round, the teacher first sends a message $m_t \in \mathcal{M}$ to the student, followed by an action $a_t \in \mathcal{A} = \{1, 2, ..., |O|, \Xi\}$ taken by the student, where number 1 to $|O|$ represent "identify a certain instance as the target," Ξ means "wait for next message" and t stamps the t-th round. Every message comes with a message cost, c_m, and the total gain for both the teacher and the student, given a_T the first non Ξ action, is $R = \sum_{t=1}^{T} -c_{m_t} + \mathbf{1}(o^\star = O[a_T])$. Notice that the game ends when the student performs a non Ξ action. We define a protocol between A and B as a set of policies

$$\Pi = \langle \pi_A : \mathbb{P}(\Omega) \times O \times \mathcal{M}^* \times \mathcal{M} \to [0, 1], \pi_B : \mathbb{P}(\Omega) \times \mathcal{M}^* \times \mathcal{M} \times \mathcal{A} \to [0, 1] \rangle$$

$\mathbb{P}(\Omega)$ is the power set of Ω, where O sampled from, and $*$ is the kleene star, standing for the history of message. Intuitively, the teacher selects a message based on the distractors, the target and the communication history. The student chooses an action according to the candidates, the communication history and latest message. The goal for both of the agents is to maximize the



(a) Referential Game

(b) ToM Agents Interaction Pipeline

Figure 13.6: (a) An example referential game. (b) First, the teacher chooses a message according to the context and her prediction of the student's reaction (**blue** arrows). After a message is sent, the student updates his belief and the teacher updates her estimation of student's belief (**purple** and **orange** arrows). Then, the student either waits or selects a candidate (**red** arrows). Only in the training phase, the actual student belief will be returned to the teacher (**gray** arrow). Bold arrows stand for the whole message space being passed. Notice that ϕ_A is part of the Q_A. O and b^{t-1}_{BinA} are passed in ϕ_A twice, for message selection and teacher's new belief estimation. Empty boxes are game and time variants, while shadowed boxes are agents' constant mental structure.

expected gain:

$$E_{O\sim\mathbb{P}(\Omega),o^\star\sim O,m_{1:T}\sim\pi_A,a_{1:T}\sim\pi_B}\left[-\sum_{t=1}^{T}c_{m_t}+\mathbf{1}(o^\star=O[a_T])\right] \tag{13.17}$$

Algorithm

Emergence of ToM Protocol: Our goal is to learn a protocol for agent A and B so that they can communicate with ToM capability. To avoid tracking the message history, which scales exponentially with the time, we use beliefs as sufficient statistics for the past. Hence, ToM can be embodied as estimating partner's current and future belief, then choose the most ideal action to manipulate them as needed. In the referential game, since the teacher knows the target, only the student holds a belief, b_B, about the target. Utilizing the obverter technique, we let the teacher holds a belief b_{BinA} as her estimation of student's belief. b_{BinA} is still a distribution over the candidates. We didn't use a distribution over distribution to model this nested belief because the belief update process is deterministic for rational agents following Bayesian rule. Given b^0_B a uniform distribution over candidates, $P(b^t_B)$ is unimodal with uncertainty merely from the likelihood and can be approximated with a single point.

Before speaking, teacher traverses all messages and predicts the student's new belief after receiving each message. She then sends the message leading to the most optimal student's new belief. Hearing the message, student updates his belief and takes action. This process is visualized in Fig. 13.6b and formalized in algorithm Algorithm 5 Line 11 to Line 20. The recursive mutual modeling in ToM is integrated within the belief update process. $\phi_{\theta_i}, i \in \{A, B\}$ are belief update functions parameterized by θ_i, taking in candidates, current belief, message and returning a new belief. The beliefs in our model are semantically meaningful hidden variables in teacher's Q-function and student's policy network, as the student directly samples an action according to his belief. The evolving of the belief update function reflects the protocol dynamics between the agents.

Within ϕ, we align the candidates' embedding into a $1 \times |O| \times D$ tensor and apply 1×1 convolution to every candidate, where D is the candidate embedding dimension. We sum the candidates

embedding as the context embedding and concatenate it after each candidate's embedding, followed by another 1×1 convolution. In each phase, we first train the teacher for a fixed student. Next, the student is trained to adapt to the teacher.

Teacher: The teacher selects messages according to her Q-values and belief update function. We use $\phi_{\theta_A}(O, b, m)$ to denote teacher's belief update function, which takes in the candidates set, current belief estimation and a message. The return value of this function is a new belief estimation $b' \in \Delta O$. ΔO represents all probabilistic distributions over the candidates. This function can be parameterized as a neural network with weighted candidates encoding and messages as inputs and softmax as the output layer. The return value of the belief update function is directly fed into the Q-function. In practice, we implement it as a submodule of the Q-net. That is, the output of the belief update function is used in A's Q-function and to predict student's belief in next step during testing. The teacher chooses messages according to her Q-function.

$$\pi_{\theta_A}(m|O, o^\star, b) = \frac{\exp\left(\beta Q_{\theta_A}(O, o^\star, b, m)\right)}{\sum_{m' \in \mathcal{M}} \exp\left(\beta Q_{\theta_A}(O, o^\star, b, m')\right)} \tag{13.18}$$

$$\begin{aligned} Q_{\theta_A}(O, o^\star, b, m) = &-c_m + E_{a \sim \pi_{\theta_B}\left(\phi_{\theta_B}(O,b,m)\right)}\big[\mathbf{1}(O[a] = o^*) \\ &+ \mathbf{1}(a = \Xi) \max_{m'} Q_{\theta_A}\big(O, o^\star, \phi_{\theta_B}(O, b, m), m'\big)\big] \end{aligned} \tag{13.19}$$

By definition, the teacher's Q-function relies on student's policy and belief update function. She has no access to these student's functions, but since we never train the teacher and student simultaneously, the expectation can be approximated through Monte-Carlo (MC) sampling. To form a protocol, agent A needs to learn two functions, her belief update function ϕ_{θ_A} and Q_{θ_A}. In the training phase, every time the student receives a message, he returns his new belief b_B^t to the teacher. During testing, she needs to use the output of ϕ_{θ_A} to approximate student's new belief. We train ϕ_{θ_A} by minimizing the cross-entropy, H, between b_B^t and teacher's prediction, denoted as L^{Obv}, the obverter loss. Teacher's Q-function is learned with Q-learning [689]. The λ in line 33 controls the scale of the two losses.

Student: We directly learn the belief update function and policy of the student through the REINFORCE algorithm [690]. In the referential game, student's policy is quite simple. If his belief is certain enough, he will choose the target based on his belief; otherwise, wait for further messages. The output of the policy network is a distribution with $|O| + 1$ dimensions. The last dimension is a function of the entropy of the original belief. If the belief is uncertain, this value will be dominant after normalization. ϕ_{θ_B} has the same structure as ϕ_{θ_A}. ϕ_{θ_B} and π_{θ_B} can be parameterized as an end-to-end trainable neural network, with the candidates encoding, original belief and received a message as the input and returning an action distribution.

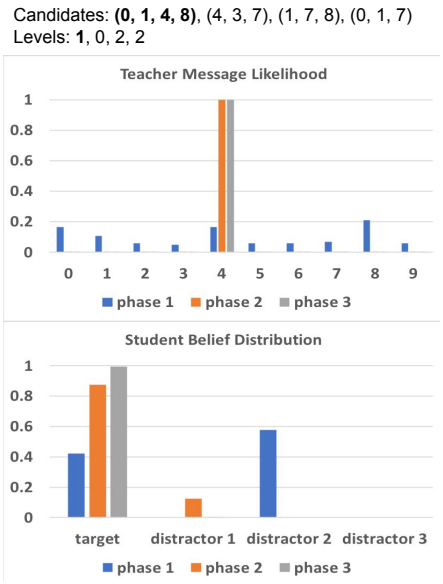
Adaptive Training: The whole training process can then be summarized as Algorithm 5. Both the teacher and student are trained in adaptive manner to maximize their expected gain defined in Eq. (13.17). The training details for teacher are illustrated in Line 28-34 of Algorithm 5, while the training details of student are in Line 35-39 of Algorithm 5.

Results

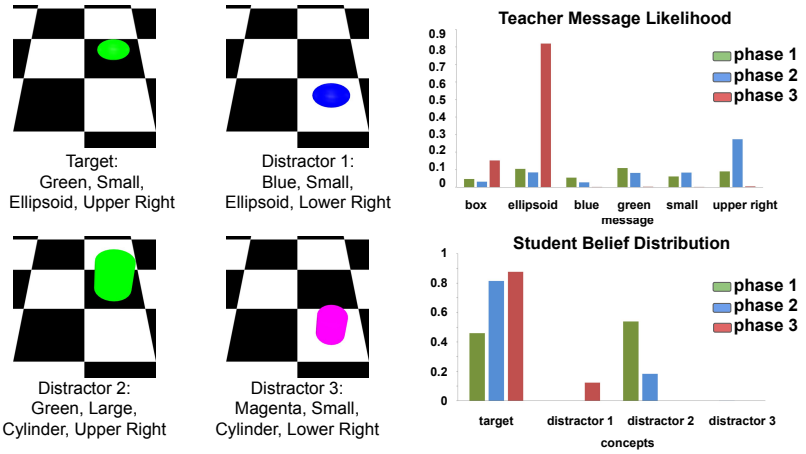
We evaluated our algorithm with two datasets, number set and 3D objects, and played referential games with four or seven candidates. The number set is a symbolic dataset, with an instance as a set of categorical numbers. For example, $[(1, 2, 3, 9), (1, 2, 4), (2, 3), (3, 4, 5)]$ consists a referential game with four candidates. Notice that the numbers are merely symbols without numerical order. If there are four candidates, we randomly choose numbers from 0 to 9, with maximum four numbers in a set; if seven candidates, we choose from 0 to 11, with maximum five numbers in a set. Each set

Algorithm 5: Iterative Adaption Protocol Emergence

Initialize θ_A, θ_B
No. candidates K
Learning rate η, Batch size N
for each phase **do**
  for $i \in \{A, B\}$ **do**
    Initialize replay buffer $\mathcal{D} \leftarrow \varnothing$
    while train agent i **do**
      $t = 1$
      Initialize $\mathcal{E} \leftarrow \varnothing$
      repeat
        if $t = 1$ **then**
          Sample $O = \{\omega_1, ..., \omega_K\}$
          Random select $o^\star = \omega_j$
          Initialize b_B^0, b_{BinA}^0 as uniform distribution
        end if
        $m_t \sim \pi_{\theta_A}(m|O, o^\star, b_{BinA}^{t-1})$
        $b_B^t = \phi_{\theta_B}(O, b_B^{t-1}, m_t)$
        $a_t \sim \pi_{\theta_B}(b_B^t)$
        $r_t = -c_{m_t} + \mathbf{1}(O[a_t] = o^\star)$
        $b_{BinA}^t = b_B^t$
        if $i = A$ **then**
          $\mathcal{D} \leftarrow \mathcal{D} \cup \{(O, o^\star, b_{BinA}^{t-1}, m_t, b_B^t, r)\}$
        else
          $\mathcal{E} \leftarrow \mathcal{E} \cup \{(O, b_B^{t-1}, m_t, a_t, r)\}$
        end if
        $t \leftarrow t + 1$
      until $a_t \neq \Xi$
      if $i = A$ **then**
        Sample $\{(O, o^\star, b_{BinA}^{t-1}, m_t, b_{BinA}^t, r)\}_N \sim \mathcal{D}$
        $\xi = r + \gamma \arg\max_m Q_{\theta'_A}(O, o^\star, b_{BinA}^t, m)$
        $L^Q = \frac{1}{N} \sum_N ||\xi - Q_{\theta_A}(O, o^\star, b_{BinA}^{t-1}, m_t)||^2$
        $L^{Obv} = \frac{1}{N} \sum_N H(b_{BinA}^t, \phi_{\theta_A}(O, b_{BinA}^{t-1}, m_t))$
        $\theta_A \leftarrow \theta_A - \eta \nabla_{\theta_A}(L^Q + \lambda L^{Obv})$
        Update $\theta'_A \leftarrow \theta_A$ periodically
      else
        Compute $R_t = \sum_{k=t}^T \gamma^{k-t} r_k$ for t in $\mathcal{E}$
        $J = \frac{1}{N} \sum_N \log \pi_{\theta_B}(a_t|\phi_{\theta_B}(O, b_B^{t-1}, m_t)) R_t$
        $\theta_B \leftarrow \theta_B + \eta \nabla_{\theta_B} J$
      end if
    end while
  end for
end for



(a) Number Set Example



(b) Geometric Shape Example

Figure 13.7: 4 distractors referential game example. Number set on the left (candidates listed in the title with the target in bold fonts) and 3D objects on the right. Due to the space limit, we only show the message distribution for the target and student's new belief after receiving the most probable message. As for the teacher's message distribution for distractors, all probability weights concentrate on the unique identifiers after the first phase of training. Student's belief illustrates that teacher's most probable message, though consistent with multiple candidates, can successfully indicate the target with more confidence as training goes. In general, both agents' behavior becomes more certain, and the certainty coordinates.

is encoded by multi-hot encoding. There are 385 and 1585 different possible number sets, consisting up to 9.0×10^9 and 4.9×10^{18} different games with four and seven candidates. Number sets make a generic referential game prototype, where each instance can be disentangled into independent attributes perfectly. To verify the generality of our algorithm on more complicated candidates, we used MoJoCo physical engine to synthesize RGB images of resolution 128×128 depicting single 3D object scenes. For each object, we pick one of six colors (blue, red, yellow, green, cyan, magenta), six shapes (box, sphere, cylinder, pyramid, cone, ellipsoid), two sizes and four locations, resulting in 288 combinations. In every game, candidates are uniformly sampled from the instances space. We use a message space with the same size as the number of attributes appeared in the dataset, *i.e.*, 10 or 12 for number set, and 18 for 3D objects. In every game, we only allow one round of communication with one message. To prevent collusion using trivial position indicator, candidates are presented to the agents in different orders. We show an example for each type of data in Fig. 13.7.

13.4 Communication Problem Definition

In a communication problem, there is a world state ω sampled from the state space Ω, holding all possible states. Suppose there are two agents, **A** and **B**. Both agents know the state space, but they may not know the exact state. Each agent receives a perception from the state according to their observation function. Based on their perception and their model θ they can form a belief of the state. The goal is for the agents to communicate and exchange their information/knowledge about the state so that the exact state or some attributes of the state can be realized by both of the agents (certain attributes come to the common mind). See Fig. 13.8 as an illustration. To accomplish this goal, we assume that the two agents share the state space, speaks the same language (same message space). We call the language between the agents a protocol. More formal mathematical definitions will be discussed in later sections.

This definition of communication problem can be utilized to formalize many concrete communication examples in real life and can be easily generalized to scenarios involving more than two agents. Now that if we narrow down a little and look into a special case, in which agent **A** knows strictly more than agent **B** does. That is the information accommodated by I_B is a subset of that by I_B. Then, the general communication problem becomes a pedagogical problem, where **A** is the teacher and **B** is the student.

13.4.1 Insight from Human Pedagogy

As we compared machine learning with human learning, one might think of human solution of the communication problem defined above. The most important characteristic of human communication is that the two people will simulate their partner's reaction and act accordingly [691]. In the pedagogical scenario, for example, the teacher will consider student's reaction after receiving different messages, then selecting the one with the most ideal outcome. Similarly, by modeling the teacher, the student can usually infer teacher's intention behind the message, and absorb more information from the message than the content of the message per se, an ability known as **reading between lines**.

To develop this capability, agents need to accommodate more complicated structures than their own beliefs. The teacher needs to have a value function evaluating student's mental status so that she can have preference over messages. The student, on the other hand, should have an estimation of teacher's message usage given different intentions, so that a counterfactual reasoning can be conducted. Moreover, both agents needs to estimate their partner's current mental status, namely other's belief in one's own mind. We summarize the new mind structures in Fig. 13.9. The ability to

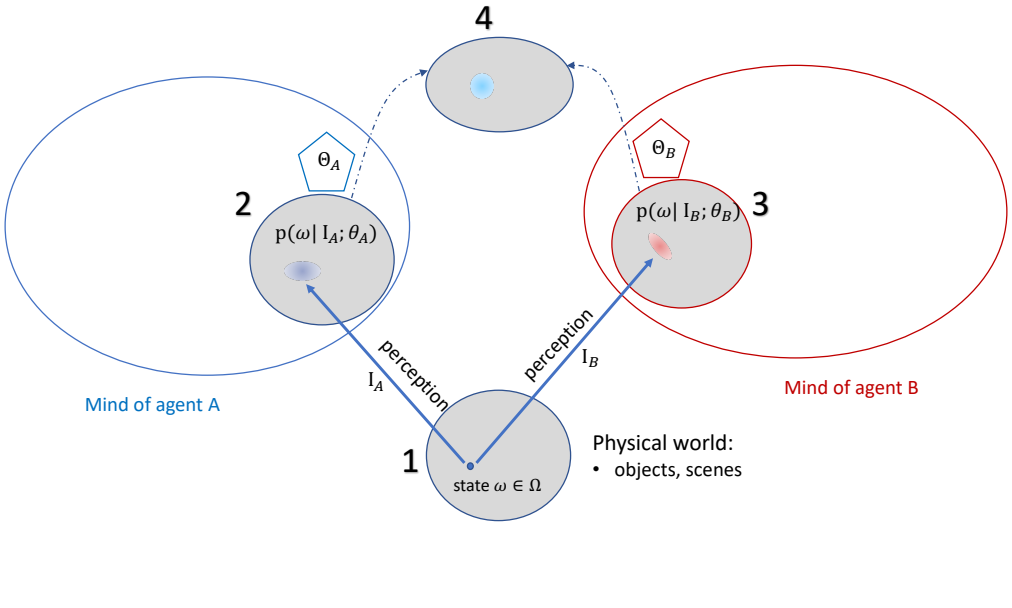


Figure 13.8: Illustration for general communication problem. 1) State space commonly known by both agents. 2) and 3) agents' belief about the actual state, a distribution over the state space. 4) common mind holds by the agents, also a distribution over the state space, usually with larger entropy than individual beliefs.

model other's mind even when the mind is different one's own is known as theory of mind (ToM). In later sections, we'll show detailed definition of CL with ToM and examples revealing its advantage over methods without it. Prior to that, let's briefly review the mainstream categories of machine learning algorithms, which we later show as special cases of CL.

13.5 Classic Learning Paradigms

13.5.1 Passive Learning

Passive learning is the most popular learning setting whereby a learner passively receives data from the outside world and tries to figure out the underlying regularity based exclusively on its input. See Fig. 13.10 for an illustration. The Probably Approximately Correct (PAC) model was introduced by Valiant [671] for its analysis. Under this model, learning an unknown concept from data is forming, with high probability, a good approximation of it. Moreover, the model requires a learning algorithm that is efficient. The main difference between passive learning and human learning is that the learner receives samples from a fixed distribution without initiatives. To disambiguate concepts, the size of the samples needs to be large.

13.5.2 Active Learning

Active learning aims to address a crucial issue in the era of big data: instances are often abundantly available at little cost whereas their labeling requires human effort and can thus be expensive. In contrast to the (agnostic) PAC model where a learning algorithm is fed with randomly sampled instances and their labels, in the active setting, a learner is only provided with unlabeled instances.

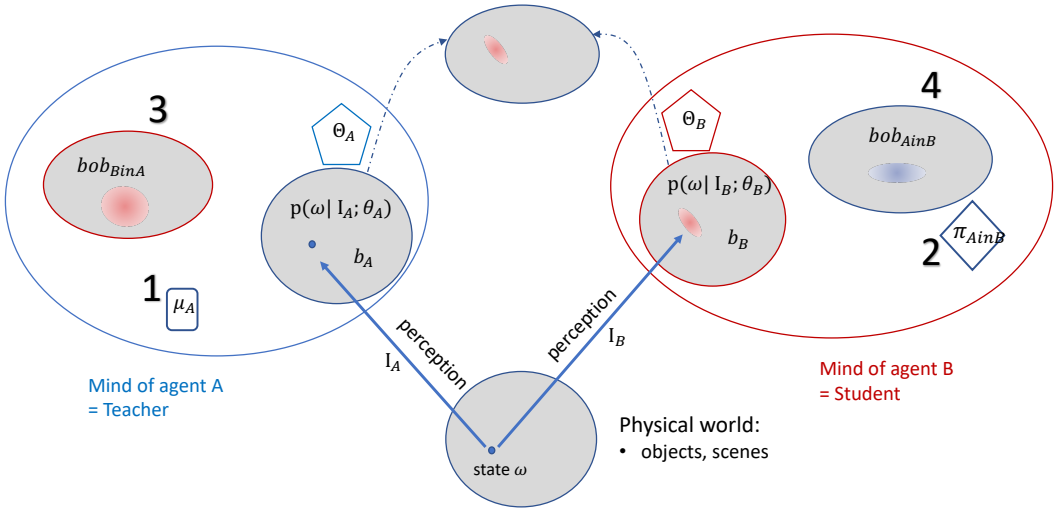


Figure 13.9: Mind representations for CL. We have four new structures with the rest same as in Fig. 13.8. 1) teacher's value function. 2) student's estimation of teacher's teaching schema. 3) and 4) agents' belief over other agent's belief. Every point represents a belief vector with length $|\omega|$.

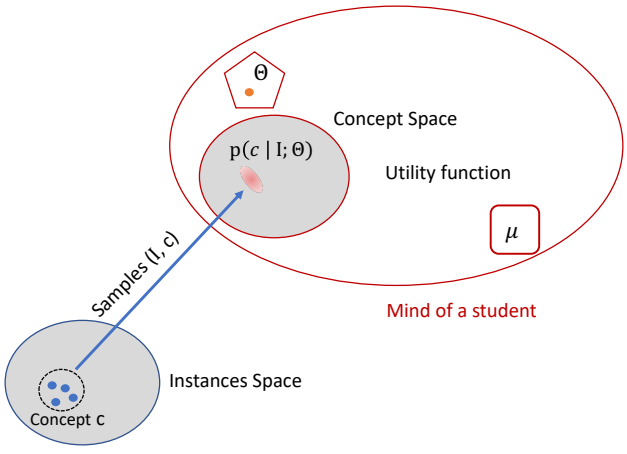


Figure 13.10: Illustration for passive learning.

Their labels are not revealed unless explicitly requested. See Fig. 13.11 for an illustration, where the learner **B** queries the teacher **A** about the labels of the selected examples.

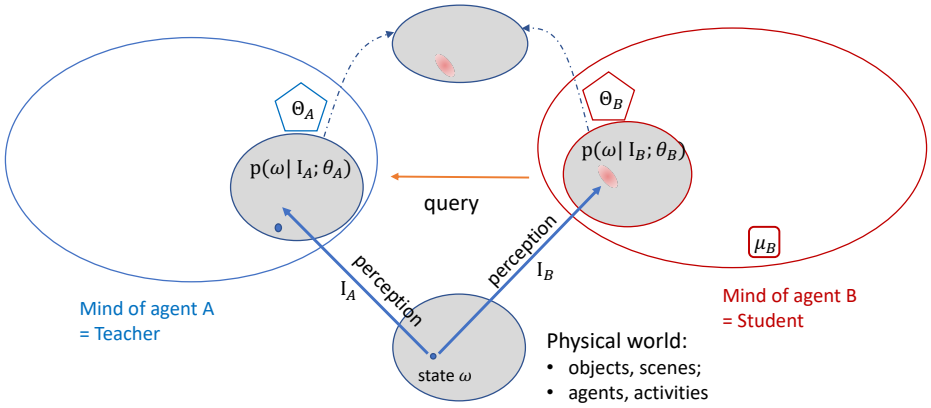


Figure 13.11: Illustration for active learning.

Despite this paradigm shift, an active learner is required to fulfill the same learning objective as its passive counterpart. But the two learning modes differ in terms of how they query labels: a passive algorithm is seen as implementing a trivial query strategy by requesting the label of every instance in its training data. On the contrary, an active learner may ask for significantly less labels than instances. As a consequence, label complexity, namely the minimum number of label queries required to achieve same certainty and accuracy as PAC learning, becomes a natural substitute to sample complexity for measuring active learning efficiency.

13.5.3 Algorithmic Teaching

In the settings considered so far, a learner has the goal of finding a hypothesis which best describes its training data, sampled from a distribution over which it has no control. Despite their solid theoretical guarantees, both passive and active learners require a rather large data set (*i.e.*, exceeding the sample complexity to make up for not knowing the distribution) to fulfill its goal reliably.

Algorithmic teaching takes a different approach to learning by involving a teacher in the process. A teacher is the counterpart to a learner, who shares the same hypothesis space and knows the target hypothesis for a given learning task. However, unable or disallowed to communicate the hypothesis directly, she exerts influence on the learner through its training data and aims to accelerate its learning. See Fig. 13.12 for an illustration, where the teacher **A** provides customized training examples to teach the learner **B**. To produce teaching materials suited to her audience, she has to possess some knowledge as to how the student reacts to data.

In general, teaching can thus be posed as an inverse problem to learning [692] and its goal is to find the smallest set of examples based on which an intended learner can output a desired hypothesis. This idea was first formalized in [676, 684] in which they consider teaching an arbitrary consistent learner *i.e.*, Empirical error minimization (ERM) and introduce a new way of measuring the complexity of a hypothesis space: the teaching dimension. There are multiple variations of

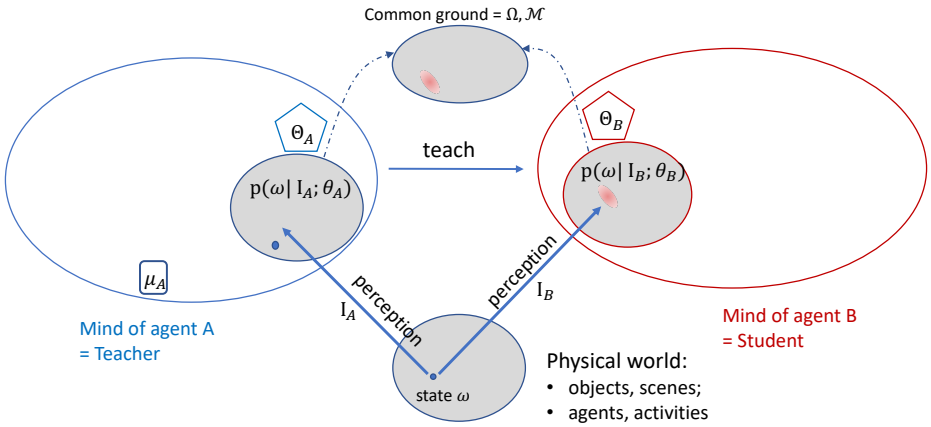


Figure 13.12: Illustration for algorithmic teaching.

teaching dimension corresponds to different teaching and learning paradigms [693]. In later chapters, we'll later argue that the complexity of CL can be represented with the recursive teaching dimension (RTD) [694, 695].

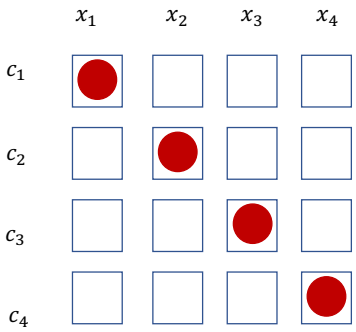
None of the listed learning paradigms assign full initiatives to both the teacher and the student, eliminating the possibility of human-like cooperative pedagogy from emerging. In next section, we'll introduce CL, in which both agents intentionally select their actions and show the generality of this learning paradigm.

13.6 Communicative Learning as A General Learning Paradigm

13.6.1 Motivation

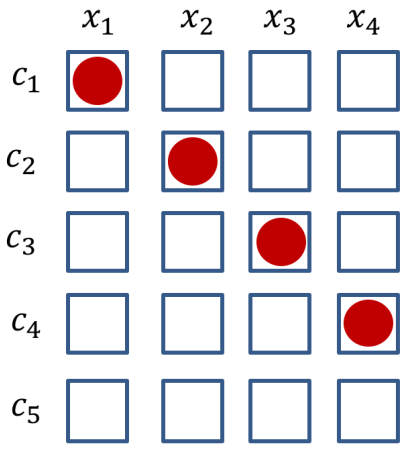
Before we actually start defining CL, let's take a look at an example and compare different learning paradigms. Suppose we have an instance space consisting of n elements $\{x_1, x_2, ..., x_n\}$ and a concept space $\Omega = \{c_1, c_2, ..., c_n\}$. Fig. 13.13 shows an example in the case of $n = 4$. As we can see that in this example, algorithmic teaching can identify any target concept with only one example by using the most representative one. In the example above, $(x_i, +)$ for c_i.

Nevertheless, in the next example shown in Fig. 13.14, all existing paradigms have high complexity. The new concept c_5 doesn't have any positive labeled instance, nor any unique identifier (a labeled instance that only consistent with one concept). Given this concept class, even algorithmic teaching cannot teach c_5 without sending all negative labeled instances. The advantage of CL reveals itself by utilizing the cooperative fact between the agents and pinpoint all concepts including c_5 with one labeled instance. Because the teacher is helpful, she will teach c_{1-5} using unique identifiers $(x_{1-5}, +)$. Thus, any negative labeled instances will suggest c_5. In next section we'll show how to mathematically formalize this process.



Paradigm	Complexity
passive statistical learning	$O(n \log n)$
active learning by label request	$O(n)$
algorithmic teaching	1

Figure 13.13: Left: A stylized case for $n = 4$, red dots means certain element is in a concept. Right: The teaching/learning complexity for different learning paradigms. Examples for each set are labeled instances such as $(x_1, +), (x_3, -)$ for c_1.



Paradigm	Complexity
passive statistical learning	$O(n\log(n))$
active learning by label request	$O(n)$
algorithmic teaching	$O(n)$
Communicative learning	*1*

Figure 13.14: As there is a concept without unique identifier, the teaching/learning complexity for all paradigms in section Section 13.5 cannot identify c_5 easily. CL, on the other hand, can pinpoint c_5 with any negative labeled example by taking advantage of the cooperative attributes between agents.

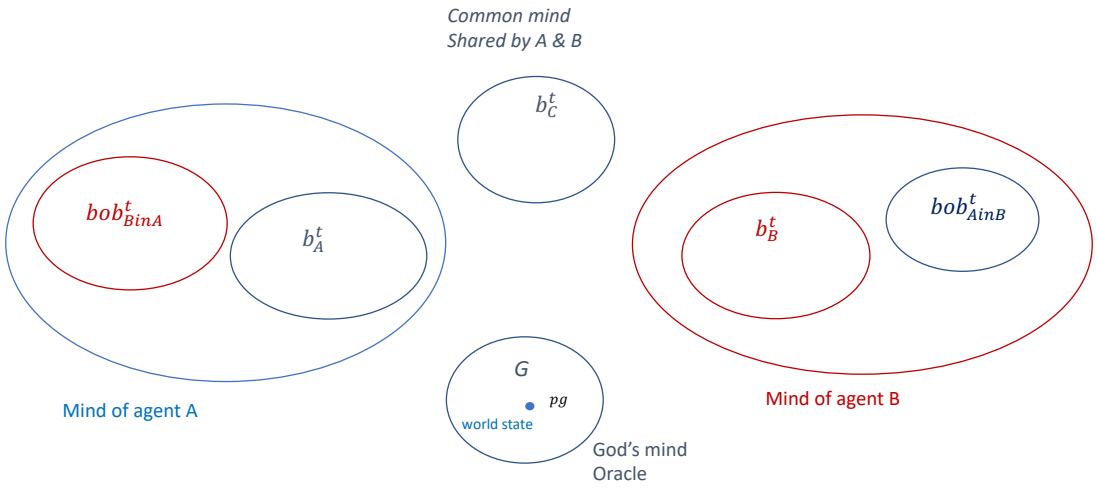


Figure 13.15: Diagrams for communicative learning agents' minds. "bob" stands for belief over belief.

13.6.2 Framework of Communicative Learning

We review the components requiring for CL in Fig. 13.15. Each agent has its own belief and an belief over its partner's belief. In addition, a common belief and the ground truth God's belief are included. In its turn of speaking, the agent will select a message using its value function taking its accessible mental structures as inputs. The listener, once receives the message, will update related mental structures as a reaction. Thus, we have:

$$m_{A\to B}^{t+1} = \arg\max_{m\in\mathcal{M}} \mu_A(b_A^t, bob_{BinA}^t, b_C^t, m) \tag{13.20}$$

$$m_{B\to A}^{t+1} = \arg\max_{m\in\mathcal{M}} \mu_B(b_B^t, bob_{AinB}^t, b_C^t, m) \tag{13.21}$$

$$b_A^{t+1} = \phi_A(b_A^t, b_C^t, bob_{BinA}^t, m_{B\to A}^{t+1}) \tag{13.22}$$

$$b_B^{t+1} = \phi_B(b_B^t, b_C^t, bob_{AinB}^t, m_{A\to B}^{t+1}) \tag{13.23}$$

$$bob_{BinA}^{t+1} = \phi_{BinA}(b_A^t, b_C^t, bob_{BinA}^t, m_{B\to A}^{t+1}) \tag{13.24}$$

$$bob_{AinB}^{t+1} = \phi_{AinB}(b_B^t, b_C^t, bob_{AinB}^t, m_{A\to B}^{t+1}) \tag{13.25}$$

where μs are score functions and ϕs are belief update functions, possibly different for different agents; $\mathcal{M}$ is the message space, shared by both agents. Formulas above only captures the intuition behind CL and needs to be designed for different tasks. However, the essence is to include "my estimate of your mental states" into my decision function, so that the cooperativeness between agents is considered and ToM can emerge. The goal of CL is to learn μs and ϕs for various tasks. Concrete examples will be provided in section Section 13.3.

To be noticed that all learning paradigms mentioned in Section 13.5 are in fact special cases of CL we just defined. For example, passive learning includes only one agent without score function for message selection and update its belief using consistent samples; active learning omits student's

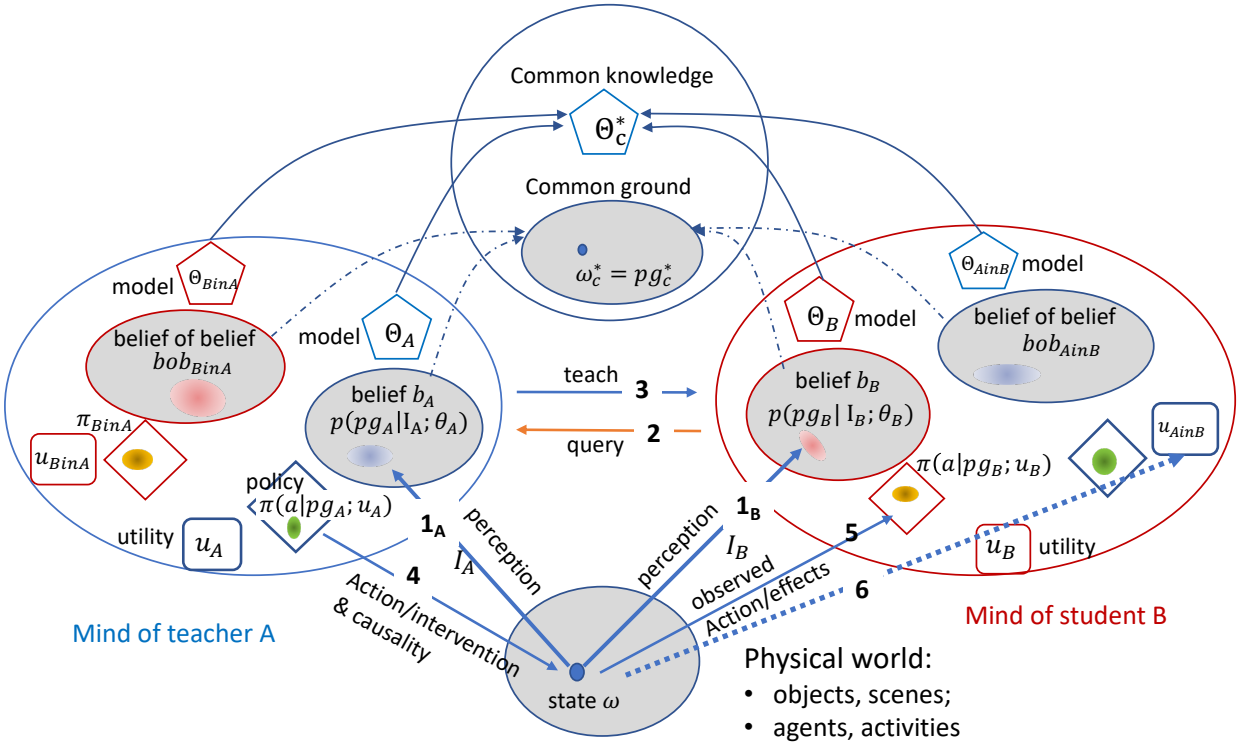


Figure 13.16: A zoomed-in view of CL: unifying all existing learning protocols and beyond. Each mind contains four spaces: i) pentagon for hypothesis space $\mathcal{H}$ of model Θ; ii) ellipse for the situate $\omega = pg$ in state space Ω where the belief is represented by a cloud; iii) diamond for policy π in action space; and iv) square for utility μ in a fluent space. The arrows illustrates the dynamics: observation, intervention, and messages.

belief during teacher's message selection and student's belief update function only takes in message and its current belief; algorithmic teaching lacks student's estimation of the teacher and simplifies student's belief update function. In next section, we'll show a comprehensive picture of learning established on top of the CL framework.

13.6.3 General Framework of Learning

To summarize the representation discussed in previous section, we illustrate the CL representations in Fig. 13.16 in a zoomed-in view. In the case of two agents, the representations include:

- CL $= (G, P_n, Q_n, \hat{P}_n, \hat{Q}_n, C_n)$ includes six minds shown by the 6 big ellipses. We can add higher level nested mental states, which will generate more advanced learning protocols.
- Each of $P_n, Q_n, \hat{P}_n, \hat{Q}_n, C_n$ has 4 representations $(\theta, \omega, \pi, \mu)$, whose spaces are represented by pentagon for hypothesis space $\theta \in \mathcal{H}$, ellipse for state space $\omega \in \Omega$, diamond for action space $a \in \Omega_a$, and square for the utility defined on fluent space Ω_F respectively.
- The uncertainty of state ω is represented by the belief b_A and b_B. The two probabilities b_A and b_B are illustrated by the blue and pink clouds in the perceived state space Ω.
- Belief-of-belief bob_{BinA} and bob_{AinB} are illustrated by larger clouds in new bob-space.

The arrows in Fig. 13.16 show the various dynamics and information flows, including 3 types:

- Observations I_A and I_B from the physical state to perceived state space Ω
- Actions or interventions that cause fluent changes in the physical state (not discussed here)
- Messages between the two agents to exchange information. Depending on the learning modes, these messages are for inference, learning, demonstration, confirmation *etc.*

For clarity, we omit arrows for other dynamics: for example, some of the message may be generated from a bob-space to probe what the other agent is thinking, like "I think your state estimate ω is ..." or "what do you know about the state ω?" Some arrows are second-order, for example, A learns the policy π_{BinA} from observing how B conducts a task, *i.e.*, learning-from-demonstration [222], or learning the utility μ_{BinA} by watching B's decision or choice [696]. In CL, the communication of A and B converges at three levels (see curved arrows in Fig. 13.16):

- When the inference process converges, they reach a common ground or situation ω_c^*.
- When the learning process converges, they reach a common model knowledge θ_c^*
- When their policy & utility converges, they reach a common social norm π_c^* and ethics μ_c^*.

Depending on the learning protocols and characteristics (*i.e.*, capacity of generating and interpreting messages) of the agents, the convergences may have different equilibria which decide the limits of learning. In CL, we assume the agents are cooperative and not deceptive, and their utility functions are aligned through learning.

CL is a unifying framework where all existing learning methods can be shown as special cases, *i.e.*, being part of the diagram in Fig. 13.16. Furthermore, CL will create more effective and advanced learning protocols. We elaborate their relations in the following.

- Shannon's communication: CL extends Shannon's communication setting by including the mental states, the bob-space, utility functions, and a common mind C_n which all evolve over time. This will allow more sophisticated messages, and enable agents to "read between lines."
- Valiant/Vapnik's theory [671] is a passive inductive statistical learning, supervised or unsupervised, from random sampled examples. This is shown by arrow 1 in Fig. 13.16. In contrast, in CL, messages are deliberated based on reflecting the mental states and utility functions.
- Active learning is arrow 2, where B can ask A for labeling certain examples selected by B. The example is selected to gain the most information in optimizing B's utility/loss function.
- Algorithmic teaching [697, 676] is a protocol complementary to active learning. Teacher A chooses best examples to teach a learner B for efficiency. A must consider what the B knows, and selects critical examples to B, *e.g.*, support vectors for classification.
- Learning-by-demonstration [698] is a typical learning protocol in robotics, and is an important component for commonsense acquisition. This learning method is shown by arrows 4 and 5 in Fig. 13.16, agent A teaches a task by a sequence of actions on objects and shows the outcomes The learner observes the actions directly, and learn the action policy from the learner.
- Causal learning is represented by arrows 1 and 4, where an agent applies actions to change the fluents of objects and scenes, and learns the causal effects of its action in terms of changed object fluents, including appearance changes (*e.g.*, painting a wall, mopping a floor), geometry changes (*e.g.*, blow a balloon) and topology changes (*e.g.*, cutting a fruit).

The CL can create new learning methods or protocols which are not well-known. For example:

- Perceptual causality learning. In contrast to causal learning [291] where the experiment / intervention requests A to perform action (arrow 4) and observe the effects of her own actions (arrow 1). We propose a new protocol named perceptual causality learning in [302]. Here B can learn causality by watching (arrows 1 and 5) of the actions of A (arrow 4), under the assumption that A is not performing magic (*i.e.*, not cheating) and B has the capability of inferring and mirroring the actions of A. This is called "perceived causality." We have shown in [302] that this is far more effective learning causality, and opens the door for learning causality from observations. This is a key aspect of human intelligence.
- Utility learning is shown by arrows 4 and 6. B infers the utility function of A by observing her decisions and choices in actions. Economics theory says that rational agents make decisions and take actions for utility maximization. By observing the actions taken by A, B can infer A utility, denoted by μ_{AinB} in CL. For example, we have demonstrated in [696] an example of

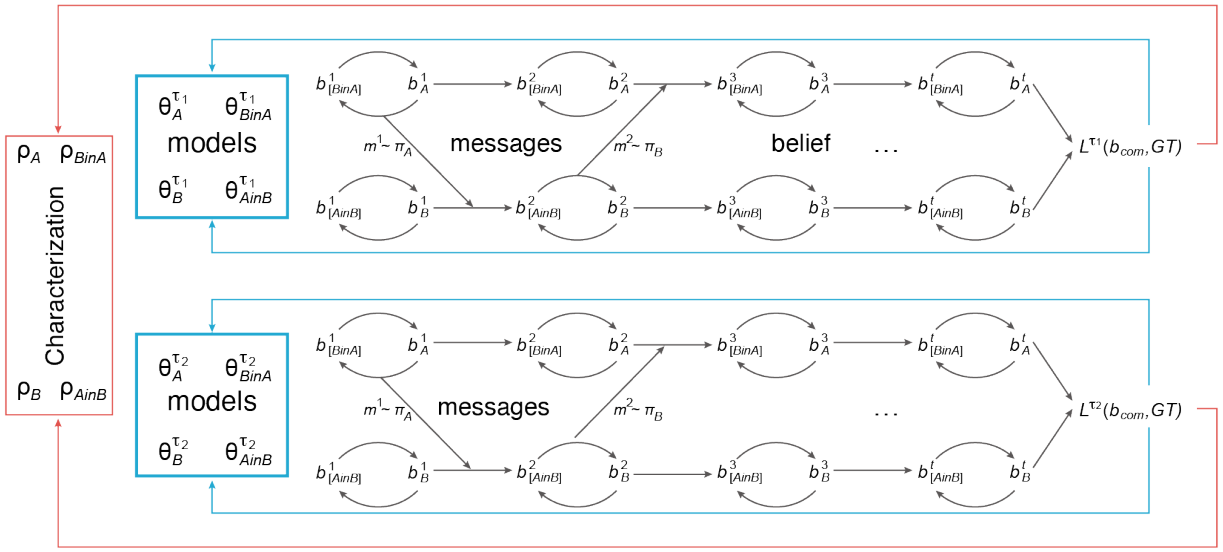


Figure 13.17: CL includes 3 nested loops: i) The reflection loop in black for deliberation of inferential messages to achieve common ground; ii) The learning loop in blue to achieve a common model; and iii) The characterization loop in red to achieve better characterization and capacity of the learners, such ToM, utility, hypothesis space *etc*.

commonsense acquisition, like folding T-shirt. By watching A folding T-shirts, B can not only learn the causality and policy π_A, but also learn an utility function μ_{AinB} for aesthetics – what states of the T-shirt have relatively higher value to A. B may choose to adopt a similar utility function $\mu_B \leftarrow \mu_{AinB}$. In CL, agents will update and align to common utility.

- Learning by analogy is a powerful learning mode used by humans [139], but missing in current popular machine learning methods. It requests shared knowledge (ω_C, Θ_C) between two agents, and the capabilities of abstraction and projections to transfer knowledge across domains using abstract graphical representation. Abstraction and projection are key intelligent capabilities in classic Raven's IQ tests, but are missing in current statistical learning. The shared mind C_n will facilitate learning-by-analogy. As C_n grows, the two agents will be more and more effective.

13.7 Halting Problem of Learning

Fig. 13.17 summarizes CL learning in three nested loops, and thus convergence occurs at three levels.

1. Reflection loop. The messages at this level communicate about a state, a cell, or an event (set) in state space Ω to achieve a common ground and common belief. Although we only discuss messages as projection on linear neurons and cells as partitions in Ω, this can be extended to nodes in a parse graph or logic predicates which correspond to compositions of the atomic cells or events.
2. Learning Loop. The messages at this level communicate about statistical summaries of data and information projections to achieve a common model $\theta \in \mathcal{H}$ in the hypothesis space. This includes updating models θ_{AinB} for $\hat{P}_n$, θ_{BinA} for $\hat{Q}_n$ in the *bob*-space for the nested minds.
3. Characterization loop. This loop will update the hyper-parameters $\rho_A, \rho_B, \rho_{AinB}, \rho_{BinA}$ that characterize the agents and their capacity of learning, including the hypothesis space and bob-space, *e.g.*, the number of neurons, protocols, and utility functions. The goal is to achieve

common characterization or mutual understanding of each other's characterization which decides the "IQ-of-learner and teacher" and efficiency of communication and learning.

The three CL loops terminate when certain halting conditions occur. By analogy to the halting problem of computing [699], we formulate a halting problem of learning, *i.e.*, whether and how a CL learning process terminates, and reach its limits of learning. The ideal halting occurs when the six minds converge to one, and validated by oracle (God's mind):

$$P_n = Q_n = \hat{P}_n = \hat{Q}_n = C_n = G. \tag{13.26}$$

When agents have diverse, and often conflicting, utility functions as stated in social choice theory, such convergence is not reachable. This project will focus on learning commonsense concepts in daily tasks for which the utility functions are not conflicting, and convergence is feasible. We will investigate some pre-mature halting conditions. For example, let $KL(P_n \| Q_n)$ denote the discrepancy between two agent's minds. When $KL(\hat{Q}_n \| \hat{P}_n) = 0$, learner Bob will think he knows what teacher Alice knows, and quits prematurely. Similarly if $KL(\hat{P}_n \| \hat{Q}_n) = 0$, Alice mistakenly thinks the Bob has already known what she knows, and stops teaching, and so on.

We assume CL agents are cooperative not deceptive, and they are also sincere: not sending fake messages, do not ignore the messages sent by others, and are willing to align their utility functions. We will study how the various CL protocols achieve game theoretical equilibrium [700, 701].

Chapter 14

Discussion: Path to General AI

Robots are mechanically capable of performing a wide range of complex activities; however, in practice, they do very little that is useful for humans. Today's robots fundamentally lack physical and social common sense; this limitation inhibits their capacity to aid in our daily lives. In this article, we have reviewed five concepts that are the crucial building blocks of common sense: functionality, physics, intent, causality, and utility (FPICU). We argued that these cognitive abilities have shown potential to be, in turn, the building blocks of cognitive AI, and should therefore be the foundation of future efforts in constructing this cognitive architecture. The positions taken in this article are not intended to serve as *the* solution for the future of cognitive AI. Rather, by identifying these crucial concepts, we want to call attention to pathways that have been less well explored in our rapidly developing AI community. There are indeed many other topics that we believe are also essential AI ingredients; for example:

- *A physically realistic VR/MR platform: from big data to big tasks.* Since FPICU is "dark"—meaning that it often does not appear in the form of pixels—it is difficult to evaluate FPICU in traditional terms. Here, we argue that the ultimate standard for validating the effectiveness of FPICU in AI is to examine whether an agent is capable of (i) accomplishing the very same task using different sets of objects with different instructions and/ or sequences of actions in different environments; and (ii) rapidly adapting such learned knowledge to entirely new tasks. By leveraging state-of-the-art game engines and physics-based simulations, we are beginning to explore this possibility on a large scale; see Section 14.1.
- *Social system: the emergence of language, communication, and morality.* While FPICU captures the core components of a single agent, modeling interaction among and within agents, either in collaborative or competitive situations [702], is still a challenging problem. In most cases, algorithms designed for a single agent would be difficult to generalize to a multiple-agent systems (MAS) setting [591, 703, 704]. We provide a brief review of three related topics in Section 14.2.
- *Measuring the limits of an intelligence system: IQ tests.* Studying FPICU opens a new direction of analogy and relational reasoning [705]. Apart from the four-term analogy (or proportional analogy), John C. Raven [706] proposed the raven's prograssive matrices test (RPM) in the image domain. The RAVEN dataset [707] was recently introduced in the computer vision community, and serves as a systematic benchmark for many visual reasoning models. Empirical studies show that abstract-level reasoning, combined with effective feature-extraction models, could notably improve the performance of reasoning, analogy, and generalization. However, the performance gap between human and computational models calls for future research in this field; see Section 14.3.

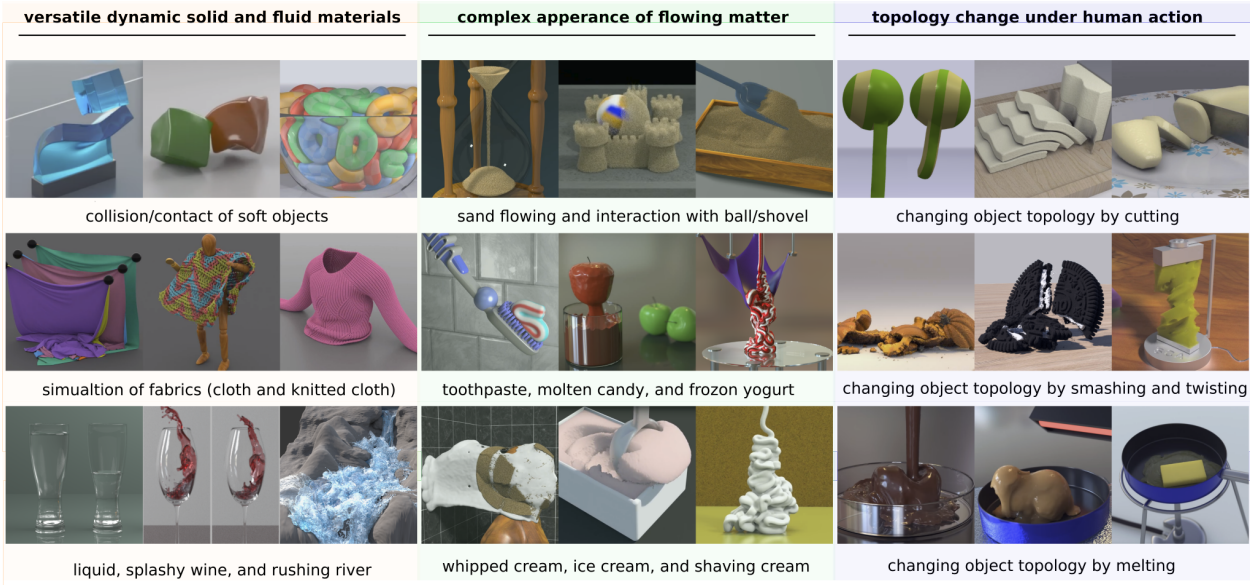


Figure 14.1: Diverse physical phenomena simulated using the material point method (MPM).

14.1 Physically-Realistic VR/MR Platform: From Big-Data to Big-Tasks

A hallmark of machine intelligence is the capability to rapidly adapt to new tasks and "achieve goals in a wide range of environments" [364]. To reach this goal, we have seen the increasing use of synthetic data and simulation platforms for indoor scenes in recent years by leveraging state-of-the-art game engines and free, publicly available 3D content [708, 709, 710, 711], including MINOR [712], HoME [713], Gibson [714], House3D [715], AI-THOR [716], VirtualHome [717], VRGym [718], and VRKitchen [719]. In addition, the AirSim [720] open-source simulator was developed for outdoor scenarios. Such synthetic data could be relatively easily scaled up compared with traditional data collection and labeling processes. With increasing realism and faster rendering speeds built on dedicated hardware, synthetic data from the virtual world is becoming increasingly similar to data collected from the physical world. In these realistic virtual environments, it is possible to evaluate any AI method or system from a much more holistic perspective. Using a holistic evaluation, whether a method or a system is intelligent or not is no longer measured by the successful performance of a single narrow task; rather, it is measured by the ability to perform well across various tasks: the perception of environments, planning of actions, predictions of other agents' behaviors, and ability to rapidly adapt learned knowledge to new environments for new tasks.

To build this kind of task-driven evaluation, physics-based simulations for multi-material, multi-physics phenomena (Fig. 14.1) will play a central role. We argue that cognitive AI needs to accelerate the pace of its adoption of more advanced simulation models from computer graphics, in order to benefit from the capability of highly predictive forward simulations, especially graphics processing unit (GPU) optimizations that allow real-time performance [721]. Here, we provide a brief review of the recent physics-based simulation methods, with a particular focus on the material point method (MPM).

The accuracy of physics-based reasoning greatly relies on the fidelity of a physics-based simulation. Similarly, the scope of supported virtual materials and their physical and interactive properties directly determine the complexity of the AI tasks involving them. Since the pioneering work of Terzopoulos*et al.* [722, 723] for solids and that of Foster and Metaxas [724] for fluids, many

mathematical and physical models in computer graphics have been developed and applied to the simulation of solids and fluids in a 3D virtual environment.

For decades, the computer graphics and computational physics community sought to increase the robustness, efficiency, stability, and accuracy of simulations for cloth, collisions, deformable, fire, fluids, fractures, hair, rigid bodies, rods, shells, and many other substances. Computer simulation-based engineering science plays an important role in solving many modern problems as an inexpensive, safe, and analyzable companion to physical experiments. The most challenging problems are those involving extreme deformation, topology change, and interactions among different materials and phases. Examples of these problems include hypervelocity impact, explosion, crack evolution, fluid-structure interactions, climate simulation, and ice-sheet movements. Despite the rapid development of computational solid and fluid mechanics, effectively and efficiently simulating these complex phenomena remains difficult. Based on how the continuous physical equations are discretized, the existing methods can be classified into the following categories:

1. Eulerian grid-based approaches, where the computational grid is fixed in space, and physical properties advect through the deformation flow. A typical example is the Eulerian simulation of free surface incompressible flow [725, 252]. Eulerian methods are more error-prone and require delicate treatment when dealing with deforming material interfaces and boundary conditions, since no explicit tracking of them is available.
2. Lagrangian mesh-based methods, represented by FEM [410, 726, 727], where the material is described with and embedded in a deforming mesh. Mass, momentum, and energy conservation can be solved with less effort. The main problem of acfem is mesh distortion and lack of contact during large deformations [458, 728] or topologically changing events [729].
3. Lagrangian mesh-free methods, such as smoothed particle hydrodynamics (SPH) [234] and the reproducing kernel particle method (RKPM) [730]. These methods allow arbitrary deformation but require expensive operations such as neighborhood searching [731]. Since the interpolation kernel is approximated with neighboring particles, these methods also tend to suffer from numerical instability issues.
4. Hybrid Lagrangian–Eulerian methods, such as the arbitrary Lagrangian–Eulerian (ALE) methods [732] and the MPM. These methods (particularly the MPM) combine the advantages of both Lagrangian methods and Eulerian grid methods by using a mixed representation.

In particular, as a generalization of the hybrid fluid implicit particle (FLIP) method [733, 236] from computational fluid dynamics to computational solid mechanics, the MPM has proven to be a promising discretization choice for simulating many solid and fluid materials since its introduction two decades ago [734, 235]. In the field of visual computing, existing work includes snow [735, 736], foam [737, 738, 739], sand [237, 740], rigid body [741], fracture [742, 743], cloth [744], hair [745], water [746], and solid-fluid mixtures [747, 748, 749]. In computational engineering science, this method has also become one of the most recent and advanced discretization choices for various applications. Due to its many advantages, it has been successfully applied to tackling extreme deformation events such as fracture evolution [750], material failure [751, 752], hyper-velocity impact [753, 754], explosion [755], fluid-structure interaction [756, 757], biomechanics [758], geomechanics [759], and many other examples that are considerably more difficult when addressed with traditional, non-hybrid approaches. In addition to experiencing a tremendously expanding scope of application, the MPM's discretization scheme has been extensively improved [760]. To alleviate numerical inaccuracy and stability issues associated with the original MPM formulation, researchers have proposed different variations of the MPM, including the generalized interpolation material point (GIMP) method [761, 762], the convected particle domain interpolation (CPDI) method [763], and the dual domain material point (DDMP) method [764].

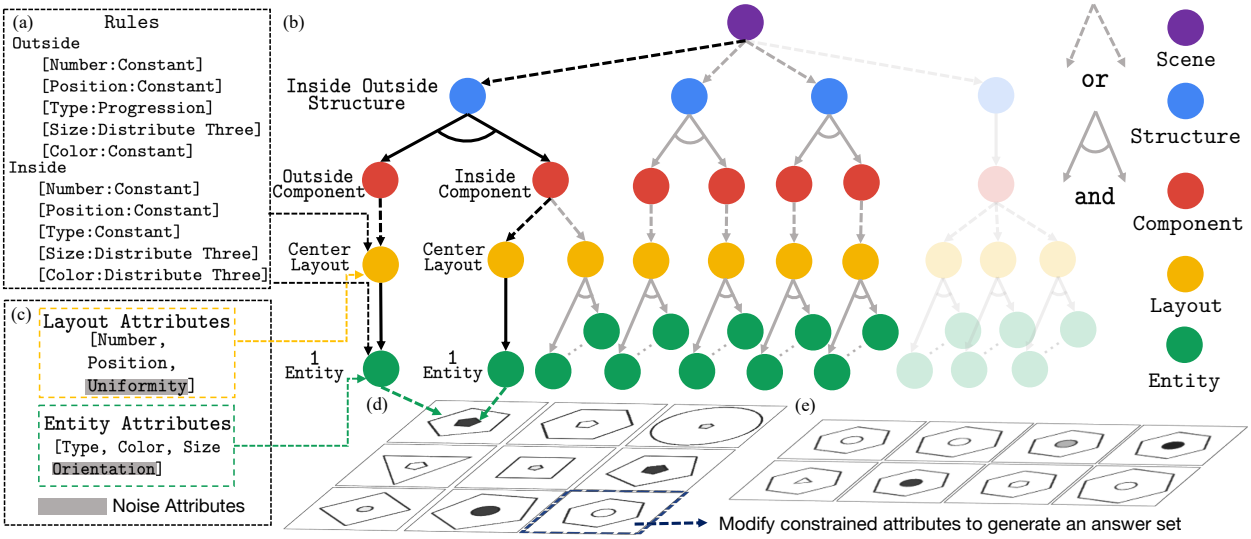


Figure 14.2: The RAVEN creation process proposed in Ref. [707]. A graphical illustration of (a) the grammar production rules used in (b) A-SIG. (c) Note that Layout and Entity have associated attributes. (d) A sample problem matrix and (e) a sample candidate set. Reproduced from Ref. [707] with permission of the authors, © 2019.

14.2 Social System: Emergence of Language, Communication, and Morality

Being able to communicate and collaborate with other agents is a crucial component of AI. In classic AI, a multi-agent communication strategy is modeled using a predefined rule-based system (*e.g.*, adaptive learning of communication strategies in MAS [702]). To scale up from rule-based systems, decentralized partially observable Markov decision processes were devised to model multi-agent interaction, with communication being considered as a special type of action [765, 766]. As with the success of RL in single-agent games [617], generalizing Q-learning [767, 704] and actor-critic [591, 592]-based methods from single-agent system to MAS have been a booming topic in recent years.

The emergence of language is also a fruitful topic in multi-agent decentralized collaborations. By modeling communication as a particular type of action, recent research [703, 768, 609] has shown that agents can learn how to communicate with continuous signals that are only decipherable within a group. The emergence of more realistic communication protocols using discrete messages has been explored in various types of communication games [769, 688, 770, 686], in which agents need to process visual signals and attach discrete tokens to attributes or semantics of images in order to form effective protocols. By letting groups of agents play communication games spontaneously, several linguistic phenomena in emergent communication and language have been studied [771, 772, 773].

Morality is an abstract and complex concept composed of common principles such as fairness, obligation, and permissibility. It is deeply rooted in the tradeoffs people make every day when these moral principles come into conflict with one another [774, 775]. Moral judgment is extremely complicated due to the variability in standards among different individuals, social groups, cultures, and even forms of violation of ethical rules. For example, two distinct societies could hold opposite views on preferential treatment of kin: one might view it as corrupt, the other as a moral obligation [776]. Indeed, the same principle might be viewed differently in two social groups with distinct cultures [777]. Even within the same social group, different individuals might have different

standards on the same moral principle or event that triggers moral judgment [778, 779, 780]. Many works have proposed theoretical accounts for categorizing the different measures of welfare used in moral calculus, including "base goods" and "primary goods" [781, 782], "moral foundations" [783], and the feasibility of value judgment from an infant's point of view [784]. Despite its complexity and diversity, devising a computational account of morality and moral judgment is an essential step on the path toward building humanlike machines. One recent approach to moral learning combines utility calculus and Bayesian inference to distinguish and evaluate different principles [776, 785, 786].

14.3 Measuring the Limits of Intelligence System: IQ tests

In the literature, we call two cases analogous if they share a common *relationship*. Such a relationship does not need to be among entities or ideas that use the same label across disciplines, such as computer vision and AI; rather, "analogous" emphasizes commonality on a more abstract level. For example, according to Ref. [787], the earliest major scientific discovery made through analogy can be dated back to imperial Rome, when investigators analogized waves in water and sound. They posited that sound waves and water waves share similar behavioral properties; for example, their intensities both diminish as they propagate across space. To make a successful analogy, the key is to understand *causes and their effects* [788].

The history of analogy can be categorized into three streams of research; see Ref. [705] for a capsule history and review of the literature. One stream is the psychometric tradition of four-term or "proportional" analogies, the earliest discussions of which can be traced back to Aristotle [789]. An example in AI is the *word2vec* model [790, 791], which is capable of making a four-term word analogy; for example, [king:queen::man:woman]. In the image domain, a similar test was invented by John C. Raven [706]—the raven's prograssive matrices test (RPM).

RPM has been widely accepted and is believed to be highly correlated with real intelligence [792]. Unlike visual question answering (VQA) [793], which lies at the periphery of the cognitive ability test circle [792], RPM lies directly at the center: it is diagnostic of abstract and structural reasoning ability [794], and captures the defining feature of high-level cognition—that is, *fluid intelligence* [795]. It has been shown that RPM is more difficult than existing visual reasoning tests in the following ways [707]:

- Unlike VQA, where natural language questions usually imply what the agent should pay attention to in an image, RPM relies merely on visual clues provided in the matrix. The *correspondence problem* itself, that is, the ability to find corresponding objects across frames to determine their relationship, is already a major factor distinguishing populations of different intelligence [792].
- While current visual reasoning tests only require spatial and semantic understanding, RPM needs joint spatial-temporal reasoning in the problem matrix and the answer set. The limit of *short-term memory*, the ability to understand *analogy*, and the grasp of *structure* must be taken into consideration in order to solve an RPM problem.
- Structures in RPM make the compositions of rules much more complicated. Problems in RPM usually include more sophisticated logic with recursions. Combinatorial rules composed at various levels also make the reasoning process extremely difficult.

The RAVEN dataset [707] was created to push the limit of current vision systems' reasoning and analogy-making ability, and to promote further research in this area. The dataset is designed to focus on reasoning and analogizing instead of only visual recognition. It is unique in the sense that it builds a semantic link between the visual reasoning and structural reasoning in RPM by grounding each problem into a sentence derived from an attributed stochastic image grammar attributed stochastic image grammar (A-SIG): each instance is a sentence sampled from a predefined A-SIG,

and a rendering engine transforms the sentence into its corresponding image. (See Fig. 14.2 [707] for a graphical illustration of the generation process.) This semantic link between vision and structure representation opens new possibilities by breaking down the problem into image understanding and abstract-level structure reasoning. Zhang *et al*. [707] empirically demonstrated that models using a simple structural reasoning module to incorporate both vision-level understanding and abstract-level reasoning and analogizing notably improved their performance in RPM, whereas a variety of prior approaches to relational learning performed only slightly better than a random guess.

Analogy consists of more than mere spatiotemporal parsing and structural reasoning. For example, the *contrast effect* [796] has been proven to be one of the key ingredients in relational and analogical reasoning for both human and machine learning [797, 798, 799, 800, 801]. Originating from perceptual learning [802, 803], it is well established in the field of psychology and education [804, 805, 806, 807, 808] that teaching new concepts by comparing noisy examples is quite effective. Smith and Gentner [809] summarized that comparing cases facilitates transfer learning and problem-solving, as well as the ability to learn relational categories. In his structure-mapping theory, Gentner [810] postulated that learners generate a structural alignment between two representations when they compare two cases. A later article [811] firmly supported this idea and showed that finding the individual difference is easier for humans when similar items are compared. A more recent study from Schwartz *et al*. [812] also showed that contrasting cases helps to foster an appreciation of deep understanding. To retrieve this missing treatment of contrast in machine learning, computer vision and, more broadly, in AI, Zhang *et al*. [813] proposed methods of learning perceptual inference that explicitly introduce the notion of contrast in model training. Specifically, a contrast module and a contrast loss are incorporated into the algorithm at the model level and at the objective level, respectively. The permutation-invariant contrast module summarizes the common features from different objects and distinguishes each candidate by projecting it onto its residual on the common feature space. The final model, which comprises ideas from contrast effects and perceptual inference, achieved state-of-the-art performance on major RPM datasets.

Parallel to work on RPM, work on *number sense* [814] bridges the induction of symbolic concepts and the competence of problem-solving; in fact, number sense could be regarded as a mathematical counterpart to the visual reasoning task of RPM. A recent work approaches the analogy problem from this perspective of strong mathematical reasoning [815]. Zhang *et al*. [815] studied the machine number-sense problem and proposed a dataset of visual arithmetic problems for abstract and relational reasoning, where the machine is given two figures of numbers following hidden arithmetic computations and is tasked to work out a missing entry in the final answer. Solving machine number-sense problems is non-trivial: the system must both recognize a number and interpret the number with its contexts, shapes, and relationships (*e.g*., symmetry), together with its proper operations. Experiments show that the current neural-network-based models do not acquire mathematical reasoning abilities after learning, whereas classic search-based algorithms equipped with an additional perception module achieve a sharp performance gain with fewer search steps. This work also sheds some light on how machine reasoning could be improved: the fusing of classic search-based algorithms with modern neural networks in order to discover essential number concepts in future research would be an encouraging development.

Bibliography

[1] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information.* Cambridge, Massachusetts, 1982.

[2] M. Mishkin, L. G. Ungerleider, and K. A. Macko, "Object vision and spatial vision: two cortical pathways," *Trends in Neurosciences*, vol. 6, pp. 414–417, 1983.

[3] B. Julesz, "Visual pattern discrimination," *IRE transactions on Information Theory*, vol. 8, no. 2, pp. 84–92, 1962.

[4] S.-C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling," *International Journal of Computer Vision (IJCV)*, vol. 27, no. 2, pp. 107–126, 1998.

[5] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, no. 5802, p. 91, 1981.

[6] S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu, "What are textons?," *International Journal of Computer Vision (IJCV)*, vol. 62, no. 1-2, pp. 121–143, 2005.

[7] C.-e. Guo, S.-C. Zhu, and Y. N. Wu, "Towards a mathematical theory of primal sketch and sketchability," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2003.

[8] C.-e. Guo, S.-C. Zhu, and Y. N. Wu, "Primal sketch: Integrating structure and texture," *Computer Vision and Image Understanding (CVIU)*, vol. 106, no. 1, pp. 5–19, 2007.

[9] M. Nitzberg and D. Mumford, "The 2.1-d sketch," in *ICCV*, 1990.

[10] J. Y. Wang and E. H. Adelson, "Layered representation for motion analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1993.

[11] J. Y. Wang and E. H. Adelson, "Representing moving images with layers," *Proceedings of Transactions on Image Processing (TIP)*, vol. 3, no. 5, pp. 625–638, 1994.

[12] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 200, no. 1140, pp. 269–294, 1978.

[13] I. Binford, "Visual perception by computer," in *IEEE Conference of Systems and Control*, 1971.

[14] R. A. Brooks, "Symbolic reasoning among 3-d models and 2-d images," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 285–348, 1981.

[15] T. Kanade, "Recovery of the three-dimensional shape of an object from a single view," *Artificial intelligence*, vol. 17, no. 1-3, pp. 409–460, 1981.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] J. M. Coughlan and A. L. Yuille, "Manhattan world: Orientation and outlier detection by bayesian inference," *Neural Computation*, 2003.

[19] A. Yuille and D. Kersten, "Vision as bayesian inference: Analysis by synthesis?," *Trends in cognitive sciences*, vol. 10, no. 7, pp. 301–308, 2006.

[20] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt, "A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization," *Psychological bulletin*, vol. 138, no. 6, p. 1172, 2012.

[21] J. Wagemans, J. Feldman, S. Gepshtein, R. Kimchi, J. R. Pomerantz, P. A. Van der Helm, and C. Van Leeuwen, "A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations," *Psychological bulletin*, vol. 138, no. 6, p. 1218, 2012.

[22] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, J. B. Tenenbaum, and S.-C. Zhu, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[23] K. Ikeuchi and M. Hebert, "Task-oriented vision," in *Exploratory vision*, pp. 257–277, Springer, 1996.

[24] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.

[25] F. Fang and S. He, "Cortical responses to invisible objects in the human dorsal and ventral pathways," *Nature Neuroscience*, vol. 8, no. 10, p. 1380, 2005.

[26] S. H. Creem-Regehr and J. N. Lee, "Neural representations of graspable objects: are tools special?," *Cognitive Brain Research*, vol. 22, no. 3, pp. 457–469, 2005.

[27] M. C. Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, 1975.

[28] M. C. Potter, "Short-term conceptual memory for pictures," *Journal of experimental psychology: human learning and memory*, vol. 2, no. 5, p. 509, 1976.

[29] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition," *Psychological science*, vol. 5, no. 4, pp. 195–200, 1994.

[30] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, p. 520, 1996.

[31] M. R. Greene and A. Oliva, "The briefest of glances: The time course of natural scene understanding," *Psychological Science*, vol. 20, no. 4, pp. 464–472, 2009.

[32] M. R. Greene and A. Oliva, "Recognition of natural scenes from global properties: Seeing the forest without representing the trees," *Cognitive Psychology*, vol. 58, no. 2, pp. 137–176, 2009.

[33] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *Journal of Vision*, vol. 7, no. 1, pp. 10–10, 2007.

[34] G. Rousselet, O. Joubert, and M. Fabre-Thorpe, "How long to get to the "gist" of real-world natural scenes?," *Visual Cognition*, vol. 12, no. 6, pp. 852–877, 2005.

[35] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision (IJCV)*, vol. 42, no. 3, pp. 145–175, 2001.

[36] A. Delorme, G. Richard, and M. Fabre-Thorpe, "Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans," *Vision Research*, vol. 40, no. 16, pp. 2187–2200, 2000.

[37] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 104, no. 15, pp. 6424–6429, 2007.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012.

[39] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2010.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[41] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," *Journal of Neuroscience*, vol. 38, no. 33, pp. 7255–7269, 2018.

[42] A. Oliva and P. G. Schyns, "Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli," *Cognitive Psychology*, vol. 34, no. 1, pp. 72–107, 1997.

[43] P. G. Schyns, "Diagnostic recognition: task constraints, object information, and their interactions," *Cognition*, vol. 67, no. 1-2, pp. 147–179, 1998.

[44] G. L. Malcolm, A. Nuthmann, and P. G. Schyns, "Beyond gist: Strategic and incremental information accumulation for scene categorization," *Psychological science*, vol. 25, no. 5, pp. 1087–1097, 2014.

[45] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[46] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2011.

[47] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision research*, vol. 41, no. 17, pp. 2261–2271, 2001.

[48] K. Ikeuchi and M. Hebert, "Task-oriented vision," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 1992.

[49] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[50] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An invitation to 3-d vision: from images to geometric models*. Springer Science & Business Media, 2012.

[51] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei, "Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2010.

[52] A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun, "Box in the box: Joint 3d layout and object reasoning from single images," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.

[53] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Understanding indoor scenes using 3d geometric phrases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[54] Y. Zhao and S.-C. Zhu, "Scene parsing by integrating function, geometry and appearance models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[55] X. Liu, Y. Zhao, and S.-C. Zhu, "Single-view 3d scene reconstruction and parsing by attribute grammar," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 3, pp. 710–725, 2018.

[56] S. Huang, S. Qi, Y. Zhu, Y. Xiao, Y. Xu, and S.-C. Zhu, "Holistic 3d scene parsing and reconstruction from a single rgb image," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[57] Y. Chen, S. Huang, T. Yuan, Y. Zhu, S. Qi, and S.-C. Zhu, "Holistic++ scene understanding with human-object interaction and physical commonsense," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[58] S. Huang, Y. Chen, T. Yuan, S. Qi, Y. Zhu, and S.-C. Zhu, "Perspectivenet: 3d object detection from a single rgb image via perspective points," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2019.

[59] E. C. Tolman, "Cognitive maps in rats and men," *Psychological review*, vol. 55, no. 4, p. 189, 1948.

[60] R. F. Wang and E. S. Spelke, "Comparative approaches to human navigation," *The Neurobiology of Spatial Behaviour*, pp. 119–143, 2003.

[61] J. J. Koenderink, A. J. van Doorn, A. M. Kappers, and J. S. Lappin, "Large-scale visual frontoparallels under full-cue conditions," *Perception*, vol. 31, no. 12, pp. 1467–1475, 2002.

[62] W. H. Warren, D. B. Rothman, B. H. Schnapp, and J. D. Ericson, "Wormholes in virtual space: From cognitive maps to cognitive graphs," *Cognition*, vol. 166, pp. 152–163, 2017.

[63] S. Gillner and H. A. Mallot, "Navigation and acquisition of spatial knowledge in a virtual maze," *Journal of Cognitive Neuroscience*, vol. 10, no. 4, pp. 445–463, 1998.

[64] P. Foo, W. H. Warren, A. Duchon, and M. J. Tarr, "Do humans integrate routes into a cognitive map? map-versus landmark-based navigation of novel shortcuts," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 31, no. 2, p. 195, 2005.

[65] E. R. Chrastil and W. H. Warren, "From cognitive maps to cognitive graphs," *PloS one*, vol. 9, no. 11, p. e112544, 2014.

[66] R. W. Byrne, "Memory for urban geography," *The Quarterly Journal of Experimental Psychology*, vol. 31, no. 1, pp. 147–154, 1979.

[67] B. Tversky, "Distortions in cognitive maps," *Geoforum*, vol. 23, no. 2, pp. 131–138, 1992.

[68] K. N. Ogle, *Researches in binocular vision*. WB Saunders, 1950.

[69] J. M. Foley, "Binocular distance perception," *Psychological review*, vol. 87, no. 5, p. 411, 1980.

[70] R. K. Luneburg, *Mathematical analysis of binocular vision*. Princeton University Press, 1947.

[71] T. Indow, "A critical review of luneburg's model with regard to global structure of visual space," *Psychological review*, vol. 98, no. 3, p. 430, 1991.

[72] W. C. Gogel, "A theory of phenomenal geometry and its applications," *Perception & Psychophysics*, vol. 48, no. 2, pp. 105–123, 1990.

[73] A. Glennerster, L. Tcheang, S. J. Gilson, A. W. Fitzgibbon, and A. J. Parker, "Humans ignore motion and stereo cues in favor of a fictional stable world," *Current Biology*, vol. 16, no. 4, pp. 428–432, 2006.

[74] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, p. 801, 2005.

[75] N. J. Killian, M. J. Jutras, and E. A. Buffalo, "A map of visual space in the primate entorhinal cortex," *Nature*, vol. 491, no. 7426, p. 761, 2012.

[76] J. O'keefe and L. Nadel, *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.

[77] J. Jacobs, C. T. Weidemann, J. F. Miller, A. Solway, J. F. Burke, X.-X. Wei, N. Suthana, M. R. Sperling, A. D. Sharan, I. Fried, *et al.*, "Direct recordings of grid-like neuronal activity in human spatial navigation," *Nature neuroscience*, vol. 16, no. 9, p. 1188, 2013.

[78] M. Fyhn, T. Hafting, M. P. Witter, E. I. Moser, and M.-B. Moser, "Grid cells in mice," *Hippocampus*, vol. 18, no. 12, pp. 1230–1238, 2008.

[79] C. F. Doeller, C. Barry, and N. Burgess, "Evidence for grid cells in a human memory network," *Nature*, vol. 463, no. 7281, p. 657, 2010.

[80] M. M. Yartsev, M. P. Witter, and N. Ulanovsky, "Grid cells without theta oscillations in the entorhinal cortex of bats," *Nature*, vol. 479, no. 7371, p. 103, 2011.

[81] R. Gao, J. Xie, S.-C. Zhu, and Y. N. Wu, "Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion," in *International Conference on Learning Representations (ICLR)*, 2019.

[82] J. Xie, R. Gao, E. Nijkamp, S.-C. Zhu, and Y. N. Wu, "Representation learning: A statistical perspective," *Annual Review of Statistics and Its Application*, vol. 7, 2019.

[83] L. Gootjes-Dreesbach, L. C. Pickup, A. W. Fitzgibbon, and A. Glennerster, "Comparison of view-based and reconstruction-based models of human navigational strategy," *Journal of vision*, vol. 17, no. 9, pp. 11–11, 2017.

[84] J. Vuong, A. Fitzgibbon, and A. Glennerster, "Human pointing errors suggest a flattened, task-dependent representation of space," *bioRxiv*, p. 390088, 2018.

[85] H. Choi and B. J. Scholl, "Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception," *Perception*, vol. 35, no. 3, pp. 385–399, 2006.

[86] B. J. Scholl and K. Nakayama, "Illusory causal crescents: Misperceived spatial relations due to perceived causality," *Perception*, vol. 33, no. 4, pp. 455–469, 2004.

[87] B. J. Scholl and T. Gao, "Perceiving animacy and intentionality: Visual processing or higher-level judgment," *Social perception: Detection and interpretation of animacy, agency, and intention*, vol. 4629, 2013.

[88] B. J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, 2001.

[89] E. Vul, G. Alvarez, J. B. Tenenbaum, and M. J. Black, "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009.

[90] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 110, no. 45, pp. 18327–18332, 2013.

[91] J. Hamrick, P. Battaglia, and J. B. Tenenbaum, "Internal physics models guide probabilistic judgments about object dynamics," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2011.

[92] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu, "Learning and inferring "dark matter" and predicting human intents and trajectories in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 7, pp. 1639–1652, 2018.

[93] T. Ullman, A. Stuhlmüller, N. Goodman, and J. B. Tenenbaum, "Learning physics from dynamical scenes," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2014.

[94] T. Gerstenberg and J. B. Tenenbaum, "Intuitive theories," in *Oxford handbook of causal reasoning*, pp. 515–548, Oxford University Press New York, NY, 2017.

[95] I. Newton and J. Colson, *The method of fluxions and infinite series; with its application to the geometry of curve-lines.* Henry Woodfall; and sold by John Nourse, 1736.

[96] C. Maclaurin, *A treatise of fluxions: in two books. 1.* Ruddimans, 1742.

[97] E. T. Mueller, *Commonsense reasoning: an event calculus based approach.* Morgan Kaufmann, 2014.

[98] E. T. Mueller, *Daydreaming in humans and machines: a computer model of the stream of thought.* Intellect Books, 1990.

[99] A. Michotte, *The perception of causality.* London, England: Methuen & Co, 1963.

[100] S. Carey, *The origin of concepts.* Oxford University Press, 2009.

[101] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[102] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2011.

[103] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[104] B. Yao and S.-C. Zhu, "Learning deformable action templates from cluttered videos," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.

[105] B. Z. Yao, B. X. Nie, Z. Liu, and S.-C. Zhu, "Animated pose templates for modeling and detecting human actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 3, pp. 436–452, 2013.

[106] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[107] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[108] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[109] R. Fleming, M. Barnett-Cowan, and H. Bülthoff, "Perceived object stability is affected by the internal representation of gravity," *PLoS One*, vol. 6, no. 4, 2010.

[110] M. Zago and F. Lacquaniti, "Visual perception and interception of falling objects: A review of evidence for an internal model of gravity," *Journal of Neural Engineering*, vol. 2, no. 3, p. S198, 2005.

[111] P. J. Kellman and E. S. Spelke, "Perception of partly occluded objects in infancy," *Cognitive psychology*, vol. 15, no. 4, pp. 483–524, 1983.

[112] R. Baillargeon, E. S. Spelke, and S. Wasserman, "Object permanence in five-month-old infants," *Cognition*, vol. 20, no. 3, pp. 191–208, 1985.

[113] S. P. Johnson and R. N. Aslin, "Perception of object unity in 2-month-old infants," *Developmental Psychology*, vol. 31, no. 5, p. 739, 1995.

[114] A. Needham, "Factors affecting infants' use of featural information in object segregation," *Current Directions in Psychological Science*, vol. 6, no. 2, pp. 26–33, 1997.

[115] R. Baillargeon, "Infants' physical world," *Current directions in psychological science*, vol. 13, no. 3, pp. 89–94, 2004.

[116] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu, "Beyond point clouds: Scene understanding by reasoning geometry and physics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[117] B. Zheng, Y. Zhao, C. Y. Joey, K. Ikeuchi, and S.-C. Zhu, "Detecting potential falling objects by inferring human action and natural disturbance," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2014.

[118] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, and S.-C. Zhu, "Scene understanding by reasoning stability and safety," *International Journal of Computer Vision (IJCV)*, pp. 221–238, 2015.

[119] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S.-C. Zhu, "Human-centric indoor scene synthesis using stochastic grammar," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[120] S. Huang, S. Qi, Y. Xiao, Y. Zhu, Y. N. Wu, and S.-C. Zhu, "Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2018.

[121] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert, "From 3d scene geometry to human workspace," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[122] M. Iacoboni, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, and G. Rizzolatti, "Grasping the intentions of others with one's own mirror neuron system," *PLoS biology*, vol. 3, no. 3, p. e79, 2005.

[123] G. Csibra and G. Gergely, "'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans," *Acta psychologica*, vol. 124, no. 1, pp. 60–78, 2007.

[124] C. L. Baker, J. B. Tenenbaum, and R. R. Saxe, "Goal inference as inverse planning," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2007.

[125] C. L. Baker, N. D. Goodman, and J. B. Tenenbaum, "Theory-based social goal inference," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2008.

[126] M. Hoai and F. De la Torre, "Max-margin early event detectors," *International Journal of Computer Vision (IJCV)*, vol. 107, no. 2, pp. 191–202, 2014.

[127] M. W. Turek, A. Hoogs, and R. Collins, "Unsupervised learning of functional categories in video scenes," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[128] H. Grabner, J. Gall, and L. Van Gool, "What makes a chair a chair?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[129] Z. Jia, A. Gallagher, A. Saxena, and T. Chen, "3d-based reasoning with blocks, support, and stability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[130] Y. Jiang, H. Koppula, and A. Saxena, "Hallucinated humans as the hidden context for labeling 3d scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[131] T. Shu, S. M. Thurman, D. Chen, S.-C. Zhu, and H. Lu, "Critical features of joint actions that signal human interaction," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2016.

[132] T. Shu, Y. Peng, L. Fan, H. Lu, and S.-C. Zhu, "Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations," *Topics in cognitive science*, vol. 10, no. 1, pp. 225–241, 2018.

[133] T. Shu, Y. Peng, H. Lu, and S.-C. Zhu, "Partitioning the perception of physical and social events within a unified psychological space," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2019.

[134] C. Baker, R. Saxe, and J. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2011.

[135] Y. Zhao, S. Holtzen, T. Gao, and S.-C. Zhu, "Represent and infer human theory of mind for human-robot interaction," in *AAAI fall symposium series*, 2015.

[136] N. Nisan and A. Ronen, "Algorithmic mechanism design," *Games and Economic behavior*, vol. 35, no. 1-2, pp. 166–196, 2001.

[137] J. Bentham, "An introduction to the principles of morals," *London: Athlone*, 1789.

[138] N. Shukla, *Utility learning, non-Markovian planning, and task-oriented programming language*. PhD thesis, UCLA, 2019.

[139] K. J. Holyoak, K. J. Holyoak, and P. Thagard, *Mental leaps: Analogy in creative thought*. MIT press, 1995.

[140] M. Tomasello, *Origins of human communication*. MIT press, 2010.

[141] J. Gibson, "The ecological approach to visual perception," *Houghton Mifflin Comp*, 1979.

[142] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *International Journal of Robotics Research (IJRR)*, vol. 34, no. 4-5, pp. 705–724, 2015.

[143] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[144] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[145] S. P. Boyd and B. Wegbreit, "Fast computation of optimal contact forces," *IEEE Transactions on Robotics*, vol. 23, no. 6, pp. 1117–1132, 2007.

[146] L. Han, J. C. Trinkle, and Z. X. Li, "Grasp analysis as linear matrix inequality problems," *IEEE Transactions on Robotics and Automation*, vol. 16, no. 6, pp. 663–674, 2000.

[147] Y. Zheng and C.-M. Chew, "Distance between a point and a convex cone in n-dimensional space: Computation and applications," *Transactions on Robotics (T-RO)*, vol. 25, no. 6, pp. 1397–1412, 2009.

[148] H. Dai, A. Majumdar, and R. Tedrake, "Synthesis and optimization of force closure grasps via sequential semi-definite programming," in *Robotics Research*, pp. 285–305, Springer, 2018.

[149] Y.-H. Liu, "Qualitative test and force optimization of 3d frictional form-closure grasps using linear programming," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 1, pp. 163–173, 1999.

[150] M. Hill, E. Nijkamp, and S.-C. Zhu, "Building a telescope to look into high-dimensional image spaces," *Quarterly of Applied Mathematics*, vol. 77, no. 2, pp. 269–321, 2019.

[151] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[152] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[153] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.

[154] S.-C. Zhu, D. Mumford, *et al.*, "A stochastic grammar of images," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2007.

[155] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[156] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[157] D. Xie, S. Todorovic, and S.-C. Zhu, "Inferring "dark matter" and "dark energy" from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[158] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.

[159] A. Fridman, "Mixed markov models," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 100, no. 14, 2003.

[160] S. M. Lavalle, "Rapidly-exploring random trees: A new tool for path planning," tech. rep., Computer Science Department, Iowa State University, 1998.

[161] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[162] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[163] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[164] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," in *AI Stats*, 2005.

[165] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[166] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[167] M. McCloskey, "Intuitive physics," *Scientific American*, vol. 248, no. 4, pp. 122–131, 1983.

[168] M. McCloskey, A. Caramazza, and B. Green, "Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects," *Science*, vol. 210, no. 4474, pp. 1139–1141, 1980.

[169] A. A. DiSessa, "Unlearning aristotelian physics: A study of knowledge-based learning," *Cognitive science*, vol. 6, no. 1, pp. 37–75, 1982.

[170] M. K. Kaiser, J. Jonides, and J. Alexander, "Intuitive reasoning about abstract and familiar physics problems," *Memory & Cognition*, vol. 14, no. 4, pp. 308–312, 1986.

[171] K. A. Smith, P. Battaglia, and E. Vul, "Consistent physics underlying ballistic motion prediction," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2013.

[172] M. K. Kaiser, D. R. Proffitt, S. M. Whelan, and H. Hecht, "Influence of animation on dynamical judgments," *Journal of experimental Psychology: Human Perception and performance*, vol. 18, no. 3, p. 669, 1992.

[173] M. K. Kaiser, D. R. Proffitt, and K. Anderson, "Judgments of natural and anomalous trajectories in the presence and absence of motion," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 11, no. 4, p. 795, 1985.

[174] I.-K. Kim and E. S. Spelke, "Perception and understanding of effects of gravity and inertia on object motion," *Developmental Science*, vol. 2, no. 3, pp. 339–362, 1999.

[175] J. Piaget and M. Cook, *The origins of intelligence in children*. International Universities Press New York, 1952.

[176] J. Piaget and M. T. Cook, *The construction of reality in the child*. Basic Books, 1954.

[177] S. J. Hespos and R. Baillargeon, "Décalage in infants' knowledge about occlusion and containment events: Converging evidence from action tasks," *Cognition*, vol. 99, no. 2, pp. B31–B41, 2006.

[178] S. J. Hespos and R. Baillargeon, "Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings," *Cognition*, vol. 107, no. 1, pp. 304–316, 2008.

[179] T. G. Bower, *Development in infancy*. WH Freeman, 1974.

[180] A. M. Leslie and S. Keeble, "Do six-month-old infants perceive causality?," *Cognition*, vol. 25, no. 3, pp. 265–288, 1987.

[181] Y. Luo, R. Baillargeon, L. Brueckner, and Y. Munakata, "Reasoning about a hidden object after a delay: Evidence for robust representations in 5-month-old infants," *Cognition*, vol. 88, no. 3, pp. B23–B32, 2003.

[182] R. Baillargeon, J. Li, W. Ng, and S. Yuan, "An account of infants' physical reasoning," in *Learning and the Infant Mind*, pp. 66–116, Oxford University Press, 2008.

[183] R. Baillargeon, "The acquisition of physical knowledge in infancy: A summary in eight lessons," *Blackwell handbook of childhood cognitive development*, vol. 1, no. 46-83, p. 1, 2002.

[184] P. Achinstein, *The nature of explanation*. Oxford University Press on Demand, 1983.

[185] J. Fischer, J. G. Mikhael, J. B. Tenenbaum, and N. Kanwisher, "Functional neuroanatomy of intuitive physical inference," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 113, no. 34, pp. E5072–E5081, 2016.

[186] T. D. Ullman, E. Spelke, P. Battaglia, and J. B. Tenenbaum, "Mind games: Game engines as an architecture for intuitive physics," *Trends in Cognitive Sciences*, vol. 21, no. 9, pp. 649–665, 2017.

[187] C. Bates, P. Battaglia, I. Yildirim, and J. B. Tenenbaum, "Humans predict liquid dynamics using probabilistic simulation," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2015.

[188] J. Kubricht, C. Jiang, Y. Zhu, S.-C. Zhu, D. Terzopoulos, and H. Lu, "Probabilistic simulation predicts human performance on viscous fluid-pouring problem," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2016.

[189] J. Kubricht, Y. Zhu, C. Jiang, D. Terzopoulos, S.-C. Zhu, and H. Lu, "Consistent probabilistic simulation underlying human judgment in substance dynamics," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2017.

[190] J. R. Kubricht, K. J. Holyoak, and H. Lu, "Intuitive physics: Current research and controversies," *Trends in Cognitive Sciences*, vol. 21, no. 10, pp. 749–759, 2017.

[191] D. Mumford and A. Desolneux, *Pattern theory: the stochastic analysis of real-world signals*. AK Peters/CRC Press, 2010.

[192] D. Mumford, "Pattern theory: a unifying perspective," in *First European congress of mathematics*, pp. 187–224, 1994.

[193] D. Broadbent, *A question of levels: Comment on McClelland and Rumelhart*. American Psychological Association, 1985.

[194] D. Lowe, *Perceptual organization and visual recognition*. Springer Science & Business Media, 2012.

[195] A. P. Pentland, "Perceptual organization and the representation of natural form," in *Readings in Computer Vision*, pp. 680–699, Elsevier, 1987.

[196] M. Wertheimer, "Experimentelle studien uber das sehen von bewegung [experimental studies on the seeing of motion]," *Zeitschrift fur Psychologie*, vol. 61, pp. 161–265, 1912.

[197] W. Köhler, *Die physischen Gestalten in Ruhe und im stationärenZustand. Eine naturphilosophische Untersuchung [The physical Gestalten at rest and in steady state]*. Braunschweig, Germany: Vieweg und Sohn., 1920.

[198] W. Köhler, "Physical gestalten," in *A source book of Gestalt psychology*, pp. 17–54, London, England: Routledge & Kegan Paul, 1938.

[199] M. Wertheimer, "Untersuchungen zur lehre von der gestalt, ii. [investigations in gestalt theory: Ii. laws of organization in perceptual forms]," *Psychologische Forschung*, vol. 4, pp. 301–350, 1923.

[200] M. Wertheimer, "Laws of organization in perceptual forms," in *A source book of Gestalt psychology*, pp. 71–94, London, England: Routledge & Kegan Paul, 1938.

[201] K. Koffka, *Principles of Gestalt psychology*. Routledge, 2013.

[202] D. Waltz, "Understanding line drawings of scenes with shadows," in *The psychology of computer vision*, 1975.

[203] H. G. Barrow and J. M. Tenenbaum, "Interpreting line drawings as three-dimensional surfaces," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 75–116, 1981.

[204] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, no. 3, pp. 355–395, 1987.

[205] R. L. Solso, M. K. MacLin, and O. H. MacLin, *Cognitive psychology*. Pearson Education New Zealand, 2005.

[206] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The helmholtz machine," *Neural computation*, vol. 7, no. 5, pp. 889–904, 1995.

[207] L. G. Roberts, *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[208] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive psychology*, pp. 143–177, 1982.

[209] M. Blum, A. Griffith, and B. Neumann, "A stability test for configurations of blocks," tech. rep., Massachusetts Institute of Technology, 1970.

[210] M. Brand, P. Cooper, and L. Birnbaum, "Seeing physics, or: Physics is for prediction," in *Proceedings of the Workshop on Physics-based Modeling in Computer Vision*, 1995.

[211] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[212] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.

[213] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[214] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering free space of indoor scenes from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[215] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[216] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Efficient structured prediction for 3d indoor scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[217] R. Guo and D. Hoiem, "Support surface prediction in indoor scenes," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.

[218] T. Shao, A. Monszpart, Y. Zheng, B. Koo, W. Xu, K. Zhou, and N. J. Mitra, "Imagining the unseen: Stability-based cuboid arrangements for scene understanding," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, 2014.

[219] Y. Du, Z. Liu, H. Basevi, A. Leonardis, B. Freeman, J. Tenenbaum, and J. Wu, "Learning to exploit stability for 3d scene parsing," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2018.

[220] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

[221] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman, "Physics 101: learning physical object properties from unlabeled videos," in *Proceedings of British Machine Vision Conference (BMVC)*, 2016.

[222] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[223] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, "Inferring forces and learning human utilities from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[224] M. A. Brubaker and D. J. Fleet, "The kneed walker for human pose tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[225] M. A. Brubaker, L. Sigal, and D. J. Fleet, "Estimating contact dynamics," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.

[226] M. A. Brubaker, D. J. Fleet, and A. Hertzmann, "Physics-based person tracking using the anthropomorphic walker," *International Journal of Computer Vision (IJCV)*, vol. 87, no. 1-2, p. 140, 2010.

[227] T.-H. Pham, A. Kheddar, A. Qammaz, and A. A. Argyros, "Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[228] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, and J. Chai, "Video-based hand manipulation capture through composite motion control," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 43, 2013.

[229] W. Zhao, J. Zhang, J. Min, and J. Chai, "Robust realtime physics-based motion control for human grasping," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 207, 2013.

[230] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, "How, whether, why: Causal judgments as counterfactual contrasts," in *Proceedings of the 37th annual conference of the cognitive science society*, 2015.

[231] J. B. Hamrick, P. W. Battaglia, T. L. Griffiths, and J. B. Tenenbaum, "Inferring mass in complex scenes by mental simulation," *Cognition*, vol. 157, pp. 61–76, 2016.

[232] A. N. Sanborn, "Testing bayesian and heuristic predictions of mass judgments of colliding objects," *Frontiers in psychology*, vol. 5, p. 938, 2014.

[233] A. N. Sanborn, V. K. Mansinghka, and T. L. Griffiths, "Reconciling intuitive physics and newtonian mechanics for colliding objects," *Psychological review*, vol. 120, no. 2, p. 411, 2013.

[234] J. J. Monaghan, "Smoothed particle hydrodynamics," *Annual review of astronomy and astrophysics*, vol. 30, no. 1, pp. 543–574, 1992.

[235] D. Sulsky, S.-J. Zhou, and H. L. Schreyer, "Application of a particle-in-cell method to solid mechanics," *Computer physics communications*, vol. 87, no. 1-2, pp. 236–252, 1995.

[236] C. Jiang, C. Schroeder, A. Selle, J. Teran, and A. Stomakhin, "The affine particle-in-cell method," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, p. 51, 2015.

[237] G. Klar, T. Gast, A. Pradhana, C. Fu, C. Schroeder, C. Jiang, and J. Teran, "Drucker-prager elastoplasticity for sand animation," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 103, 2016.

[238] C. Jiang, C. Schroeder, J. Teran, A. Stomakhin, and A. Selle, "The material point method for simulating continuum materials," in *ACM SIGGRAPH 2016 Courses*, ACM SIGGRAPH, 2016.

[239] P. McCullagh and J. A. Nelder, *Generalized linear models*. Routledge, 2019.

[240] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[241] D. L. Gilden and D. R. Proffitt, "Heuristic judgment of mass ratio in two-body collisions," *Perception & Psychophysics*, vol. 56, no. 6, pp. 708–720, 1994.

[242] S. Runeson, P. Juslin, and H. Olsson, "Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence," *Psychological review*, vol. 107, no. 3, p. 525, 2000.

[243] J. Clement, "Use of physical intuition and imagistic simulation in expert problem solving," in *Implicit and explicit knowledge: An educational approach*, Ablex Publishing, 1994.

[244] M. Hegarty, "Mechanical reasoning by mental simulation," *Trends in cognitive sciences*, vol. 8, no. 6, pp. 280–285, 2004.

[245] D. L. Schwartz and J. B. Black, "Analog imagery in mental model reasoning: Depictive models," *Cognitive Psychology*, vol. 30, no. 2, pp. 154–219, 1996.

[246] K. A. Smith and E. Vul, "Sources of uncertainty in intuitive physics," *Topics in cognitive science*, vol. 5, no. 1, pp. 185–199, 2013.

[247] C. K. Batchelor and G. Batchelor, *An introduction to fluid dynamics*. Cambridge university press, 2000.

[248] E. A. McAfee and D. R. Proffitt, "Understanding the surface orientation of liquids," *Cognitive Psychology*, vol. 23, no. 3, pp. 483–514, 1991.

[249] D. L. Schwartz and T. Black, "Inferences through imagined actions: Knowing by simulated doing," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 1, p. 116, 1999.

[250] I. P. Howard, "Recognition and knowledge of the water-level principle," *Perception*, vol. 7, no. 2, pp. 151–160, 1978.

[251] H. Krist, E. L. Fieberg, and F. Wilkening, "Intuitive physics in action and judgment: The development of knowledge about projectile motion," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, no. 4, p. 952, 1993.

[252] R. Bridson, *Fluid simulation for computer graphics*. CRC Press, 2015.

[253] B. Tversky, J. B. Morrison, and M. Betrancourt, "Animation: Can it facilitate?," *International journal of human-computer studies*, vol. 57, no. 4, pp. 247–262, 2002.

[254] T. Kawabe, K. Maruya, R. W. Fleming, and S. Nishida, "Seeing liquids from visual motion," *Vision research*, vol. 109, pp. 125–138, 2015.

[255] Y. Zhu and R. Bridson, "Animating sand as a fluid," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 965–972, 2005.

[256] C. Jiang, *The material point method for the physics-based simulation of solids and fluids*. University of California, Los Angeles, 2015.

[257] W. Liang, Y. Zhao, Y. Zhu, and S.-C. Zhu, "Evaluating human cognition of containing relations with physical simulation," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2015.

[258] K. Museth, J. Lait, J. Johanson, J. Budsberg, R. Henderson, M. Alden, P. Cucka, D. Hill, and A. Pearce, "Openvdb: An open-source data structure and toolkit for high-resolution volumes," in *Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2013.

[259] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[260] J. De Kleer and J. S. Brown, "A qualitative physics based on confluences," *Artificial intelligence*, vol. 24, no. 1-3, pp. 7–83, 1984.

[261] R. Grzeszczuk, D. Terzopoulos, and G. Hinton, "Neuroanimator: Fast neural network emulation and control of physics-based models," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998.

[262] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

[263] E. S. Spelke, "Principles of object perception," *Cognitive science*, vol. 14, no. 1, pp. 29–56, 1990.

[264] L. J. Rips and S. J. Hespos, "Divisions of the physical world: Concepts of objects and substances," *Psychological bulletin*, vol. 141, no. 4, p. 786, 2015.

[265] A. Lerer, S. Gross, and R. Fergus, "Learning physical intuition of block towers by example," in *Proceedings of International Conference on Machine Learning (ICML)*, 2016.

[266] J. K. Witt and D. R. Proffitt, "Perceived slant: A dissociation between perception and action," *Perception*, vol. 36, no. 2, pp. 249–257, 2007.

[267] J. K. Stefanucci and D. R. Proffitt, "The roles of altitude and fear in the perception of height," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, no. 2, p. 424, 2009.

[268] Q.-Y. Shi and K.-S. Fu, "Parsing and translation of (attributed) expansive graph languages for scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 5, pp. 472–485, 1983.

[269] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *International Journal of Computer Vision (IJCV)*, vol. 63, no. 2, pp. 113–140, 2005.

[270] DARPA, "Robots rescue people," 2014.

[271] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *International Symposium on Mixed and Augmented Reality*, 2011.

[272] A. Barbu and S.-C. Zhu, "Generalizing swendsen-wang to sampling arbitrary posterior probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 27, no. 8, pp. 1239–1253, 2005.

[273] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3d object dataset: Putting the kinect to work," in *Consumer depth cameras for computer vision*, pp. 141–165, Springer, 2013.

[274] M. Attene, B. Falcidieno, and M. Spagnuolo, "Hierarchical mesh segmentation based on fitting primitives," *The Visual Computer*, vol. 22, no. 3, pp. 181–193, 2006.

[275] J. Poppinga, N. Vaskevicius, A. Birk, and K. Pathak, "Fast plane detection and polygonalization in noisy 3d range images," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2008.

[276] Y. Zhao and S.-C. Zhu, "Image parsing with stochastic scene grammar," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2011.

[277] V. Hedau, D. Hoiem, and D. Forsyth, "Thinking inside the box: Using appearance models and context based on room geometry," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[278] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Manhattan-world stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[279] M. Phillips and M. Likhachev, "Sipp: Safe interval path planning for dynamic environments," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2011.

[280] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic, "People watching: Human actions as a cue for single view geometry," *International Journal of Computer Vision (IJCV)*, vol. 110, no. 3, pp. 259–274, 2014.

[281] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros, "Scene semantics from long-term observation of people," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[282] Y. Jiang and A. Saxena, "Infinite latent conditional random fields for modeling environments through humans," in *Proceedings of Robotics: Science and Systems (RSS)*, 2013.

[283] D. Wales *et al.*, *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

[284] M. M. Blane, Z. Lei, H. Civi, and D. B. Cooper, "The 3l algorithm for fitting implicit polynomial curves and surfaces to data," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, no. 3, pp. 298–313, 2000.

[285] B. Zheng, J. Takamatsu, and K. Ikeuchi, "An adaptive and stable method for fitting implicit polynomial curves and surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 3, pp. 561–568, 2009.

[286] R. Sagawa, K. Nishino, and K. Ikeuchi, "Adaptively merging large-scale range data with reflectance properties," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 27, no. 3, pp. 392–405, 2005.

[287] D. J. Kriegman, "Let them fall where they may: Capture regions of curved objects and polyhedra," *International Journal of Robotics Research (IJRR)*, vol. 16, no. 4, pp. 448–472, 1997.

[288] X. Chen, A. Golovinskiy, and T. Funkhouser, "A benchmark for 3d mesh segmentation," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, pp. 1–12, 2009.

[289] A. Falcon, "Aristotle's method of inquiry in eudemian ethics 1 and 2," in *Thinking, Knowing, Acting: Epistemology and Ethics in Plato and Ancient Platonism*, pp. 186–206, Brill, 2019.

[290] J. L. Mackie, *The cement of the universe: A study of causation*. Clarendon Press, 1974.

[291] J. Pearl, *Causality: Models, reasoning and inference*. Cambridge University Press, 2000.

[292] D. R. Shanks and A. Dickinson, "Associative accounts of causality judgment," in *Psychology of learning and motivation*, vol. 21, pp. 229–261, Elsevier, 1988.

[293] D. R. Shanks, "Categorization by a connectionist network," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 17, no. 3, p. 433, 1991.

[294] R. A. Rescorla and A. R. Wagner, "A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," *Classical conditioning II: Current research and theory*, vol. 2, pp. 64–99, 1972.

[295] K. Holyoak and P. W. Cheng, "Causal learning and inference as a rational process: The new synthesis," *Annual Review of Psychology*, vol. 62, pp. 135–163, 2011.

[296] M. R. Waldmann and K. J. Holyoak, "Predictive and diagnostic learning within causal models: asymmetries in cue competition," *Journal of Experimental Psychology: General*, vol. 121, no. 2, pp. 222–236, 1992.

[297] P. W. Cheng, "From covariation to causation: a causal power theory," *Psychological Review*, vol. 104, no. 2, pp. 367–405, 1997.

[298] T. L. Griffiths and J. B. Tenenbaum, "Structure and strength in causal induction," *Cognitive psychology*, vol. 51, no. 4, pp. 334–384, 2005.

[299] T. L. Griffiths and J. B. Tenenbaum, "Theory-based causal induction," *Psychological review*, vol. 116, no. 4, p. 661, 2009.

[300] H. Lu, A. L. Yuille, M. Liljeholm, P. W. Cheng, and K. J. Holyoak, "Bayesian generic priors for causal learning," *Psychological Review*, vol. 115, no. 4, pp. 955–984, 2008.

[301] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in Cognitive Sciences*, vol. 4, no. 8, pp. 299–309, 2000.

[302] A. Fire and S.-C. Zhu, "Learning perceptual causality from video," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 2, p. 23, 2016.

[303] W.-k. Ahn, C. W. Kalish, D. L. Medin, and S. A. Gelman, "The role of covariation versus mechanism information in causal attribution," *Cognition*, vol. 54, no. 3, pp. 299–352, 1995.

[304] P. Wolff, "Representing causation," *Journal of experimental psychology: General*, vol. 136, no. 1, p. 82, 2007.

[305] B. Schölkopf, "Causality for machine learning," *arXiv preprint arXiv:1911.10500*, 2019.

[306] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson, *Causation, prediction, and search*. MIT press, 2000.

[307] R. A. Fisher, "Design of experiments," *Br Med J*, vol. 1, no. 3923, pp. 554–554, 1936.

[308] F. Bacon, *Novum organum*. Clarendon press, 1878.

[309] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[310] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[311] A. Fire and S.-C. Zhu, "Using causal induction in humans to learn and infer causality from video," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2013.

[312] S.-C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural computation*, vol. 9, no. 8, pp. 1627–1660, 1997.

[313] N. R. Bramley, P. Dayan, T. L. Griffiths, and D. A. Lagnado, "Formalizing neurath's ship: Approximate algorithms for online causal learning," *Psychological review*, vol. 124, no. 3, p. 301, 2017.

[314] F. Heider, "The psychology of interpersonal relations hillsdale," *New Jersey: LEA*, 1958.

[315] J. R. Kubricht, H. Lu, and K. J. Holyoak, "Individual differences in spontaneous analogical transfer," *Memory & Cognition*, vol. 45, no. 4, pp. 576–588, 2017.

[316] E. Catto, "Box2d: A 2d physics engine for games," *URL: http://www. box2d. org*, 2011.

[317] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[318] J. B. Tenenbaum, T. L. Griffiths, and C. Kemp, "Theory-based bayesian models of inductive learning and reasoning," *Trends in cognitive sciences*, vol. 10, no. 7, pp. 309–318, 2006.

[319] M. Thielscher, "Introduction to the fluent calculus," *Computer and Information Science*, vol. 3, no. 14, 1998.

[320] M. Edmonds, J. Kubricht, C. Summers, Y. Zhu, B. Rothrock, S.-C. Zhu, and H. Lu, "Human causal transfer: Challenges for deep reinforcement learning," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2018.

[321] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[322] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.

[323] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of International Conference on Machine Learning (ICML)*, 2016.

[324] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

[325] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[326] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

[327] D. A. Van Valen, T. Kudo, K. M. Lane, D. N. Macklin, N. T. Quach, M. M. DeFelice, I. Maayan, Y. Tanouchi, E. A. Ashley, and M. W. Covert, "Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments," *PLoS computational biology*, vol. 12, no. 11, p. e1005177, 2016.

[328] F. Osiurak, C. Jarry, and D. Le Gall, "Grasping the affordances, understanding the reasoning: toward a dialectical theory of human tool use," *Psychological review*, vol. 117, no. 2, p. 517, 2010.

[329] B. B. Beck, *Animal tool behavior: The use and manufacture of tools by animals*. Garland STPM Press New York, 1980.

[330] R. W. Shumaker, K. R. Walkup, and B. B. Beck, *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

[331] C. Baber, *Cognition and tool use: Forms of engagement in human and animal use of tools*. CRC Press, 2003.

[332] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[333] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2011.

[334] S. Nakano, G. Ueno, and T. Higuchi, "Merging particle filter for sequential data assimilation," *Nonlinear Processes in Geophysics*, vol. 14, no. 4, pp. 395–408, 2007.

[335] J. Goodall, *The chimpanzees of Gombe: patterns of behavior*. Bellknap Press of the Harvard University Press, 1986.

[336] A. Whiten and R. W. Byrne, *Machiavellian monkeys: Cognitive evolution and the social world of primates*. Clarendon Press/Oxford University Press, 1989.

[337] A. Whiten, J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, C. E. Tutin, R. W. Wrangham, and C. Boesch, "Cultures in chimpanzees," *Nature*, vol. 399, no. 6737, p. 682, 1999.

[338] G. Sabbatini, H. M. Manrique, C. Trapanese, A. D. B. Vizioli, J. Call, and E. Visalberghi, "Sequential use of rigid and pliable tools in tufted capuchin monkeys," *Animal Behaviour*, vol. 87, pp. 213–220, 2014.

[339] L. R. Santos, H. M. Pearson, G. M. Spaepen, F. Tsao, and M. D. Hauser, "Probing the limits of tool competence: Experiments with two non-tool-using species (cercopithecus aethiops and saguinus oedipus)," *Animal cognition*, vol. 9, no. 2, pp. 94–109, 2006.

[340] A. A. Weir, J. Chappell, and A. Kacelnik, "Shaping of hooks in new caledonian crows," *Science*, vol. 297, no. 5583, pp. 981–981, 2002.

[341] W. McGrew, "Tool-use by free-ranging chimpanzees: the extent of diversity," *Journal of Zoology*, vol. 228, no. 4, pp. 689–694, 1992.

[342] S. H. Frey, "What puts the how in where? tool use and the divided visual streams hypothesis," *Cortex*, vol. 43, no. 3, pp. 368–375, 2007.

[343] K. R. Gibson, K. R. Gibson, and T. Ingold, *Tools, language and cognition in human evolution.* Cambridge University Press, 1993.

[344] K. Vaesen, "The cognitive bases of human tool use," *Behavioral and Brain Sciences*, vol. 35, no. 4, pp. 203–218, 2012.

[345] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[346] J. W. Lewis, "Cortical networks related to human use of tools," *The neuroscientist*, vol. 12, no. 3, pp. 211–231, 2006.

[347] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele, "Functional object class detection based on learned affordance cues," in *International conference on computer vision systems*, 2008.

[348] K. M. Varadarajan and M. Vincze, "Afrob: The affordance network ontology for robots," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[349] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[350] A. Stoytchev, "Behavior-grounded representation of tool affordances," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2005.

[351] A. Pieropan, C. H. Ek, and H. Kjellström, "Functional object descriptors for human activity modeling," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2013.

[352] V. Jain and M. Varma, "Learning to re-rank: Query-dependent image re-ranking using click data," in *Proceedings of the 20th international conference on World wide web*, 2011.

[353] K. M. Varadarajan and M. Vincze, "Affordance based part recognition for grasping and manipulation," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2011.

[354] A. Myers, A. Kanazawa, C. Fermuller, and Y. Aloimonos, "Affordance of object parts from geometric features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[355] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2015.

[356] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[357] S.-B. Ho, *Representing and using functional definitions for visual recognition.* PhD thesis, The University of Wisconsin-Madison, 1987.

[358] L. Stark and K. Bowyer, "Achieving generalized object recognition through reasoning about association of function to structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 13, no. 10, pp. 1097–1104, 1991.

[359] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Computer Vision and Image Understanding*, vol. 115, no. 1, pp. 81–90, 2011.

[360] Y. Zhu, A. Fathi, and L. Fei-Fei, "Reasoning about object affordances in a knowledge base representation," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[361] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3d object detection with rgbd cameras," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.

[362] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser, "Shape2pose: Human-centric shape analysis," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–12, 2014.

[363] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," in *Proceedings of British Machine Vision Conference (BMVC)*, 2011.

[364] S. Legg and M. Hutter, "Universal intelligence: A definition of machine intelligence," *Minds and machines*, vol. 17, no. 4, pp. 391–444, 2007.

[365] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[366] P. Kormushev, S. Calinon, and D. G. Caldwell, "Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input," *Advanced Robotics*, vol. 25, no. 5, pp. 581–603, 2011.

[367] A. Montebelli, F. Steinmetz, and V. Kyrki, "On handing down our tools to robots: Single-phase kinesthetic teaching for dynamic in-contact tasks," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2015.

[368] S. Manschitz, M. Gienger, J. Kober, and J. Peters, "Probabilistic decomposition of sequential force interaction tasks into movement primitives," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[369] M. Racca, J. Pajarinen, A. Montebelli, and V. Kyrki, "Learning in-contact control strategies from demonstration," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[370] K. Kukliński, K. Fischer, I. Marhenke, F. Kirstein, V. Maria, N. Krüger, T. R. Savarimuthu, *et al.*, "Teleoperation for learning by demonstration: Data glove versus object manipulation for intuitive robot control," in *2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2014.

[371] G. M. Hayes and J. Demiris, *A robot controller using learning by imitation.* University of Edinburgh, Department of Artificial Intelligence, 1994.

[372] M. Muhlig, M. Gienger, S. Hellbach, J. J. Steil, and C. Goerick, "Task-level imitation learning using variance-based movement optimization," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2009.

[373] S. Ross, G. J. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011.

[374] C. Paxton, F. Jonathan, M. Kobilarov, and G. D. Hager, "Do what i want, not what i did: Imitation of skills by planning sequences of actions," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[375] C. Xiong, N. Shukla, W. Xiong, and S.-C. Zhu, "Robot learning with a spatial, temporal, and causal and-or graph," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2016.

[376] T. Shu, X. Gao, M. S. Ryoo, and S.-C. Zhu, "Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2017.

[377] J. Kober and J. R. Peters, "Policy search for motor primitives in robotics," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009.

[378] F. Guenter, M. Hersch, S. Calinon, and A. Billard, "Reinforcement learning for imitating constrained reaching movements," *Advanced Robotics*, vol. 21, no. 13, pp. 1521–1544, 2007.

[379] E. Theodorou, J. Buchli, and S. Schaal, "Reinforcement learning of motor skills in high dimensions: A path integral approach," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2010.

[380] U. Prieur, V. Perdereau, and A. Bernardino, "Modeling and planning high-level in-hand manipulation actions from human knowledge and active learning from demonstration," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[381] A. Gupta, C. Eppner, S. Levine, and P. Abbeel, "Learning dexterous manipulation for a soft robotic hand from human demonstrations," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[382] S. Levine and P. Abbeel, "Learning neural network policies with guided policy search under unknown dynamics," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

[383] S. Levine, N. Wagener, and P. Abbeel, "Learning contact-rich manipulation skills with guided policy search," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2015.

[384] A. Y. Ng, S. J. Russell, *et al.*, "Algorithms for inverse reinforcement learning," in *Proceedings of International Conference on Machine Learning (ICML)*, 2000.

[385] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of International Conference on Machine Learning (ICML)*, 2004.

[386] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[387] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[388] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of International Conference on Machine Learning (ICML)*, 1999.

[389] J. MacGlashan and M. L. Littman, "Between imitation and intention learning," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[390] K. Dautenhahn and C. L. Nehaniv, *Imitation in animals and artifacts*. MIT Press Cambridge, MA, 2002.

[391] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive brain research*, vol. 3, no. 2, pp. 131–141, 1996.

[392] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nature reviews neuroscience*, vol. 2, no. 9, p. 661, 2001.

[393] M. Iacoboni, R. P. Woods, M. Brass, H. Bekkering, J. C. Mazziotta, and G. Rizzolatti, "Cortical mechanisms of human imitation," *Science*, vol. 286, no. 5449, pp. 2526–2528, 1999.

[394] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annual Review of Neuroscience*, vol. 27, pp. 169–192, 2004.

[395] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers, "The anthropomorphic brain: The mirror neuron system responds to human and robotic actions," *Neuroimage*, vol. 35, no. 4, pp. 1674–1684, 2007.

[396] L. M. Oberman, J. P. McCleery, V. S. Ramachandran, and J. A. Pineda, "Eeg evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots," *Neurocomputing*, vol. 70, no. 13-15, pp. 2194–2203, 2007.

[397] M. J. Rochat, F. Caruana, A. Jezzini, I. Intskirveli, F. Grammont, V. Gallese, G. Rizzolatti, M. A. Umiltà, *et al.*, "Responses of mirror neurons in area f5 to hand and tool grasping observation," *Experimental brain research*, vol. 204, no. 4, pp. 605–616, 2010.

[398] M. Umiltà, I. Intskirveli, F. Grammont, M. Rochat, F. Caruana, A. Jezzini, V. Gallese, G. Rizzolatti, *et al.*, "When pliers become fingers in the monkey motor system," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 2209–2213, 2008.

[399] S. Thill, D. Caligiore, A. M. Borghi, T. Ziemke, and G. Baldassarre, "Theories and computational models of affordance and mirror systems: an integrative review," *Neuroscience & Biobehavioral Reviews*, vol. 37, no. 3, pp. 491–521, 2013.

[400] M. Iacoboni, "Imitation, empathy, and mirror neurons," *Annual review of psychology*, vol. 60, pp. 653–670, 2009.

[401] G. Hickok, "Eight problems for the mirror neuron theory of action understanding in monkeys and humans," *Journal of cognitive neuroscience*, vol. 21, no. 7, pp. 1229–1243, 2009.

[402] T.-H. Pham, N. Kyriazis, A. A. Argyros, and A. Kheddar, "Hand-object contact force estimation from markerless visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 12, pp. 2883–2896, 2018.

[403] Y. Gu, W. Sheng, M. Liu, and Y. Ou, "Fine manipulative action recognition through sensor fusion," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[404] F. L. Hammond, Y. Mengüç, and R. J. Wood, "Toward a modular soft sensor-embedded glove for human hand motion and tactile pressure measurement," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[405] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, and S.-C. Zhu, "A glove-based system for studying hand-object manipulation via joint pose and force sensing," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[406] H. Liu, C. Zhang, Y. Zhu, C. Jiang, and S.-C. Zhu, "Mirroring without overimitation: Learning functionally equivalent manipulation actions," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[407] S. H. Johnson-Frey, F. R. Maloof, R. Newman-Norlund, C. Farrer, S. Inati, and S. T. Grafton, "Actions or hand-object interactions? human inferior frontal cortex and action observation," *Neuron*, vol. 39, no. 6, pp. 1053–1058, 2003.

[408] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 7, pp. 881–892, 2002.

[409] C. W. Macosko, *Rheology: Principles, measurements, and applications*. Wiley-vch, 1994.

[410] J. Bonet and R. D. Wood, *Nonlinear continuum mechanics for finite element analysis*. Cambridge university press, 1997.

[411] H. Si, "Tetgen, a delaunay-based quality tetrahedral mesh generator," *ACM Transactions on Mathematical Software*, vol. 41, no. 2, p. 11, 2015.

[412] M. T. Mason, "Toward robotic manipulation," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 1–28, 2018.

[413] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

[414] S. M. LaValle, J. J. Kuffner, B. Donald, *et al.*, "Rapidly-exploring random trees: Progress and prospects," *Algorithmic and computational robotics: new directions*, vol. 5, pp. 293–308, 2001.

[415] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2000.

[416] S. Karaman and E. Frazzoli, "Incremental sampling-based algorithms for optimal motion planning," *Robotics Science and Systems VI*, vol. 104, no. 2, 2010.

[417] A. Petrovskaya and A. Y. Ng, "Probabilistic mobile manipulation in dynamic environments, with application to opening doors," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[418] S. Chitta, B. Cohen, and M. Likhachev, "Planning for autonomous door opening with a mobile manipulator," in *2010 IEEE International Conference on Robotics and Automation*, 2010.

[419] K. Gochev, A. Safonova, and M. Likhachev, "Planning with adaptive dimensionality for mobile manipulation," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2012.

[420] D. J. Webb and J. Van Den Berg, "Kinodynamic rrt*: Asymptotically optimal motion planning for robots with linear dynamics," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2013.

[421] M. Stilman, "Global manipulation planning in robot joint space with task constraints," *IEEE Transactions on Robotics*, vol. 26, no. 3, pp. 576–584, 2010.

[422] N. Ratliff, M. Zucker, J. A. Bagnell, and S. Srinivasa, "Chomp: Gradient optimization techniques for efficient motion planning," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2009.

[423] M. Kalakrishnan, S. Chitta, E. Theodorou, P. Pastor, and S. Schaal, "Stomp: Stochastic trajectory optimization for motion planning," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2011.

[424] J. Schulman, J. Ho, A. X. Lee, I. Awwal, H. Bradlow, and P. Abbeel, "Finding locally optimal, collision-free trajectories with sequential convex optimization," in *Proceedings of Robotics: Science and Systems (RSS)*, 2013.

[425] D. M. Bodily, T. F. Allen, and M. D. Killpack, "Motion planning for mobile robots using inverse kinematics branching," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2017.

[426] D. Berenson, J. Kuffner, and H. Choset, "An optimization approach to planning for mobile manipulation," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2008.

[427] B. Magyar, N. Tsiogkas, J. Deray, S. Pfeiffer, and D. Lane, "Timed-elastic bands for manipulation motion planning," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3513–3520, 2019.

[428] K. Shankar, *Kinematics and Local Motion Planning for Quasi-static Whole-Body Mobile Manipulation*. PhD thesis, California Institute of Technology, 2016.

[429] M. Stuede, K. Nuelle, S. Tappe, and T. Ortmaier, "Door opening and traversal with an industrial cartesian impedance controlled mobile robot," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2019.

[430] N. Likar, B. Nemec, and L. Žlajpah, "Virtual mechanism approach for dual-arm manipulation," *Robotica*, vol. 32, no. 6, 2014.

[431] J. Pratt, C.-M. Chew, A. Torres, P. Dilworth, and G. Pratt, "Virtual model control: An intuitive approach for bipedal locomotion," *The International Journal of Robotics Research*, vol. 20, no. 2, pp. 129–143, 2001.

[432] X. Wang and G.-H. Yang, "Cooperative adaptive fault-tolerant tracking control for a class of multi-agent systems with actuator failures and mismatched parameter uncertainties," *IET Control Theory & Applications*, vol. 9, no. 8, pp. 1274–1284, 2015.

[433] J. Schulman, Y. Duan, J. Ho, A. Lee, I. Awwal, H. Bradlow, J. Pan, S. Patil, K. Goldberg, and P. Abbeel, "Motion planning with sequential convex optimization and convex collision checking," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1251–1270, 2014.

[434] T. Yoshikawa, "Manipulability of robotic mechanisms," *The international journal of Robotics Research*, vol. 4, no. 2, pp. 3–9, 1985.

[435] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

[436] N. Ratliff, M. Toussaint, and S. Schaal, "Understanding the geometry of workspace obstacles in motion optimization," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2015.

[437] S. C. Zhu and D. Mumford, "Prior learning and gibbs reaction-diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1236–1250, 1997.

[438] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, 1977.

[439] T. Hermans, J. M. Rehg, and A. Bobick, "Affordance prediction via learned object attributes," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2011.

[440] G. Fritz, L. Paletta, R. Breithaupt, E. Rome, and G. Dorffner, "Learning predictive features in affordance based robotic perception systems," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2006.

[441] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, "Scenegrok: Inferring action maps in 3d environments," *ACM transactions on graphics (TOG)*, vol. 33, no. 6, pp. 1–10, 2014.

[442] Y. Jiang and A. Saxena, "Hallucinating humans for learning robotic placement of objects," in *Experimental Robotics*, 2013.

[443] Y. Jiang, M. Lim, and A. Saxena, "Learning object arrangements in 3d scenes using human context," in *Proceedings of International Conference on Machine Learning (ICML)*, 2012.

[444] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[445] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2014.

[446] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.

[447] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray, "Very high frame rate volumetric integration of depth images on mobile devices," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 11, pp. 1241–1250, 2015.

[448] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, pp. 1–11, 2013.

[449] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald, "Kintinuous: Spatially extended kinectfusion," tech. rep., MIT-CSAIL, 2012.

[450] R. Bridson, "Fast poisson disk sampling in arbitrary dimensions," *SIGGRAPH sketches*, vol. 10, p. 1, 2007.

[451] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[452] B. K. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Josa a*, vol. 4, no. 4, pp. 629–642, 1987.

[453] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 286–298, 2007.

[454] C. Sylvain, "Robot programming by demonstration: A probabilistic approach," 2009.

[455] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[456] N. Molino, R. Bridson, J. Teran, and R. Fedkiw, "A crystalline, red green strategy for meshing highly deformable objects with tetrahedra," in *IMR*, 2003.

[457] A. Stomakhin, R. Howes, C. Schroeder, and J. M. Teran, "Energetically consistent invertible elasticity," in *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, 2012.

[458] T. F. Gast, C. Schroeder, A. Stomakhin, C. Jiang, and J. M. Teran, "Optimization integrator for large time steps," *Proceedings of IEEE Transactions on Visualization & Computer Graph (TVCG)*, vol. 21, no. 10, pp. 1103–1115, 2015.

[459] E. W. Dijkstra *et al.*, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[460] G. S. Becker and W. M. Landes, "Front matter, essays in the economics of crime and punishment," in *Essays in the Economics of Crime and Punishment*, pp. 20–0, NBER, 1974.

[461] L. E. Blume and D. Easley, "Rationality," *The new Palgrave dictionary of economics*, vol. 6, pp. 884–893, 2008.

[462] P. Hedström and C. Stern, "Rational choice and sociology," *The new Palgrave dictionary of economics*, vol. 2, 2008.

[463] S. Lohmann, "Rational choice and political science," *The new Palgrave dictionary of economics*, vol. 2, 2008.

[464] K.-U. Hoffgen, H.-U. Simon, and K. S. Vanhorn, "Robust trainability of single neurons," *Journal of Computer and System Sciences*, vol. 50, no. 1, pp. 114–125, 1995.

[465] Q.-Y. Zhou, S. Miller, and V. Koltun, "Elastic fragments for dense scene reconstruction," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.

[466] S.-H. Lee, E. Sifakis, and D. Terzopoulos, "Comprehensive biomechanical modeling and simulation of the upper body," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 4, pp. 1–17, 2009.

[467] Y. Lee, M. S. Park, T. Kwon, and J. Lee, "Locomotion control for many-muscle humanoids," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 1–11, 2014.

[468] M. L. Patterson, *Nonverbal behavior: A functional perspective*. Springer Science & Business Media, 2012.

[469] R. M. Krauss, Y. Chen, and P. Chawla, "Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?," *Advances in experimental social psychology*, vol. 28, pp. 389–450, 1996.

[470] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[471] N. Tinbergen, ""derived" activities; their causation, biological significance, origin, and emancipation during evolution," *The Quarterly review of biology*, vol. 27, no. 1, pp. 1–32, 1952.

[472] D. Sperber and D. Wilson, *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA, 1986.

[473] J. W. Schooler and T. Y. Engstler-Schooler, "Verbal overshadowing of visual memories: Some things are better left unsaid," *Cognitive psychology*, vol. 22, no. 1, pp. 36–71, 1990.

[474] J. W. Schooler, S. Ohlsson, and K. Brooks, "Thoughts beyond words: When language overshadows insight," *Journal of experimental psychology: General*, vol. 122, no. 2, p. 166, 1993.

[475] T. D. Wilson, D. J. Lisle, J. W. Schooler, S. D. Hodges, K. J. Klaaren, and S. J. LaFleur, "Introspecting about reasons can reduce post-choice satisfaction," *Personality and Social Psychology Bulletin*, vol. 19, no. 3, pp. 331–339, 1993.

[476] T. D. Wilson and J. W. Schooler, "Thinking too much: Introspection can reduce the quality of preferences and decisions," *Journal of personality and social psychology*, vol. 60, no. 2, p. 181, 1991.

[477] H. P. Grice, "Meaning," *The philosophical review*, vol. 66, no. 3, pp. 377–388, 1957.

[478] H. P. Grice, P. Cole, J. Morgan, *et al.*, "Logic and conversation," *1975*, pp. 41–58, 1975.

[479] J. R. Searle, S. Willis, *et al.*, *The construction of social reality*. Simon and Schuster, 1995.

[480] M. E. Bratman, "Shared cooperative activity," *The philosophical review*, vol. 101, no. 2, pp. 327–341, 1992.

[481] M. Gilbert, *On social facts*. Princeton University Press, 1992.

[482] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and brain sciences*, vol. 28, no. 5, pp. 675–691, 2005.

[483] M. I. Posner, Y. Cohen, *et al.*, "Components of visual orienting," *Attention and performance X: Control of language processes*, vol. 32, pp. 531–556, 1984.

[484] R. Parasuraman, "Vigilance, monitoring, and search," in *Handbook of perception and human performance*, John Wiley & Sons, 1986.

[485] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *perception*, vol. 28, no. 9, pp. 1059–1074, 1999.

[486] D. J. Simons and D. T. Levin, "Failure to detect changes to people during a real-world interaction," *Psychonomic Bulletin & Review*, vol. 5, no. 4, pp. 644–649, 1998.

[487] N. J. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[488] R. V. Exline, "Visual interaction: The glances of power and preference," in *Nebraska symposium on motivation*, 1971.

[489] M. Argyle and M. Cook, *Gaze and mutual gaze*. Cambridge U Press, 1976.

[490] R. Exline, D. Gray, and D. Schuette, "Visual behavior in a dyad as affected by interview content and sex of respondent," *Journal of Personality and Social Psychology*, vol. 1, no. 3, p. 201, 1965.

[491] N. F. Russo, "Eye contact, interpersonal distance, and the equilibrium theory," *Journal of Personality and Social Psychology*, vol. 31, no. 3, p. 497, 1975.

[492] B. Butterworth and G. Beattie, "Gesture and silence as indicators of planning in speech," in *Recent advances in the psychology of language*, pp. 347–360, Springer, 1978.

[493] S. Duncan, L. J. Brunner, and D. W. Fiske, "Strategy signals in face-to-face interaction," *Journal of Personality and Social Psychology*, vol. 37, no. 2, p. 301, 1979.

[494] L. J. Brunner, "Smiles can be back channels," *Journal of Personality and Social Psychology*, vol. 37, no. 5, p. 728, 1979.

[495] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[496] M. M. Haith, T. Bergman, and M. J. Moore, "Eye contact and face scanning in early infancy," *Science*, vol. 198, no. 4319, pp. 853–855, 1977.

[497] R. J. Itier and M. Batty, "Neural bases of eye and gaze processing: the core of social cognition," *Neuroscience & Biobehavioral Reviews*, vol. 33, no. 6, pp. 843–863, 2009.

[498] M. Jording, A. Hartz, G. Bente, M. Schulte-Rüther, and K. Vogeley, "The "social gaze space": A taxonomy for gaze-based communication in triadic interactions," *Frontiers in psychology*, vol. 9, p. 226, 2018.

[499] H. Kobayashi and S. Kohshima, "Unique morphology of the human eye," *Nature*, vol. 387, no. 6635, pp. 767–768, 1997.

[500] A. Seemann, *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*. MIT Press, 2011.

[501] Z. Wei, Y. Yan, L. Huang, and J. Nie, "Inferring intrinsic correlation between clothing style and wearers' personality," *Multimedia Tools and Applications*, vol. 76, no. 19, pp. 20273–20285, 2017.

[502] C.-J. Kim, "Dynamic linear models with markov-switching," *Journal of Econometrics*, vol. 60, no. 1-2, pp. 1–22, 1994.

[503] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.

[504] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.

[505] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[506] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[507] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? jointly inferring human attention and intentions in complex tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[508] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental brain research*, vol. 139, no. 3, pp. 266–277, 2001.

[509] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[510] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1165–1179, 2016.

[511] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

[512] M. S. Gazzaniga, *The new cognitive neurosciences*. MIT press, 2000.

[513] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[514] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[515] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2017.

[516] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[517] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

[518] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[519] C. L. Kleinke, "Gaze and eye contact: a research review," *Psychological bulletin*, vol. 100, no. 1, p. 78, 1986.

[520] J. K. Burgoon, K. Floyd, and L. K. Guerrero, "Nonverbal communication theories of interpersonal adaptation," in *The new SAGE handbook of communication science*, pp. 93–110, Sage, 2010.

[521] C. Yu, P. Schermerhorn, and M. Scheutz, "Adaptive eye gaze patterns in interactions with human and artificial agents," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 2, p. 13, 2012.

[522] M. Staudte and M. W. Crocker, "Investigating joint attention mechanisms through spoken human–robot interaction," *Cognition*, vol. 120, no. 2, pp. 268–291, 2011.

[523] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[524] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

[525] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[526] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *International Conference on Machine Learning*, pp. 2688–2697, PMLR, 2018.

[527] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018.

[528] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 18, no. 11, pp. 1473–1488, 2008.

[529] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[530] F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American journal of psychology*, vol. 57, no. 2, pp. 243–259, 1944.

[531] D. A. Baldwin and J. A. Baird, "Discerning intentions in dynamic human action," *Trends in Cognitive Sciences*, vol. 5, no. 4, pp. 171–178, 2001.

[532] G. Gergely, Z. Nádasdy, G. Csibra, and S. Bíró, "Taking the intentional stance at 12 months of age," *Cognition*, vol. 56, no. 2, pp. 165–193, 1995.

[533] G. Csibra, "Teleological and referential understanding of action in infancy," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1431, pp. 447–458, 2003.

[534] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[535] G. Sukthankar, C. Geib, H. H. Bui, D. Pynadath, and R. P. Goldman, *Plan, activity, and intent recognition: Theory and practice*. Newnes, 2014.

[536] A. J. Brizard, *An introduction to Lagrangian mechanics*. World Scientific Publishing Company, 2014.

[537] J. Kwon and K. M. Lee, "Wang-landau monte carlo-based tracking methods for abrupt motions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 1011–1024, 2013.

[538] Z. Tu and S.-C. Zhu, "Image segmentation by data-driven markov chain monte carlo," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 657–673, 2002.

[539] S. Qi, B. Jia, and S.-C. Zhu, "Generalized earley parser: Bridging symbolic grammars and sequence data for future prediction," in *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

[540] B. Hayes and B. Scassellati, "Autonomously constructing hierarchical task networks for planning and human-robot collaboration," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2016.

[541] W. H. Dittrich and S. E. Lea, "Visual perception of intentional motion," *Perception*, vol. 23, no. 3, pp. 253–268, 1994.

[542] T. Gao, G. E. Newman, and B. J. Scholl, "The psychophysics of chasing: A case study in the perception of animacy," *Cognitive psychology*, vol. 59, no. 2, pp. 154–179, 2009.

[543] T. Shu, Y. Peng, L. Fan, H. Lu, and S.-C. Zhu, "Inferring human interaction from motion trajectories in aerial videos," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2017.

[544] T. D. Ullman, C. L. Baker, O. Macindoe, O. Evans, N. D. Goodman, and J. B. Tenenbaum, "Help or hinder: Bayesian models of social goal inference," tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF BRAIN AND COGNITIVE SCIENCES, 2009.

[545] P. C. Pantelis, C. L. Baker, S. A. Cholewiak, K. Sanik, A. Weinstein, C.-C. Wu, J. B. Tenenbaum, and J. Feldman, "Inferring the intentional states of autonomous virtual agents," *Cognition*, vol. 130, no. 3, pp. 360–379, 2014.

[546] E. Farnioli, M. Gabiccini, and A. Bicchi, "Optimal contact force distribution for compliant humanoid robots in whole-body loco-manipulation tasks," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2015.

[547] J. Xie, W. Hu, S.-C. Zhu, and Y. N. Wu, "Learning sparse frame models for natural image patterns," *International Journal of Computer Vision*, vol. 114, no. 2, pp. 91–112, 2015.

[548] H. M. Wellman and S. A. Gelman, "Cognitive development: Foundational theories of core domains," *Annual review of psychology*, vol. 43, no. 1, pp. 337–375, 1992.

[549] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, 1998.

[550] J. K. Hamlin and K. Wynn, "Young infants prefer prosocial to antisocial others," *Cognitive development*, vol. 26, no. 1, pp. 30–39, 2011.

[551] Y. Luo, "Three-month-old infants attribute goals to a non-human agent," *Developmental science*, vol. 14, no. 2, pp. 453–460, 2011.

[552] K. H. Onishi and R. Baillargeon, "Do 15-month-old infants understand false beliefs?," *Science*, vol. 308, no. 5719, pp. 255–258, 2005.

[553] J. W. Weibull, *Evolutionary game theory*. MIT press, 1997.

[554] C. Camerer, *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2003.

[555] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[556] T. Rusch, S. Steixner-Kumar, P. Doshi, M. Spezio, and J. Gläscher, "Theory of mind and decision science: towards a typology of tasks and computational models," *Neuropsychologia*, vol. 146, p. 107488, 2020.

[557] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.

[558] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing," *Nature Human Behaviour*, vol. 1, no. 4, p. 0064, 2017.

[559] L. Zhang and J. Gläscher, "A network supporting social influences in human decision-making," *bioRxiv*, p. 551614, 2019.

[560] E. D. Boorman, J. P. O'Doherty, R. Adolphs, and A. Rangel, "The behavioral and neural mechanisms underlying the tracking of expertise," *Neuron*, vol. 80, no. 6, pp. 1558–1571, 2013.

[561] S. Collette, W. M. Pauli, P. Bossaerts, and J. O'Doherty, "Neural computations underlying inverse reinforcement learning in the human brain," *Elife*, vol. 6, p. e29718, 2017.

[562] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[563] J.-J. Rousseau, *A discourse on inequality*. Penguin, 1984.

[564] P. Doshi, X. Qu, and A. Goodie, "Decision-theoretic planning in multi-agent settings with application to behavioral modeling," *Plan, Activity, and Intent Recognition: Theory and Practice*, pp. 205–224, 2014.

[565] J. Call and M. Tomasello, "Does the chimpanzee have a theory of mind? 30 years later," *Human Nature and Self Design*, pp. 83–96, 2011.

[566] B. F. d'Arc, M. Devaine, and J. Daunizeau, "A reverse turing-test for predicting social deficits in people with autism," *BioRxiv*, p. 414540, 2018.

[567] P. J. Gmytrasiewicz and E. H. Durfee, "Rational coordination in multi-agent environments," *Autonomous Agents and Multi-Agent Systems*, vol. 3, no. 4, pp. 319–350, 2000.

[568] P. J. Gmytrasiewicz and P. Doshi, "A framework for sequential planning in multi-agent settings," *Journal of Artificial Intelligence Research*, vol. 24, pp. 49–79, 2005.

[569] H. L. Gallagher and C. D. Frith, "Functional imaging of 'theory of mind'," *Trends in cognitive sciences*, vol. 7, no. 2, pp. 77–83, 2003.

[570] M. Schurz, J. Radua, M. Aichhorn, F. Richlan, and J. Perner, "Fractionating theory of mind: A meta-analysis of functional brain imaging studies," *Neuroscience & Biobehavioral Reviews*, vol. 42, pp. 9–34, 2014.

[571] R. Saxe and N. Kanwisher, "People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind"," *Neuroimage*, vol. 19, no. 4, pp. 1835–1842, 2003.

[572] N. Tang, S. Stacy, M. Zhao, G. Marquez, and T. Gao, "Bootstrapping an imagined we for cooperation," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020.

[573] S. Stacy, Q. Zhao, M. Kleiman-Weiner, and T. Gao, "Intuitive visual communication through physical-social commonsense," *Journal of Vision*, vol. 20, no. 11, pp. 1517–1517, 2020.

[574] K. Quennesson, E. Ioup, and C. L. Isbell Jr, "Wavelet statistics for human motion classification," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2006.

[575] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature, 2019.

[576] S. S. Sundar, "Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii)," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 74–88, 2020.

[577] G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, "Memor: A dataset for multimodal emotion reasoning in videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[578] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[579] N. Ajmeri, H. Guo, P. K. Murukannaiah, and M. P. Singh, "Elessar: Ethics in norm-aware agents," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020.

[580] I. Chalkidis and D. Kampas, "Deep learning in law: Early adaptation and legal word embeddings trained on large corpora," *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 171–198, 2019.

[581] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.

[582] X. Yongqiang, G. Baojiao, and G. Jianfeng, "The theory of thermodynamics for chemical reactions in dispersed heterogeneous systems," *Journal of colloid and interface science*, vol. 191, no. 1, pp. 81–85, 1997.

[583] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 3, pp. 115–135, 2016.

[584] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 2007.

[585] I. Bakker, T. Van der Voordt, P. Vink, and J. de Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Current Psychology*, vol. 33, no. 3, pp. 405–421, 2014.

[586] G. Karakurt and T. Cumbie, "The relationship between egalitarianism, dominance, and violence in intimate relationships," *Journal of Family Violence*, vol. 27, no. 2, pp. 115–122, 2012.

[587] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[588] J. Urbanek, A. Fan, S. Karamcheti, S. Jain, S. Humeau, E. Dinan, T. Rocktäschel, D. Kiela, A. Szlam, and J. Weston, "Learning to speak and act in a fantasy text adventure game," *arXiv preprint arXiv:1903.03094*, 2019.

[589] A. Adhikari, X. Yuan, M.-A. Côté, M. Zelinka, M.-A. Rondeau, R. Laroche, P. Poupart, J. Tang, A. Trischler, and W. L. Hamilton, "Learning dynamic knowledge graphs to generalize on text-based games," *arXiv preprint arXiv:2002.09127*, 2020.

[590] C. Zhang and V. Lesser, "Multi-agent learning with policy prediction," in *Twenty-fourth AAAI conference on artificial intelligence*, 2010.

[591] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2017.

[592] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[593] J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, "Learning with opponent-learning awareness," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

[594] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, *et al.*, "The hanabi challenge: A new frontier for ai research," *Artificial Intelligence*, vol. 280, p. 103216, 2020.

[595] S. V. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artificial Intelligence*, vol. 258, pp. 66–95, 2018.

[596] P. Doshi and P. J. Gmytrasiewicz, "A particle filtering based approach to approximating interactive pomdps," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

[597] P. Doshi and D. Perez, "Generalized point based value iteration for interactive pomdps," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[598] P. Doshi and P. J. Gmytrasiewicz, "Monte carlo sampling methods for approximating interactive pomdps," *Journal of Artificial Intelligence Research*, vol. 34, pp. 297–337, 2009.

[599] P. Doshi, Y. Zeng, and Q. Chen, "Graphical models for interactive pomdps: representations and solutions," *Autonomous Agents and Multi-Agent Systems*, vol. 18, no. 3, pp. 376–416, 2009.

[600] E. Sonu and P. Doshi, "Scalable solutions of interactive pomdps using generalized and bounded policy iteration," *Autonomous Agents and Multi-Agent Systems*, vol. 29, no. 3, pp. 455–494, 2015.

[601] A. Panella and P. Gmytrasiewicz, "Bayesian learning of other agents' finite controllers for interactive pomdps," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[602] Y. Han and P. Gmytrasiewicz, "Learning others' intentional models in multi-agent settings using interactive pomdps," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2018.

[603] T. N. Hoang and K. H. Low, "Interactive pomdp lite: Towards practical planning to predict and exploit intentions for interacting with self-interested agents," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[604] S. Sarkka and J. Hartikainen, "Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[605] R. S. Sutton, "Reinforcement learning: Past, present and future," in *Asia-Pacific Conference on Simulated Evolution and Learning*, 1998.

[606] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proceedings of International Conference on Machine Learning (ICML)*, 2009.

[607] H. R. Maei, C. Szepesvari, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton, "Convergent temporal-difference learning with arbitrary smooth function approximation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009.

[608] H. Seijen and R. Sutton, "True online td (lambda)," in *Proceedings of International Conference on Machine Learning (ICML)*, 2014.

[609] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[610] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, "Machine theory of mind," in *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

[611] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of International Conference on Machine Learning (ICML)*, 2014.

[612] F. B. Von Der Osten, M. Kirley, and T. Miller, "The minds of many: Opponent modeling in a stochastic game," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[613] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind?"," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[614] E. T. Chancey, J. P. Bliss, A. B. Proaps, and P. Madhavan, "The role of trust as a mediator between system characteristics and response behaviors," *Human factors*, vol. 57, no. 6, pp. 947–958, 2015.

[615] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[616] J. B. Lyons, M. A. Clark, A. R. Wagner, and M. J. Schuelke, "Certifiable trust in autonomous systems: Making the intractable tangible," *AI Magazine*, vol. 38, no. 3, 2017.

[617] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[618] Z. C. Lipton, "The mythos of model interpretability," in *Proceedings of International Conference on Machine Learning (ICML)*, 2016.

[619] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[620] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.

[621] S. Yang, Q. Gao, S. Saba-Sadiya, and J. Chai, "Commonsense justification for action explanation," in *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[622] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *34th International Conference on Machine Learning*, 2017.

[623] R. Ramprasaath, D. Abhishek, V. Ramakrishna, C. Michael, P. Devi, and B. Dhruv, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CVPR 2016*, 2016.

[624] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[625] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[626] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

[627] R. Hoffman, "A taxonomy of emergent trusting in the human–machine relationship," *Cognitive systems engineering: The future for a changing world*, 2017.

[628] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[629] Q. Zhang, Y. Nian Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[630] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[631] H. H. Clark and E. F. Schaefer, "Contributing to discourse," *Cognitive science*, vol. 13, no. 2, pp. 259–294, 1989.

[632] S. Devin and R. Alami, "An implemented theory of mind to improve human-robot shared plans execution," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[633] A. I. Goldman, *Theory of mind.* The Oxford handbook of philosophy of cognitive science, 2012.

[634] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Technical report, University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[635] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[636] R. Greiner, B. Poulin, P. Lu, J. Anvik, Z. Lu, C. Macdonell, D. Wishart, R. Eisner, and D. Szafron, "Explaining naive bayes classifications," tech. rep., Department of Computing Science, University of Alberta, 2003.

[637] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[638] H. Strobelt, S. Gehrmann, B. Huber, H. Pfister, and A. M. Rush, "Visual analysis of hidden state dynamics in recurrent neural networks," *arXiv preprint arXiv:1606.07461*, 2016.

[639] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[640] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *ICCV*, 2017.

[641] A. Datta, A. Datta, A. D. Procaccia, and Y. Zick, "Influence in classification via cooperative game theory," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[642] C. Brinton, "A framework for explanation of machine learning decisions," in *IJCAI-17 workshop on explainable AI (XAI)*, 2017.

[643] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[644] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8827–8836, 2018.

[645] A. Stone, H. Wang, M. Stark, Y. Liu, D. S. Phoenix, and D. George, "Teaching compositionality to cnns," *CVPR*, 2017.

[646] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

[647] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2018.

[648] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proceedings of International Conference on Machine Learning (ICML)*, 2019.

[649] D. J. Hilton, "Conversational processes and causal explanation," *Psychological Bulletin*, vol. 107, no. 1, p. 65, 1990.

[650] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[651] S. Park, B. X. Nie, and S.-C. Zhu, "Attribute and-or grammar for joint parsing of human attributes, part and pose," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1555–1569, 2018.

[652] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, no. 1, pp. 1–39, 1986.

[653] R. R. Hoffman, P. A. Hancock, and J. M. Bradshaw, "Metrics, metrics, metrics, part 2: Universal metrics?," *IEEE Intelligent Systems*, vol. 25, no. 6, pp. 93–97, 2010.

[654] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.

[655] C. Liu, S. Yang, S. Saba-Sadiya, N. Shukla, Y. He, S.-C. Zhu, and J. Chai, "Jointly learning grounded task structures from language instruction and visual demonstration," in *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[656] T. Wu and S.-C. Zhu, "A numerical study of the bottom-up and top-down inference processes in and-or graphs," *International journal of computer vision*, vol. 93, no. 2, pp. 226–252, 2011.

[657] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *IEEE MultiMedia*, vol. 21, no. 2, pp. 42–70, 2014.

[658] A. P. Witkin, "Scale-space filtering," in *Readings in Computer Vision*, pp. 329–332, Elsevier, 1987.

[659] L. Carlson, D. Marcu, and M. E. Okurowski, "Building a discourse-tagged corpus in the framework of rhetorical structure theory," in *Current and new directions in discourse and dialogue*, pp. 85–112, Springer, 2003.

[660] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[661] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," *Proceedings of the 5th International Conference on Learning Representations, ICLR*, 2017.

[662] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2000.

[663] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proceedings of British Machine Vision Conference (BMVC)*, 2010.

[664] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[665] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Science Robotics*, vol. 4, no. 37, 2019.

[666] K. Tu, M. Pavlovskaia, and S. C. Zhu, "Unsupervised structure learning of stochastic and-or grammars," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2013.

[667] M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, and S.-C. Zhu, "Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[668] T. Lombrozo, *Explanation and abductive inference*, ch. 14. Oxford University Press, 2012.

[669] J. E. Laird, *The Soar cognitive architecture*. MIT press, 2012.

[670] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[671] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[672] M.-F. Balcan and A. Blum, "An augmented pac model for semi-supervised learning," in *Semi-Supervised Learning*, MIT Press, 2006.

[673] A. Barbu, M. Pavlovskaia, and S. C. Zhu, "Rates for inductive learning of compositional models," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[674] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Localized rademacher complexities," in *International Conference on Computational Learning Theory*, 2002.

[675] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, 1989.

[676] S. A. Goldman and M. J. Kearns, "On the complexity of teaching," *Journal of Computer and System Sciences*, vol. 50, no. 1, pp. 20–31, 1995.

[677] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[678] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[679] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[680] A. C.-C. Yao, "Some complexity questions related to distributive computing," in *Proceedings of the eleventh annual ACM symposium on Theory of computing*, 1979.

[681] A. A. Sherstov, "Lower bounds in communication complexity and learning theory via analytic methods," in *UT Electronic Theses and Dissertations*, Citeseer, 2009.

[682] K. Tuyls, K. Tumer, and G. Weiss, "Multiagent learning," *Multiagent Systems*, p. 423, 2013.

[683] R. Fagin, J. Halpern, Y. Moses, and M. Vardi, "Reasoning about knowledge, paperback edn," 2004.

[684] A. Shinohara and S. Miyano, "Teachability in computational learning," *New Generation Computing*, vol. 8, no. 4, pp. 337–347, 1991.

[685] R. J. Aumann, "Agreeing to disagree," *The annals of statistics*, vol. 4, no. 6, pp. 1236–1239, 1976.

[686] A. Lazaridou, K. M. Hermann, K. Tuyls, and S. Clark, "Emergence of linguistic communication from referential games with symbolic and pixel input," in *International Conference on Learning Representations (ICLR)*, 2018.

[687] D. Lewis, *Convention: A philosophical study*. John Wiley & Sons, 2008.

[688] S. Havrylov and I. Titov, "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2017.

[689] C. J. C. H. Watkins, *Learning from delayed rewards*. PhD thesis, King's College, Cambridge United Kingdom, 1989.

[690] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[691] P. Shafto, N. D. Goodman, and T. L. Griffiths, "A rational account of pedagogical reasoning: Teaching by, and learning from, examples," *Cognitive psychology*, vol. 71, pp. 55–89, 2014.

[692] X. Zhu, "Machine teaching: An inverse problem to machine learning and an approach toward optimal education," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[693] S. Zilles, S. Lange, R. Holte, M. Zinkevich, and N. Cesa-Bianchi, "Models of cooperative teaching and learning," *Journal of Machine Learning Research*, vol. 12, no. 2, 2011.

[694] X. Chen, Y. Cheng, and B. Tang, "On the recursive teaching dimension of vc classes," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016.

[695] T. Doliwa, G. Fan, H. U. Simon, and S. Zilles, "Recursive teaching dimension, vc-dimension and sample compression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3107–3131, 2014.

[696] N. Shukla, Y. He, F. Chen, and S.-C. Zhu, "Learning human utility from video demonstrations for deductive planning in robotics," in *Conference on Robot Learning*, 2017.

[697] F. J. Balbach, "Measuring teachability using variants of the teaching dimension," *Theoretical Computer Science*, vol. 397, no. 1-3, pp. 94–113, 2008.

[698] M. K. Ho, M. Littman, J. MacGlashan, F. Cushman, and J. L. Austerweil, "Showing versus doing: Teaching by demonstration," *Advances in neural information processing systems*, vol. 29, pp. 3027–3035, 2016.

[699] A. Turing, "On computable numbers, with an application to the entscheidungs problem," *The essential Turing: seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life, plus the secrets of Enigma*, p. 58, 2004.

[700] J. F. Nash *et al.*, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[701] T. Roughgarden, "Algorithmic game theory," *Communications of the ACM*, vol. 53, no. 7, pp. 78–86, 2010.

[702] M. Kinney and C. Tsatsoulis, "Learning communication strategies in multiagent systems," *Applied intelligence*, vol. 9, no. 1, pp. 71–91, 1998.

[703] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016.

[704] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

[705] K. J. Holyoak, "Analogy and relational reasoning," in *The Oxford Handbook of Thinking and Reasoning*, pp. 234–259, Oxford University Press, 2012.

[706] J. C. e. a. Raven, "Raven's progressive matrices," *Western Psychological Services*, 1938.

[707] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu, "Raven: A dataset for relational and analogical visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[708] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[709] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[710] C. Jiang, S. Qi, Y. Zhu, S. Huang, J. Lin, L.-F. Yu, D. Terzopoulos, and S.-C. Zhu, "Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars," *International Journal of Computer Vision (IJCV)*, pp. 920–941, 2018.

[711] T. Feng, L.-F. Yu, S.-K. Yeung, K. Yin, and K. Zhou, "Crowd-driven mid-scale layout design," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 132–1, 2016.

[712] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "Minos: Multimodal indoor simulator for navigation in complex environments," *arXiv preprint arXiv:1712.03931*, 2017.

[713] S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. Courville, "Home: A household multimodal environment," *arXiv preprint arXiv:1711.11017*, 2017.

[714] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[715] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, "Building generalizable agents with a realistic and rich 3d environment," *arXiv preprint arXiv:1801.02209*, 2018.

[716] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[717] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[718] X. Xie, H. Liu, Z. Zhang, Y. Qiu, F. Gao, S. Qi, Y. Zhu, and S.-C. Zhu, "Vrgym: A virtual testbed for physical and interactive ai," in *Proceedings of the ACM TURC*, 2019.

[719] X. Gao, R. Gong, T. Shu, X. Xie, S. Wang, and S.-C. Zhu, "Vrkitchen: An interactive 3d virtual environment for task-oriented learning," *arXiv preprint arXiv:1903.05757*, 2019.

[720] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*, 2018.

[721] M. Gao, X. Wang, K. Wu, A. Pradhana, E. Sifakis, C. Yuksel, and C. Jiang, "Gpu optimization of material point methods," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, 2019.

[722] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer, "Elastically deformable models," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 4, pp. 205–214, 1987.

[723] D. Terzopoulos and K. Fleischer, "Modeling inelastic deformation: viscolelasticity, plasticity, fracture," *ACM Transactions on Graphics (TOG)*, vol. 22, no. 4, pp. 269–278, 1988.

[724] N. Foster and D. Metaxas, "Realistic animation of liquids," *Graphical models and image processing*, vol. 58, no. 5, pp. 471–483, 1996.

[725] J. Stam, "Stable fluids," in *ACM Transactions on Graphics (TOG)*, 1999.

[726] S. Blemker, J. Teran, E. Sifakis, R. Fedkiw, and S. Delp, "Fast 3d muscle simulations using a new quasistatic invertible finite-element algorithm," in *International Symposium on Computer Simulation in Biomechanics*, 2005.

[727] J. Hegemann, C. Jiang, C. Schroeder, and J. M. Teran, "A level set method for ductile fracture," in *Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2013.

[728] M. Li, M. Gao, T. Langlois, C. Jiang, and D. M. Kaufman, "Decomposed optimization time integrator for large-step elastodynamics," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, p. 70, 2019.

[729] Y. Wang, C. Jiang, C. Schroeder, and J. Teran, "An adaptive virtual node algorithm with robust mesh cutting," in *Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2014.

[730] W. K. Liu, S. Jun, and Y. F. Zhang, "Reproducing kernel particle methods," *International journal for numerical methods in fluids*, vol. 20, no. 8-9, pp. 1081–1106, 1995.

[731] S. Li and W. K. Liu, "Meshfree and particle methods and their applications," *Applied Mechanics Reviews*, vol. 55, no. 1, pp. 1–34, 2002.

[732] J. Donea, S. Giuliani, and J.-P. Halleux, "An arbitrary lagrangian-eulerian finite element method for transient dynamic fluid-structure interactions," *Computer methods in applied mechanics and engineering*, vol. 33, no. 1-3, pp. 689–723, 1982.

[733] J. U. Brackbill and H. M. Ruppel, "Flip: A method for adaptively zoned, particle-in-cell calculations of fluid flows in two dimensions," *Journal of Computational physics*, vol. 65, no. 2, pp. 314–343, 1986.

[734] D. Sulsky, Z. Chen, and H. L. Schreyer, "A particle method for history-dependent materials," *Computer methods in applied mechanics and engineering*, vol. 118, no. 1-2, pp. 179–196, 1994.

[735] A. Stomakhin, C. Schroeder, L. Chai, J. Teran, and A. Selle, "A material point method for snow simulation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 102, 2013.

[736] J. Gaume, T. Gast, J. Teran, A. van Herwijnen, and C. Jiang, "Dynamic anticrack propagation in snow," *Nature communications*, vol. 9, no. 1, p. 3047, 2018.

[737] D. Ram, T. Gast, C. Jiang, C. Schroeder, A. Stomakhin, J. Teran, and P. Kavehpour, "A material point method for viscoelastic fluids, foams and sponges," in *Proceedings of ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2015.

[738] Y. Yue, B. Smith, C. Batty, C. Zheng, and E. Grinspun, "Continuum foam: A material point method for shear-dependent flows," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 5, p. 160, 2015.

[739] Y. Fang, M. Li, M. Gao, and C. Jiang, "Silly rubber: an implicit material point method for simulating non-equilibrated viscoelastic and elastoplastic solids," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, p. 118, 2019.

[740] G. Daviet and F. Bertails-Descoubes, "A semi-implicit material point method for the continuum simulation of granular materials," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 102, 2016.

[741] Y. Hu, Y. Fang, Z. Ge, Z. Qu, Y. Zhu, A. Pradhana, and C. Jiang, "A moving least squares material point method with displacement discontinuity and two-way rigid body coupling," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 150, 2018.

[742] S. Wang, M. Ding, T. F. Gast, L. Zhu, S. Gagniere, C. Jiang, and J. M. Teran, "Simulation and visualization of ductile fracture with the material point method," *ACM Transactions on Graphics (TOG)*, vol. 2, no. 2, p. 18, 2019.

[743] J. Wolper, Y. Fang, M. Li, J. Lu, M. Gao, and C. Jiang, "Cd-mpm: continuum damage material point methods for dynamic fracture animation," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, p. 119, 2019.

[744] C. Jiang, T. Gast, and J. Teran, "Anisotropic elastoplasticity for cloth, knit and hair frictional contact," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 152, 2017.

[745] X. Han, T. F. Gast, Q. Guo, S. Wang, C. Jiang, and J. Teran, "A hybrid material point method for frictional contact with diverse materials," *ACM Transactions on Graphics (TOG)*, vol. 2, no. 2, p. 17, 2019.

[746] C. Fu, Q. Guo, T. Gast, C. Jiang, and J. Teran, "A polynomial particle-in-cell method," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 222, 2017.

[747] A. Stomakhin, C. Schroeder, C. Jiang, L. Chai, J. Teran, and A. Selle, "Augmented mpm for phase-change and varied materials," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 138, 2014.

[748] A. P. Tampubolon, T. Gast, G. Klár, C. Fu, J. Teran, C. Jiang, and K. Museth, "Multi-species simulation of porous sand and water mixtures," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 105, 2017.

[749] M. Gao, A. Pradhana, X. Han, Q. Guo, G. Kot, E. Sifakis, and C. Jiang, "Animating fluid sediment mixture in particle-laden flows," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 149, 2018.

[750] J. A. Nairn, "Material point method calculations with explicit cracks," *Computer Modeling in Engineering and Sciences*, vol. 4, no. 6, pp. 649–664, 2003.

[751] Z. Chen, L. Shen, Y.-W. Mai, and Y.-G. Shen, "A bifurcation-based decohesion model for simulating the transition from localization to decohesion with the mpm," *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, vol. 56, no. 5, pp. 908–930, 2005.

[752] H. Schreyer, D. Sulsky, and S.-J. Zhou, "Modeling delamination as a strong discontinuity with the material point method," *Computer Methods in Applied Mechanics and Engineering*, vol. 191, no. 23, pp. 2483–2507, 2002.

[753] D. Sulsky and H. L. Schreyer, "Axisymmetric form of the material point method with applications to upsetting and taylor impact problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 139, no. 1-4, pp. 409–429, 1996.

[754] P. Huang, X. Zhang, S. Ma, and H. Wang, "Shared memory openmp parallelization of explicit mpm and its application to hypervelocity impact," *CMES: Computer Modelling in Engineering & Sciences*, vol. 38, no. 2, pp. 119–148, 2008.

[755] W. Hu and Z. Chen, "Model-based simulation of the synergistic effects of blast and fragmentation on a concrete wall using the mpm," *International journal of impact engineering*, vol. 32, no. 12, pp. 2066–2096, 2006.

[756] A. R. York, D. Sulsky, and H. L. Schreyer, "Fluid–membrane interaction based on the material point method," *International Journal for Numerical Methods in Engineering*, vol. 48, no. 6, pp. 901–924, 2000.

[757] S. Bandara and K. Soga, "Coupling of soil deformation and pore fluid flow using material point method," *Computers and geotechnics*, vol. 63, pp. 199–214, 2015.

[758] J. E. Guilkey, J. B. Hoying, and J. A. Weiss, "Computational modeling of multicellular constructs with the material point method," *Journal of biomechanics*, vol. 39, no. 11, pp. 2074–2086, 2006.

[759] P. HUANG, *Material point method for metal and soil impact dynamics problems.* Tsinghua University, 2010.

[760] Y. Fang, Y. Hu, S.-M. Hu, and C. Jiang, "A temporally adaptive material point method with regional time stepping," in *Computer Graphics Forum*, 2018.

[761] S. Bardenhagen and E. Kober, "The generalized interpolation material point method," *Computer Modeling in Engineering and Sciences*, vol. 5, no. 6, pp. 477–496, 2004.

[762] M. Gao, A. P. Tampubolon, C. Jiang, and E. Sifakis, "An adaptive generalized interpolation material point method for simulating elastoplastic materials," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 223, 2017.

[763] A. Sadeghirad, R. M. Brannon, and J. Burghardt, "A convected particle domain interpolation technique to extend applicability of the material point method for problems involving massive deformations," *International Journal for numerical methods in Engineering*, vol. 86, no. 12, pp. 1435–1456, 2011.

[764] D. Z. Zhang, X. Ma, and P. T. Giguere, "Material point method enhanced by modified gradient of shape function," *Journal of Computational Physics*, vol. 230, no. 16, pp. 6379–6398, 2011.

[765] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, 2002.

[766] C. V. Goldman and S. Zilberstein, "Optimizing information exchange in cooperative multi-agent systems," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2003.

[767] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PloS one*, vol. 12, no. 4, p. e0172395, 2017.

[768] S. Sukhbaatar, R. Fergus, *et al.*, "Learning multiagent communication with backpropagation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016.

[769] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-agent cooperation and the emergence of (natural) language," in *International Conference on Learning Representations (ICLR)*, 2017.

[770] K. Evtimova, A. Drozdov, D. Kiela, and K. Cho, "Emergent language in a multi-modal, multi-step referential game," *arXiv preprint arXiv:1705.10369*, 2017.

[771] K. Wagner, J. A. Reggia, J. Uriagereka, and G. S. Wilkinson, "Progress in the simulation of emergent communication and language," *Adaptive Behavior*, vol. 11, no. 1, pp. 37–69, 2003.

[772] R. Ibsen-Jensen, J. Tkadlec, K. Chatterjee, and M. A. Nowak, "Language acquisition with communication between learners," *Journal of The Royal Society Interface*, vol. 15, no. 140, p. 20180073, 2018.

[773] L. Graesser, K. Cho, and D. Kiela, "Emergent linguistic phenomena in multi-agent communication games," *arXiv preprint arXiv:1901.08706*, 2019.

[774] E. Dupoux and P. Jacob, "Universal moral grammar: a critical appraisal," *Trends in Cognitive Sciences*, vol. 11, no. 9, pp. 373–378, 2007.

[775] J. Mikhail, *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press, 2011.

[776] M. Kleiman-Weiner, R. Saxe, and J. B. Tenenbaum, "Learning a commonsense moral theory," *cognition*, vol. 167, pp. 107–123, 2017.

[777] P. Blake, K. McAuliffe, J. Corbit, T. Callaghan, O. Barry, A. Bowie, L. Kleutsch, K. Kramer, E. Ross, H. Vongsachang, *et al.*, "The ontogeny of fairness in seven societies," *Nature*, vol. 528, no. 7581, p. 258, 2015.

[778] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath, "In search of homo economicus: Behavioral experiments in 15 small-scale societies," *American Economic Review*, vol. 91, no. 2, pp. 73–78, 2001.

[779] B. R. House, J. B. Silk, J. Henrich, H. C. Barrett, B. A. Scelza, A. H. Boyette, B. S. Hewlett, R. McElreath, and S. Laurence, "Ontogeny of prosocial behavior across diverse societies," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 110, no. 36, pp. 14586–14591, 2013.

[780] J. Graham, P. Meindl, E. Beall, K. M. Johnson, and L. Zhang, "Cultural differences in moral judgment and behavior, across and within societies," *Current Opinion in Psychology*, vol. 8, pp. 125–130, 2016.

[781] T. Hurka, *Virtue, vice, and value*. Oxford University Press, 2000.

[782] J. Rawls, *A theory of justice*. Harvard university press, 1971.

[783] J. Haidt, "The new synthesis in moral psychology," *Science*, vol. 316, no. 5827, pp. 998–1002, 2007.

[784] J. K. Hamlin, "Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core," *Current Directions in Psychological Science*, vol. 22, no. 3, pp. 186–193, 2013.

[785] R. Kim, M. Kleiman-Weiner, A. Abeliuk, E. Awad, S. Dsouza, J. B. Tenenbaum, and I. Rahwan, "A computational model of commonsense moral decision making," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[786] M. Kleiman-Weiner, T. Gerstenberg, S. Levine, and J. B. Tenenbaum, "Inference of intention and permissibility in moral decision making," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2015.

[787] K. J. Holyoak and P. Thagard, "The analogical mind," *American psychologist*, vol. 52, no. 1, p. 35, 1997.

[788] P. W. Cheng and M. J. Buehner, "Causal learning," in *The Oxford Handbook of Thinking and Reasoning*, pp. 210–233, Oxford University Press, 2012.

[789] M. B. Hesse, *Models and analogies in science*. Notre Dame University Press, 1966.

[790] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2013.

[791] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[792] P. A. Carpenter, M. A. Just, and P. Shell, "What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test," *Psychological review*, vol. 97, no. 3, p. 404, 1990.

[793] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[794] R. E Snow, P. Kyllonen, and B. Marshalek, "The topography of ability and learning correlations," *Advances in the psychology of human intelligence*, pp. 47–103, 1984.

[795] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig, "Improving fluid intelligence with training on working memory," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 105, no. 19, pp. 6829–6833, 2008.

[796] G. H. Bower, "A contrast effect in differential conditioning," *Journal of Experimental Psychology*, vol. 62, no. 2, p. 196, 1961.

[797] D. R. Meyer, "The effects of differential rewards on discrimination reversal learning by monkeys," *Journal of Experimental Psychology*, vol. 41, no. 4, p. 268, 1951.

[798] A. M. Schrier and H. F. Harlow, "Effect of amount of incentive on discrimination learning by monkeys," *Journal of comparative and physiological psychology*, vol. 49, no. 2, p. 117, 1956.

[799] R. M. Shapley and J. D. Victor, "The effect of contrast on the transfer properties of cat retinal ganglion cells," *The Journal of physiology*, vol. 285, no. 1, pp. 275–298, 1978.

[800] R. Lawson, "Brightness discrimination performance and secondary reward strength as a function of primary reward amount," *Journal of Comparative and Physiological Psychology*, vol. 50, no. 1, p. 35, 1957.

[801] A. Amsel, "Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension," *Psychological review*, vol. 69, no. 4, p. 306, 1962.

[802] J. J. Gibson and E. J. Gibson, "Perceptual learning: Differentiation or enrichment?," *Psychological review*, vol. 62, no. 1, p. 32, 1955.

[803] J. J. Gibson, *The ecological approach to visual perception: Classic edition*. Psychology Press, 2014.

[804] R. Catrambone and K. J. Holyoak, "Overcoming contextual limitations on problem-solving transfer," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 15, no. 6, p. 1147, 1989.

[805] D. Gentner and V. Gunn, "Structural alignment facilitates the noticing of differences," *Memory & Cognition*, vol. 29, no. 4, pp. 565–577, 2001.

[806] R. Hammer, G. Diesendruck, D. Weinshall, and S. Hochstein, "The development of category learning strategies: What makes the difference?," *Cognition*, vol. 112, no. 1, pp. 105–119, 2009.

[807] M. L. Gick and K. Paterson, "Do contrasting examples facilitate schema acquisition and analogical transfer?," *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 46, no. 4, p. 539, 1992.

[808] E. Haryu, M. Imai, and H. Okada, "Object similarity bootstraps young children to action-based verb extension," *Child Development*, vol. 82, no. 2, pp. 674–686, 2011.

[809] L. Smith and D. Gentner, "The role of difference-detection in learning contrastive categories," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2014.

[810] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive science*, vol. 7, no. 2, pp. 155–170, 1983.

[811] D. Gentner and A. B. Markman, "Structural alignment in comparison: No difference without similarity," *Psychological science*, vol. 5, no. 3, pp. 152–158, 1994.

[812] D. L. Schwartz, C. C. Chase, M. A. Oppezzo, and D. B. Chin, "Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer," *Journal of Educational Psychology*, vol. 103, no. 4, p. 759, 2011.

[813] C. Zhang, B. Jia, F. Gao, Y. Zhu, H. Lu, and S.-C. Zhu, "Learning perceptual inference by contrasting," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2019.

[814] S. Dehaene, *The number sense: How the mind creates mathematics*. OUP USA, 2011.

[815] W. Zhang, C. Zhang, Y. Zhu, and S.-C. Zhu, "Machine number sense: A dataset of visual arithmetic problems for abstract and relational reasoning," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020.