Evaluating Physical Quantities and Learning Human Utilities From RGBD Videos

Yixin Zhu[1*] Chenfanfu Jiang[2*] Yibiao Zhao[1] Demetri Terzopoulos[2] Song-Chun Zhu[1]

[1] UCLA Center for Vision, Cognition, Learning and Autonomy
[2] UCLA Computer Graphics & Vision Laboratory *

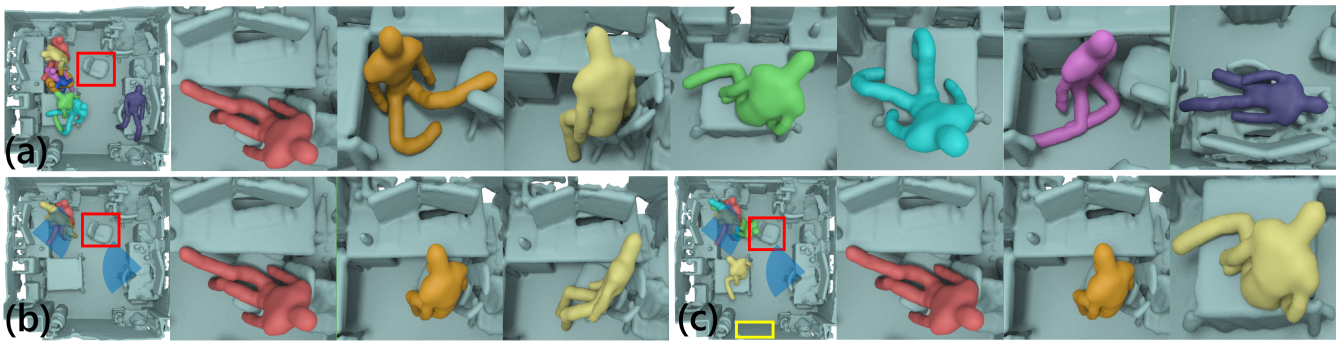

Figure 1: *(a) The top 7 human poses using physical quantities $\phi_p(\mathcal{G})$. The algorithm seeks physically comfortable sitting poses, resulting in casual sitting styles; e.g., lying on the desk. (b) Improved results after adding spatial features $\phi_s(\mathcal{G})$ to restrict the human-object relative orientations and distances. Further including temporal features $\phi_t(\mathcal{G})$ 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. The lack of chair arms leads to low scores.*

Abstract

We propose a notion of affordance that takes into account physical quantities generated when the human body interacts with real-world objects, and introduce a learning framework that incorporates the concept of human utilities, which in our opinion provides a deeper and finer-grained account not only of object affordance but also of people's interaction with objects. Rather than defining affordance in terms of the geometric compatibility between body poses and 3D objects, we devise algorithms that employ physics-based simulation to infer the relevant forces/pressures acting on body parts. By observing the choices people make in videos (particularly in selecting a chair in which to sit) our system learns the comfort intervals of the forces exerted on body parts (while sitting). We account for people's preferences in terms of human utilities, which transcend comfort intervals to account also for meaningful tasks within scenes and spatiotemporal constraints in motion planning, such as for the purposes of robot task planning.

Keywords: Forces, Physical Quantities, Human Utilities

Concepts: •**Computing methodologies** → **Physical simulation;** *Learning from demonstrations;*

1 Introduction

In recent years, there has been growing interest in studying object affordance in computer vision and graphics. We propose to go beyond visible *geometric compatibility* to infer, through physics-based simulation, the forces/pressures on various body parts as people interact with objects. By observing people's choices in videos—for example, in selecting a specific chair in a scene (Fig. 2)—we 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. We 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.

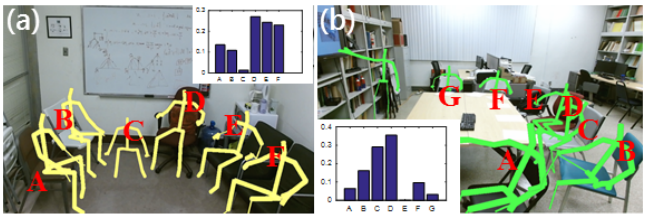

Figure 2: *Sitting activities in (a) an office and (b) a meeting room.*

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. 2(a)) or participating in a teleconference in a lab (Fig. 2(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, we demonstrate that human utilities can be used to analyze human activities, such as in the context of robot task planning. A longer version of this paper was reported in [Zhu et al. 2016].

2 Related Work

The concept of affordance was first introduced by [Gibson 1977]. Later, researchers incorporated affordance cues in shape recognition by observing people interacting with 3D scenes [Delaitre et al. 2012; Fouhey et al. 2014; Wei et al. 2013]. Adding geometric con-

*Y. Zhu and C. Jiang contributed equally to this work.

SA '16, December 05-08 2016, Macao
ISBN: 978-1-4503-4548-4/16/12
DOI: http://dx.doi.org/10.1145/2992138.2992144

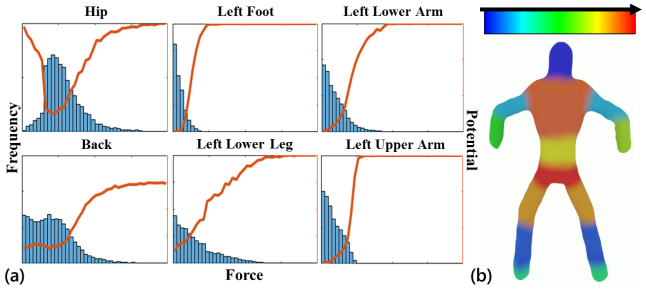

Figure 3: *(a) The final force histograms of 6 (out of 14) body parts. The x axis indicates the magnitudes of the forces, the y axis their frequencies and potential energy. (b) The average forces of each body part normalized and remapped to a T pose.*

straints, several researchers computed alignments of a small set of discrete poses [Grabner et al. 2011; Gupta et al. 2011; Jiang and Saxena 2013]. More recently, Savva et al. [Savva et al. 2014] predicted regions in 3D scenes where actions may take place. A closely related topic is to infer the stability and the supporting relations in a scene [Jia et al. 2013; Zheng et al. 2014; Liang et al. 2015].

3 Learning and Inferring Human Utilities

Extracting Features. We craft features $\boldsymbol{\phi}(\mathcal{G})$ of three types: (i) spatial features $\phi_s(\mathcal{G})$ encoding spatial relations, (ii) temporal features $\phi_t(\mathcal{G})$ associated with plan cost, and (iii) physical quantities $\phi_p(\mathcal{G})$ produced during human interactions with scenes. **Spatial features** $\phi_s(\mathcal{G})$ are defined as human-object / object-object relative distances and orientations. **Temporal features** $\phi_t(\mathcal{G})$ are defined as the plan cost from a given initial position to a goal position. **Physical quantities** $\phi_p(\mathcal{G})$ produced by people interacting with scenes are computed using the FEM [Gast et al. 2015].

Learning Human Utilities. The goal in the learning phase is to find the proper coefficient vector $\boldsymbol{\omega}$ of the feature space $\boldsymbol{\phi}(\mathcal{G})$ that best separates the positive examples of people interacting with scenes from the negative examples.

Under the rational choice assumption, we consider the *observed* rational person interacting with the scenes $\mathcal{G}^\star$ a positive example, and the *imagined* random configurations $\{\mathcal{G}_i\}$ as negative examples. Here, we formulate the learning phase as a ranking problem [Joachims 2002]—the *observed* rational person interaction $\mathcal{G}^\star$ should have lower cost than any *imagined* random configurations $\{\mathcal{G}_i\}$ with respect to the correct coefficient vector $\boldsymbol{\omega}$ of $\boldsymbol{\phi}(\mathcal{G})$.

Learning the ranking function is equivalent to finding the coefficient vector $\boldsymbol{\omega}$ such that the maximum number of the following inequalities are satisfied: $\langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^\star)\rangle > \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i)\rangle, \ \ \forall i \in \{1, 2, \cdots, n\}$, 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 [Hoffgen et al. 1995], we introduce non-negative slack variables ξ_i [Cortes and Vapnik 1995]: $\min \ \frac{1}{2}\langle \boldsymbol{\omega}, \boldsymbol{\omega}\rangle + \lambda \sum_i^n \xi_i^2, \forall i \in \{1, \cdots, n\}$, s.t. $\xi_i \geq 0, \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^\star)\rangle - \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i)\rangle > 1 - \xi_i^2$, 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 $\{\mathcal{G}_i\}$ in the solution space, the best configuration $\mathcal{G}^\star$ that receives the highest score: $\mathcal{G}^\star = \arg\max_{\mathcal{G}_i} \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i)\rangle$.

4 Conclusion

We have taken a step further from the current stream of studies on object affordance by inferring the invisible physical quantities and learning human utilities from 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 account for object affordance as well as for human behaviors.

References

CORTES, C., AND VAPNIK, V. 1995. Support-vector networks. *Machine learning 20*, 3, 273–297.

DELAITRE, V., FOUHEY, D. F., LAPTEV, I., SIVIC, J., GUPTA, A., AND EFROS, A. A. 2012. Scene semantics from long-term observation of people. In *Computer Vision–ECCV 2012*. Springer, 284–298.

FOUHEY, D. F., DELAITRE, V., GUPTA, A., EFROS, A. A., LAPTEV, I., AND SIVIC, J. 2014. People watching: Human actions as a cue for single view geometry. *International Journal of Computer Vision 110*, 3, 259–274.

GAST, T., SCHROEDER, C., STOMAKHIN, A., JIANG, C., AND TERAN, J. 2015. Optimization integrator for large time steps. *Visualization and Computer Graphics, IEEE Transactions on 21*, 10, 1103–1115.

GIBSON, J. J. 1977. The theory of affordances. *Hilldale, USA*.

GRABNER, H., GALL, J., AND VAN GOOL, L. 2011. What makes a chair a chair? In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, 1529–1536.

GUPTA, A., SATKIN, S., EFROS, A., HEBERT, M., ET AL. 2011. From 3D scene geometry to human workspace. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, 1961–1968.

HOFFGEN, K.-U., SIMON, H.-U., AND VANHORN, K. S. 1995. Robust trainability of single neurons. *Journal of Computer and System Sciences 50*, 1, 114–125.

JIA, Z., GALLAGHER, A., SAXENA, A., AND CHEN, T. 2013. 3D-based reasoning with blocks, support, and stability. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, 1–8.

JIANG, Y., AND SAXENA, A. 2013. Infinite latent conditional random fields for modeling environments through humans. In *Robotics: Science and Systems*.

JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *Knowledge Discovery and Data Mining, Proceedings of the ACM SIGKDD International Conference on*, 133–142.

LIANG, W., ZHAO, Y., ZHU, Y., AND ZHU, S.-C. 2015. Evaluating human cognition of containing relations with physical simulation. In *Proceedings of the 37th Annual Cognitive Science Conference (CogSci)*.

SAVVA, M., CHANG, A. X., HANRAHAN, P., FISHER, M., AND NIESNER, M. 2014. Scenegrok: Inferring action maps in 3D environments. *ACM Transactions on Graphics (TOG) 33*, 6, 212.

WEI, P., ZHAO, Y., ZHENG, N., AND ZHU, S.-C. 2013. Modeling 4D human-object interactions for event and object recognition. In *Computer Vision (ICCV), Proceedings of the IEEE International Conference on*, 3272–3279.

ZHENG, B., ZHAO, Y., YU, J. C., IKEUCHI, K., AND ZHU, S.-C. 2014. Detecting potential falling objects by inferring human action and natural disturbance. In *Robotics and Automation (ICRA), Proceedings of the IEEE International Conference on*, 3417–3424.

ZHU, Y., JIANG, C., ZHAO, Y., TERZOPOULOS, D., AND ZHU, S.-C. 2016. Inferring forces and learning human utilities from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3823–3833.