Inferring Forces and Learning Human Utilities From 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

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.

1. Introduction

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 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. [19] designed an "affordance detector" for chairs by fitting typical human sitting poses to 3D objects.

In this paper, we 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. 1)—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, providing a deeper and finer-grained account not only of object affordance but also of people's behaviors observed in videos.

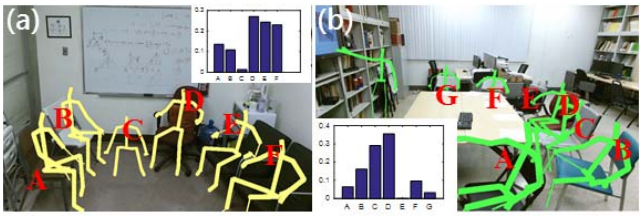

Figure 1. Examples of sitting activities in (a) an office and (b) a meeting room. In addition to geometry and appearance, people also consider other important factors including comfortability, reaching cost, and social goals when choosing a chair. The histograms indicate human preferences for different candidate chairs.

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. 1(a)) or participating in a teleconference in a lab (Fig. 1(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.

* Y. Zhu and C. Jiang contributed equally to this work.
Email: yixin.zhu@ucla.edu, cffjiang@cs.ucla.edu, ybz@mit.edu, dt@cs.ucla.edu, sczhu@stat.ucla.edu

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.415

IEEE computer society

1.1. Related Work

Modeling affordances: The concept of affordance was first introduced by Gibson [18]. Hermans et al. [24] and Fritz et al. [15] predicted action maps for autonomous robots. Later, researchers incorporated affordance cues in shape recognition by observing people interacting with 3D scenes [11, 14, 64]. Adding geometric constraints, several researchers computed alignments of a small set of discrete poses [19, 20, 32]. By searching a continuous pose parameter space of shapes, Kim et al. [37] obtained accurate alignments between shapes and human skeletons. More recently, Savva et al. [53] predicted regions in 3D scenes where actions may take place. Applications that use affordance in scene labeling and object placement are reported in [31, 30, 29]. A closely related topic is to infer the stability and the supporting relations in a scene [28, 70, 41].

Inferring forces from videos: For pose tracking, Brubaker et al. [5, 7, 6] estimate contact forces and internal joint torques using a mass-spring system. More recently, Zhu et al. and Pham et al. [72, 51] 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 [67, 63].

Task planning in robotics: Robotics has a rich history in seeking to understand human motion through synthesized trajectories. Hierarchical task planning through 2D human motion synthesis is explored in [73], but these models are constrained to 2D motion plans and relatively simplistic location-oriented goals. More complex models such as [36] seek to understand task-oriented human motion on a musculoskeletal level, but they do not take into account the context of an entire 3D environment. To synthesize logical trajectories, we rely on robust planning algorithms developed for robotics control applications (e.g. [17]) and we apply these forward planning engines to scene understanding by synthesizing rational human trajectories, a well-studied robotics problem [38].

Physics-based human simulation in graphics: Physics-based techniques for simulating deformable objects have been widely employed in computer graphics after the pioneering work on the topic [61, 60]. Popular methods for simulating elastoplastic material include mass-spring-damper systems [45, 62], the Finite Element Method (FEM) [59, 27, 44, 23], and the Material Point Method (MPM) [56, 57]. We adopt the FEM as it is physically accurate, robust, and computationally efficient. Among various deformable solids, the human body has received much attention due to its importance in character animation for movies and games. Significant prior work models human anatomical structure as a biomechanical musculoskeletal system including adipose tissues [39, 40, 54, 52]. For efficiency, our human body model is simply a single isotropic elastic body. This enables us to run a large number of simulations in a reasonable time limit and still achieve useful results.

1.2. Contributions

This paper makes five major contributions:

1. We 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 paper 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.
2. 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.
3. By incorporating a conventional robotics path planner, our proposed framework can generalize a static sitting pose to extend over a *dynamic* moving sequence.
4. 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.
5. We propose a method to robustly generate *volumetric* human models from the widely-used stick-man models acquired using Kinect sensors [50], 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.

1.3. Overview

The remainder of this paper is organized as follows: In Sec. 2, we introduce our representation, which incorporates physical quantities into the spatiotemporal spaces of interest. In Sec. 3, we describe the pipeline for calculating the relevant physical quantities, which makes use of the Finite Element Method (FEM). In Sec. 4, we formulate the problem as a ranking task, and introduce a learning and inference algorithm under the assumption of rational choice. Sec. 5 demonstrates that our proposed framework can be easily generalized to challenging new situations. Sec. 6 concludes the paper by discussing limitations and future work.

2. Representation

2.1. Spatial Entities and Relations in 3D Spaces

We represent sitting behaviors and associated relations in a parse graph $\mathcal{G}$, which includes (i) spatial entities—objects

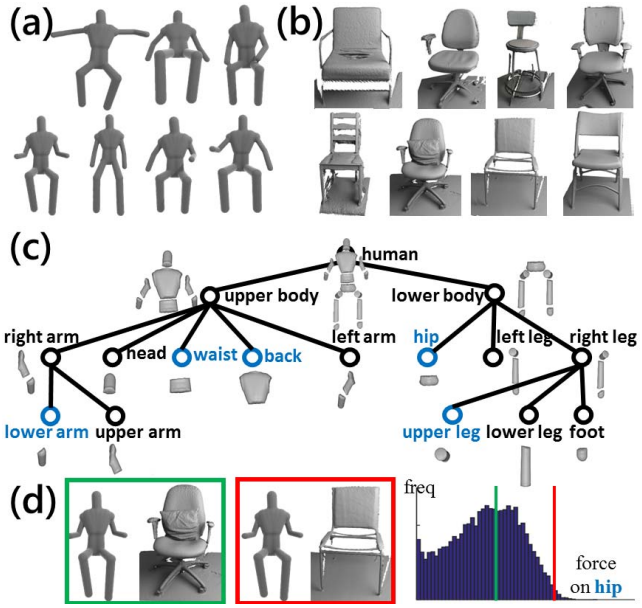

Figure 2. (a) We collect a set of human poses and cluster them into 7 average poses. (b) Various chairs extracted from scanned scenes. (c) Each human pose is decomposed into 14 body parts. When a human interacts with a chair, we infer the forces on each body part using FEM simulations. (d) Examples illustrating human preferences; green indicates a comfortable sitting activity, red an uncomfortable one.

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 $\mathcal{G}$ 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. 2(b)). In this paper 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. 2(a)). For each average pose, we first convert the Kinect stick-man models (Fig. 3(a)) into tetrahedralized human models (Fig. 3(b)). These are then discretized into 14 pre-defined human body parts (Fig. 3(c)) for simulations, as shown in Fig. 3(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. Figs. 6(d)(e) show an example of spatial relations. For the purposes of this paper, we define these two spatial relations as spatial features $\phi_s(\mathcal{G})$ that encode the relative spatial distances and orientations. At a higher level, human-object relations also encode visual attention and social goals.

2.2. Physical Quantities of Human Utilities

To date, researchers have mostly generated affordance maps by evaluating the geometric compatibility between people and objects [37, 30, 14, 29, 53, 64]. We employ a more meaningful and quantifiable metric—forces (including pressures) as physical quantities $\phi_p(\mathcal{G})$ 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. 2(d)).

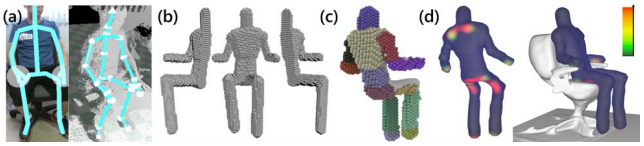

Figure 3. The stick-man model (a) captured using a Kinect is converted into a tetrahedralized human model (b) and then segmented into 14 body parts (c). Using FEM simulation the physical quantities $\phi_p(\mathcal{G})$ are estimated at each vertex of the FEM mesh; the forces at each vertex are visualized in (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. 3(d). In this paper, 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.

2.3. Human Utilities in Time

To model the human utility, a plan cost $\phi_t(\mathcal{G})$ 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 [35] 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.

3. Estimating the Forces in 3D Scenes

3.1. 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., [9, 21, 66, 53]) 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

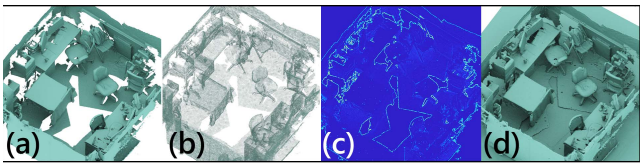

Figure 4. From a reconstructed 3D indoor scene (a) [9, 53], we uniformly sample vertices in the input mesh with Poisson disk sampling (b) [4], then convert them into a watertight mesh [43, 47] with well-defined interior and exterior regions. Differences (c) between the input mesh and the converted watertight mesh. By adding a ground geometry, we obtain a detailed, watertight reconstruction (d) of the 3D scene, which is inputted to the simulation.

objects and human poses, as shown in Fig. 2(a)(b), as well as more challenging and cluttered scenes.

3.2. Reconstructing Watertight Scenes

Reconstructing closed-loop scenes: Reconstruction methods that use purely geometric registration [48, 34, 49, 65] 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 [9], resulting in detailed reconstructions with loop closures, as shown in Fig. 4(a).

Converting to watertight scenes: Collision detection and resolution in the simulation requires a watertight scene mesh. We first use Poisson disk sampling [4] to generate uniformly distributed vertices from the input triangle mesh, as illustrated in Fig. 4(b). Each vertex is then replaced with a fixed-radius sphere level set [47]. 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 [43] is applied to the level set in order to generate the watertight surface, as shown in Fig.4(d). The resulting scene has the well-defined interior and exterior regions required by the simulation.

3.3. 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 [26]; i.e., finding the optimal rotation and translation that maps one collection of vertices to another in a least squares sense:

$$\min \sum_i ||\mathbf{R}\mathbf{A}(:,i) + \mathbf{t} - \mathbf{B}(:,i)||^2, \qquad (1)$$

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 [8, 58, 12] is then applied to cluster the resting poses into 7 categories, as shown in Fig. 2(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 [47]; its zero isocontour approximates the skin [43].

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 [46]. This results in a tetrahedralized human shape geometry as shown in Fig. 3(b).

3.4. 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 $\phi_p(\mathcal{G})$; e.g., forces and pressures.

Elasticity: The human body is modeled as an elastic material. The total elastic potential energy is defined as

$$\Phi^E(\mathbf{x}) = \int_\Omega \Psi^E(\mathbf{x}) d\mathbf{x} \approx \sum_e V_e^0 \Psi^E(\mathbf{F}(\mathbf{x})), \qquad (2)$$

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 [55] 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:

$$\Phi^C(\mathbf{x}_i) = \frac{1}{2} k_c (\mathbf{x}_i - \mathcal{P}(\mathbf{x}_i))^2, \qquad (3)$$

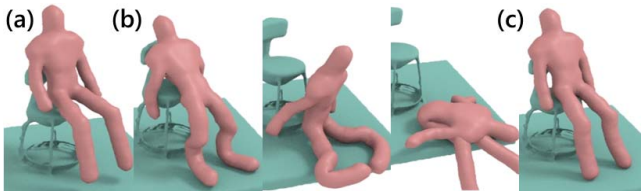

Figure 5. Given an initial human pose in a 3D scene subject to gravity (a), without adequate damping (b) the human body is too energetic and produces unnaturally bouncy motion. With proper damping, the simulation converges to a physically stable rest pose (c) in a small number of timesteps.

Table 1. Physical simulation parameters

Timestep: $1 \times 10^{-3} s$	Density: $1000 kg/m^3$	Young's modulus: $0.15 kPa$	Poisson's ratio: 0.3
Collision stiffness: $1 \times 10^{4} kg/s^2$	Friction coeff: 1×10^{-3}	Damping coeff: $50 kg/s$	Gravity: $9.81 m/s^2$

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

$$\mathbf{M}\frac{\mathbf{v}^{n+1}-\mathbf{v}^{n}}{\Delta t} = \mathbf{f}(\mathbf{x}^{n+1}, \mathbf{v}^{n+1}) + \mathbf{M}g, \quad (4)$$

$$\mathbf{f}(\mathbf{x}^{n+1}, \mathbf{v}^{n+1}) = \mathbf{f}^{E}(\mathbf{x}^{n+1}) + \mathbf{f}^{C}(\mathbf{x}^{n+1}) + \mathbf{f}^{D}(\mathbf{v}^{n+1}), \quad (5)$$

$$\mathbf{x}^{n+1} - \mathbf{x}^{n} = \mathbf{v}^{n+1}\Delta t. \quad (6)$$

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, $g = 9.8 m/s$ is gravity, and $\mathbf{f}^D = -\nu \mathbf{v}$ is an additional force to dampen the velocities, where ν is the damping coefficient. Fig. 5(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 [16].

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.

4. Learning and Inferring Human Utilities

4.1. 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.

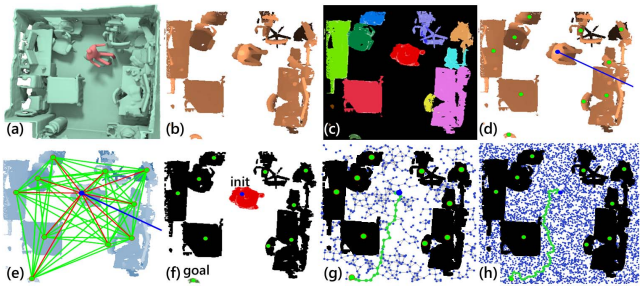

Figure 6. **Data pre-processing.** Given a reconstructed 3D scene (a), we project it down onto a planar map (b), and segment 3D objects from the scene (c). (d) visualizes 3D object positions (green dots), human head position (blue dot), and orientation (blue line). (e) **Spatial features** $\phi_s(\mathcal{G})$ are defined as human-object (red lines) and object-object (green lines) relative distances and orientations. (f) **Temporal features** $\phi_t(\mathcal{G})$ 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).

Data pre-processing is illustrated in Fig. 6(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 [1] and then grouped into object segments as in [68, 69]. 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 $\phi_s(\mathcal{G})$ are defined as human-object / object-object relative distances and orientations as shown in Fig. 6(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 $\phi_t(\mathcal{G})$ 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 [13]. Fig. 6(f)–(h) show two solutions using different numbers of nodes in the planner graph.

Physical quantities $\phi_p(\mathcal{G})$ 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. 2(c). The net force on each body part is obtained by summing up the forces at all its vertices. The net force di-

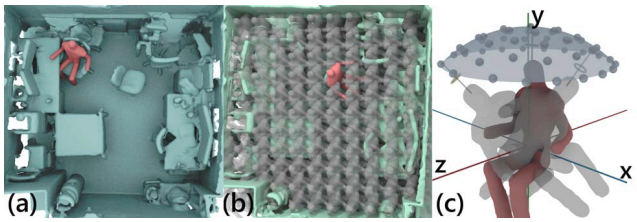

Figure 7. In the learning phase, based on rational choice theory, we assume that the observed demonstration is optimal, and therefore regard it a positive example. (a) In this example, a person is sitting on an armchair facing a desk with a monitor. The learning algorithm then imagines different configurations $\{\mathcal{G}_i\}$ 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 $\{\mathcal{G}_i\}$ are regarded negative examples. In the inference phase, the inference algorithm performs the same sampling process (b)(c), and finds the optimal configuration $\mathcal{G}^\star$ with the highest score.

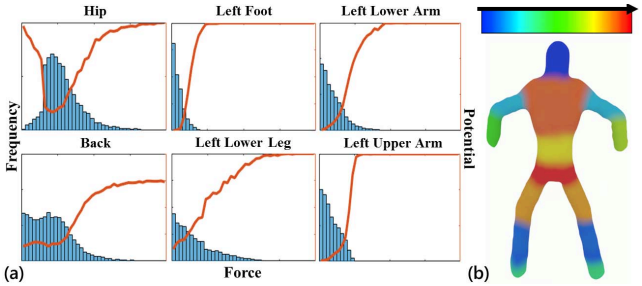

Figure 8. (a) The final force histograms of 6 (out of 14) body parts. The x axis indicates the magnitudes of the forces, the y axis their frequencies and potential energy. Histogram areas reflect the number of cases with non-zero forces. (b) The average forces of each body part normalized and remapped to a T pose.

vided by the number of contributing vertices yields the local pressure. Fig. 3(d) illustrates a force heatmap for sitting.

4.2. 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 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 [2, 3, 22, 42]. 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 $\phi_s(\mathcal{G})$, (ii) the plan cost from the current position to a goal position $\phi_t(\mathcal{G})$, and (iii) the physical quantities $\phi_p(\mathcal{G})$ 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 $\mathcal{G}^\star$ is assumed to be optimal; hence, this is regarded a positive example. If we *imagine* the same person making random choices $\{\mathcal{G}_i\}$ by randomly sitting on other objects (e.g., the ground), the rational choice assumption implies that the costs of the imagined configurations $\{\mathcal{G}_i\}$ 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 $\mathcal{G}^\star$ a positive example, and the *imagined* random configurations $\{\mathcal{G}_i\}$ as negative examples. However, the random generated configurations $\{\mathcal{G}_i\}$ may be similar or even identical to the observed optimal configuration $\mathcal{G}^\star$. 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 [33]—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 $\boldsymbol{\phi}(\mathcal{G})$, which includes spatial relations $\phi_s(\mathcal{G})$, plan cost $\phi_t(\mathcal{G})$, and physical quantities $\phi_p(\mathcal{G})$. Each coefficient ω_i reflects the importance of its corresponding feature. The ranking function is defined as

$$R(\mathcal{G}) = \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}) \rangle. \tag{7}$$

Learning the ranking function is equivalent to finding the coefficient vector $\boldsymbol{\omega}$ such that the maximum number of the following inequalities are satisfied:

$$\langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^\star) \rangle > \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle, \ \ \forall i \in \{1, 2, \cdots, n\}, \tag{8}$$

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 [25], we introduce non-negative slack variables ξ_i [10]:

$$\min \frac{1}{2}\langle \boldsymbol{\omega}, \boldsymbol{\omega} \rangle + \lambda \sum_i^n \xi_i^2, \ \ \forall i \in \{1, \cdots, n\} \tag{9}$$

$$\text{s.t. } \xi_i \geq 0, \ \ \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^\star) \rangle - \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle > 1 - \xi_i^2, \tag{10}$$

where λ is the trade-off parameter between maximizing the margin and satisfying the pairwise relative constraints.

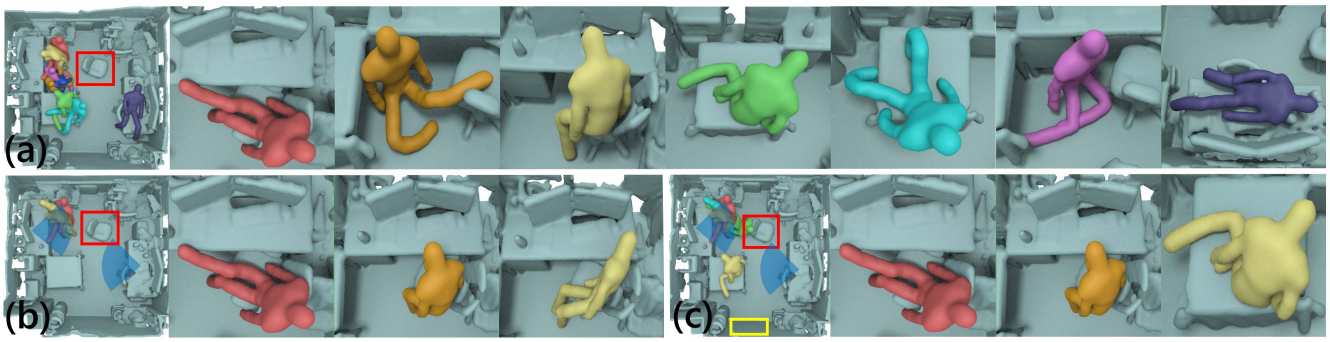

Figure 9. (a) The top 7 human poses using physical quantities $\phi_p(\mathcal{G})$. The algorithm seeks physically comfortable sitting poses, resulting in casual sitting styles; e.g., lying on the desk. (b) Improved results after adding spatial features $\phi_s(\mathcal{G})$ to restrict the human-object 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 (Fig. 7(a)). The lack of chair arms leads to low scores.

4.3. 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:

$$\mathcal{G}^\star = \arg\max_{\mathcal{G}_i} \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle. \tag{11}$$

4.4. Sampling the Solution Space

Without observing a human interacting with the scenes, the inference algorithm must sample the solution space by imagining different configurations $\{\mathcal{G}_i\}$. 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. 2(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(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 Sec. 4.1 are then used to extract the features $\boldsymbol{\phi}(\mathcal{G}_i)$.

In the learning phase, the $\boldsymbol{\phi}(\mathcal{G}_i)$ are then used to learn the ranking function(7). In the inference phase, the extracted features are then evaluated by (11). The configuration with the highest score is taken as the optimal configuration $\mathcal{G}^\star$.

5. Experiments

5.1. 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(a). The observed demonstrations were then used as positive training examples. For each 3D scene, we further generated over 4,000 different configurations $\mathcal{G}_i$ by enumerating all poses and randomly sampling different initial human translations and rotations in the solution space, as shown in Fig. 7(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 (7) learned the coefficient vector $\boldsymbol{\omega}$ of the ranking function under three different settings: (i) physical quantities $\phi_p(\mathcal{G})$, (ii) with additional spatial relations $\phi_s(\mathcal{G})$, and (iii) with all features $\phi_p(\mathcal{G})$, $\phi_s(\mathcal{G})$, and $\phi_t(\mathcal{G})$.

Fig. 8(a) shows the final force histograms of 6 (out of 14) body parts. Unsurprisingly when 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. 8(b).

5.2. Inferring Optimal Affordance in Static Scenes

Next, we tested the learned models on our dataset as well as on prior 3D datasets [53, 9] 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. 9 shows examples of the top ranked human poses. Although using physical quantities $\phi_p(\mathcal{G})$ produced physically plausible sitting poses (Fig. 9(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 $\phi_s(\mathcal{G})$, the relative orientations and distances between the human model and objects in the scene, improved the results, as shown in Fig. 9(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. 9,

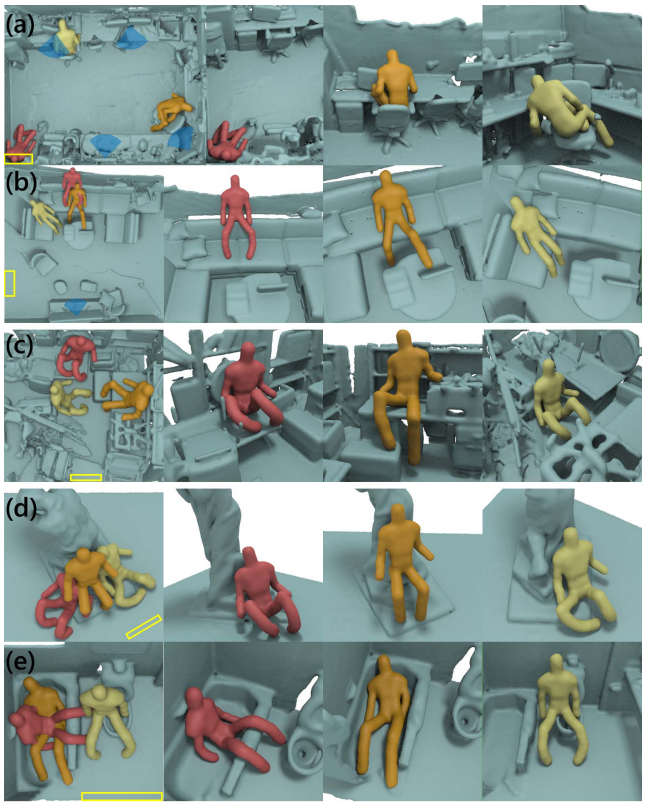

Figure 10. Top 3 poses in (a)(b) canonical scenarios, (c) cluttered scenarios, and (d)(e) novel scenarios. All the features $\boldsymbol{\phi}(\mathcal{G})$ are used in (a) and (b). Both physical quantities $\phi_p(\mathcal{G})$ and plan costs $\phi_t(\mathcal{G})$ are used in (c)–(e). The initial position for the path planner is indicated by the yellow bounding box.

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 $\phi_t(\mathcal{G})$ 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 [53, 66, 9, 71] are shown in Fig. 10.

Evaluations: We asked 4 subjects to rank the highest-scored sitting poses. Fig. 11 plots the correlations between their rankings and our system's output.

6. Discussion and Future Work

The current stream of studies on object affordance [11, 14, 64, 19, 20, 32, 37, 53, 72] 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.

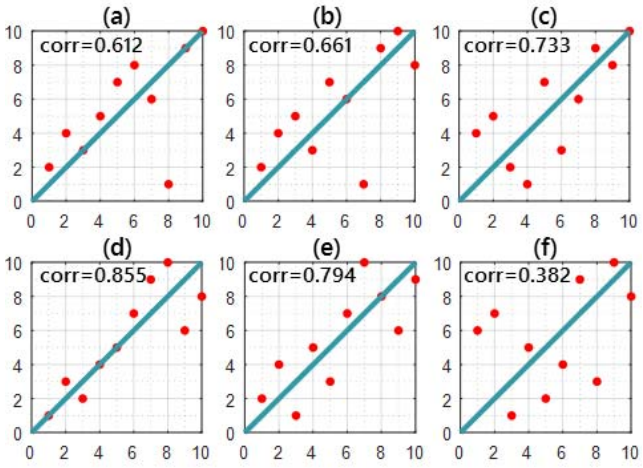

Figure 11. Correlations of the ranking by human subjects (x-axis) and our system's output (y-axis). The closer the plotted points fall to the diagonal lines the better our proposed method matches the performance of the human subjects. Plots (a)–(e) correspond to Fig. 10(a)–(e). Plot (f) corresponds to Fig. 9(c).

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., [39, 54]) 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.

Acknowledgments

We thank Steven Holtzen, Siyuan Qi, Mark Edmonds, and Nishant Shukla for proofreading drafts and assistance with reviewing related work, and Chuyuan Fu for video voice overs. We also thank Professor Joseph Teran of the UCLA Math Department for useful discussions. The work reported herein was supported by DARPA SIMPLEX grant N66001-15-C-4035, ONR MURI grant N00014-16-1-2007, and DoD CDMRP grant W81XWH-15-1-0147.

References

[1] M. Attene, B. Falcidieno, and M. Spagnuolo. Hierarchical mesh segmentation based on fitting primitives. *The Visual Computer*, 22(3):181–193, 2006.

[2] G. S. Becker. Crime and punishment: An economic approach. In *Essays in the Economics of Crime and Punishment*, pages 1–54. NBER, 1974.

[3] L. E. Blume and D. Easley. Rationality. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.

[4] R. Bridson. Fast Poisson disk sampling in arbitrary dimensions. In *ACM SIGGRAPH 2007 Sketches*, SIGGRAPH '07, 2007.

[5] M. A. Brubaker and D. J. Fleet. The kneed walker for human pose tracking. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 1–8, 2008.

[6] M. A. Brubaker, D. J. Fleet, and A. Hertzmann. Physics-based person tracking using the anthropomorphic walker. *International Journal of Computer Vision*, 87(1-2):140–155, 2010.

[7] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating contact dynamics. In *Computer Vision (ICCV), Proceedings of the IEEE International Conference on*, pages 2389–2396, 2009.

[8] S. Calinon, F. Guenter, and A. Billard. On learning, representing, and generalizing a task in a humanoid robot. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(2):286–298, 2007.

[9] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 5556–5565, 2015.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. In *Computer Vision–ECCV 2012*, pages 284–298. Springer, 2012.

[12] 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*, pages 1–38, 1977.

[13] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[14] 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*, 110(3):259–274, 2014.

[15] G. Fritz, L. Paletta, R. Breithaupt, E. Rome, and G. Dorffner. Learning predictive features in affordance based robotic perception systems. In *Intelligent Robots and Systems (IROS), Proceedings of the IEEE/RSJ International Conference on*, pages 3642–3647, 2006.

[16] T. Gast, C. Schroeder, A. Stomakhin, C. Jiang, and J. Teran. Optimization integrator for large time steps. *Visualization and Computer Graphics, IEEE Transactions on*, 21(10):1103–1115, 2015.

[17] R. Geraerts and M. H. Overmars. A comparative study of probabilistic roadmap planners. In *Algorithmic Foundations of Robotics V*, pages 43–57. Springer, 2004.

[18] J. J. Gibson. The theory of affordances. *Hilldale, USA*, 1977.

[19] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 1529–1536, 2011.

[20] A. Gupta, S. Satkin, A. Efros, M. Hebert, et al. From 3D scene geometry to human workspace. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 1961–1968, 2011.

[21] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *Robotics and Automation (ICRA), IEEE International Conference on*, pages 1524–1531, 2014.

[22] P. Hedström and C. Stern. Rational choice and sociology. In *The New Palgrave Dictionary of Economics*, pages 872–877. Palgrave Macmillan Basingstoke, UK, 2008.

[23] J. Hegemann, C. Jiang, C. Schroeder, and J. M. Teran. A level set method for ductile fracture. In *Proceedings of the ACM SIGGRAPH/EG Symposium on Computer Animation*, pages 193–201, 2013.

[24] T. Hermans, J. M. Rehg, and A. Bobick. Affordance prediction via learned object attributes. In *IEEE International Conference on Robotics and Automation (ICRA) Workshop on Semantic Perception, Mapping, and Exploration*. Citeseer, 2011.

[25] K.-U. Hoffgen, H.-U. Simon, and K. S. Vanhorn. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50(1):114–125, 1995.

[26] B. K. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987.

[27] G. Irving, J. Teran, and R. Fedkiw. Invertible finite elements for robust simulation of large deformation. In *Proceedings of the ACM SIGGRAPH/EG Symposium on Computer Animation*, pages 131–140, 2004.

[28] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3D-based reasoning with blocks, support, and stability. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 1–8, 2013.

[29] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3D scenes. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 2993–3000, 2013.

[30] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3D scenes using human context. *arXiv preprint arXiv:1206.6462*, 2012.

[31] Y. Jiang and A. Saxena. Hallucinating humans for learning robotic placement of objects. In *Experimental Robotics*, pages 921–937. Springer, 2013.

[32] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. In *Robotics: Science and Systems*, 2013.

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

[34] O. Kahler, V. Prisacariu, C. 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*, 21:1241–1250, 2015.

[35] L. E. Kavraki, P. Švestka, J.-C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *Robotics and Automation, IEEE Transactions on*, 12(4):566–580, 1996.

[36] O. Khatib, E. Demircan, V. De Sapio, L. Sentis, T. Besier, and S. Delp. Robotics-based synthesis of human motion. *Journal of Physiology Paris*, 103(3):211–219, 2009.

[37] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Transactions on Graphics (TOG)*, 33(4):120, 2014.

[38] S. M. LaValle. *Planning algorithms*. Cambridge University Press, 2006.

[39] S.-H. Lee, E. Sifakis, and D. Terzopoulos. Comprehensive biomechanical modeling and simulation of the upper body. *ACM Transactions on Graphics (TOG)*, 28(4):99:1–99:17, 2009.

[40] Y. Lee, M. S. Park, T. Kwon, and J. Lee. Locomotion control for many-muscle humanoids. *ACM Transactions on Graphics (TOG)*, 33(6):218:1–218:11, 2014.

[41] W. Liang, Y. Zhao, Y. Zhu, and S.-C. Zhu. Evaluating human cognition of containing relations with physical simulation. In *Proceedings of the 37th Annual Cognitive Science Conference (CogSci)*, 2015.

[42] S. Lohmann. Rational choice and political science. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.

[43] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics (Proceedings of ACM SIGGRAPH)*, 21(4):163–169, 1987.

[44] A. McAdams, Y. Zhu, A. Selle, M. Empey, R. Tamstorf, J. Teran, and E. Sifakis. Efficient elasticity for character skinning with contact and collisions. *ACM Transactions on Graphics (TOG)*, 30(4):37:1–37:12, 2011.

[45] G. S. Miller. The motion dynamics of snakes and worms. *Computer Graphics (Proceedings of ACM SIGGRAPH)*, 22(4):169–173, 1988.

[46] N. Molino, R. Bridson, J. Teran, and R. Fedkiw. A crystalline, red green strategy for meshing highly deformable objects with tetrahedra. In *Proceedings of the 12th International Meshing Roundtable*, pages 103–114, 2003.

[47] 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 *ACM SIGGRAPH 2013 Courses*, page 19, 2013.

[48] 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 *Mixed and Augmented Reality (ISMAR), Proceedings of the IEEE International Symposium on*, pages 127–136, 2011.

[49] M. Niesner, M. Zollhofer, S. Izadi, and M. Stamminger. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 32(6):169, 2013.

[50] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3D tracking of hand articulations using kinect. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 1, page 3, 2011.

[51] 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 *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, 2015.

[52] S. Saito, Z.-Y. Zhou, and L. Kavan. Computational bodybuilding: Anatomically-based modeling of human bodies. *ACM Transactions on Graphics (TOG)*, 34(4), 2015.

[53] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Niesner. Scenegrok: Inferring action maps in 3D environments. *ACM Transactions on Graphics (TOG)*, 33(6):212, 2014.

[54] W. Si, S.-H. Lee, E. Sifakis, and D. Terzopoulos. Realistic biomechanical simulation and control of human swimming. *ACM Transactions on Graphics (TOG)*, 34(1):10:1–15, Nov. 2014.

[55] A. Stomakhin, R. Howes, C. Schroeder, and J. M. Teran. Energetically consistent invertible elasticity. In *Proceedings of the ACM SIGGRAPH/EG Symposium on Computer Animation*, SCA '12, pages 25–32, 2012.

[56] A. Stomakhin, C. Schroeder, L. Chai, J. Teran, and A. Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4):102:1–102:10, 2013.

[57] 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)*, 33(4):138:1–138:11, 2014.

[58] C. Sylvain. *Robot programming by demonstration: A probabilistic approach*. EPFL Press, 2009.

[59] J. Teran, S. Blemker, V. N. T. Hing, and R. Fedkiw. Finite volume methods for the simulation of skeletal muscle. In *Proceedings of the ACM SIGGRAPH/EG Symposium on Computer Animation*, pages 68–74, 2003.

[60] D. Terzopoulos and K. Fleischer. Deformable models. *The Visual Computer*, 4(6):306–331, 1988.

[61] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer. Elastically deformable models. *Computer Graphics (Proceedings of ACM SIGGRAPH)*, 21(4):205–214, 1987.

[62] X. Tu and D. Terzopoulos. Artificial fishes: Physics, locomotion, perception, behavior. In *Computer Graphics (Proceedings of ACM SIGGRAPH)*, pages 43–50, 1994.

[63] 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)*, 32(4):43, 2013.

[64] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4D human-object interactions for event and object recognition. In *Computer Vision (ICCV), Proceedings of the IEEE International Conference on*, pages 3272–3279, 2013.

[65] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended kinectfusion. Technical Report MIT-CSAIL-TR-2012-020, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, 2012.

[66] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *Computer Vision (ICCV), Proceedings of the IEEE International Conference on*, pages 1625–1632, 2013.

[67] W. Zhao, J. Zhang, J. Min, and J. Chai. Robust realtime physics-based motion control for human grasping. *ACM Transactions on Graphics (TOG)*, 32(6):207, 2013.

[68] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, and S.-C. Zhu. Scene understanding by reasoning stability and safety. *International Journal of Computer Vision*, 112(2):221–238, 2015.

[69] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 3127–3134, 2013.

[70] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Detecting potential falling objects by inferring human action and natural disturbance. In *Robotics and Automation (ICRA), Proceedings of the IEEE International Conference on*, pages 3417–3424, 2014.

[71] Q.-Y. Zhou, S. Miller, and V. Koltun. Elastic fragments for dense scene reconstruction. In *Computer Vision (ICCV), Proceedings of the IEEE International Conference on*, pages 473–480, 2013.

[72] Y. Zhu, Y. Zhao, and S.-C. Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, pages 2855–2864, 2015.

[73] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. In *Intelligent Robots and Systems (IROS), Proceedings of the IEEE/RSJ International Conference on*, pages 3931–3936, 2009.