Noname manuscript No.
(will be inserted by the editor)

Scene Understanding by Reasoning Stability and Safety

Bo Zheng · Yibiao Zhao · Joey Yu · Katsushi Ikeuchi · Song-Chun Zhu

Received: date / Accepted: date

Abstract This paper 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.

We adopt a novel intuitive physics model and represent the energy landscape of each primitive and object in the scene by a disconnectivity graph (DG). We construct a contact graph with nodes being 3D volumetric primitives and edges representing the supporting relations. Then we 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, we 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. We also compare the safety prediction from the intuitive mechanics model with human ratings.

B. Zheng and K. Ikeuchi
The University of Tokyo, Japan
E-mail: {zheng,ki}@cvl.iis.u-tokyo.ac.jp

Yibiao Zhao, Joey Yu and S.-C. Zhu
University of California, Los Angeles (UCLA), USA
E-mail: {ybzhao,chengchengyu}@ucla.edu
E-mail: sczhu@stat.ucla.edu

1 Introduction

1.1 Motivation and Objectives

Traditional approaches, *e.g.*, (Shi and Fu 1983; Tu et al 2005), 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.

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

ii) *Human assistant robots.* Objects have the potential to fall onto or hit people at workplaces, as the warning sign shows in Fig.1 (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.

Fig. 1 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.

As the example shows in Fig.1 (b), we can 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.

iii) *Disaster rescue robots.* Fig.1 (c) shows a demonstration of a HDR-IAI Multi-Arm robot rescuing people during a mock disaster in the DARPA robot challenge (DARPA 2014). 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 paper, we 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(Newcombe et al 2011). We 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.

1.2 Overview

As Fig. 2 shows, given the input point cloud, our method consists of two main steps: stability reasoning and safety reasoning.

1) **Stability reasoning**: hierarchically pursuing a physically stable scene understanding in two sub-steps:

- *Geometric preprocessing*: recovering solid 3D volumetric primitives from a defective point cloud. Firstly we segment and fit the input $2\frac{1}{2}$D depth map or point cloud to small simple (*e.g.*, planar) surfaces; secondly, we merge convexly connected segments into shape primitives; and thirdly, we 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. We build a contact graph for the neighborhood relations of the primitives. For example, as shown in Fig.2(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. 2(d)), but becomes stable when they are group into one object – the lamp. So is the computer screen grouped with its base.

2) **Safety reasoning** – Given a static scene consisting of stable objects, our method first infers hidden and situated causes (disturbance field, red arrows in Fig. 2(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. 2(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, we

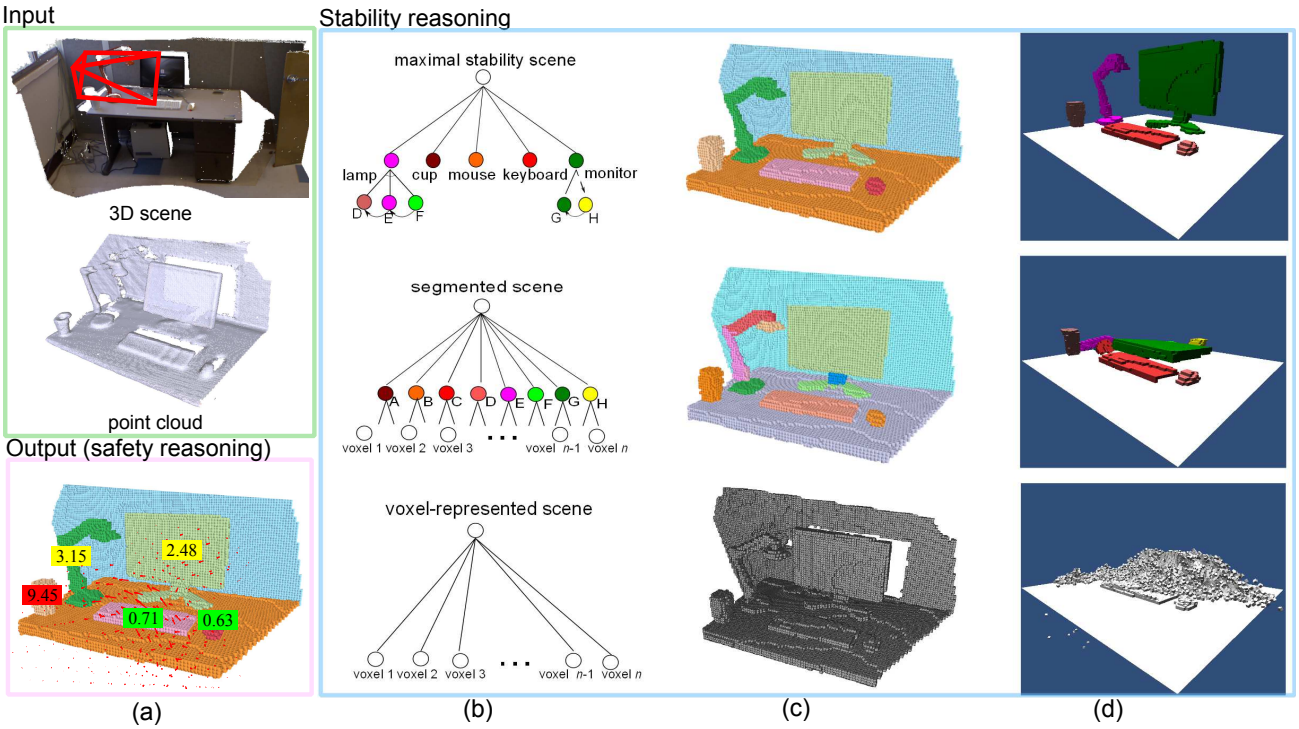


Fig. 2 Overview of our method. (a) Input: 3D scene reconstructed by SLAM technique and Output: parsed 3D scene as stable objects with supporting relations. The number are unsafety scores for each object under the disturbance field (in red arrows), (b) scene parsing graphs corresponding to 3 bottom-up 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.

propose to construct a contact graph and adopt the cluster sampling method, Swendsen-Wang Cut, introduced in image segmentation (Barbu and Zhu 2005). 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, we 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, we 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, we explore two natural disturbances: wind and earthquakes. We then reason the "effects" (*e.g.*, falling) of each possible disturbance by our intuitive physics model. In this case, we 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 we 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, we 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 (Nathan Silberman and Fergus 2012). We evaluate the accuracy of potentially unsafe object detection by ranking the falling risk w.r.t. human judgements.

1.3 Related Work

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

1. Geometric segmentation and grouping. Our approach for geometric pre-processing is related to a set of segmentation methods, *e.g.*, (Felzenszwalb and Huttenlocher 2004; Janoch et al 2011; Attene et al 2006; Poppinga et al 2008). 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, Attene et al (2006) 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 (Zhao and Zhu 2011; Lee et al 2009, 2010; Hedau et al 2010). Furukawa et al (2009) 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.*

2. Physical reasoning. The vision communities have studied the physical properties based on a single image for the "block world" in the past three decades (Biederman et al 1982; Gupta et al 2010, 2011; Zhao and Zhu 2011; Lee et al 2009, 2010)). E.g. Biederman et al.(Biederman et al 1982) studied human sensitivity of objects that violate certain physical relations. Our goal of inferring physical relations is most closely related to Gupta et al (2010) who infer volumetric shapes, occlusion, and support relations in outdoor scenes inspired by physical reasoning from a 2D image, and Silberman et al. (Nathan Silberman and Fergus 2012; Jiang et al 2013; Guo and Hoiem 2013) 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.

3. 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. (Jia et al 2013) adopts an intuitive physics model in (McCloskey 1983), however this model lacks deep consideration on complex physical relations. In our recent work (Zheng et al 2013, 2014), we propose a novel intuitive physics model based on gravity potential energy transfer. In this paper, we 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.

4. Safe Motion Planning. As motion planning is a classic problem in robotics, (Petti and Fraichard 2005; Phillips and Likhachev 2011) 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. We first argue that a robot needs to be aware of potential dangers even in a static environment due to possible incoming disturbances.

5. 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". Grabner et al (2011) recognized chairs by hallucinating a "sitting" actor interacting with the scene. Gupta et al (2011) predicted the "workspace" of a human given an estimated 3D scene geometry. Fouhey et al (2012) and Delaitre et al (2012) demonstrated that observing people performing different actions can significantly improve estimates of scene geometry and scene semantics. Jiang et al (2013) and Jiang and Saxena (2013) proposed scene labeling algorithms by considering humans as the hidden context.

6. Cognitive studies. Recent psychology studies suggested that approximate Newtonian principles underlie human judgements about dynamics and stability (Fleming et al 2010; Hamrick et al 2011) Hamrick et al.(Hamrick et al 2011) 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.

7. Semantic Labeling. Recently many semantic labeling methods for 3D point clouds play an important role in robotics: Koppula et al (2011); Anand et al (2012); Wu et al (2014), *etc*; in graphics: *e.g.*, Nan et al (2012); Shao et al (2012, 2014); Savva et al (2014), *etc*; in 3D shape recognition: Karpathy et al (2013), *etc.* In this paper, however we only focus on the stability and safety reasoning and show its influence on scene understanding.

1.4 Contributions

This paper makes the following contributions.

1. 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.

2. It introduces a novel disconnectivity graph (DG) from physics (Wales 2004) to represent the energy landscapes of objects.

3. It solves the complex optimization problem by applying the cluster sampling method Swendsen-Wang cut used in image segmentation (Barbu and Zhu 2005) to physical reasoning.

4. It proposes an intuitive physics model for safety prediction.

5. 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 (Zheng et al 2013), we extend it to a safety model by introducing various disturbance fields.

The rest of this paper 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.

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, we first compute a 3D volumetric representation for each object part. We proceed in two steps: 1) point cloud segmentation, and 2) volumetric completion.

2.1 Segmentation with Implicit Algebraic Models

We adopt a segmentation method using implicit algebraic models (IAMs) (Blane et al 2000) which fits IAMs to point clouds with simple geometry.

$$f_i(\mathbf{p}) \approx 0, \tag{1}$$

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

$$f_i(\mathbf{p}) = \sum_{0 \le i,j,k; i+j+k \le n} a_{ijk} x^i y^j z^k, \tag{2}$$

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 (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.

2.1.1 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 (Blane et al 2000) As shown in Figure 3(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.

We adopt a region growing scheme (Poppinga et al 2008) 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. We 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., (Zheng et al 2010) to select the models in a degree-increasing manner.

2.1.2 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, we 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 (b), we first sample the points on a line which connects two adjacent regions (the circle lines in Fig. 3 (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, we take a condition as the judgment:

$$\frac{\#\{\mathbf{p} | \mathbf{p}_l \in L \wedge f_i(\mathbf{p}_l) < 0 \wedge f_j(\mathbf{p}_l) < 0\}}{\#\{\mathbf{p} | \mathbf{p}_l \in L\}} > \delta_2, \tag{3}$$

where the ratio threshold δ_2 is set as 0.6. As illustrated in Fig 3 (b), since the circular points drawn between f_2 and f_3 are negative, the segments should be merged. Fig. 4 (a) and (b) shows the difference before and after merging the convexly connected regions.

2.2 Volumetric Space Completion

The primitives output from the above method are still insufficient to reason the physical properties, *e.g.*, in Fig. 4 (b), the wall and table have hollow surfaces with holes and the cup has missing volume. To overcome this, we 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, we

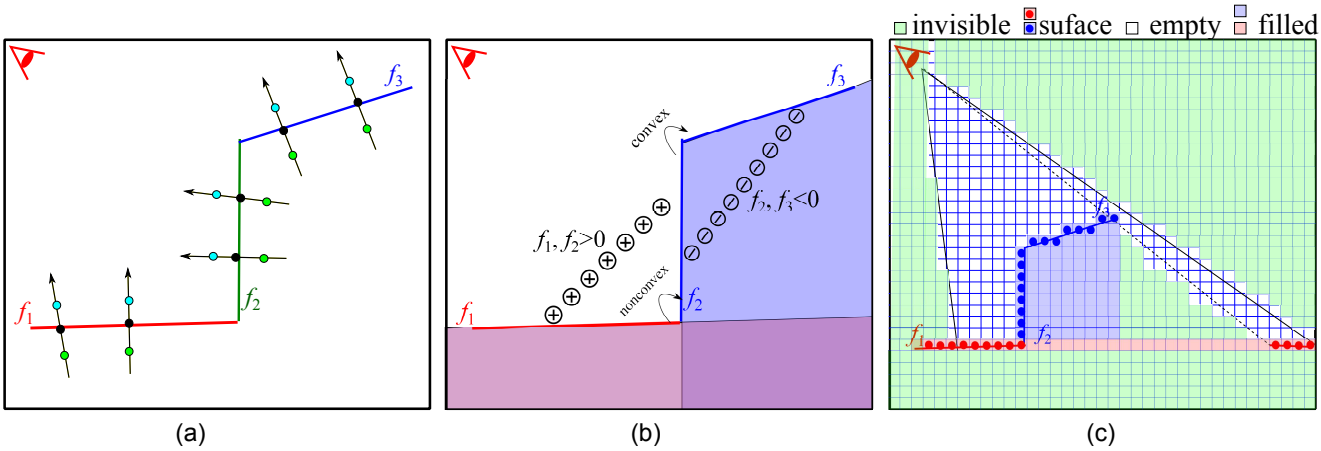


Fig. 3 (a)Splitting. Two 1-degree IAMs f_1, f_2 and f_3 (in red, green and blue lines respectively) are fitted to the 3-Layer point cloud. Points in green and blue are the extra layer points generated from original points in black. (b)Merging. the segments fitted by f_2 and f_3 are merged together, because 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).

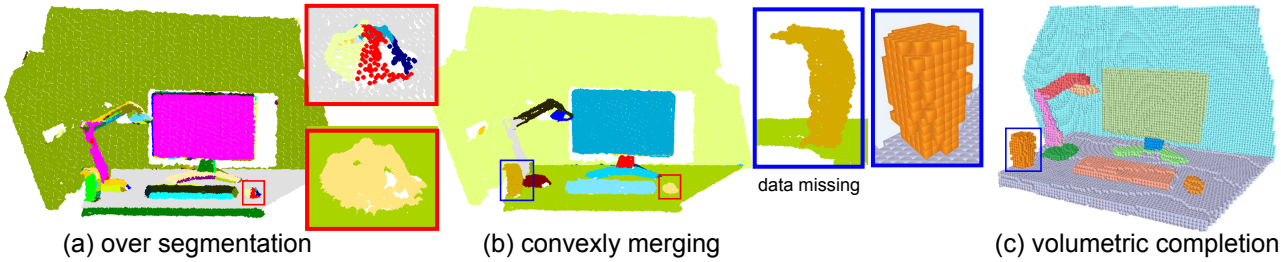


Fig. 4 (a) Over-segmentation result obtained by splitting with IAMs. (b) Result after merging the convexly connected faces. (see the difference on "mouse" object). (c) Result after volumetric completion. (see the difference on "cup" object and hole on the back wall).

fill out the hidden voxels for each incomplete volumetric primitive obtained by the segmentation result above.

2.2.1 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. (Sagawa et al 2005), after which the point cloud is regularized into the coordinate system under the Manhattan world assumption (Furukawa et al 2009), supposing many visible surfaces orient along one of three orthogonal directions. To detect gravity direction, 1) we first calculate the distributions of the principal orientations of the 3D scene by clustering the surface normals into K ($K > 3$) clusters; 2) Then we extract three biggest clusters and take their corresponding normals as three main orientations; 3) After the orthogonalization of these three orientations, we choose the one with smallest angle to the Y-axis of camera plane as the gravity direction.

2.2.2 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 (Furukawa et al 2009), the Manhattan space is carved by the point cloud into three parts, as shown in Figure 3(c): Object surface $\mathbb{S}$ (colored-dots voxels), Invisible space $\mathbb{U}$ (light green voxels) and Visible space $\mathbb{E}$ (white voxels).

2.2.3 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 Figure 3 (b) illustrates, the algorithm can be described as:

Loop: for each invisible voxel $v_i \in \mathbb{U}$, $i = 1, 2, \ldots$

1. 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.
2. Checking the labels of v_js, if there are more than two same labels exist, then assign this label to current voxel.

Fig. 4 (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.

3 Modeling Physical Stability and Safety

3.1 Energy Landscapes

Since any object (or primitive) has potential energy determined by its mass and height to the ground, we 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 5 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 we will discuss later. Without loss of generality, we divide the change into two cases.

Case I: pose change. In Fig. 5 (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. We define the energy change needed for the state change $\mathbf{x}_0 \to \mathbf{x}_1$ by

$$\mathcal{E}_r(\mathbf{x}_0 \to \mathbf{x}_1) = (R\mathbf{c} - \mathbf{t}_1) \cdot m\mathbf{g}, \tag{4}$$

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 we can visualize the energy landscape in a spherical coordinate system (ϕ, θ): $S^2 \to \mathbb{R}$ with two pose angles $\{\phi \in [-\pi\ \pi], \theta \in [-\pi/2, \pi/2]\}$. In Fig. 5 (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. We refer to the early work of Kriegman (Kriegman 1995) for further details.

Case II: position change. We 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. 5 (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.

$$\mathcal{E}_t(\mathbf{x}_0 \to \mathbf{x}_2) = -(\mathbf{c} - \mathbf{t}) \cdot m\mathbf{g} + W_f, \tag{5}$$

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, we 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 $\mathbb{R}^3 \to \mathbb{R}$.

3.2 Disconnectivity Graph Representation

The energy map is continuously defined over the object position and pose. For our purpose, we 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, we represent the energy landscape by a so-called disconnectivity graph (DG) which has been used in studying spin-glass models in physics (Wales 2004). As Fig. 6 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 (Wales 2004).

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

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

For the box on the desk in Fig.5, 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.

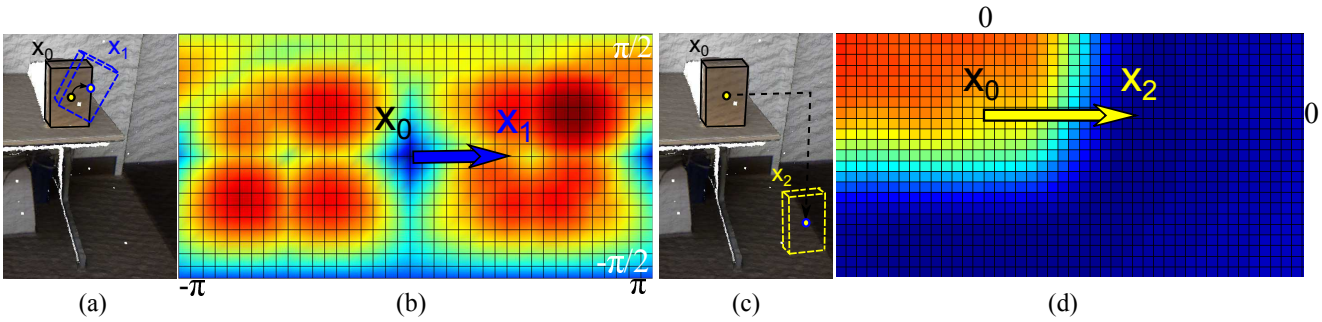


Fig. 5 An example of potential energy map determined by pose and position changes: (a) the box on desk changes pose from state x_0 to x_1. Mass center trajectory is shown as black arrow. (b) the energy map of changing box poses in arbitrary directions. State x_0 is at local minimum on the map. (c) the 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.

Definition 2 *Energy release* $\Delta\mathcal{E}(\widetilde{\mathbf{x}} \rightarrow \mathbf{x}'_o)$ *is the potential energy released when an object moves from its unstable equilibrium* $\widetilde{\mathbf{x}}$ *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.

3.3 Definition of Stability

With DG, we 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*.*

$$\begin{aligned} &\mathrm{S}(a, \mathbf{x}_0, W) \\ &= \max_{\mathbf{x}'_0} \triangle\mathcal{E}(\widetilde{\mathbf{x}} \rightarrow \mathbf{x}'_o)\delta([\min_{\widetilde{\mathbf{x}}} \triangle\mathcal{E}(\mathbf{x}_o \rightarrow \widetilde{\mathbf{x}})] \leq W), \end{aligned} \quad (6)$$

where $\delta()$ is an indicator function and $\delta(z) = 1$ if condition z is satisfied, otherwise $\delta(z) = 0$. $\triangle\mathcal{E}(\mathbf{x}_o \rightarrow \widetilde{\mathbf{x}})$ is the energy absorbed, if it is overcome by W, then $\delta() = 1$, and thus the energy $\triangle\mathcal{E}(\widetilde{\mathbf{x}} \rightarrow \mathbf{x}'_o)$ is released. We 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.

3.4 Definition of Safety

We 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.*

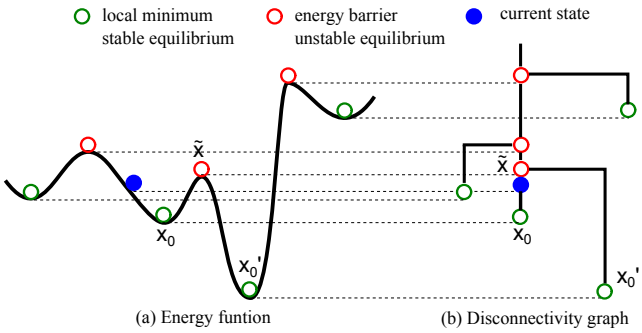


Fig. 6 (a) Energy landscapes and its corresponding disconnectivity graph (b).

$$R(a, \mathbf{x}_0) = \int p(W, \mathbf{x}_0) S(a, \mathbf{x}_0, W) dW, \quad (7)$$

For example, it is more unsafe if there exist a disturbance that makes the box in Fig. 5 fall off from the desk than just fall down on the desk.

With the definition of the instability and risk, we first present the algorithm for static scene understanding by reasoning the stability in Sec.4, and then we introduce the inference of the disturbance field in Sect.5.1 and the calculation of potential energy and initial kinetic energy given a disturbance in Sect.5.2

4 Reasoning Stability

4.1 Stability Optimization

Given a list of 3D volumetric primitives obtained by our geometric reasoning step, we first construct the contact graph, and then the task of physical reasoning can be posed as a well-known graph labelling or partition problem, through which the unstable primitives can be grouped together and assigned the same label to achieve

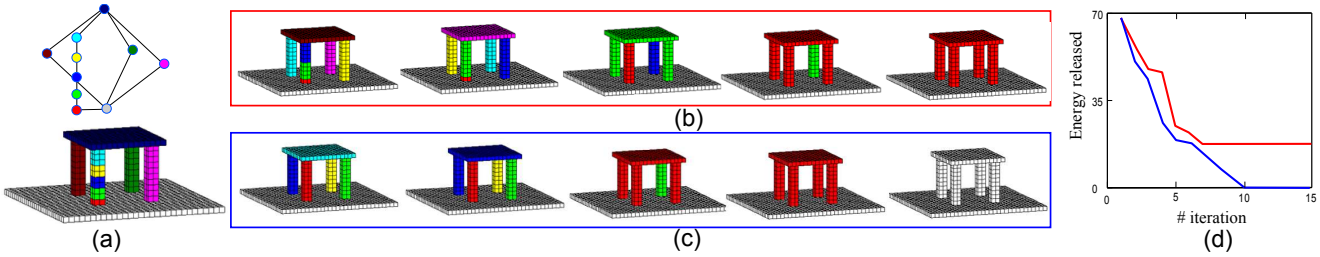


Fig. 7 Example of illustrating the Swendsen-Wang cut sampling process. (a) Initial state with corresponding contact graph. (b) shows the grouping proposals accepted by SWC at different iterations. (c) convergence under increasingly (from left to right) larger disturbance W and consequently the table 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).

global stability of the whole scene at a certain disturbance level W.

4.2 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.7 (a) top where each node corresponds to a primitive in Fig. 7 (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

$$E(L|G;W) = \sum_{O_i \in L} (\mathrm{S}(O_i, \mathbf{x}(O_i), W) - \mathcal{F}(O_i)), \qquad (8)$$

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.

$$\mathcal{F}(O_i) = \lambda_1 f_1(O_i) + \lambda_2 f_2(O_i) + \lambda_3 f_3(O_i), \qquad (9)$$

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. 2. Therefore under the stable conditions, objects should have maximal freedom of movement.

4.3 Inference of Maximum Stability

As the labels of primitives are coupled with each other, we adopt the graph partition algorithm Swendsen-Wang Cut (SWC) (Barbu and Zhu 2005) for efficient MCMC inference. To obtain the globally optimal L^* by the SWC, the next 3 main steps work iteratively until convergence.

Algorithm 1: SWC Inference

Input: $G =< V, E >$, discriminative probabilities q_e, $e \in E$ and generative posterior probability $p(L|G;W)$
Output: Samples $L^* \sim p(L^*|G;W)$

```
Initialize a graph partition π : G = ∪_{l=1}^{n} G_l
for current state L = π do
    for subgraph G_l =< V_l, E_l >, l = 1, 2, ..., n in L do
        for e ∈ E_l turn e = "on" with probability q_e in Eq. (10) do
            Partition G_l into n_l connected components: g_li =< V_{li,E_li} >
        end
        Collect all the connected components in CPP= {Π_li : l = 1, ..., n, i = 1, ..., n_l}.
    end
    Select a connected component Π ∈ CPP randomly
    Propose to reassign Π to a subgraph G_l'
    Accept the move with probability α(L → L') in Eq.(11)
end
```

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

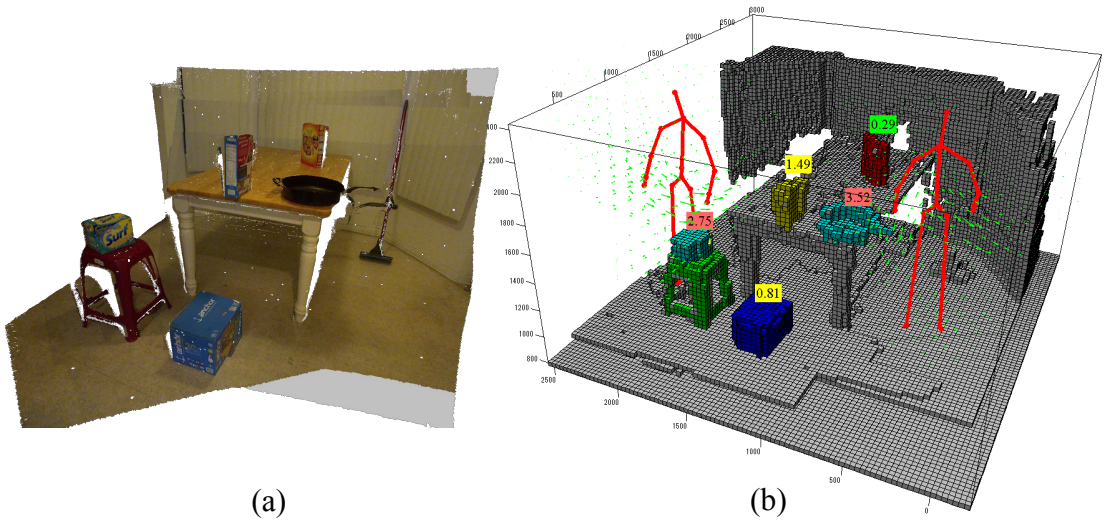


Fig. 8 (a) The input point cloud; (b) Hallucinated human action field and detected potential falling objects with red tags.

probability as:

$$q_e = p(\mu_e = on|v_i, v_j) = \exp\left(-(F(v_i, v_j)/T\right), \qquad (10)$$

where T is temperature factor and $F(\cdot,\cdot)$ denotes the feature between two connected primitives. Here we 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.

(ii) *Graph Clustering.* Given the current label map, it removes all edges between nodes with different labels. Then all the remaining edges are turned on independently with probability q_e. Thus, we have a set of connected components (CCPs) Π's, in which all nodes have the same category label.

(iii) *Graph Flipping.* It randomly selects a CCP Π from the set formed in step (ii) with a uniform probability, and then flips the labels of all nodes in Π to a label $l' \in \{1, 2, ..., L\}$. The flip is accepted with probability (Barbu and Zhu 2005):

$$\alpha(L \to L') = \min\left(1, \frac{\prod_{e\in\mathcal{C}(V_o, V_{L'} \setminus V_o)}(1-q_e)}{\prod_{e\in\mathcal{C}(V_o, V_L - V_o)}(1-q_e)} \cdot \frac{p(L'|G;W)}{p(L|G;W)}\right), \qquad (11)$$

where $p = \frac{1}{z}\exp(-E)$. Fig. 7 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 (11) shown in (b) and (c), resulting in the energy v.s. iterations in (d). It is worth noticing that i) in case of Fig. 7 (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.

We summarize the main three steps above in Algorithm 1. Here we adopt an annealing scheme in the MCMC sampling process, when the temperature is low, the algorithm will converge to a global optimal solution, i.e. partition, with very high probability. Fortunately, the solution space in our algorithm is quite small after geometric processing. For example, there are only 12 geometric entities as graph nodes in the table scene in Fig. 4 (c), and the algorithm converges in several seconds.

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.8 (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, we detect the potential falling objects by calculating its expected falling risk given a disturbance field in Fig.8 (b).

5.1 Safety Under Different Disturbances

5.1.1 Natural Disturbance Field

Aside from the gravity applying a constant downward force to all the voxels, other natural disturbances such

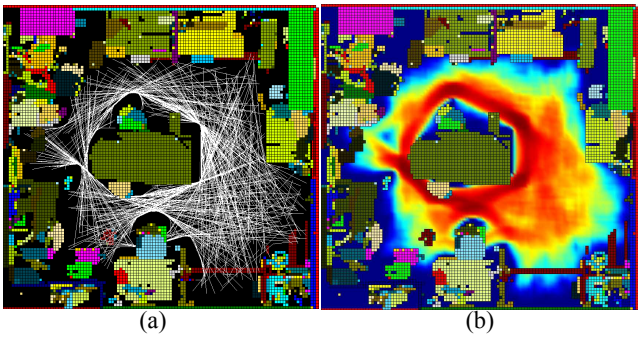


Fig. 9 Primary motion field: (a) The hallucinated human trajectories (white lines); (b) The distribution of the primary motion space. The red represents high probability to be visited.

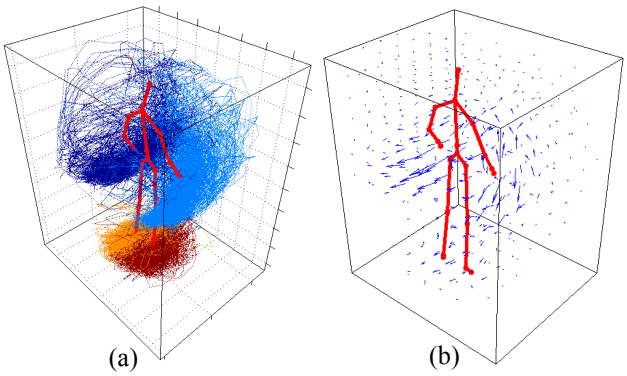


Fig. 10 Secondary motion field: (a) Secondary motion trajectories from motion capture data; (b) Distribution of the secondary motion field. Long vectors represent large velocity of body movement.

as earthquakes and winds are also present in a natural scene.

1) **Earthquake** transmits energy by forces of interactions between contacting surfaces, typically by the frictions in our scenes. Here, we estimate the disturbance field by generating random horizontal forces to the voxels along the contacting surfaces. We use a certain constant to simulate the strength of the earthquake and the work W it generates.

2) **Wind** applies fluid forces to exposed voxels in the space. A precise simulation needs to simulate the fluid flow in the space. Here, we simplify it as a uniformly distributed field over the space.

5.1.2 Human Action Disturbance Field

In order to generate a meaningful disturbance field of human actions, we decompose the human actions into the primary motions *i.e.*the center of mass movements in Fig.9 and the secondary motions *i.e.*the body parts' movements in Fig.10 We 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. We estimate the distribution of primary human motion space by synthesizing human motion trajectories following two simple observations:

1) A rational agent mostly walks along a shortest path with minimal effort;

2) An agent has a basic need to travel between any two walkable positions in the scene.

Therefore, we randomly pick 500 pairs of positions in the walkable space, we calculate the shortest path connecting these two positions as shown in Fig.9 (a), and we calculate the walking frequency as well as walking directions based on the synthesized trajectories. Fig.9 (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.9 (b), we 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. We also the Kinect camera to collect human motion capture data Fig.10 (a), and then calculate the distribution of moving velocities as shown in Fig.10 (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 above the table center, or any unreachable points, is zero. Five typical cases in the integrated field is demonstrated in Fig.11

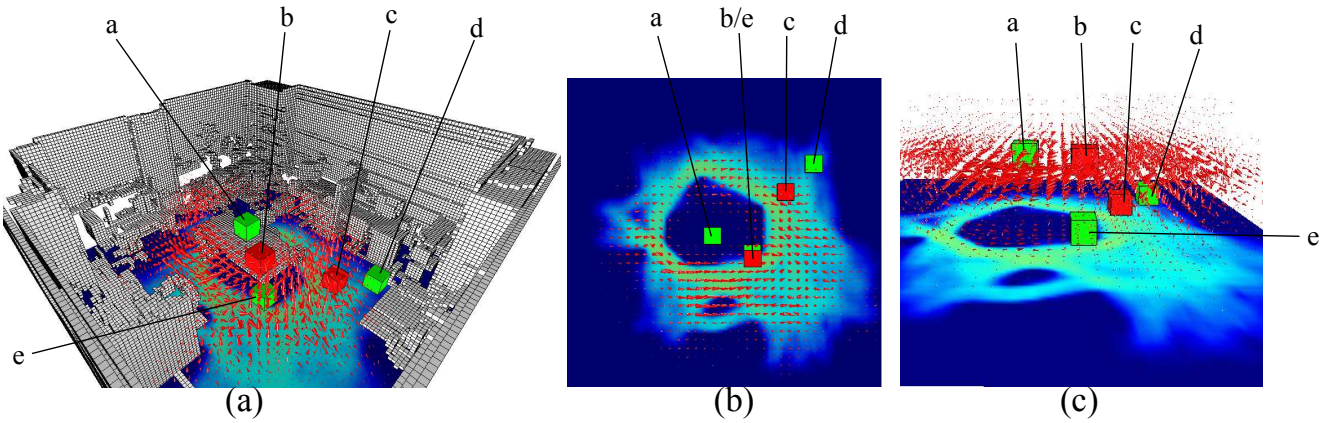


Fig. 11 The integrated human action field by convolving primary motions with secondary motions. The objects **a-e** are five typical cases in the disturbance field: the object **b** on edge of table and the object **c** along the passway exhibit more disturbances (accidental collisions) than other objects such as **a** in the center of the table, **e** below the table and **d** in a concave corner of space.

5.2 Calculating the Physical Energy

Given the disturbance field, in this section we 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 we 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 (Hamrick et al 2011)

In this paper, for simplicity, we 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, we intuitively calculate the kinetic energy of human motion, as the input work. Here, we 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, we can simply calculate out the kinetic energy distribution by multiplying half of the velocity squares.

	Office	Living room	desk	total
Scenes	5	4	4	13

Table 1 Summary of the Dataset. Some samples are shown in Fig. 12

6 Experimental Result

We 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 (Nathan Silberman and Fergus 2012) including 1449 RGBD images with manually labeled ground truth.
- synthesized depth map and volumetric images simulated from CAD scene data.
- 13 reconstructed 3D scene data captured by Kinect Fusion (Fig. 12) gathered from office and residential rooms with ground truth labeled by a dense mesh coloring.

6.1 Evaluating Single Depth Image Segmentation

Two evaluation criteria: "Cut Discrepancy" and "Hamming Distance" mentioned in Chen et al (2009) 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. 14, our segmentation by physical reasoning has a lower error rate than the other two: region growing segmentation Poppinga et al (2008), and our geometric reasoning.

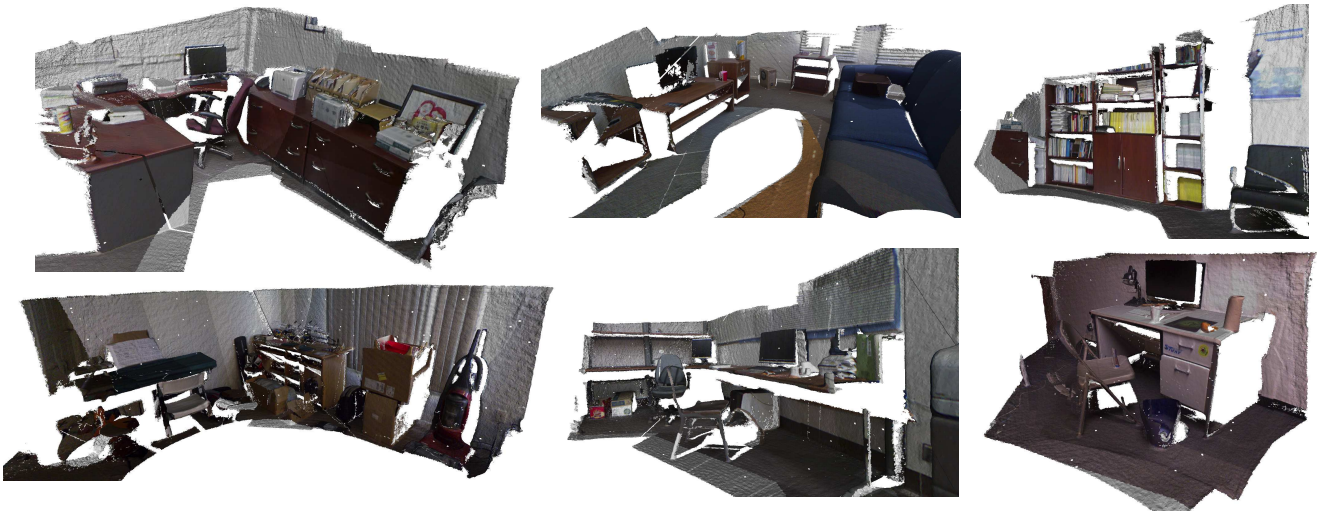


Fig. 12 Samples of our dataset.

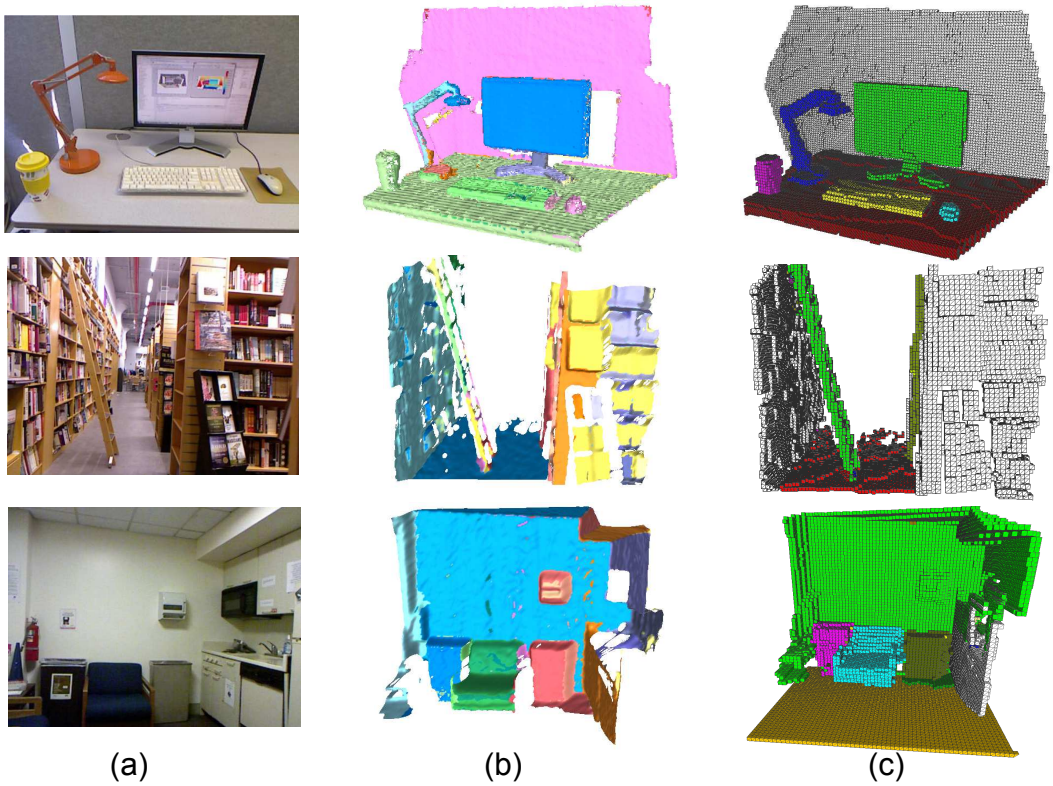


Fig. 13 Segmentation result for single depth images. (a) RBG images for reference. (b) segmentation result by region growing Poppinga et al (2008). (c) stable volumetric objects by physical reasoning.

Fig. 13 shows some examples of comparing another point cloud segmentation result Poppinga et al (2008) 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, stabilities, *etc.*

Fig. 15 shows a qualitative comparison with the method proposed in Jia et al (2013, 2014). However we would like to clarify that our method is not designed for segmentation but for understanding the physical relations such as unfixed support or fixed in the scene. As shown in Fig. 13 (c), our method fixed a kitchen cabinet onto the wall. From the view of segmentation, this is viewed as a miss-merged segment, but it reflects the truth of physics that kitchen cabinet seems to be

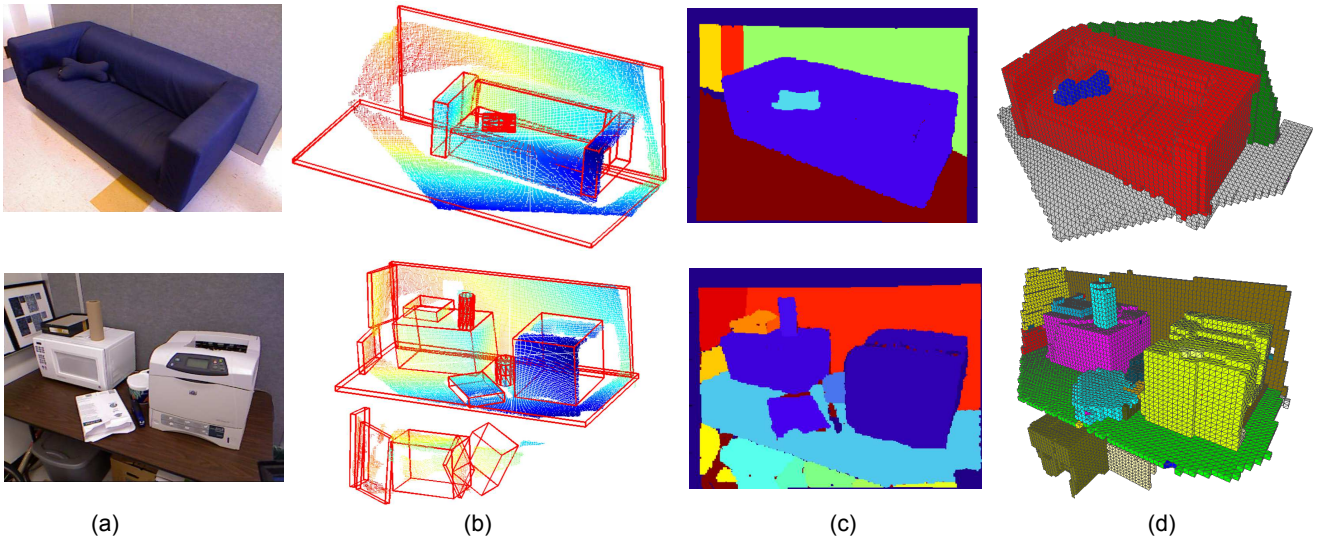


Fig. 15 Intuitive result comparison: (a) original RGB iamges for reference, (b) 3D point cloud and boxes calculated with the method proposed by Jia et al (2013), (c) the corresponding segmentation result of (b), and (d) our result.

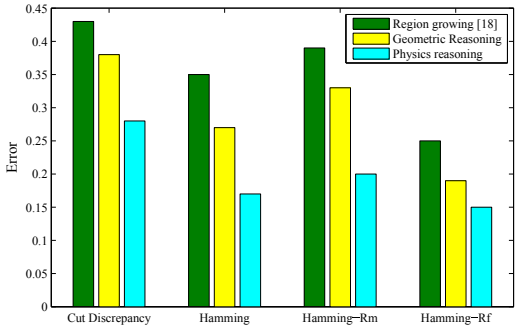


Fig. 14 Segmentation accuracy comparison of three methods: Region growing method Poppinga et al (2008), result of our geometric reasoning and physical reasoning by one "Cut Discrepancy" and three "Hamming Distance".

a part of the wall and it is difficult to move it on the wall.

6.2 Evaluating Volumetric Completion

For evaluating the accuracy of volumetric completion, we densely sample point clouds from a set of CAD data including 3 indoor scenes. We simulate the volumetric data (as ground truth) and depth images from a certain view (as test images). We calculate the precision and recall which evaluates voxel overlapping between ground truth and the volumetric completion of testing data. Tab. 6.2 shows the result that our method has much better accuracy than traditional Octree methods such as Sagawa et al (2005). Fig. 16 intuitively illustrates the completed objects (bottom row) by our method have more overlaps with ground truth planes (top row in red) than the original sample point clouds (top row in blue).

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

Table 2 Precision and recall of Volumetric completion. Comparison of three method: 1) voxel-based representation generated by Octree algorithm (Sagawa et al 2005), 2) voxels in surface and invisible space (sec. 2.2), and 3) our volumetric completion.

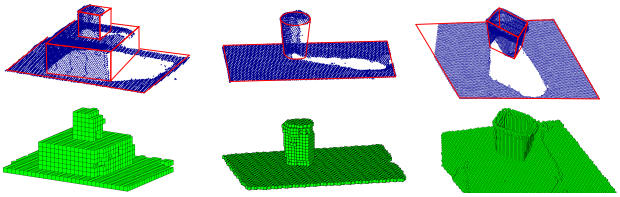


Fig. 16 Examples of volumetric completion. Top row: densely sampled point clouds (in blue) in a view direction with missing parts referring to the original shape guide lines (in red). Bottom row: volumetric completions of the objects in top row.

6.3 Evaluating Physical Inference Accuracy

Because the physical relations are defined in terms of our contact graph, we map the ground-truth labels to the nodes of contact graphs obtained by geometric reasoning. Than we evaluate our physical reasoning against two baselines: discriminative methods using 3D feature priors similar to the method in Nathan Silberman and Fergus (2012), and greedy inference methods such as

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

Table 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.

the marching pursuit algorithm for physical inference. The result shown in Tab. 6.3 is evaluated by the average over 13 scene data captured by Kinect Fusion.

Figure 17 (a)-(d) and (e)-(j) show two examples from the results. Here we discuss some irregular cases illustrated by close-ups of the figures.

Case I: Figure 17 (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 Eq. (6). The unstable state is calculated 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: Figure 17 (d) the "air pump" unstably stands on floor but is an independent object, because although its stability is very low, the penalty designed in Eq.(7) penalized it to be fixed onto the floor. The lamp is not affixed for the same reason, as shown in Figure 17 (h).

Case III: Figure 17 (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: Figure 17 (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 Figure 17 (j) misses more data.

6.4 Running system and time

All the experiments were implemented in Matlab 2012a with a modern PC having an Intel core i7 CPU, 3.4 GHz, and 16 GB memory. For dealing with one single image, such as shown in Fig. 13, the running time is around 2 minutes. For large scene data, such as the cases shown in Fig. 17 and 12 , the running time is around 7-12 minutes.

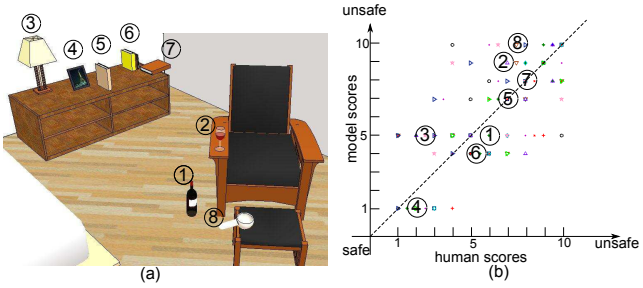


Fig. 20 Scoring object unsafeness in a scene (a) with 8 objects. We show the correlation graph (b) with human score against our measurement $R(a, \mathbf{x})$ which is normalized from 1 to 10. Color/shape points show human vs. model scores corresponding to different persons. Circle points with numbers inside show the average of human vs. model scores for each object corresponding to (a).

6.5 Evaluating Safety Ratings

First we provide a selected qualitative result shown in Fig. 18. We compare the potential falling objects under three different disturbance fields: 1) The human action field in Fig. 18 (b,e); 2) The wind field (a uniform directional field) in Fig. 18 (c,f) and 3) earthquake (random forces on contacting object surface) in Fig. 18 (d,g). As we 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 we report selected results in different 3D scenes, as shown in Fig. 19 top row: vending machine room and bottom row: copy machine room. We 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.

6.6 Discussion

For evaluating safety ratings, we rank object unsafeness in a scene in comparison with human subjects. Fig. 20 (a) shows a 3D scene (constructed in CAD design), from which we pick 8 objects and ask 14 participants to rank the unsafeness of these objects considering gravity, common life activity and the risk of falling. We compare the human ranking with our unsafeness function $R(a, \mathbf{x})$ in Fig. 20 (b). We found that 1) humans got big variations

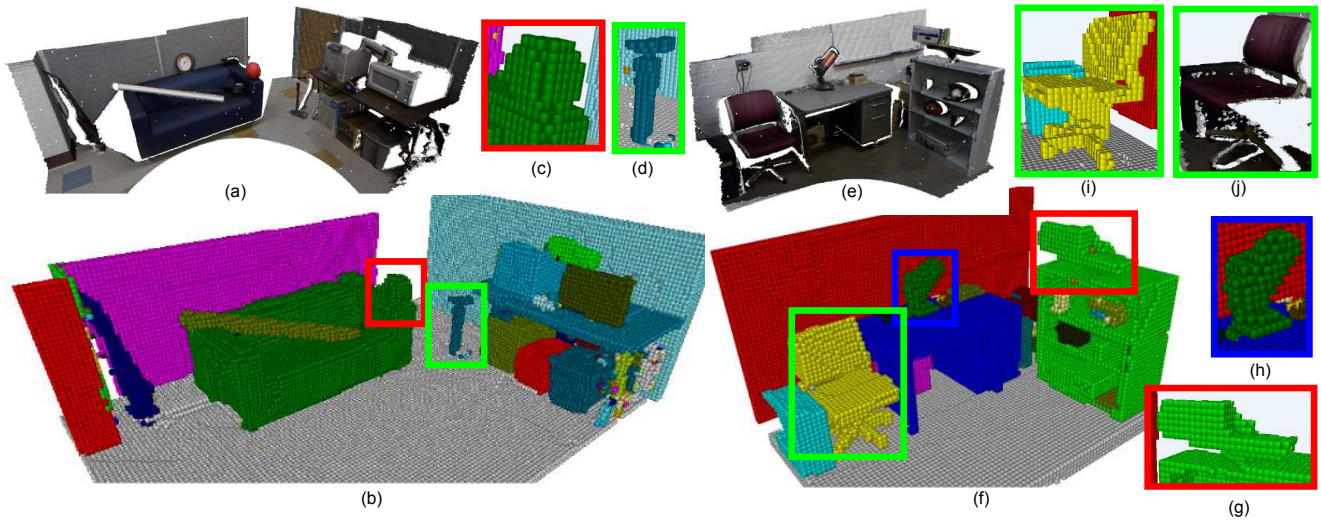


Fig. 17 Example result. (a) and (e): data input. (b) and (f): volumetric representation of stable objects. (c): the ball is fixed onto the handle of sofa. (d): the "pump" is unstable (see text). (i): a irregular case of (g). (j): hidden voxels under chair compared to (h).

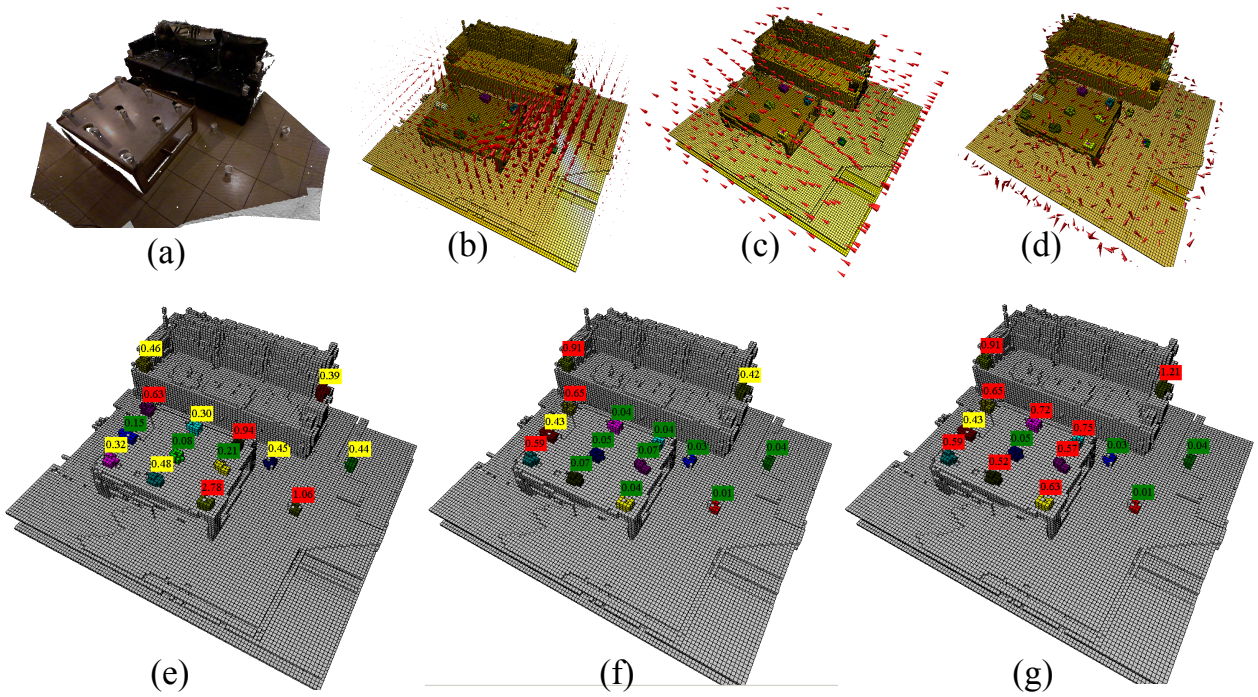


Fig. 18 The potential falling objects (with red tags) under the human action field (b,e), the wind field (c,f) and the earthquake field (d,g) respectively. The results match with human perception: (i) objects around table corner are not safe w.r.t human walking action; (ii) object along the edge of wind direction are not safe w.r.t wind disturbance; and (iii) object along all the edges are not safe w.r.t earthquake disturbance.

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. 20 (b), the average of human vs. model scores for each object lies near to the diagonal line.

7 Conclusion

We 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). We demonstrated its feasibility in experiments and show that this provides a new method for object grouping when it is hard to pre-define all possible object shapes and appearance in an object category.

This paper also presents a novel approach for detecting potential unsafe objects. We 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. 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, we plan to explore several directions: i) Connecting our work to human psychology models like the one in (Hamrick et al 2011), 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 we 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.

Acknowledgment

This work is supported by 1) MURI ONR N00014-10-1-0933 and DARPA MSEE grant FA 8650-11-1-7149, USA, 2) Next-generation Energies for Tohoku Recovery (NET) and SCOPE Program of Ministry of Internal Affairs and Communications, Japan, 3) and the 10-th core project grant of Microsoft Japan.

References

Anand A, Koppula H, Joachims T, Saxena A (2012) Contextually guided semantic labeling and search for 3d point clouds. IJRR

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

Barbu A, Zhu SC (2005) Generalizing swendsen-wang to sampling arbitrary posterior probabilities. TPAMI 27:1239–1253

Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception: Detecting and judging objects undergoing relational violations. Cog Psy 14(2):143 – 177

Blane M, Lei ZB, Cooper DB (2000) The 3L Algorithm for Fitting Implicit Polynomial Curves and Surfaces to Data. TPAMI 22(3):298–313

Chen X, Golovinskiy A, Funkhouser T (2009) A benchmark for 3D mesh segmentation. In: SIGGRAPH

DARPA (2014) Robots Rescue People. http://www.i-programmer.info/news/169-robotics/6857-robots-rescue-people.html

Delaitre V, Fouhey D, Laptev I, Sivic J, Gupta A, Efros A (2012) Scene semantics from long-term observation of people. In: ECCV

Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comput Vision 59(2):167–181

Fleming R, Barnett-Cowan M, Bülthoff H (2010) Perceived object stability is affected by the internal representation of gravity. Perception 39:109

Fouhey D, Delaitre V, Gupta A, Efros A, Laptev I, Sivic J (2012) People watching: Human actions as a cue for single-view geometry. In: ECCV

Furukawa Y, Curless B, Seitz SM, Szeliski R (2009) Manhattan-world stereo. In: CVPR

Grabner H, Gall J, Van GL (2011) What makes a chair a chair? In: CVPR

Guo R, Hoiem D (2013) Support surface prediction in indoor scenes. In: ICCV

Gupta A, Efros A, Hebert M (2010) Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: ECCV

Gupta A, Satkin S, Efros A, Hebert M (2011) From 3D Scene Geometry to Human Workspace. In: CVPR

Hamrick J, Battaglia P, Tenenbaum J (2011) Internal physics models guide probabilistic judgments about object dynamics. In: Conf. Cog. Sc.

Hedau V, Hoiem D, Forsyth D (2010) Thinking inside the box: Using appearance models and context based on room geometry. In: ECCV

Janoch A, Karayev S, Jia Y, Barron JT, Fritz M, Saenko K, Darrell T (2011) A category-level 3-d object dataset: Putting the kinect to work. In: ICCV Workshop

Jia Z, Gallagher A, Saxena A, Chen T (2013) 3d-based reasoning with blocks, support, and stability. In: CVPR

Jia Z, Gallagher A, Saxena A, Chen T (2014) 3d reasoning from blocks to stability. PAMI

Jiang Y, Saxena A (2013) Infinite latent conditional random fields for modeling environments through humans. In: In Robotics: Science and Systems (RSS)

Jiang Y, Koppula HS, Saxena A (2013) Hallucinated humans as the hidden context for labeling 3d scenes. In: CVPR

Karpathy A, Miller S, Fei-Fei L (2013) Object discovery in 3d scenes via shape analysis. In: International Conference on Robotics and Automation (ICRA)

Koppula H, Anand A, Joachims T, Saxena A (2011) Semantic labeling of 3d point clouds for indoor scenes. In: NIPS

Kriegman DJ (1995) Let them fall where they may: Capture regions of curved objects and polyhedra. International Journal of Robotics Research 16:448–472

Lee D, Hebert M, Kanade T (2009) Geometric reasoning for single image structure recovery. In: CVPR

Lee D, Gupta A, Hebert M, Kanade T (2010) Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces advances in neural information processing systems. Cambridge: MIT Press pp 609–616

McCloskey M (1983) Intuitive physics. Scientific American 248(4):114–122

Nan L, Xie K, Sharf A (2012) A search-classify approach for cluttered indoor scene understanding. ACM Trans on Graph (TOG) 31(6)

Nathan Silberman PK Derek Hoiem, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: ECCV

Newcombe R, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A, Kohli P, Shotton J, Hodges S, Fitzgibbon A (2011) Kinectfusion: Real-time dense surface mapping and tracking. In: ISMAR

Petti S, Fraichard T (2005) Safe motion planning in dynamic environments. In: IROS

Phillips M, Likhachev M (2011) Sipp: Safe interval path planning for dynamic environments. In: ICRA

Poppinga J, Vaskevicius N, Birk A, Pathak K (2008) Fast plane detection and polygonalization in noisy 3D range images. In: IROS

Sagawa R, Nishino K, Ikeuchi K (2005) Adaptively merging large-scale range data with reflectance properties. TPAMI 27:392–405

Savva M, Chang AX, Hanrahan P, Fisher M, Nießner M (2014) Scenegrok: Inferring action maps in 3d environments. ACM Trans on Graph (TOG) 33(6)

Shao T, Xu W, Zhou K, Wang J, Li D, Guo B (2012) An interactive approach to semantic modeling of indoor scenes with an rgbd camera. ACM Trans on Graph (TOG) 31(6)

Shao T, Monszpart A, Zheng Y, Koo B, Xu W, Zhou K, Mitra NJ (2014) Imagining the unseen: Stability-based cuboid arrangements for scene understanding. ACM Trans on Graph (TOG)

Shi QY, Fu Ks (1983) Parsing and translation of (attributed) expansive graph languages for scene analysis. TPAMI PAMI-5(5):472–485

Tu Z, Chen X, Yuille AL, Zhu SC (2005) Image parsing: Unifying segmentation, detection, and recognition. Int J Computer Vision (IJCV)

Wales D (2004) Energy Landscapes: Applications to Clusters, Biomolecules and Glasses. Cambridge Molecular Science, Cambridge University Press

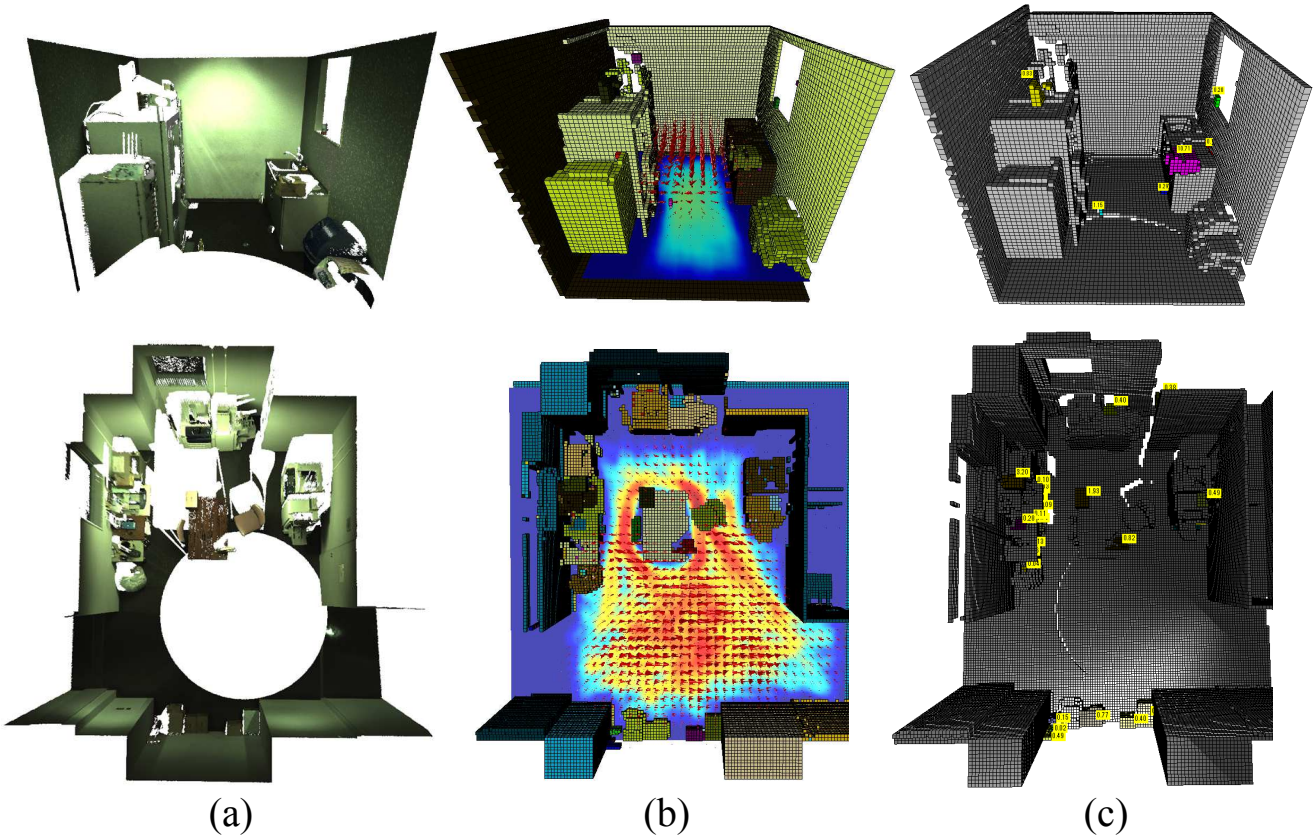
Wu C, Lenz I, Saxena A (2014) Hierarchical semantic labeling for task-relevant rgb-d perception. In: Robotics: Science and Systems (RSS)

Zhao Y, Zhu SC (2011) Image parsing via stochastic scene grammar. In: NIPS

Zheng B, Takamatsu J, Ikeuchi K (2010) An Adaptive and Stable Method for Fitting Implicit Polynomial Curves and Surfaces. PAMI 32(3):561–568

Zheng B, Zhao Y, Yu JC, Ikeuchi K, Zhu SC (2013) Beyond point cloud: Scene understanding by reasoning geometry and physics. In: CVPR

Zheng B, Zhao Y, Yu JC, Ikeuchi K, Zhu SC (2014) Detecting potential falling objects by inferring human action and natural disturbance. In: IEEE Int. Conf. on Robotics and Automation (ICRA)



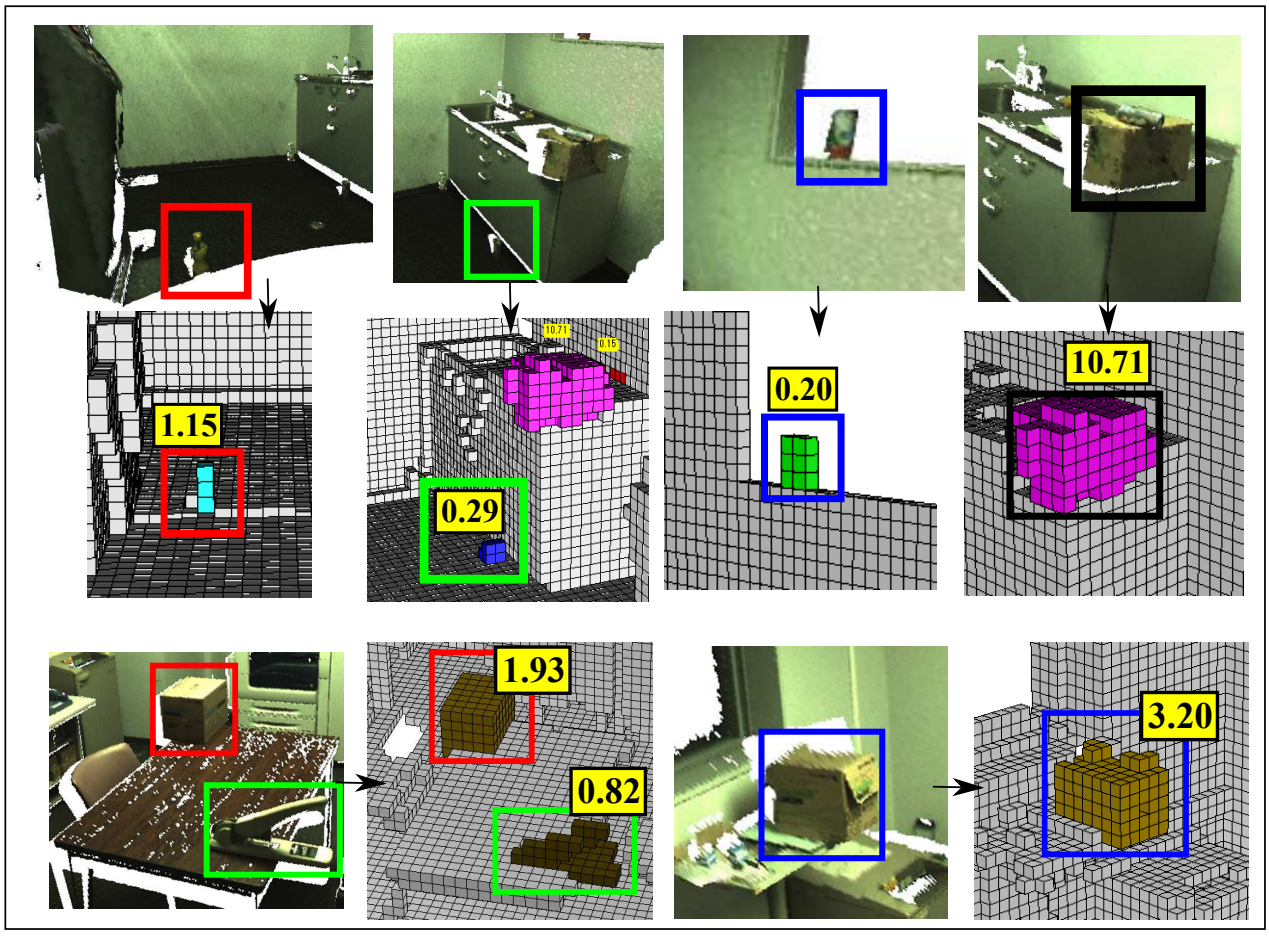


Fig. 19 (a) Input 3D scene point clouds; (b) Segmented volumetric objects in different colors and inferred disturbance fields of human activity; (c) objects with risk scores. (d) Zoom-in details of detected potential risky objects.