UCLA
UCLA Electronic Theses and Dissertations

Title
Single View 3D Reconstruction and Parsing Using Geometric Commonsense for Scene Understanding

Permalink
https://escholarship.org/uc/item/4sh1z61s

Author
Yu, Chengcheng

Publication Date
2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
Los Angeles

Single View 3D Reconstruction and Parsing Using Geometric Commonsense for Scene Understanding

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Chengcheng Yu

2017

© Copyright by

Chengcheng Yu

2017

ABSTRACT OF THE DISSERTATION

Single View 3D Reconstruction and Parsing Using Geometric Commonsense for Scene Understanding

by

Chengcheng Yu
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2017
Professor Song-Chun Zhu, Chair

Scene understanding is a fundamental problem in computer vision. My thesis studies this topic in three perspective: (1) 3D scene reconstruction to understand the 3D structure of a scene. (2) Geometry and physics reasoning to understand the relationships of objects in a scene. (3) The interaction between human action and objects in a scene.

Specifically, the 3D reconstruction builds a unified grammatical framework capable of reconstructing a variety of scene types (e.g., urban, campus, county etc.) from a single input image. The key idea of our approach is to study a novel commonsense reasoning framework that mainly exploits two types of prior knowledges: (i) prior distributions over a single dimension of objects, e.g., that the length of a sedan is about 4.5 meters; (ii) pair-wise relationships between the dimensions of scene entities, e.g., that the length of a sedan is shorter than a bus. These unary or relative geometric knowledge, once extracted, are fairly stable across different types of natural scenes, and are informative for enhancing the understanding of various scenes in both 2D images and 3D world. Methodologically, we propose to construct a hierarchical graph representation as a unified representation of the input image and related geometric knowledge. We formulate these objectives with a unified probabilistic formula and develop a data-driven Monte Carlo method to infer the optimal solution with both bottom-to-up and top-down computations. Results with comparisons on public datasets showed that our method clearly outperforms the alternative methods.

For geometry and physics reasoning, we present an approach for scene understanding by reasoning physical stability of objects from point cloud. We utilize a simple observation that, by human design, objects in static scenes should be stable with respect to gravity. This assumption is applicable to all scene categories and poses useful constraints for the plausible interpretations (parses) in scene understanding. Our method consists of two major steps: 1) geometric reasoning: recovering solid 3D volumetric primitives from defective point cloud; and 2) physical reasoning: grouping the unstable primitives to physically stable objects by optimizing the stability and the scene prior. We propose to use a novel disconnectivity graph (DG) to represent the energy landscape and use a Swendsen-Wang Cut (MCMC) method for optimization. In experiments, we demonstrate that the algorithm achieves substantially better performance for i) object segmentation, ii) 3D volumetric recovery of the scene, and iii) better parsing result for scene understanding in comparison to state-of-the-art methods in both public dataset and our own new dataset.

Detecting potential dangers in the environment is a fundamental ability of living beings. In order to endure such ability to a robot, my thesis presents an algorithm for detecting potential falling objects, i.e. physically unsafe objects, given an input of 3D point clouds captured by the range sensors. We formulate the falling risk as a probability or a potential that an object may fall given human action or certain natural disturbances, such as earthquake and wind. Our approach differs from traditional object detection paradigm, it first infers hidden and situated "causes (disturbance) of the scene, and then introduces intuitive physical mechanics to predict possible "effects (falls) as consequences of the causes. In particular, we infer a disturbance field by making use of motion capture data as a rich source of common human pose movement. We show 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.

The dissertation of Chengcheng Yu is approved.

Luminita Vese

Hongjing Lu

Ying Nian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2017

To my family.

TABLE OF CONTENTS

1 Introduction **1**

1.1 Motivation 1

1.1.1 Scene Reconstruction Based on Commonsense and Attributed Grammar 1

1.1.2 Scene Understanding by Reasoning Geometry and Physics 7

1.1.3 Scene Understanding by Detection Potential Falling Objects by Inferring Human Action and Natural 9

1.2 Related Work 10

2 Unified Single-Image 3D Scene Parsing Using Geometric Commonsense 18

2.1 Introduction 18

2.2 Our Approach 21

2.2.1 Stochastic Scene Grammar 21

2.2.2 Bayesian Formula 25

2.2.3 Bottom-up and Top-Down Inference 26

2.3 Experiments 29

2.3.1 Qualitative result 30

2.3.2 Quantitative result 31

2.3.3 Measurement result 33

2.3.4 Indoor result 35

3 Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics **40**

3.1 Introduction 40

3.2 Geometric reasoning 43

3.2.1 Segmentation with implicit algebraic models 43
3.2.2 Volumetric space completion 46
3.3 Modelling object stability 47
3.3.1 Energy landscapes 47
3.3.2 Disconnectivity graph representation 49
3.4 Physical reasoning 50
3.4.1 Contact graph and group labelling 51
3.4.2 Inference of Maximum stability 52
3.5 Experimental result 54
3.5.1 Evaluating single depth image segmentation 54
3.5.2 Evaluating volumetric completion 56

4 Detecting Potential Falling Objects by Inferring Human Action and Natural 59

4.1 Introduction 59
4.1.1 Motivation 59
4.1.2 Overview of our approach 61
4.2 Definition of the falling risk 63
4.3 Inferring the disturbance field 65
4.3.1 Natural disturbance field 66
4.3.2 Human action disturbance field 66
4.4 Calculating the physical energy 68
4.4.1 Initial kinetic energy after an elastic collision 70
4.4.2 Decomposition of the force, the velocity and the momentum 71
4.4.3 Potential energy 72

4.5 Experiments 74

4.5.1 Qualitative evaluation 74

4.5.2 Quantitative evaluation 74

5 Conclusion **79**

References **81**

LIST OF FIGURES

1.1 A typical result of our approach. (a) Input image overlaid with detected parallel lines; (b) surface normal map where each color indicates a unique normal orientation; (c) synthesized images from a novel viewpoint; and (d) depth map (darker is closer). 3

1.2 Parsing an image using attribute grammar. Left: global geometric attributes are associated with the root node (scene) of the parse graph, including focal length of camera, and Cartesian Coordination System defined by Manhattan frames. Right: parse graph augmented with local geometric attributes, such as surface normals and vanishing points (VPs) associated with a surface, or multiple vanishing points for a building. R1; :::;R5 are the five grammar rules for scene decomposition 4

1.3 1. Overview of our method. (a) 3D scene reconstructed by SLAM technique, (b) point cloud as Input. In geometric reasoning, (c) a portion is shown to be segmented by a segment-and-merge approach, with missing voxels, (d) solid primitives by volumetric completion. In physical reasoning, (e) the contact graph are labeled through stability optimization. (f). Final parsing results with stable objects. 8

1.4 The detection of potential falling objects is an essential ability of a safety-aware robot: (a) the safety surveillance robot for a construction site, (b) the human assistant robot for the baby proofing, and (c) A building where was crashed by earthquake and tsunami on March 11, 2011, Japan. 11

2.1 Single-view 3D scene reconstruction using Geometric commonsense. Top: the world is full of commonsense over geometric dimensions, e.g., that a sedan is about 4.5 meters long. Bottom: exemplar result of the proposed method, including synthesized image (left), planar segmentation (middle), and depth map (right). . 18

2.2 Attribute Parse Graph. Every graph node represents a scene entity, and is augmented with a set of geometric attributes ($\hat{l}$, line direction; d, depth;$\bar{n}$, normal orientation; w, width; h, height; f, focal length; θ, camera angle). There are five types of terminal nodes: L, line, P, planar surface; T, texture; H, human; V, vehicles. 20

2.3 Here is an example for measurement. Given the height of the reference object (e.g. a pillar), the goal is to estimate the height of the man. 23

2.4 Geometric commonsense of Type-II: a total of 21 pair-wise relationships between scene elements. 24

2.5 Exemplar results of the proposed method.Column-1: input images; Column-2: edge map; Column-3: semantic region map; Column-4&5: two novel viewpoints synthesized; Column-6: depth map; Column-7: depth map recovered by; Column-8: occurrence frequencies of equations (in blue) and occurrence frequencies of valid equations (in yellow), where the horizontal axis indicates the 21 equations shown in Fig. 2.4 28

2.6 Failure cases of the proposed method. See texts. 31

2.7 Average occurrence frequencies of commonsense equations on images of five scene categories. See Texts. 32

2.8 The measurements are used as metric. 34

2.9 An example of our measurement results. 35

2.10 A: RGB output image; B: Furniture CAD model; C: Labeling map; D: Depth map 36

2.11 Column-left: Input image; Column-Middle&right: Two novel viewpoints synthesized. 37

2.12 A&B: Input image. C&D: Reconstruction results. 38

2.13 A,B,C: Input image from different views in a single scene. D: Multi-view reconstructing result 39

3.1 1. Overview of our method. (a) 3D scene reconstructed by SLAM technique, (b) point cloud as Input. In geometric reasoning, (c) a portion is shown to be segmented by a segment-and-merge approach, with missing voxels, (d) solid primitives by volumetric completion. In physical reasoning, (e) the contact graph are labeled through stability optimization. (f). Final parsing results with stable objects. 41

3.2 (a) Two 1-degree IAMs f_1 and f_2 (in blue and red lines respectively) are fitted to the 3-Layer point cloud. The light red and blue areas denote in which functions f_1, f_2 and f_3 are minus. (b) Invisible space estimation and voxel completion. Four types of voxels are estimated: 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). 44

3.3 (a) A chair in a stable state x_0 is moved to (b) an unstable state x_1. (c) The landscape of potential energy is calculated by Eq. (3.3) over two rotation angles where x_0 is a local minimum and x_1 is a saddle point passing which, the chair will fall to a deeper energy basin (blue). 48

3.4 (a) Energy landscapes and its corresponding disconnectivity graph (b). 49

3.5 Example of illustrating the Swendsen-Wang sampling process. (a) Initial state with corresponding contact graph. (b) shows the grouping proposals accepted by SWC at different iterations. (c) convergence under 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) 53

3.6 Segmentation accuracy comparison of three methods: Region growing method [PVB08], result of our geometric reasoning and physical reasoning by one Cut Discrepancy and three Hamming Distance 55

3.7 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). 57

4.1 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 among the debris 60

4.2 (a) The input point cloud; (b) Imagined human action field and detected potential falling objects with red tags 62

4.3 An illustration of falling risk definition and other basic concepts on the potential energy curve; 64

4.4 An simple example that a cone is being knocked down. It is pushed up from the stable equilibrium B, and about to go over the energy barrier B'. The correspondent potential energy map is on the right. 65

4.5 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. 66

4.6 Secondary motion field: (a) Secondary motion trajectories of hands and feet from motion capture data; (b) Distribution of the secondary motion field. Long vectors represent large velocity of body movement. 67

4.7 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** on a concave corner of space. 69

4.8 The decomposition of action velocity. The gray polygon represents an object with its center of mass on the red dot. The action velocity V first decompose as a rotational velocity V_r and translational velocity V_t, and each velocity is further decomposed as three components along three dimensions. 71

4.9 Potential energy map for (b) the rotational movement and (c) the displacement movement of the box on a table in (a). 73

4.10 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. 75

4.11 (a) Input 3D scene point clouds; (b) Inferred human action fields and segmented objects shown in different colors; (c) Detected potential falling objects with their risk scores on the yellow labels; (d) Some zoom-in details of detected potential falling objects. See text for more explanation. 76

4.12 (a) The ROC curve of the potential falling object detection (b) The correlation between the falling risk ranking by our algorithm and the ranking by human subjects. (c) The correlation between the falling risk ranking by two different random split groups of human subjects. 77

LIST OF TABLES

2.1 Geometric commonsense of type-I: distribution over absolute dimensions. L: length; W: width; H: height. 23

2.2 Numerical comparisons on surface orientation. 30

2.3 Numerical comparisons on segmentation. 31

3.1 Precision and recall of Volumetric completion. Comparison of three method: 1) voxel-based representation generated by Octree algorithm [SNI05], 2) voxels in surface and invisible space (sec. 2.2), and 3) our volumetric completion. 56

3.2 Results of inferring the fixed joints and support relations between primitives. Accuracy is measured by nodes of contact graph whose label is correctly inferred divided by the total number of labeled nodes. 56

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisor Dr.Song-Chun Zhu for his guidance and mentorship during my PhD study at UCLA. His high-level vision, passion and deep insight are always inspiring me during my Learning journey, and has helped me improve as a devoted, professional, and rigorous person. I would also like to thank my committee Dr.Hongjing Lu, Dr. Luminita Vese as well as Dr. Ying Nian Wu for their insightful advice and discussions along the way. My appreciation also goes to Dr. Xiaobai Liu for his inspired discussions and support during our collaborations.

Many thanks also go to my many colleagues in the VCLA lab for their friendship, collaboration, and support throughout my time here. In particular, I thank Yibiao Zhao, Yang lu, Xiaohan Nie, Yang Liu, Yuanlu Xu, Siyuan Huang, for working closely with me during the past years, all the credit of my work also belongs to them. I also extend my thanks to Tianfu Wu, Wenze Hu, Ping Wei, Brandon Rothrock, Yixin Zhu, Siyuan Qi, Hang Qi, Meng Meng, Tianmin Shu, Dan xie, Bo Li, Tao Yuan, Weixin Li, Nawin Waree, Caiming Xiong, Jifeng Dai and Quanshi Zhang.

Lastly, I cannot thank my parents and my wife Shuang enough for their unwavering support. None of this would have been possible without the love and patience of my family.

VITA

2014–2017 Ph.D. Candidate (Statistics), UCLA, Los Angeles, CA

2012–2014 M.S. (Statistics), UCLA, Los Angeles, CA

2002–2006 B.S. (Computer Science), Wuhan Institute of Technology, Wuhan, China

PUBLICATIONS

Joey C. Yu, Xiaobai Liu, Song-Chun Zhu, *"Unified Single-Image 3D Scene Parsing Using Geometric Commonsense"* ,IJCAI, 2017

Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, Song-Chun Zhu, *"Scene Understanding by Reasoning Stability and Safety"* ,International Journal of Computer Vision, S.I. Scene Understanding, 2015

Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, Song-Chun Zhu, *"Detecting Potential Falling Objects by Inferring Human Action and Natural Disturbance"*, ICRA, 2014

Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, Song-Chun Zhu, *"Cross-view Action Modeling, Learning and Recognition"*, CVPR, 2013

CHAPTER 1

Introduction

Scene understanding is a fundamental problem in computer vision. The structure of a scene can help improve other computer vision tasks such as object detection and object tracking by providing occlusion information. Researchers in artificial intelligence can also obtain information of geometry and physics in a scene to help the interaction between an agent and its environment.

1.1 Motivation

1.1.1 Scene Reconstruction Based on Commonsense and Attributed Grammar

The goal of computer vision, as coined by Marr [Mar82], is to compute what and where, which correspond to the tasks of recognition and reconstruction respectively. The former is often posed as parsing an image in a hierarchical representation, e.g., from sketches, semantic regions, objects, to scene categories. The latter recovers 3D scene structures, including camera parameters [Zha00], surface normals and depth [HEH08b], and local Manhattan world [CY03]. While the recognition and reconstruction problems are usually addressed separately or sequentially in the literature, it is mutually beneficial to solve them jointly in a tightly coupled framework for two reasons.

(1) 2D image parsing is capable of providing semantic contextual knowledge for pruning the uncertainties during 3D modelling. For example, if two neighbor pixels are classified the same label (e.g. building), it is likely that they are projections of the same 3D plane. In addition, semantic region labels, e.g. building or groundplane, often provide strong prior on surface normal.

(2) 3D reconstruction can provide additional geometric information to boost recognition. In the literature, there have been a number of efforts that utilizes geometry to help region segmentation [HZ03, LZZ14], objection detection [HEH08b], visual tracking [WZZ13] or event classification [PM11], etc.

To couple the two tasks, we propose an attribute grammar as a unified representation, which augments levels of geometric attributes (e.g., camera parameters, vanish points, surface normal etc.) to the nodes in the parse graph. Thus the recognition and reconstruction tasks are solved in a joint parsing process simultaneously. Fig. 1 shows a typical parsing result with seven planar surfaces plus a sky region and a high-quality 3D scene model.

We consider outdoor urban scenes that may contain multiple local Manhattan worlds (LMW) or mixture Manhattan world [SRF14], where, for example, buildings are composed of multiple planar surfaces and touch the ground on contact lines. In contrast to the widely used Manhattan world assumption [CY03], this paper considers a more general scenario that, the adjacent surfaces of a building may not be orthogonal to each other (see the main building in Fig. 1.1). Curved surfaces are approximated by piecewise linear splines. The surface is further decomposed into super-pixels and edge elements. These representational units can be naturally organized in a hierarchical parse graph with the root node being the scene and terminal nodes being the edges and super-pixels. Fig. 1.2 illustrates a parse graph.

Different from the widely studied appearance attributes of scenes in the vision literature [WJW13, PM11, ZZ11], our interest is in the geometric attributes for all the nodes in the parse graph. An edge segment has its associated vanishing point, and a super-pixel has a surface normal, a planar facet of a building has two vanishing points and a surface normal, and a building has 3 vanishing points, and finally the whole scene shares a set of camera parameters (focal length etc.). We amount these geometric attributes to the parse graph as is shown in Fig. 2. In this attribute parse graph, attributes of a node can be inherited by its offspring, and thus impose geometric constraints in the hierarchy. These constraints are expressed as additional energy terms in the parsing algorithm so as to maintain consistency in the hierarchy. Consequently, the parsing and reconstruction problems are solved in a tightly coupled manner. This attribute parse graph is different from, and can be integrated

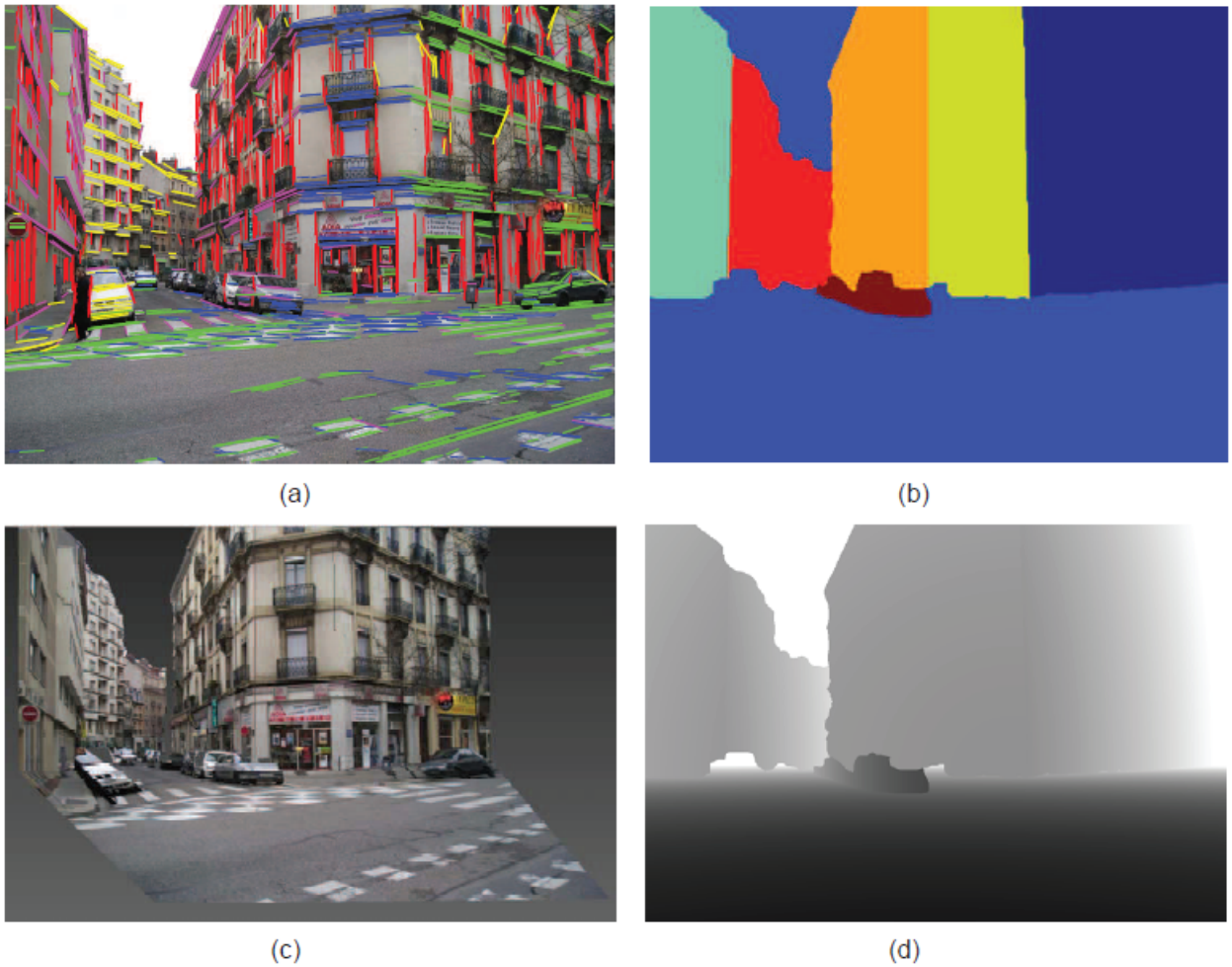


Figure 1.1: A typical result of our approach. (a) Input image overlaid with detected parallel lines; (b) surface normal map where each color indicates a unique normal orientation; (c) synthesized images from a novel viewpoint; and (d) depth map (darker is closer).

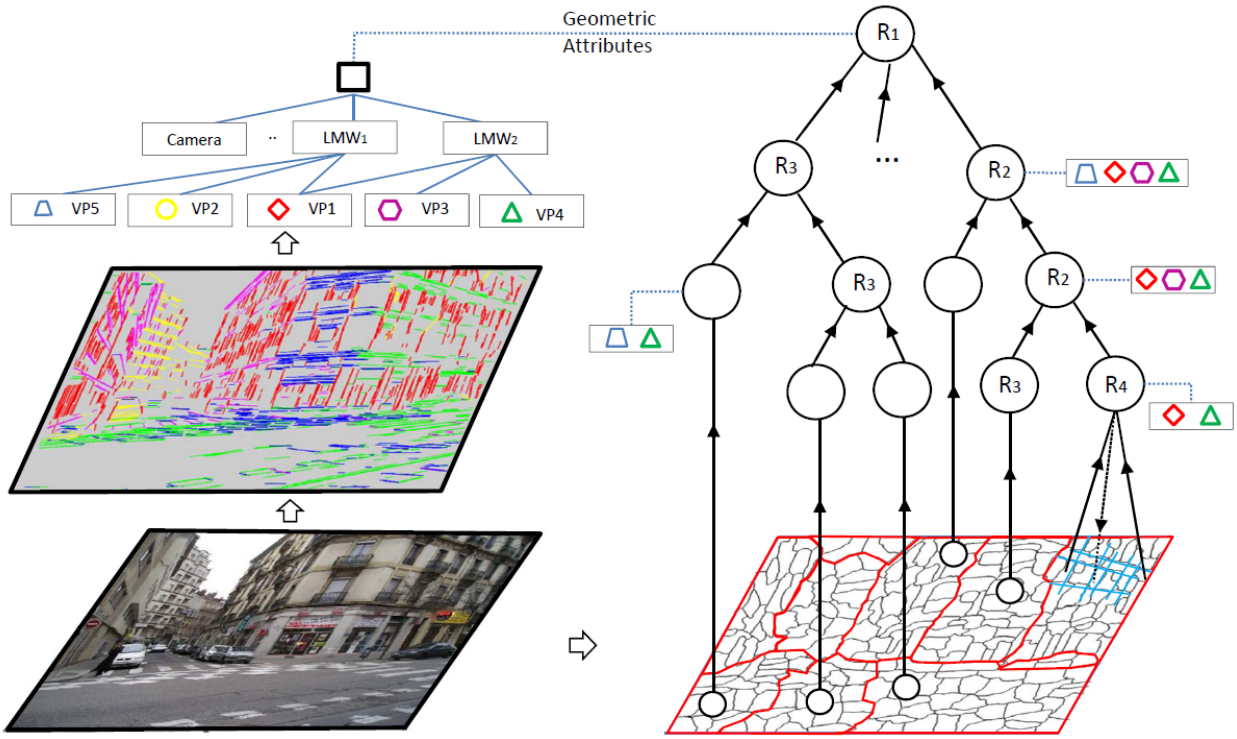


Figure 1.2: Parsing an image using attribute grammar. Left: global geometric attributes are associated with the root node (scene) of the parse graph, including focal length of camera, and Cartesian Coordination System defined by Manhattan frames. Right: parse graph augmented with local geometric attributes, such as surface normals and vanishing points (VPs) associated with a surface, or multiple vanishing points for a building. R1; :::;R5 are the five grammar rules for scene decomposition

with, other scene parsing problems, e.g., finegrained scene classification [WJW13] that uses appearance attributes cast sky, yellow field etc.

To construct the attribute parse graph, we define an attribute grammar which is a 5-tuple: $G = (\mathbf{V}_T, \mathbf{V}_N, S, \mathbf{R}, P)$. The set of terminal nodes $\mathbf{V}_T$ include surface fragments or superpixels, the non-terminal nodes $\mathbf{V}_N$ include planar surfaces, composite surfaces, building block and Manhattan world, the root node S is the scene, and $\mathbf{R}$ is the set of production rules, and P is the probability associated with the rules. Each node a 2 $\mathbf{V}_T$ (or A 2 $\mathbf{V}_N$) is associated with set of geometric attributes.

We observe that a few production rules (or compositions) are capable of explaining most of the outdoor urban scenes. We construct 5 production rules which are quite generic for urban scenes. Each rule $A \rightarrow A_1, \cdots, A_k$ represents a certain spatial arrangement between the children surfaces $A_1, \cdots, A_k$, and imposes constraints on the attributes of $X(A)$ and $X(A_1), \cdots, X(A_k)$.

These composition rules compete with each other to interpret the input image in a recursive way, which results in a parse graph as a valid interpretation of the scene. The parse graph includes both appearance models for 2D segmentation and geometric models for 3D reconstruction.

We formulate the inference of attribute parse graph from a single image in a probabilistic framework. The state space is the set of all possible attribute parse graphs with large structural variations. To efficiently sample this complex state space, we adopt the Data-Driven Markov Chain Monte Carlo paradigm [TZ02]. In particular, our inference method starts with an initial parse graph constructed by a greedy method, and then simulates a Markov Chain in the state space by a set of diffusion-jump dynamics [BZ05]. During the initialization stage, we utilize a heuristic search procedure for camera calibration, and introduce a belief propagation method to obtain region labelling which leads to an initial parse graph. During the following sampling stage, we introduce five dynamics that are paired with each other to exploit the joint solution space periodically, which can guarantee nearly global convergence [TZ02].

Commonsense or commonsense reasoning [DM15, DSS93] functions in all parties of Artificial Intelligence (AI), including language, vision, planning, etc. It basically studies the consensus reality, knowledge and reasoning available to the overwhelming majority of people and attracted a lot of attentions in the past literature. In the field of computer vision, however, it is still unclear how to formally describe visual commonsense knowledge, or how commonsense can be used to enhance the understanding of images or videos [ZP13]. This work aims to fill in this gap by studying the reasoning of geometric commonsense for 3D scene parsing. Such a parsing task aims to segment both low-level scene entities (e.g., straight edges, semantic regions) and object-level scene entities (e.g., human, vehicles) in 2D images, and estimate their geometric dimensions in the 3D world [HEH05, DBK13, WFU15a, LZZ14, MFY16]. Most existing 3D parsing algorithms [HEH08a] are designed for a particular type of scene categories, e.g., urban [LZZ14, GEH10], indoor [WFU15b]. However, a practical AI system, e.g., autonomous driving, usually needs to deal with a wide variety of scene categories.

Our solution to the above challenges is motivated by the fact that we human beings, unconsciously sometimes, utilize rich prior knowledge of the geometric dimensions of scene entities to understand the scene structures in images or videos [DSS93]. This knowledge can be roughly divided into two types: i) prior distributions over a single dimension of objects, e.g., the height of a female adult is about 1.75 meters, or that the length of a sedan is about 4.5 meters; ii) pair-wise comparisons between the dimensions of different scene entities at both object-level, e.g., human, windows, vehicles, etc., and part-level, e.g., straight edges, planar regions, etc. As illustrated in Figure 2.1, for example, the window edges on the same facade are parallel to each other and are orthogonal to the edges on the ground, a building is higher than a human, or the length of all sedans are roughly equal. These unary and pair-wise knowledges, once acquired, are valid across images and scene categories, and thus form the commonsense in geometric space, called *geometric commonsense*.

1.1.2 Scene Understanding by Reasoning Geometry and Physics

Traditional approaches for scene understanding have been mostly focused on segmentation and object recognition from 2D images. Such representations lacks important physical information, such as the 3D volume of the objects, supporting relations, stability, and affordance which are critical for robotics applications: grasping, manipulation and navigation. With the recent development of Kinect camera and the SLAM techniques, there has been growing interest in studying these properties in the literature [NIH11].

We present an approach for reasoning physical stability 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 [NIH11]. We utilize a simple observation that, by human design, objects in static scenes should be stable. 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 attached neighbors or fixed to its supporting base. In addition, while objects are stable physically, they should enjoy a movable space (freedom) for manipulation. Such assumption is applicable to all scene categories and thus pose quite powerful constraints for the plausible interpretations (parses) in scene understanding.

As Fig. 1.3 shows, our method consists of two main steps.

1) Geometric reasoning: recovering solid 3D volumetric primitives from defective point cloud. Firstly we segment and fit the input 2.5D depth map or point cloud to small simple (e.g., planar) surfaces; secondly, we merge convexly connected segments into shape primitives; and thirdly, we form 3D volumetric shape primitives by filling the missing (occluded) voxels, so that each shape primitive can own its physical properties: volume, mass and supporting areas to compute the potential energies in the scene. Fig. 1.3.(d) shows the 3D primitives in rectangular or cylindrical shapes.

2) Physical reasoning: 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 as shown in Fig. 1.3.(e), coloring this graph corresponds to grouping them

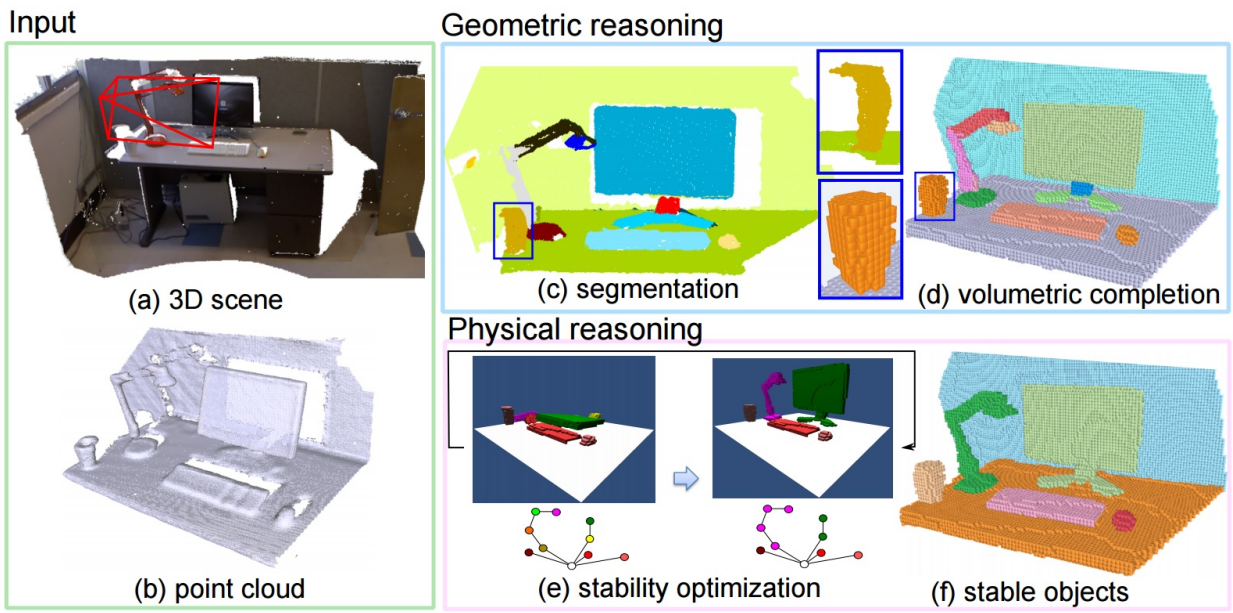


Figure 1.3: 1. Overview of our method. (a) 3D scene reconstructed by SLAM technique, (b) point cloud as Input. In geometric reasoning, (c) a portion is shown to be segmented by a segment-and-merge approach, with missing voxels, (d) solid primitives by volumetric completion. In physical reasoning, (e) the contact graph are labeled through stability optimization. (f). Final parsing results with stable objects.

into objects. For example, the lamp on the desk originally was divided in 3 primitives and will fall under gravity (see result simulated using a physics engine), and become stable when they are grouping into one object - the lamp. So is the computer screen with its base. To achieve the physical reasoning goal, we make the following novel contributions in comparison to the most recent work in dealing with physical space reasoning [GEH10, SHK12].

- We define the physical stability function explicitly by studying minimum energy (physical work) need to change the pose and position of an primitive (or object) from one equilibrium to another, and thus to release potential energy.
- We introduce disconnectivity graph (DG) from physics (Spin-glass) to represent the energy landscapes.
- We solve the complex optimization problem by the cluster sampling method Swendsen-Wang cut in image segmentation [BZ05] to maximize global stability.
- We collect a new dataset for large scenes by depth sensors for scene understanding and will release the data and annotations to the public.

In experiments, we demonstrate that the algorithm achieve a substantially better performance for i) object segmentation, ii) 3D volumetric recovery of the scene, and iii) better parsing result for scene understanding in comparison to state-of-the-art methods in both public dataset [SHK12] and our own new dataset.

1.1.3 Scene Understanding by Detection Potential Falling Objects by Inferring Human Action and Natural

The recent development of consumer-grade range cameras, such as the Kinect camera, has attracted increasing studies in the field of 3D scene understanding [AKJ13, JKJ13, SNI05, XF14]. However, most of existing work is focused on locating and naming the object in the scene, and leaves a big gap to answer human-level scene understanding questions, such as: how does a human interact with a scene? how does the scene response to the action? What and where are potential dangers in the environment?

We present an potential falling object detection algorithm, which is an essential component of a safetyaware robot. As shown in Fig.1.4, the algorithm is useful for three main scenarios:

i) Safety surveillance robots. Objects have the potential to fall onto or hit people at the construction site as the warning sign shown in Fig.1.4 (a). To prevent objects from falling freely from one level to another, the safety risk surveillance ensures that objects are being stored where a secure physical barrier provided.

ii) Human assistant robots for children, elders and people with disabilities. As the example shown in Fig.1.4 (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. The Fig.1.4 (c) showed postdisaster scene captured by a 3D range sensor. It was a extremely dangerous environment due to the M9.0 earthquake and tsunami in Japan. A robot working in such environments requires to understand the potential risks due to many objects at unstable state.

1.2 Related Work

The attributes grammar for 3D reconstruction is closely related to the following four research streams in computer vision.

Semantic scene labelling has been widely studied to deal with appearance variations, low-resolution and semantic ambiguities. A popular choice is the Conditional Random Fields [LMP01] model that describes qualitative contextual relations between region segments. Such relations are proved to be helpful in the recognition of outdoor objects. Choi et al. [CTW12] further studied a 2D context model to guide detectors and produced a semantically coherent interpretation for the given image. Felzenszwalb and Vekser [FV10] applied the dynamic programming method for pixel labelling of 2D scenes with simple tiered structure. These methods formulate scene labelling as a pixel-wise labelling problem which

Figure 1.4: The detection of potential falling objects is an essential ability of a safety-aware robot: (a) the safety surveillance robot for a construction site, (b) the human assistant robot for the baby proofing, and (c) A building where was crashed by earthquake and tsunami on March 11, 2011, Japan.

however ignores the hierarchical and recursive composition relations between regions. In contrast, our method models semantic regions using a hierarchical parse graph which can be used to understand the input image at different levels, from pixels to regions to scene layout.

Single-View 3D modelling has been extensively studied in previous literature. Han and Zhu [HZ03] studied a generative model for reconstructing objects and plants from a single-view. Hoiem et al. [HEH05] explored rich geometric features and context information to recognize normal orientation labels of 2D regions, and Heitz et al. [HGS09] further proposed to recognize geometric surface labels and semantic labels simultaneously in a supervised way. Gupta et al. [GEH10] considered 3D objects as blocks and inferred their 3D properties such as occlusion, exclusion and stability. However, these methods were built on the classification of 2D segmentation, and thus did not directly reconstruct 3D or infer depth values. Mobahi et al. [MZY11] reconstructed a single view by extracting low rank textures on building faade. Saxena et al. [SSN09] and Haene et al. [HZC13] ever studied a fully supervised model to build mappings between informative features and depth values. Schwing et al. [SU12] presented an exact inference method (i.e. branch-and-bound) for single-view indoor scene reconstruction. Pero et al. formulated the 3D reconstruction of room in a Bayesian framework and proposed a sampling method for inference [DGB11]. Ladicky et al. [ZP14] proposed a discriminatingly trained boosting method for estimating surface normal.

Joint Recognition and Reconstruction has been investigated for a number of computer vision tasks. Haene et al. [HZC13] presented a continuous-discrete formulation for jointly solving scene reconstruction and labelling of images of multiple views. Ladicky et al. [ZP14] proposed to train a depth-wise classifier for each class, used to predict semantic classes and depth maps for a single image. Their method requires groundtruth depth maps for training. Carbral et al. [CF14] tried to recover planar-wise 3D scene model from panorama images of office areas.

The other studies include jointly solving object recognition and object modelling. Haene et al. [HSP14] proposed to learn 3D shape priors from surface normals which has been proved to be very successful. Hejrati et al. [HR14] proposed to synthesize 2D object instances

from 3D models and used the instances to help solve object recognition task. Schwing et al. [HSP14] introduced a method for recovering 3D room layout and objects simultaneously. Xiao et al. presented a supervised method for localizing 3D cubois in 2D images [XRT12] . They also introduced a benchmark [XOT13] for joint Structure-from-Motion and Object Labelling. Payet and Todorovic [PT11] proposed a joint model to recognize objects and estimate scene shape. Zhang et al. [ZST14] proposed to reconstruct a room using Panoramic images by exploiting both object parsing (e.g. table detection) and scene geometry (e.g. vanishing points).

Scene grammar. Koutsourakis et al [KST09] proposed a shape grammar to reconstruct building faades. The proposed model focused on rectifying faade images but not recovering 3D geometry. Han and Zhu [HEH05], Zhao and Zhu [ZZ11] and Pero et al. [DBK13] built generative scene grammar models to model the compositionality of Manhattan structures in the indoor scenes. Furukawa et al. [FCS09] studied the reconstruction of Manhattan scenes from stereo inputs. In contrast, we relax the Manhattan assumption and generalize the scene grammar model to handle more complex and cluttered outdoor environment. We contribute a hierarchical representation for urban scene modelling and augment it with both semantic and geometric attributes.

In comparison with the literature, my work makes the following contributions: (1) We present a grammatical model with geometric attributes that tightly couples the image parsing and 3D scene reconstruction tasks. (2) We develop a stage-wise sampling inference method that is capable of exploiting the constrained space efficiently. (3) In experiments on both public datasets and our self-collected datasets, our method achieves considerably better performances than the existing methods in terms of both 2D parsing and 3D reconstruction.

The scene reconstruction by commonsense is closely related to four research streams in Computer Vision and Artificial Intelligence (AI).

Commonsense Reasoning is one of the long-standing tasks in AI [DM15, DSS93] and has recently attracted a lot of attentions in the field of computer vision. Such commonsense knowledges were used as context information to enhance visual recognition [FDG14, GGV11],

scene understanding [WJW13], activity recognition [WPP07, WPC05, RSN13] and affordance prediction [ZFF14, ZZC15, WPP07]. However, there is still no formal representation of visual commonsense in the past literature, which restricts the generalization capability of the developed techniques. Moreover, these works share the same insight that visual commonsense only functions in a high-level semantic understanding of images, rather than low-level pixel-wise understanding, which are not necessarily bold. In contrast, this paper studies geometric commonsense extracted from both low-level and high-level scene entities and defines a unified presentation for describing such knowledges.

3D scene reconstruction methods are mostly referred to Shape-from-X, where X stands shading [ZTC99], contour [TNC13], focus/defocus [NN90], texture [CZ00], motion [DST00], photometric stereo [LK81] etc. Single-view reconstruction has also been studied [HGS09, HEH05, GEH10]. Most recently, Liu et al. [LML16, LZZ14] proposed an explicit image representation and presented a joint solution of 2D recognition and 3D reconstruction. These efforts utilized various modeling assumptions to guide the reconstruction process. For example, the shape-from-texture methods [DM15] assume that the scene comprises of homogeneous texture, and the shape-from-contour methods assume that contours are known to be projections of curves on a planar surface. These methods would fail to work while dealing with complex scenes for which none of a single modeling assumption is valid. For example, the methods by Liu et al. [LML16, LZZ14] can only used to reconstruct Manhattan or near Manhattan type scenarios in urban scenes. In this work, we introduce a concept-of-proof framework for geometric commonsense reasoning and demonstrate how this knowledge can be used to enable the 3D reconstruction of a wide variety of scene types.

Stochastic Image Grammar has been applied for a number of image parsing problems in computer vision for scene understanding. Koutsourakis et al. [KST09] proposed a shape grammar to explain building facades with levels of details. Researchers have [HZ09, DGB11, ZZ11] specified generative scene grammar to model the compositional of Manhattan structures in images. Furukawa et al. [FCS09] studied the reconstruction of Manhattan scenes from stereo inputs. Liu et al. [LZZ14] proposed an attribute grammar for 3D scene modeling to enable compact representation of images. With only a few grammar rules, the

grammar model can explain most urban images and achieved state-of-the-art performance on a few public image benchmarks. In this work, we will extend the attribute grammar to model geometric commonsense knowledge of scene entities for the reconstruction of various scene categories. We formulate the above objectives in Bayesian framework in order to keep uncertainties during inference, which is critical to avoid pre-mature decision making. For inference, we develop an iterative data-driven Monte Carlo Markov Chain (DDMCMC) method [TZ02] that searches the optimal parse graph with both bottom-up and top-down computations. On the one hand, we partition images into compositional elements, extract visual features for elements (e.g., color) and measure their pair-wise similarities, and group elements in a bottom-up fashion to create upper-level graph nodes. On the other hand, we can decompose a graph node into multiple children nodes or propagate the attributes of a parent node to all its offspring in a top-down fashion. Both bottom-up and top-down computations are intelligently scheduled by the Metropolis Hasting Principle [TZ02] to guarantee convergence toward the posterior probability. To evaluate the proposed method, we collect an unbiased image dataset that covers five different categories: country, road, suburban, campus and urban, and manually annotate their semantic and geometric labels in 3D. This dataset is different from the previous datasets [LZZ14, HEH08a] which mostly include one or two types of scenes. Results with comparisons demonstrated that the proposed method clearly outperforms the alternative methods in the recent literature [LZZ14, LML16].

The study of falling objects can be traced back to an early work by Kriegman [Kri97] that first proposed an algorithm to calculate the capture regions where a 3D object may fall according to the Morse theory. There is a recent rise of related studies in following seven streams:

1. Geometric reasoning. Our approach for geometry reasoning is related to a set of segmentation methods (e.g., [JKJ13, AFS06, PVB08]). 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 reason, Attene et al. [AFS06] extracts 3D geometric primitives (planes or cylinders) from 3D mesh. In comparison, our method is more faithful to the original geometric shape of object in the point cloud data. There have been also

interesting work in constructing 3D scene layouts from 2D images for indoor scenes, such as Zhao and Zhu [ZZ11], Lee et al. [LHK09, GHK10], Hedau et al. [HHF10]. Furukawa et al. [FCS09] 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 single image for the "block world" in the past three decades [BMR82, GEH10, GSE11, ZZ11, LHK09, GHK10]). E.g. Biederman et al. [BMR82] studied human sensitivity of objects that violate certain physical relations. Our goal of inferring physical relations is most closely related to Gupta et al. [GEH10] who infer volumetric shapes, occlusion, and support relations in outdoor scenes inspired by physical reasoning from a 2D image, and Silberman et al. [SHK12] who infer the support relations between objects from 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 supervised learning process, our method is able to infer the 3D objects with maximum stability.

3. Intuitive physics model. Recent psychology studies suggested that approximate Newtonian principles underlie human judgments about dynamics and stability [FBB10, HBT11]. Hamrick et al. [HBT11] showed that knowledge of Newtonian principles and probabilistic representations are generally applied for human physical reasoning, and 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. Physics engines in graphics can accurately simulate the motion of objects under gravity, but it is computationally expensive for the purpose of measuring object stability.

4. Safe Motion Planning. As the planning is a classic problem in robotics, Petti and Fraichard [PF05], Phillips and Likhachev [PL11] tackled the problem of safe motion planning in the presence of moving obstacles. They consider the moving obstacles as the real-time constraint inherent to the dynamic environment. However, we first argue that a

robot need to be aware of potential dangers even in a static environment due to possible incoming disturbances.

5. Physics based model. Gupta et al. [GEH10] revisited the block world model and worked on labeling of the 2D image by reasoning the physical force based on a block representation of 2D image segments. Lee et al. [GHK10], Zhao and Zhu [ZZ11], has made promising progress on volumetric reasoning of 2D indoor scene. Recently, Zheng et al. [ZZY13] and Jia et al. [JGS13] proposed very interesting approaches to segment point clouds and detect 3d objects by incorporating the physics stability as a prior.

6. 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. [GGV11] recognizes chairs by imagining an "sitting" actor interacting with the scene. Gupta et al. [GSE11] predicts the "workspace" of a human given a estimated 3D scene geometry. Fouhey and Delaitre et al. [DFL12, FDG14] demonstrate that observing people performing different actions can signifi- cantly improve estimates of scene geometry and scene semantics. Jiang [JKS13, JS13] proposed scene labeling algorithms by considering humans as the hidden context.

7. Cognitive studies. Psychology studies suggested that approximate Newtonian principles underlie human judgments about dynamics and stability [FBB10, ZL05]. Hamrick et al. [HBT11] showed that knowledge of Newtonian principles and probabilistic representations are generally applied for human physical reasoning, and the intuitive physics model is an important perspective for human-level complex scene understanding.

CHAPTER 2

Unified Single-Image 3D Scene Parsing Using Geometric Commonsense

2.1 Introduction

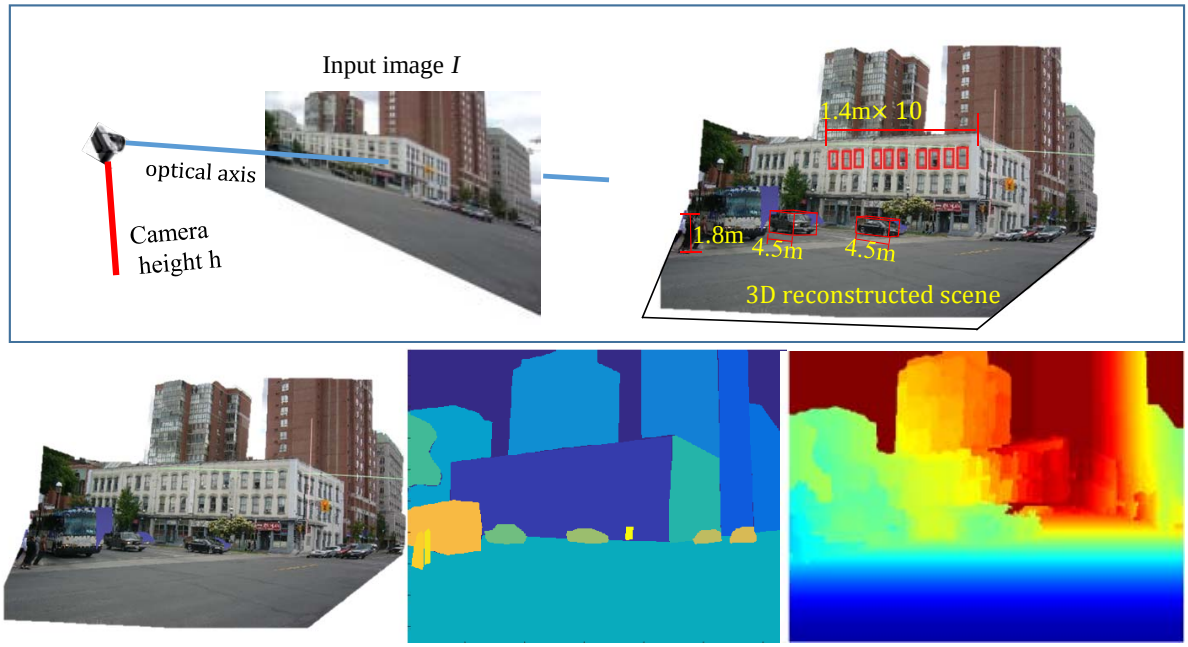


Figure 2.1: Single-view 3D scene reconstruction using Geometric commonsense. Top: the world is full of commonsense over geometric dimensions, e.g., that a sedan is about 4.5 meters long. Bottom: exemplar result of the proposed method, including synthesized image (left), planar segmentation (middle), and depth map (right).

Commonsense or commonsense reasoning [DM15, DSS93] functions in all parties of Artificial Intelligence (AI), including language, vision, planning, etc. It basically studies the

consensus reality, knowledge and reasoning available to the overwhelming majority of people and attracted a lot of attentions in the past literature. In the field of computer vision, however, it is still unclear how to formally describe visual commonsense knowledge, or how commonsense can be used to enhance the understanding of images or videos [ZP13]. This work aims to fill in this gap by studying the reasoning of geometric commonsense for 3D scene parsing. Such a parsing task aims to segment both low-level scene entities (e.g., straight edges, semantic regions) and object-level scene entities (e.g., human, vehicles) in 2D images, and estimate their geometric dimensions in the 3D world [HEH05, DBK13, WFU15a, LZZ14, MFY16]. Most existing 3D parsing algorithms [HEH08a] are designed for a particular type of scene categories, e.g., urban [LZZ14, GEH10], indoor [WFU15b]. However, a practical AI system, e.g., autonomous driving, usually needs to deal with a wide variety of scene categories.

Our solution to the above challenges is motivated by the fact that we human beings, unconsciously sometimes, utilize rich prior knowledge of the geometric dimensions of scene entities to understand the scene structures in images or videos [DSS93]. This knowledge can be roughly divided into two types: i) prior distributions over a single dimension of objects, e.g., the height of a female adult is about 1.75 meters, or that the length of a sedan is about 4.5 meters; ii) pair-wise comparisons between the dimensions of different scene entities at both object-level, e.g., human, windows, vehicles, etc., and part-level, e.g., straight edges, planar regions, etc. As illustrated in Figure 2.1, for example, the window edges on the same facade are parallel to each other and are orthogonal to the edges on the ground, a building is higher than a human, or the length of all sedans are roughly equal. These unary and pair-wise knowledges, once acquired, are valid across images and scene categories, and thus form the commonsense in geometric space, called *geometric commonsense*.

This chapter presents a stochastic attribute scene grammar for representing both visual content and geometric commonsense constraints for parsing a single image in 3D world. Our grammar model recursively decompose a scene into a small number number of scene primitives (e.g., straight lines, planar regions, vehicles) arranged by a set of spatial relations. This results in a hierarchical representation, called parse graph, which has a root node

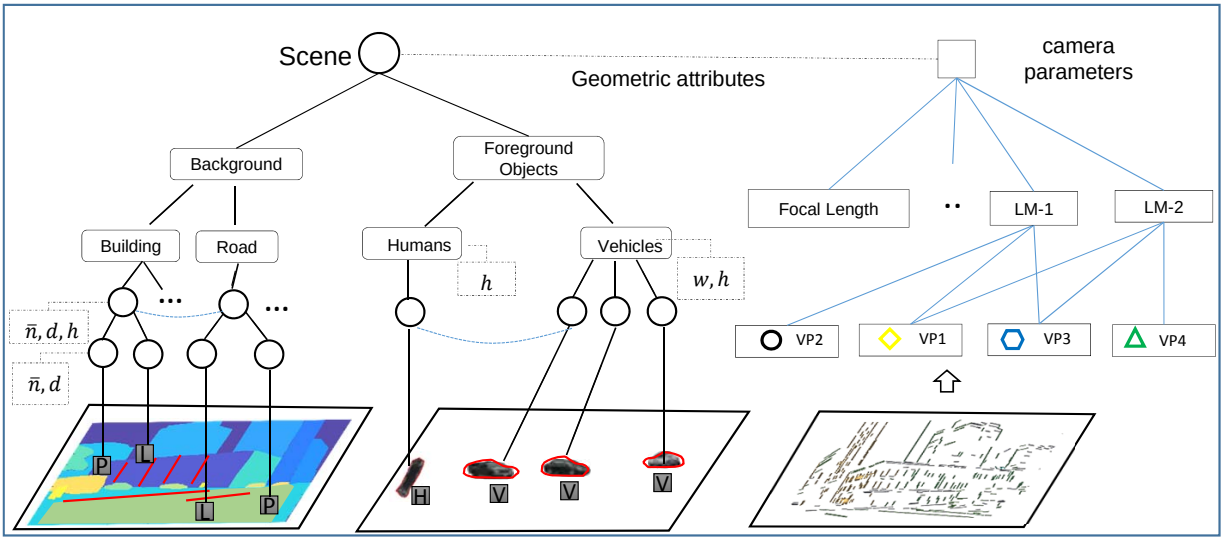


Figure 2.2: Attribute Parse Graph. Every graph node represents a scene entity, and is augmented with a set of geometric attributes ($\hat{l}$, line direction; d, depth;$\bar{n}$, normal orientation; w, width; h, height; f, focal length; θ, camera angle). There are five types of terminal nodes: L, line, P, planar surface; T, texture; H, human; V, vehicles.

for the whole scene and a terminal node for each scene primitive. Each graph node is described with a set of geometric attributes, e.g., 3D position, normal direction etc. We augment the attribute variables using a set of common sense knowledge that are either pre-defined or mined from online databases. These commonsense are defined over geometric dimensions of various categories (e.g., human height), and are widely used by human beings to understand visual inputs. Since these geometric commonsense knowledges are generic enough, our method is capable of reconstructing a wide variety of scene types, e.g., urban, suburban, campus etc. In contrast, most existing methods for single-view reconstruction were developed for a particular scene type (as reviewed in Section 1.2).

We formulate the above objectives in Bayesian framework in order to keep uncertainties during inference, which is critical to avoid pre-mature decision making. For inference, we develop an iterative data-driven Monte Carlo Markov Chain (DDMCMC) method [TZ02] that searches the optimal parse graph with both bottom-up and top-down computations. On the one hand, we partition images into compositional elements, extract visual features

for elements (e.g., color) and measure their pair-wise similarities, and group elements in a bottom-up fashion to create upper-level graph nodes. On the other hand, we can decompose a graph node into multiple children nodes or propagate the attributes of a parent node to all its offspring in a top-down fashion. Both bottom-up and top-down computations are intelligently scheduled by the Metropolis Hasting Principle [TZ02] to guarantee convergence toward the posterior probability. To evaluate the proposed method, we collect an unbiased image dataset that covers five different categories: country, road, suburban, campus and urban, and manually annotate their semantic and geometric labels in 3D. This dataset is different from the previous datasets [LZZ14, HEH08a] which mostly include one or two types of scenes. Results with comparisons demonstrated that the proposed method clearly outperforms the alternative methods in the recent literature [LZZ14, LML16].

2.2 Our Approach

The objective of this work is to parse a single image into a set of scene entities, e.g., straight lines, surfaces, objects, etc., and reason their geometric attributes in 3D, e.g., 3D positions, normal direction etc. We consider a wide variety of scene types, e.g., urban, garden, or road. In the rest of this section, we introduce image representation, Bayesian formulation, and inference algorithm in turn.

2.2.1 Stochastic Scene Grammar

An attribute scene grammar is specified by a 5-tuple: $G = (V_N, V_T, S, R, P)$ where V_N and V_T are the sets of non-terminal nodes and terminal nodes respectively, S is the initial node for the scene, R is a set of production rules for spatial relationships, P is the probability for the grammar. Our grammar model results in a hierarchical graph representation, i.e. **Parse Graph**, to represent the semantic content of an image. Every graph node represents a scene primitive, including straight lines, planar regions, homogeneous or inhomogeneous texture regions, human, and vehicles, and their composites. The first three image primitives are conventionally considered as background while the other two entities are foreground objects.

Figure 2.2 illustrates a typical parse graph used in this chapter. Single-view 3D scene parsing is equivalent to creating a plausible parse graph from the input image. Note that in this work we focus on parsing outdoor images, but the proposed technique can be easily extended to deal with indoor images as well.

For every graph node, we introduce a set of geometric **Attributes** to describe their dimensions in 3D, as summarized below.

- Attributes of a straight line include its position and orientation direction in 3D. In addition, we divide all straight lines into two categories: parallel lines and non-parallel lines [LZZ14]. Multiple orthogonal families of parallel lines forms one of the Manhattan frames [LZZ14].
- Attributes of a planar or texture region include its geometric properties, i.e., position, normal direction and size in 3D, and semantic labels, e.g., ground, building, grass, road, etc.
- Attributes of a human include one's geometric properties, i.e., positions and human height in 3D, and fine-grained semantic labels, i.e., genders, children/adult, races.
- Attributes of vehicles include positions and dimensions (length, width, height) in 3D, and catalog information, i.e., categories (sedan, car, bus).

Figure 2.2 visualizes an attribute graph where each node is augmented with a set of geometric attributes.

Geometric Commonsense are defined over the geometric attributes of graph nodes. There are two types of geometric commonsense.

- Type-I: prior distributions over dimensions of image entities. This type of knowledge includes, for instance, the average height of adult, the width of door, or the width of door. Table 2.2.1 summarizes some of these statistics (mean and standard deviation) collected by the U.S. Department of Health and Human Services[1]. These statistics can

[1] U.S. Department of Health and Human Services. Anthropometric Reference data for children and adults: united states, 2007-2010. Vital and health statistics, series 11, number 252, October, 2012

Category	Variable	Sub-Type	MEAN	STD
Human	Height	Female	1.6475	0.1934
		Male	1.7741	0.2217
Vehicle	(L, W, H)	Sedan	(4.6585, 1.8114, 1.4314)	(0.2395, 0.0434, 0.0621)
		Pickup	(6.0961, 2.0304, 1.8711)	(1.5121, 0.2487, 0.2413)
		SUV	(4.9414, 1.9574, 1.7275)	(0.4163, 0.1084, 0.0863)
Vehicle	Length	Sedan	4.6585	0.2395
		Pickup	(6.0961, 2.0304, 1.8711)	(1.5121, 0.2487, 0.2413)
		SUV	(4.9414, 1.9574, 1.7275)	(0.4163, 0.1084, 0.0863)
Window	(H, W)	-	(1.5234, 1.4647)	(0.3471, 0.8462)
Door	(H, W)	-	(2.0163, 0.8514)	(0.0689, 0.0726)

Table 2.1: Geometric commonsense of type-I: distribution over absolute dimensions. L: length; W: width; H: height.

be used to regularize the creation of the desired parse graph.For the measurement of image entities, we need the camera parameters, and project the 2D paring result into 3D space.Figure 2.3 plots the basic geometry used for measurement. In general, it's enough if we have the camera parameters and the contact information. This has been studied by Criminisi[CRZ00].

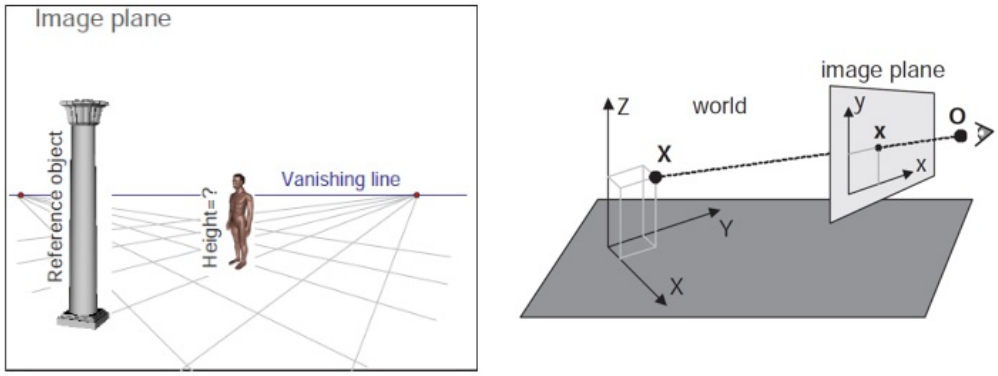


Figure 2.3: Here is an example for measurement. Given the height of the reference object (e.g. a pillar), the goal is to estimate the height of the man.

- Type-II: pair-wise relationship between scene entities. This type of knowledge is defined

over the comparisons of dimensions of two scene entities that are either from the same or different categories. This leads to a set of commonsense equations, each of which is defined as a 5-tuple: (entity-1, attribute-1, operator, entity-2, attribute-2), where "operator" represents, for example, "parallel", "orthogonal", "equal", or ">". Figure 2.4 lists the 21 pair-wise relationships used in this chapter.

1. (line, 'location', 'co-linear', line, 'location')
2. (line, 'direction', 'parallel', line, 'direction')
3. (line, 'direction', ''orthogonal, line, 'direction')
4. (line, 'location', 'co-planar', planar, 'location')
5. (line, 'direction', 'parallel', planar, 'normal')
6. (line, 'direction', 'orthogonal', planar, 'normal')
7. (planar, 'location', 'co-planar', planar, 'location')
8. (planar, 'normal', 'parallel', planar, 'normal')
9. (planar, 'normal', 'orthogonal', planar, 'normal')
10. (planar, 'normal', 'parallel', texture, 'normal')
11. (planar, 'normal', 'orthogonal', texture, 'normal')
12. (planar, 'location', 'contact', vehicle, 'cubiod')
13. (planar, 'location', 'contact', human, 'cubiod')
14. (planar, 'normal', 'parallel', human, 'direction')
15. (texture, 'location', 'contact', texture, 'location')
16. (vehicle, 'length', '=', vehicle, 'length')
17. (vehicle, 'width', '=', vehicle, 'width')
18. (vehicle, 'height', '=', vehicle, 'height')
19. (vehicle, 'height', 'parallel', vehicle, 'orientation')
20. (vehicle, 'length', '>', human, 'height')
21. (human, 'height', '=', human, 'height')

Figure 2.4: Geometric commonsense of Type-II: a total of 21 pair-wise relationships between scene elements.

2.2.2 Bayesian Formula

Given an input image, our goal is to create an attribute parse graph that is a valid interpretation of the input image. Let G denote the parse graph, we solve the optimal parse graph by maximizing a posterior (MAP): $G^* = \arg\max_G P(G|I)$ where I is the input image.

According to Bayes' rule, we rewrite the posterior distribution $P(G|I)$ as

$$P(G|I) \propto P(G)P(I|G) \tag{2.1}$$

where $P(G)$ is the prior model and $P(I|G)$ is the likelihood model.

The prior model is defined over the validness of commonsense constraints. Let $A.CSet$ denote the set of nodes which are are linked to the node A, $X(A)$ be the attributes of A. For a graph node, let i index the type-I commonsense distributions, j index the type-II commonsense equations. Then, we can define the prior model as,

$$P(G) = \frac{1}{Z}\exp\{-E(G)\} \tag{2.2}$$

where Z is the normalization constant. The energy function is defined as:

$$E(G) = \sum_{A\in G}\sum_{i}\mathbf{h}_i(X(A)) + \sum_{B\in A.CSet}\sum_{j}\mathbf{g}_j[X(A), X(B)] \tag{2.3}$$

where $\mathbf{h}_i()$ denotes the normal distribution over the i-th attribute (with mean and variance in Table- 2.2.1). The model $\mathbf{g}_j(X(A), X(B))$ is defined as

$$\mathbf{g}_j(X(A), X(B)) = C \cdot \mathbf{1}(X(A), X(B), j) \tag{2.4}$$

where C is a constant, and $\mathbf{1}(X(A), X(B), j)$ returns 1 if the j-th commonsense equation between A and B holds; otherwise 0. Note that given a parse graph G and its attributes, it is straightforward to calculate $\mathbf{h}_i()$ and check $\mathbf{g}_j()$ by definitions.

The data likelihood model $P(I|G)$ is defined over the terminal nodes of the parse graph G. In this work, there are five types of terminal nodes: line (L), planar (P), texture (T), human (H) and vehicles (V). For a graph node A, let $A.type \in \{L, P, T, H, V\}$ return the type of terminal nodes. We specify a likelihood model:

$$P(I|G) = \prod_{A\in V_T, k=A.type}\frac{Pg^k(A|I)}{Pb^k(A|I)}\prod_{B\in I-V_T}Pb^k(B|I) \tag{2.5}$$

where $Pg^k()$ and $Pb^k()$ denote the foreground distributions and background distributions. The likelihood ratio $\frac{Pg^k(A|I)}{Pb^k(A|I)}$ is defined for each of the five terminal node types. In implementation, we approximate the ratio using detection scores of terminal nodes. If A is a straight line, for example, we define $\frac{Pg^k(A|I)}{Pb^k(A|I)}$ to be the edge confidences by [LPY12]; if A represents a human or a vehicle, we use the output confidences by the object detection method [FGM10]. In addition, our method allows regions not covered by any terminal nodes, and aims to explain these regions using the background distributions. In particular, we define $Pb^k(B|I)$ using the confidences of classifying B as the k-th terminal.

2.2.3 Bottom-up and Top-Down Inference

Given an input image, our inference aims to create a plausible attribute parse graph so that all attributes satisfy the geometric commonsense constraints, including both type-I and type-II knowledge. This is an intractable problem since the optimal parse graph is defined in a joint space: discrete labels (e.g., segmentation) and continuous attributes (e.g., location or orientation). To search the optimal graph, we develop a data-driven Markov Chain Monte Carlo (DDMCMC) method [TZ02] which starts with an initial graph and then reconfigures the graph with a set of dynamics to simulate a Markov Chain in the joint solution space. Two dynamics are paired with each other to guarantee the convergence to the target distribution $P(G|I)$. Our algorithm follows the Metropolis-Hastings strategy [TZ02]. Given the current graph G, we apply a dynamic to get a new graph G', and accept it with the following probability,

$$\alpha(G \to G') = \min(1, \frac{P(G'|I)Q(G \to G')}{P(G|I)Q(G' \to G)}) \tag{2.6}$$

where $Q()$ is the proposal probability.

The initial graph includes a root node and a set of terminal nodes. Then we introduce four dynamics that are performed in either bottom-up or top-down fashion.

Dynamics 1-2: birth/death of nonterminal nodes are used to create or delete a nonterminal node and thus transition the current parse graph to a new one. To create a new graph node, we need to select two or more candidate graph nodes to group with three

Algorithm 1 Algorithm: Building Parse Graph via attribute Grammar

1: Input: Single Image I;

2: Partition I into superpixels;(image)

3: Detect families of parallel lines and VPs;

4: Initialize labeling of superpixels by minimizing E(G);

5: Initialize the parse graph G;

6: Iterate until convergence, (according to previous transition probability and likelihood we sample new pares graph descted whether accept base on the accept rate)

- Randomly select one of the four MCMC dynamic
- Make dynamic proposals accordingly to reconfigure the current parse graph
- Accept the change with a probability

criteria: i) being spatially adjacent; ii) belonging to the same semantic category; or iii) conveying type-II commonsense knowledge. These would lead to a set of candidate nodes, and we select one of them as a new graph node. The proposal probability is defined over the detection scores of these graph nodes. To delete a graph node, we need to randomly select a graph node and then remove it to reconfigure the graph. Each candidate in this list is represented by its probability. Let B_i denote the ith candidate for the terminal node detection score. The proposal probability for selecting B_i calculate as follows:

$$Q(G \to G') = \frac{P(I, X(B_i))}{\sum_j P(I, X(B_j))} \tag{2.7}$$

Similarly, we obtain another set of candidate nodes to delete based on detection score.The proposal probabilities for deleting the node D_i is calculated as follows:

$$Q(G \to G') = 1 - \frac{P(I, X(D_i))}{\sum_j P(I, X(D_j))} \tag{2.8}$$

Dynamics 3: changing attribute is used to modify the attributes of graph nodes. We will randomly select a graph node as well as an attribute, and assign a different value to this attribute, e.g., normal orientation for planar regions. The changes, as introduced, will be accepted with probability. We set the proposal probability for this dynamic to be uniform.

Dynamics 4: attribute propagation is a top-down process that assigns the attributes of parent nodes to children nodes, and used to guarantee the consistency in the hierarchy. To do so, we will randomly select a parent node, and propagate all its attributes to the offspring nodes. Note that the previous works [LZZ14] only have bottom-up computations which might get stuck during inference because its time consuming to flip the attributes of a subtree only using bottom-to-up dynamics.

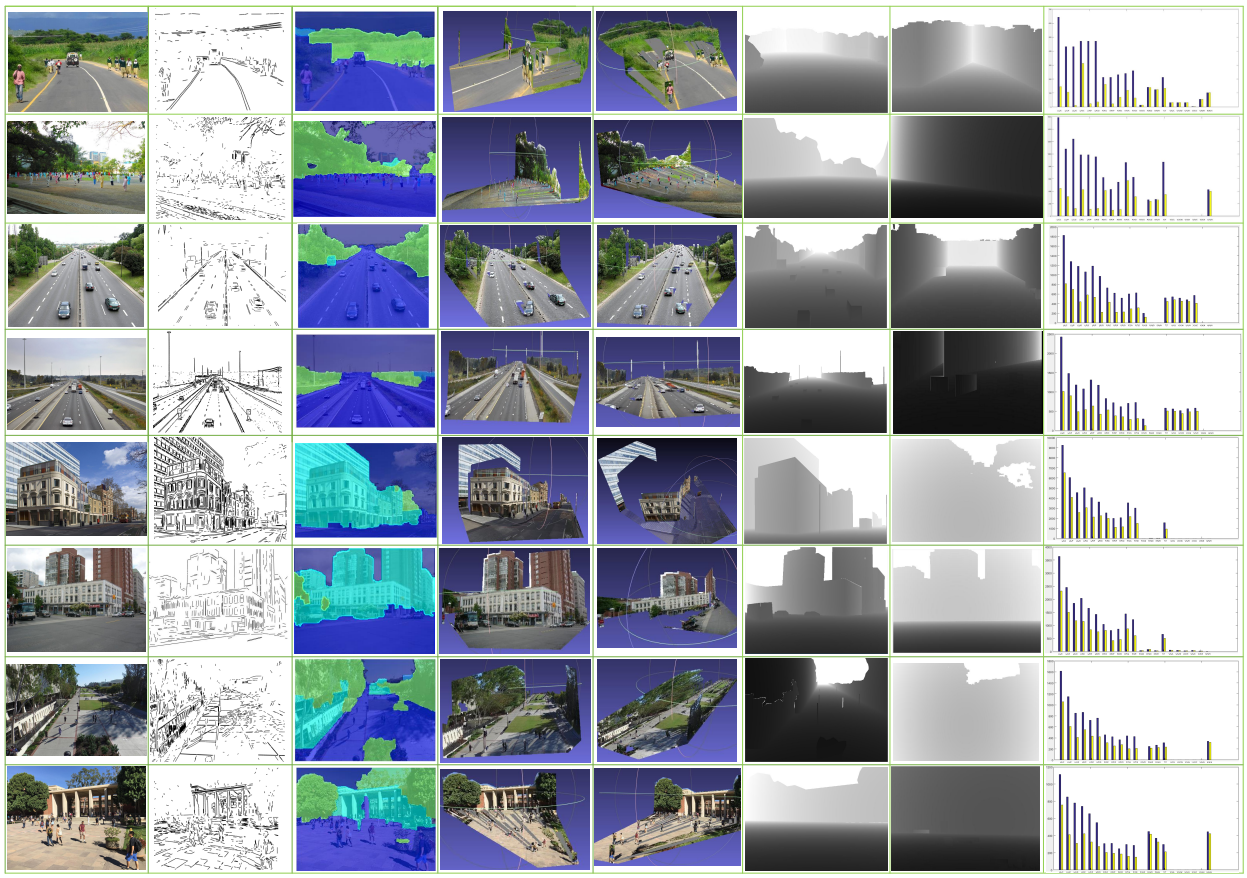


Figure 2.5: Exemplar results of the proposed method.Column-1: input images; Column-2: edge map; Column-3: semantic region map; Column-4&5: two novel viewpoints synthesized; Column-6: depth map; Column-7: depth map recovered by; Column-8: occurrence frequencies of equations (in blue) and occurrence frequencies of valid equations (in yellow), where the horizontal axis indicates the 21 equations shown in Fig. 2.4

2.3 Experiments

Dataset To evaluate the generalization capability of the proposed method, we collect 2000 images for each of the following categories: 1) country; 2) suburban; 3) road; 4) campus; and 5) urban, resulting in a collection of 10,000 images. The urban images mostly follow the Manhattan assumption [LZZ14] while the county images do not. These images are selected from existing datasets [EEV15, HEH08a, LZZ14]. For every category, we use 100 images for training and use the rest for testing. We manually annotate semantic and geometric labels for every image. To label an image, we first divide every image into three main categories: ground, sky and vertical, and further divide the vertical category into porous, solid and oriented surfaces. The number of planar orientations is equal to the number of horizontal vanishing points, as defined in [LZZ14]. For quantitative comparisons, all images are manually annotated with vanishing points (VPs), surface segmentation and surface orientation (represented by the correspondent VPs for each surface).

Baseline We compare the proposed method with three popular methods by Hoiem et al. [HEH08a], Gupta et al. [GEH10], and Liu. et al. [LZZ14]. Among these algorithms, the method in [HEH08a] only utilizes appearance models, the method in [GEH10] tries to reason block structures, and the method in [LZZ14] tries to explore the parallel or orthogonal relationships between planar regions, which are special cases of the proposed geometric commonsense equations. These baseline methods are restricted to their capability of dealing various scene types. In contrast, the proposed method can adaptively find suitable commonsense knowledge and thus are applicable to all scene types.

Implementation We use the method by Ren et al. [RM03] to partition each image into 200 -300 superpixels, the method by Li et al. [LPY12] to detect straight lines as well as vanishing points, the method [HEH08a] to identify planar regions and texture regions, and the method [FGM10] to detect human and vehicles. These algorithms are called in preprocessing steps and the results are used as inputs for the proposed parsing algorithm. We set the maximal iteration of DDMCMC to be 2000. It costs about 1-2 minutes to converge on a Dell Workstation (i7-4790 CPU@3.6GHZ with 16GB memory). We implemented two

variants of our method to evaluate the effectiveness of individual commonsense. i) *Our-I*, that only utilizes the type-II geometric commonsense and ii) *Our-II*, that utilizes both type-I and type-II commonsense equations.

2.3.1 Qualitative result

Figure 2.5 visualizes a few exemplar results by the proposed method. In Columns 1, 2 and 3, we show the input images, edge maps, and semantic region partition, respectively. We further show two synthesized viewpoints in columns 4 and 5, and the estimated depth map in column 6. For comparisons, we plot the depth map obtained by Gupta et al. [GEH10] in column 7. We can observe that the obtained 3D scene model include the variety of vivid details, contain windows, small size facades (e.g., in the first row and second row), and doors (in the third row) etc. The obtained depth map is more accurate than those by [GEH10]. In contrast to [GEH10] that needs a post-processing step to approximate the depth, our method directly optimizes the geometric attributes while respecting various commonsense constraints. Figure 2.6 shows two failure results of the proposed method. For the image in the top row the groundplane is curved but was reconstructed as a planar region since our method assumes surface regions to be planar-wise. For the image in the bottom, the building floats off the groundplane in the scene model which is not physically plausible. This is due to the limitation of our method, i.e., no physical-based constraints are enforced during modeling or inference. In addition, our method cannot accurately reconstruct plants, trees or grass which are assumed to be a texture planar.

Table 2.2: Numerical comparisons on surface orientation.

	Country	Suburban	Road	Campus	Urban
Our-II	64.4%	60.6%	74.4%	71.6%	75.2%
Our-I	59.1%	58.6%	72.3%	69.4%	72.5%
Liu et al. [LZZ14]	54.6%	57.3%	68.6%	62.3%	68.7%
Gupta et al. [GEH10]	53.8%	51.4%	55.8%	57.4%	62.2%
Hoiem et al. [HEH08a]	52.3%	50.6%	59.3%	58.6%	63.3%

Table 2.3: Numerical comparisons on segmentation.

	Country	Suburban	Road	Campus	Urban
Our-II	67.9%	65.7%	71.3%	76.1%	75.8%
Our-I	62.8%	64.8%	69.8%	74.3%	73.9%
Liu et al. [LZZ14]	56.5%	63.3%	65.3%	70.6%	67.2%
Gupta et al. [GEH10]	52.1%	57.8%	58.1%	65.7%	61.3%
Hoiem et al. [HEH08a]	53.2%	52.9%	54.4%	66.3%	64.6%

2.3.2 Quantitative result

We compare the proposed method with three baseline methods for two parsing problems: surface orientation prediction and region segmentation. For *normal orientation estimation*, we use the metric of *accuracy*, i.e., the percentage of pixels that have the correct normal orientation label, and average accuracies over test images for every dataset. On the estimation of main geometric classes, i.e., 'ground', 'vertical', and 'sky', both our method and baseline methods can achieve high-quality results with accuracy 0.98 or more. Therefore, we focus on the vertical subclasses, like [GEH10], and discard the superpixels belonging to ground and sky while calculating the accuracies of all methods.

Figure 2.6: Failure cases of the proposed method. See texts.

Table 2.2 reports the numerical comparisons on five scene categories. From the results, we can observe the following. Firstly, the proposed method clearly outperforms the other baseline methods on all five scene categories. The improvements over country images are most significant since our method can adaptively select the informative cues over planar

regions, and discard the edge/gradient cues while the other methods can not. As stated by Gupta et al. [GEH10], it is difficult to improve vertical subclass performance. Our method, however, can improve these three baselines with large margins. Secondly, The proposed method outperforms the recent grammar based method [LZZ14], which tries to create a hierarchical graph as well. As summarized, the method in [LZZ14] applies Manhattan or mixture Manhattan type assumptions, which do not always hold in images. In contrast, the proposed commonsense knowledge are effective across a wide variety of images.

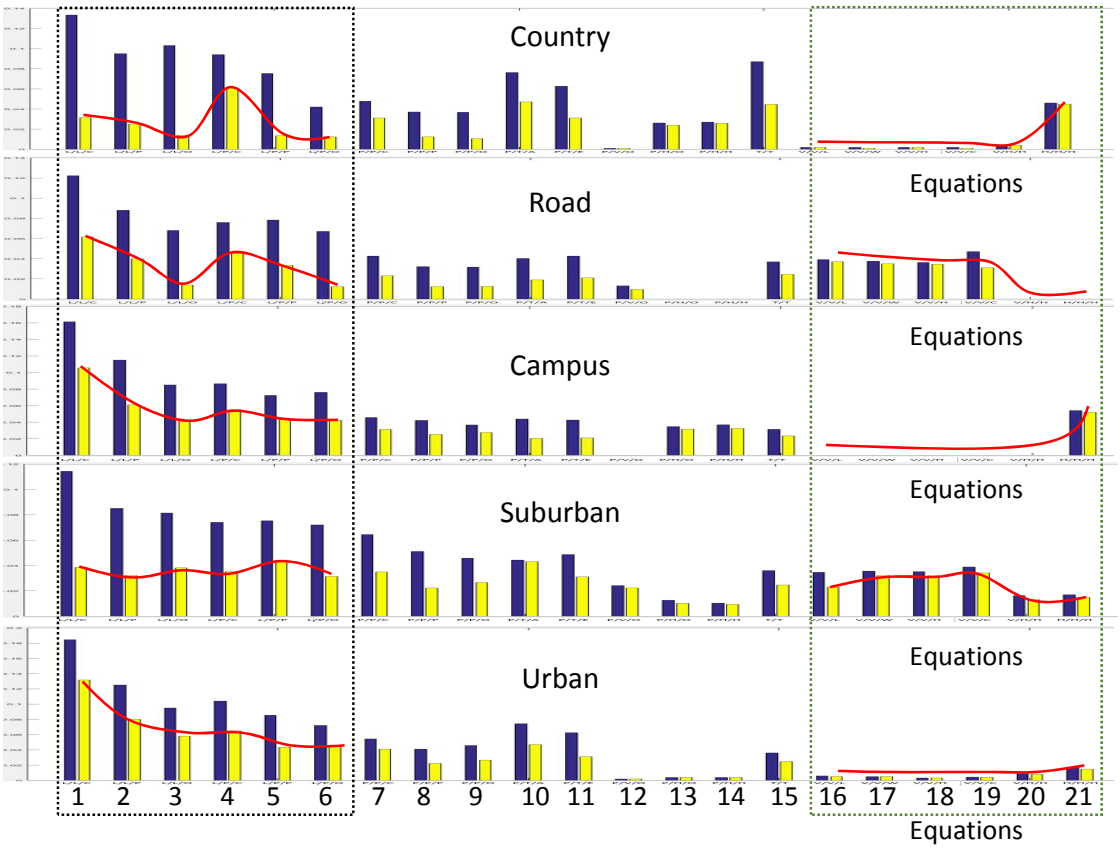


Figure 2.7: Average occurrence frequencies of commonsense equations on images of five scene categories. See Texts.

For *region segmentation*, we use the *best spatial support* metric as [GSE11], which first estimates the best overlap score of each ground truth labeling and then averages it over all ground-truth labeling. We discard the superpixels belonging to ground and sky while calculating the accuracies of all methods. Table 2.3 reports the region labeling performance on the five scene categories. Our method outperforms the method [LZZ14] with the margins

of 11.4, 2.4, 6.0, 5.5 and 8.6 percentages on the five categories, respectively. These comparisons show that jointly solving recognition and reconstruction has the ability to considerably improve recognition accuracies.

Reasoning of Valid Geometric Commonsense An interesting aspect of the proposed method is that we can identify the invalid commonsense equations and graph nodes that have inconsistent attributes with their children nodes. To do so, we introduce a heuristic test that comprises of two major steps. First, we solve the optimal parse graph as well as its attribute variables using the proposed DDMCMC algorithm. Second, with the optimal graph, we check if a constraint equation is satisfied using a simple threshold based method. In the 8-th column of the Figure 2.5, we visualize the occurrence frequencies of valid equations (yellow bars) and the total number of equations (blue bars) for every image. Figure 2.7 further plots the average occurrence frequencies of valid equations in individual categories. Note that the distributions of valid equations (red curves) vary significantly across categories, which shows that our method can adaptively find the suitable commonsense knowledge for various types of images.

2.3.3 Measurement result

Only the images of Campus are used in this test.All baseline methods can only reconstruct scene models up to a scale.We compare the proposed method with the groundtruth. As Figure 2.8 shows every image in the Campus category, we set 15-20 landmark dimensions, e.g., height of building, height of pillar, width of door, etc., each of which is defined with two endpoints in 3D. With 3D scene model recovered, we manually find every landmark (two endpoints) and compare its dimension to the groundtruth, and the Landmarks endpoints are provided manually using testing. According to the length we calculate average measurement errors are used as rate.Figure 2.9 shows an example of our measurement results. The average measurement errors in the Campus category is 16.47%.

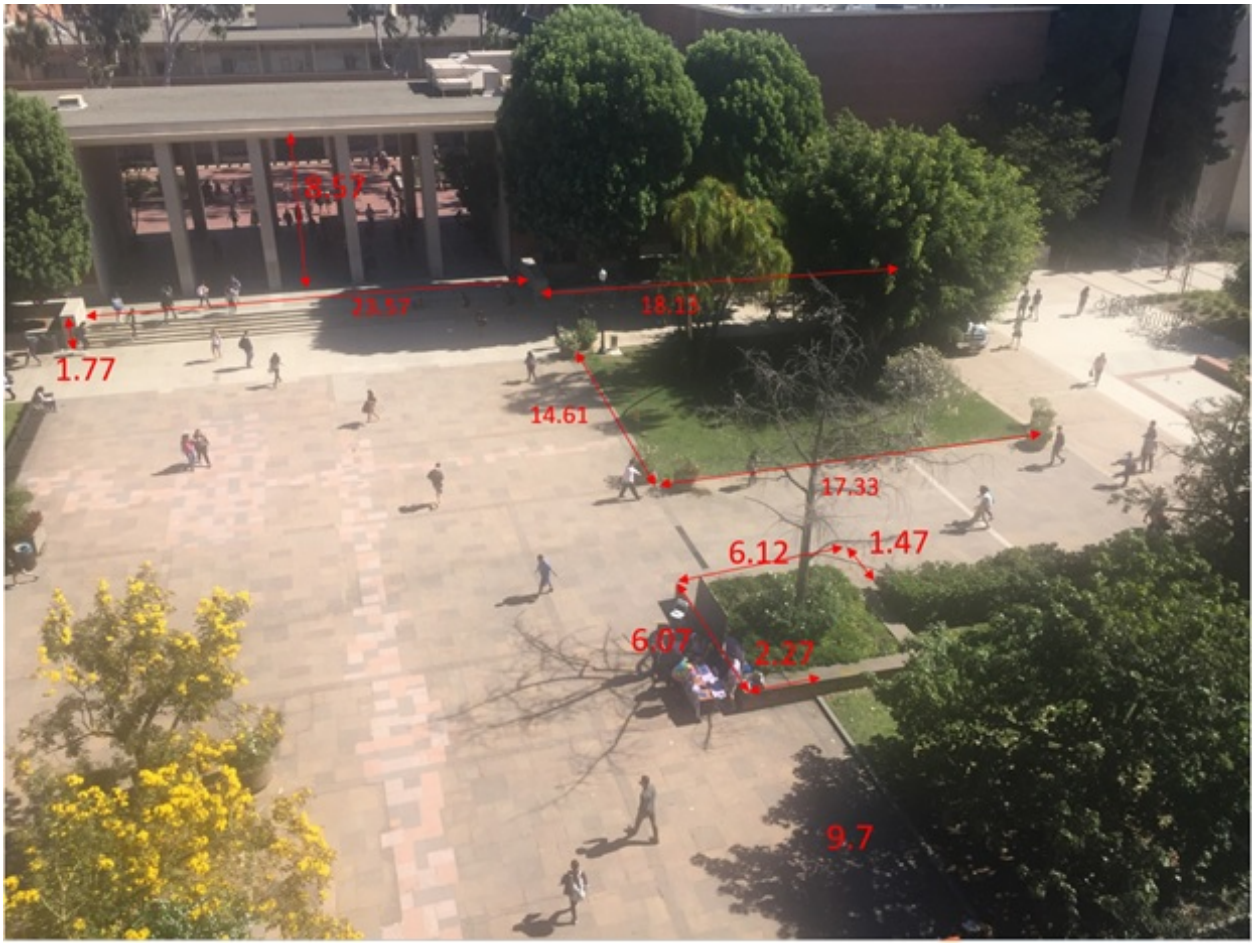
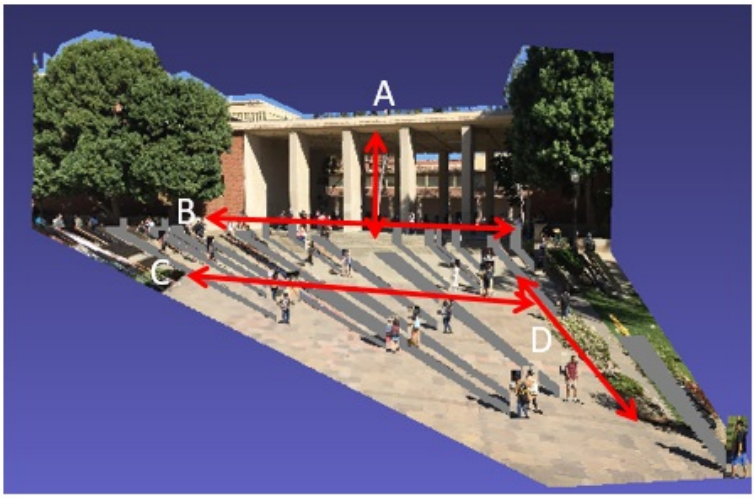


Figure 2.8: The measurements are used as metric.



	Our Method	Ground truth
A	8.17m	8.57m
B	26.73m	23.57m
C	28.29m	25.16m
D	18.45m	14.61m

	Average Error Ratio (error/ dimension)
all	16.47%

Figure 2.9: An example of our measurement results.

2.3.4 Indoor result

The proposed method can also be applied on indoor scene. The algorithm is similar to outdoor scene reconstruction except that the terminal node detection in indoor scene is replaced by cubic detection. The attributes of terminal node in indoor scene is the same as outdoor scene. To reconstruction indoor scene, we add more objects statistics as constraints such as the heights of desks and chairs and so forth. We evaluate our method on two classes of dataset. 1) The simulated dataset. The indoor scenes are generated firstly by generating 3D furnitures from a set of furniture 3D models. Textures are rendered into the generated furniture models by Houdini Engine. By rendering, we can obtain the depth and label map information so that we can avoid human labors to manually label the data. We generate more than one million scenes with various configurations and furnitures as illustrated in Figure 2.10. Figure 2.11. displays the reconstructed results.

2) Real scene dataset. We collect 100 indoor scenes and manually label the surface orientation and segmentation information. We have two experimental settings. One is single view scene as illustrated in Figure 2.12. The other is multi-view scene, which is captured

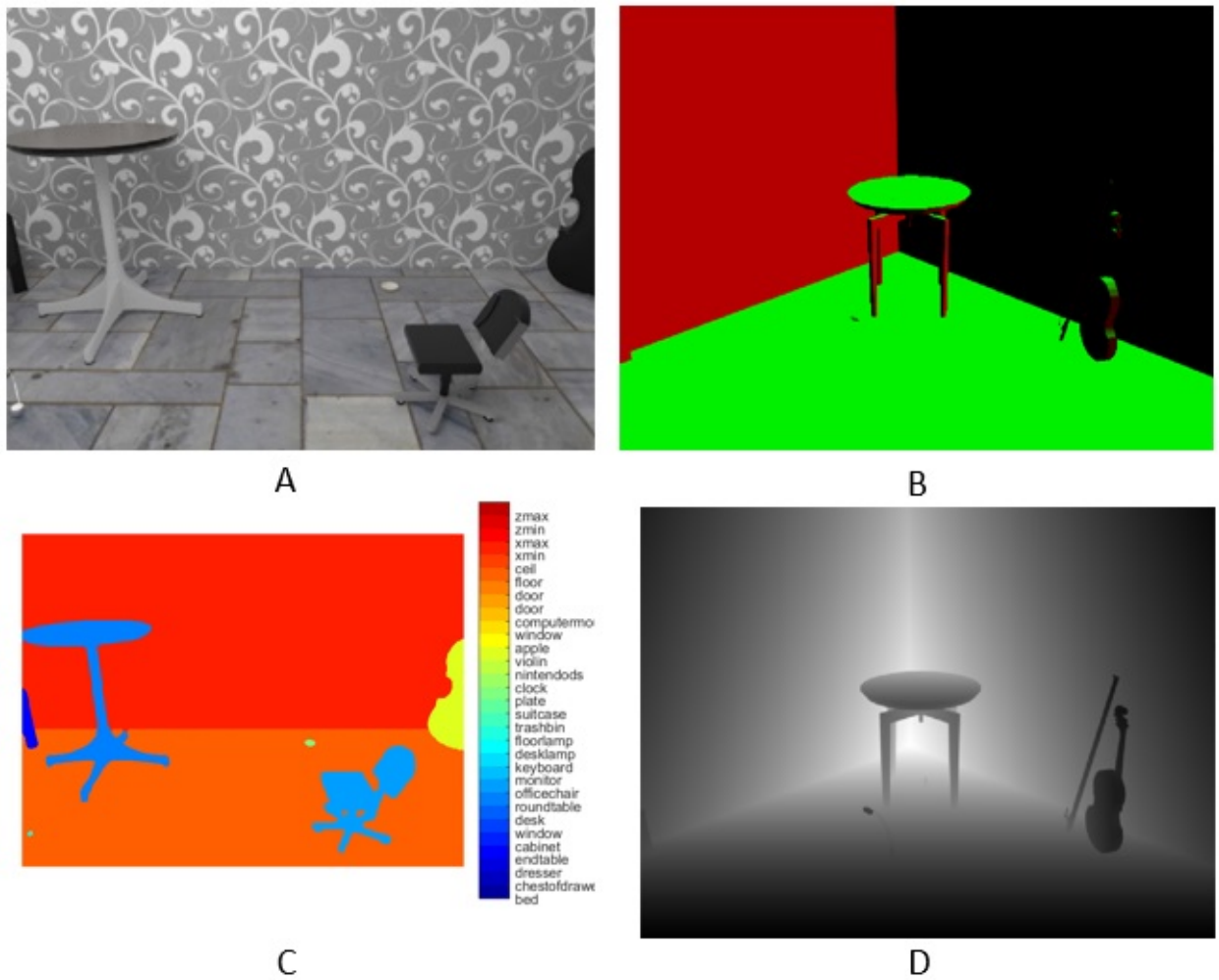


Figure 2.10: A: RGB output image; B: Furniture CAD model; C: Labeling map; D: Depth map

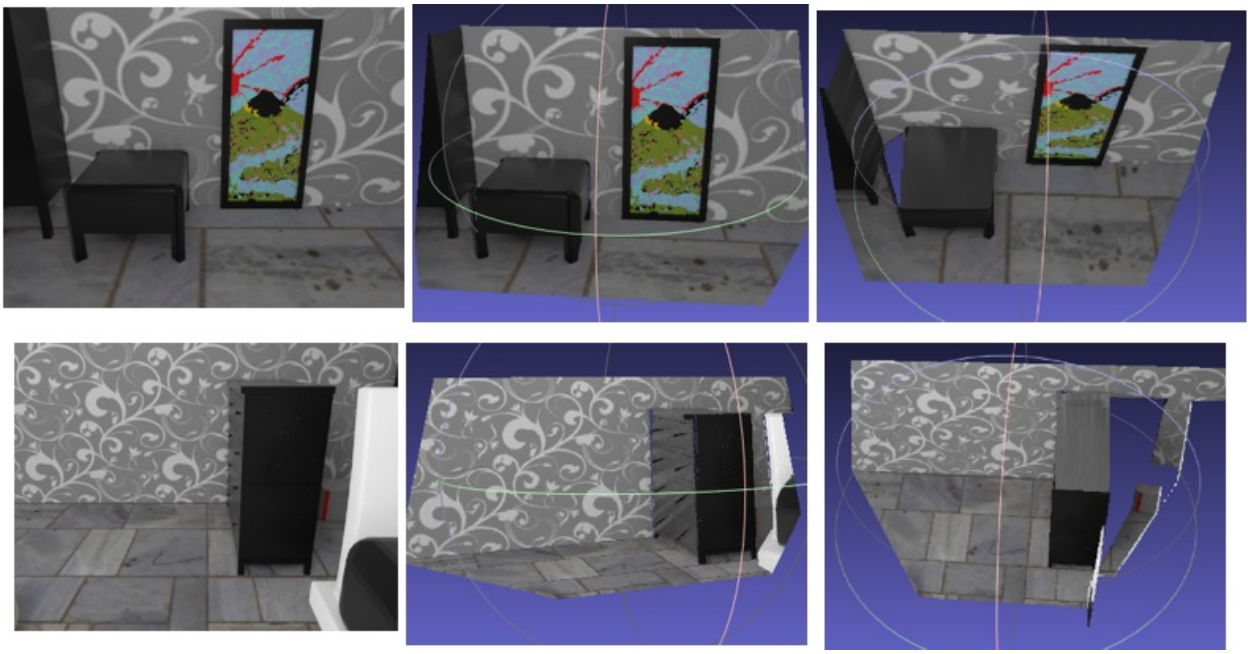


Figure 2.11: Column-left: Input image; Column-Middle&right: Two novel viewpoints synthesized.

in a single scene from different views. By reconstructing from multi-view images, we can reconstruct objects occluded in a single view. Figure displays the reconstructed results. In figure 2.13(a), the objects behind the pillar is occluded. However, its information can be obtained by figure 2.13(b).

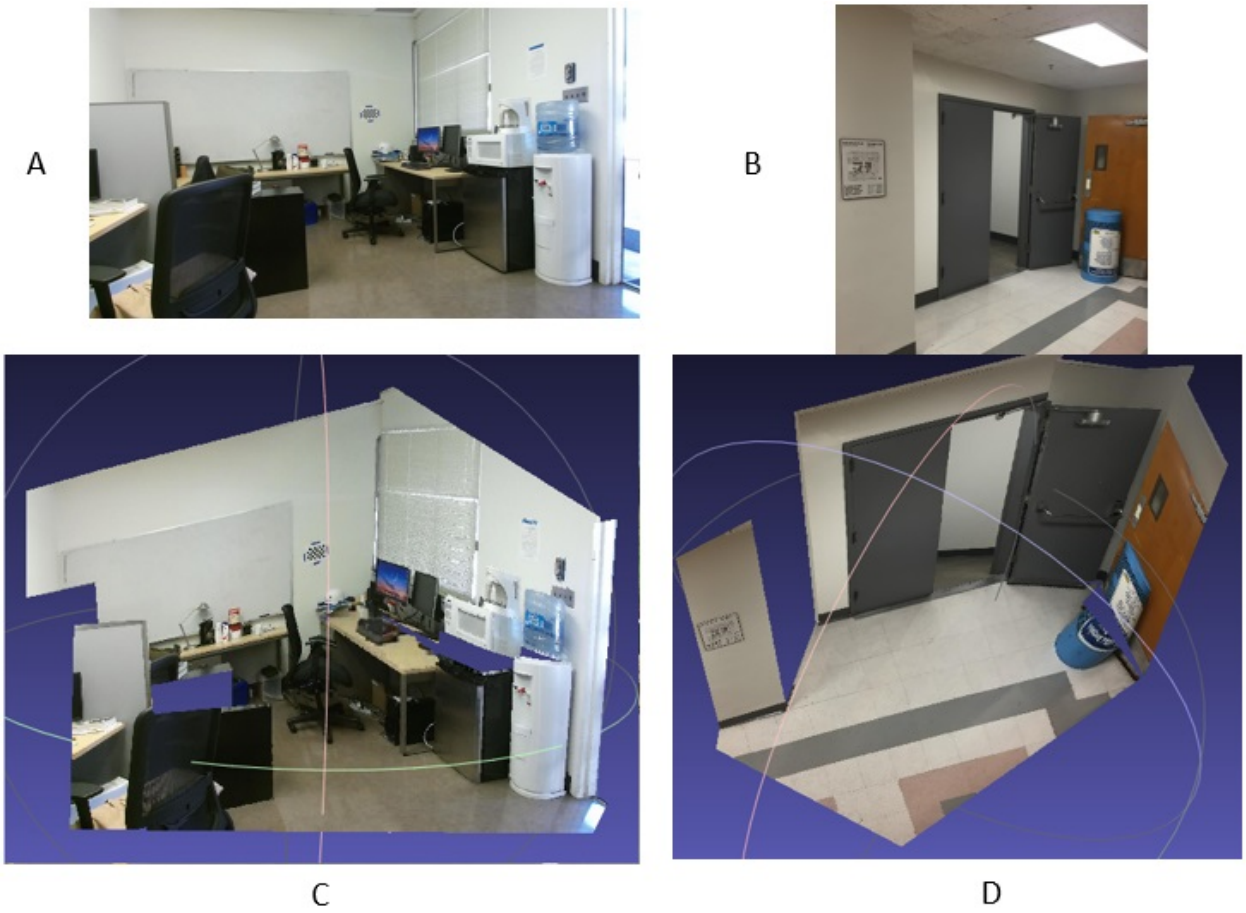


Figure 2.12: A&B: Input image. C&D: Reconstruction results.

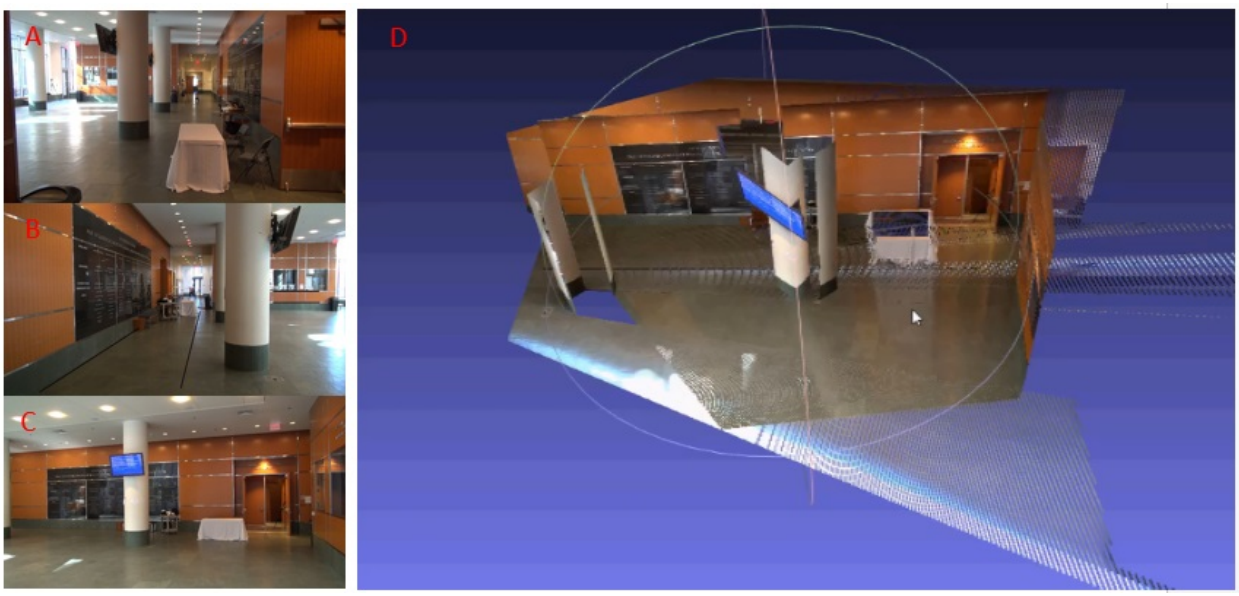


Figure 2.13: A,B,C: Input image from different views in a single scene. D: Multi-view reconstructing result

CHAPTER 3

Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics

3.1 Introduction

Motivation and Objectives. Traditional approaches for scene understanding have been mostly focused on segmentation and object recognition from 2D images. Such representations lacks important physical information, such as the 3D volume of the objects, supporting relations, stability, and affordance which are critical for robotics applications: grasping, manipulation and navigation. With the recent development of Kinect camera and the SLAM techniques, there has been growing interest in studying these properties in the literature [NIH11].

In this chapter, we present an approach for reasoning physical stability 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 [NIH11]. We utilize a simple observation that, by human design, objects in static scenes should be stable. 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 attached neighbors or fixed to its supporting base. In addition, while objects are stable physically, they should enjoy a movable space (freedom) for manipulation. Such assumption is applicable to all scene categories and thus pose quite powerful constraints for the plausible interpretations (parses) in scene understanding.

As Fig. 3.1 shows, our method consists of two main steps.

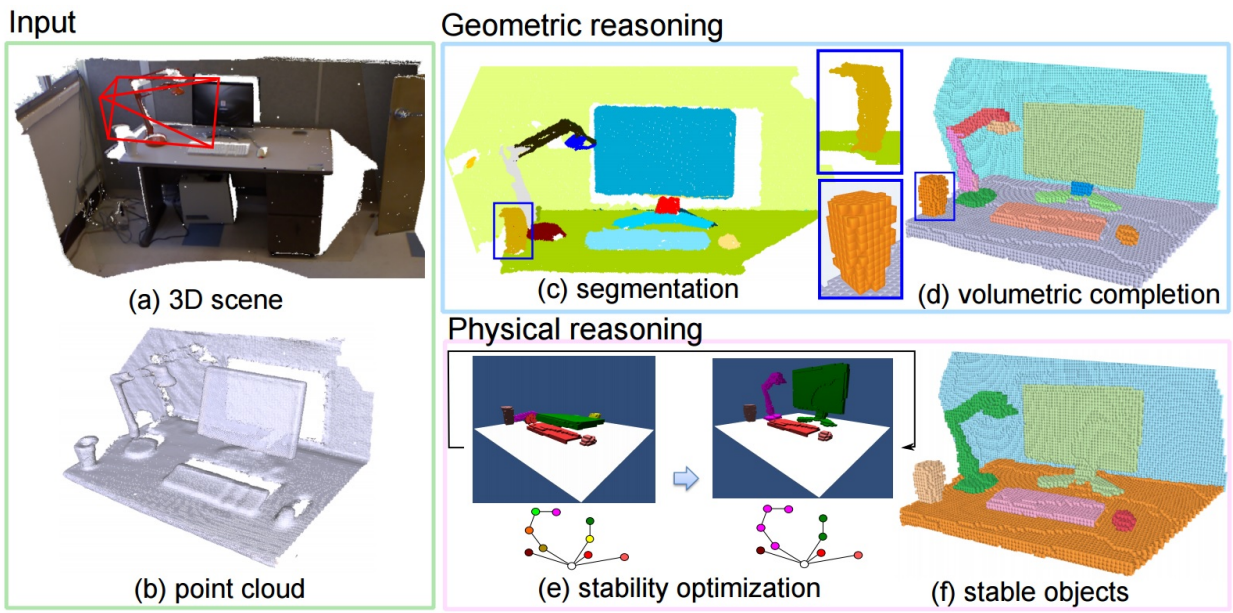


Figure 3.1: 1. Overview of our method. (a) 3D scene reconstructed by SLAM technique, (b) point cloud as Input. In geometric reasoning, (c) a portion is shown to be segmented by a segment-and-merge approach, with missing voxels, (d) solid primitives by volumetric completion. In physical reasoning, (e) the contact graph are labeled through stability optimization. (f). Final parsing results with stable objects.

1) Geometric reasoning: recovering solid 3D volumetric primitives from defective point cloud. Firstly we segment and fit the input 2.5D depth map or point cloud to small simple (e.g., planar) surfaces; secondly, we merge convexly connected segments into shape primitives; and thirdly, we form 3D volumetric shape primitives by filling the missing (occluded) voxels, so that each shape primitive can own its physical properties: volume, mass and supporting areas to compute the potential energies in the scene. Fig. 1.(d) shows the 3D primitives in rectangular or cylindrical shapes.

2) Physical reasoning: 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 as shown in Fig. 1.3.(e), coloring this graph corresponds to grouping them into objects. For example, the lamp on the desk originally was divided in 3 primitives and will fall under gravity (see result simulated using a physics engine), and become stable when they are grouping into one object - the lamp. So is the computer screen with its base. To achieve the physical reasoning goal, we make the following novel contributions in comparison to the most recent work in dealing with physical space reasoning [GEH10, SHK12].

- We define the physical stability function explicitly by studying minimum energy (physical work) need to change the pose and position of an primitive (or ob- ject) from one equilibrium to another, and thus to release potential energy.
- We introduce disconnectivity graph (DG) from physics (Spin-glass) to represent the energy landscapes.
- We solve the complex optimization problem by the cluster sampling method Swendsen-Wang cut in image segmentation [BZ05] to maximize global stability.
- We collect a new dataset for large scenes by depth sensors for scene understanding and will release the data and annotations to the public.

In experiments, we demonstrate that the algorithm achieve a substantially better performance for i) object segmentation, ii) 3D volumetric recovery of the scene, and iii) better

parsing result for scene understanding in comparison to state-of-the-art methods in both public dataset [SHK12] and our own new dataset.

In experiments, we demonstrate that the algorithm achieve a substantially better performance for i) object segmentation, ii) 3D volumetric recovery of the scene, and iii) better parsing result for scene understanding in comparison to state-of-the-art methods in both public dataset [SHK12] and our own new dataset.

3.2 Geometric reasoning

Given a point cloud of scene, the goal of geometric reasoning is to infer the object primitives (e.g., the colored objects in Fig. 1.3 (d)), such as that each primitive can own physical properties (e.g., volume, mass, supporting area, etc.). We infer the object primitives with two major steps: 1) point cloud segmentation and 2) Volumetric completion.

3.2.1 Segmentation with implicit algebraic models

We first adopt implicit algebraic models (IAMs) [BLC00] to separate point cloud into several simple surfaces. We adopt a split-and-merge strategy as: 1) splitting the point cloud into simple and smooth regions by IAM fitting, and then 2) merging the regions which are "convexly" connected each other. As a 2D example illustrated in Fig. 2.(a), suppose the 2D point cloud is first split into three line segments with first-order IAM fitting: f_1, f_2 and f_3, and then f_2 and f_3 are merged together, since they are "convexly" connected.

Splitting point cloud. The objective in this process can be considered as to find out the 3D regions, and each of them can be well fitted by an IAM.

The IAM fitting for each region can be formulated in least squares optimization using the 3-Layer method proposed by Blane et al. [BLC00]. As shown in Figure 3.2.(a), it first generate two extra point layers: Γ_{-}(green points) and Γ_{+}(light blue points) along the normals of points in the original region M (red and blue points). Then an IAM can be fit to M by linear least-squared method with linear constraints:

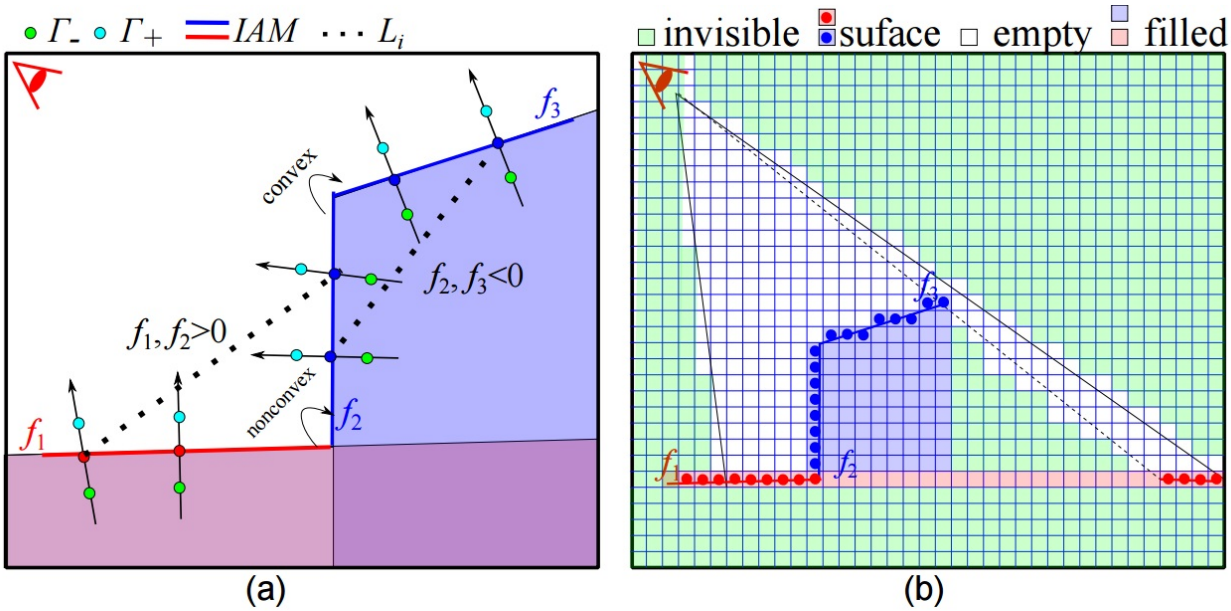


Figure 3.2: (a) Two 1-degree IAMs f_1 and f_2 (in blue and red lines respectively) are fitted to the 3-Layer point cloud. The light red and blue areas denote in which functions f_1, f_2 and f_3 are minus. (b) Invisible space estimation and voxel completion. Four types of voxels are estimated: invisible 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).

$$f(P_i) = \begin{cases} 0, & P_i \in M \\ +d_i, & P_i \in \Gamma_+ \\ -d_i, & P_i \in M \end{cases} \tag{3.1}$$

where f is an implicit polynomial,$\pm d$ is the Euclidean distance how long the two points move along the normals in opposite directions. Therefore, as shown in Fig. 3.2 (a), each IAM fit can split the space into two parts: "inside" (colored with negative value) and "outside" (uncolored (white) with positive value).

For splitting point cloud into pieces, we adopt region growing scheme [PVB08]. Our method can be described as: starting from several given seeds, the regions grow until there is no unlabeled point can be fitted by certain IAM. In this chapter, we adopt the IAM of 1 or 2 degree, i.e., planes or second order algebraic surfaces and the IAM fitting algorithm proposed by Zheng et al. [ZTI10] to select the models in a degree-increasing manner.

Merging "convexly" connect regions. The splitting strategy seems separating the points to be object faces (e.g., a box can be split into six faces). However we can further merge the "convexly" connected regions to better represent object parts (primitives). To this end, we first define "convex connection" of two regions as follow:

Definition 1. for any line segment L whose two ends are in two connected regions with IAM fits f_i and f_j respectively, if the points on this line, $\{\forall p_l | p_l \in L\}$, satisfy $f_i(p_l) < 0$ and $f_j(p_l) < 0$, then we say regions i and j are convexly connected.

To detect the convex connection, as shown in Fig. 3.2 (a), we first randomly sample several line points (in dark dot lines) between connected regions, and then check them if satisfy the convexly connected relationship defined above. In practice, we merge the convex connections when the following condition is satisfied:

$$\frac{\#\{p | p_l \in L \wedge f_i(p_l) < 0 \wedge f_j(p_l) < 0\}}{\#\{p | p_l / inL\}} > \delta \tag{3.2}$$

where the ratio threshold δ is set as 0.6 according the sensor noise. In Fig 3.2 (a), since the dark points connecting f_2 and f_3 are submerged by both minus regions of them.

3.2.2 Volumetric space completion

To obtain the physical properties for each object primitive (e.g., size, mass etc.), we need volumetric representation but not surface segments. Thus, we complete each surface segment into a volumetric (voxel-based) primitive under three assumptions:

- a) Occlusion assumption: voxels occluded by the observed point cloud could be parts of objects.
- b) Solid assumption: hollow object is not preferred (e.g., plane should not with holes, or a box should be solid).
- c) Manhattan assumption: most object shapes are aligned with Manhattan axes.

Voxel generation and gravity direction. We first generate voxels for each segment obtained by above point cloud segmentation by

- 1) detecting Manhattan axes [FCS09].
- 2) constructing voxels from point cloud along Manhattan axes by octree construction method [SNI05].
- 3) detecting gravity direction.

To detect gravity direction, we simply choose the one with smallest angle to the vertical axis of sensor coordinate system.

Invisible (occluded) space estimation. 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 occlusion. Inspired by Furukawa's method in [FCS09], the Manhattan space is carved by the point cloud into three spaces (as shown in Figure 3.2(b)): Object surface S (colored-dots voxels), Invisible space U (light green voxels) and Visible space E (white voxels).

Voxels filling. We complete an object primitive from each labeled surface segment. Suppose each convex surface segment is the visible part of a primitive, we complete invisible

part by filling voxels in a visual hull which is occluded by the surface under two assumptions: 1) as lights travel in lines, the voxels complected are behind the point clouds, as shown in Fig. 3.2.(b); 2) a primitive should be completed if it can be seen from at least two directions of Manhattan axes. Therefore our algorithm can be simply described as:

Loop: for each invisible voxel $v_i \in U, i = 1, 2, ...$

- 1) From v_i, searching the voxels along 6 directions of Manhattan axes, to collect six nearest surface voxels $\{v_j \in S\}(j \leq 6)$.
- 2) Checking the label for each v_j, if there exist more than two same labels, then assign this label to voxel v_i.

3.3 Modelling object stability

3.3.1 Energy landscapes

A 3D object (or primitive) has a potential energy de- fined by gravity and its state (pose and center) supported by neighboring object in 3D space. The object is said to be in equilibrium when its current state is a local minimum (stable) or local maximum (unstable) of this potential function (See Fig 3.4 for illustration). This equilibrium can be broken by external work (e.g., nature disturbance) and then the object moves to a new equilibrium and releases energy. Without loss of generality, we divide the change in two cases.

Case I: pose change. In Fig. 3.3, the chair in (a) 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 to the state change $x0 \rightarrow x_1$ by

$$E_r(x0 \rightarrow x_1) = (R_c - t_1) \cdot mg, \tag{3.3}$$

where R is rotation matrix; c is center of mass, $g = (0, 0, 1)^T$ is the gravity direction, t_1 is the lowest contact point on the support region (its legs). We visualize the energy landscape

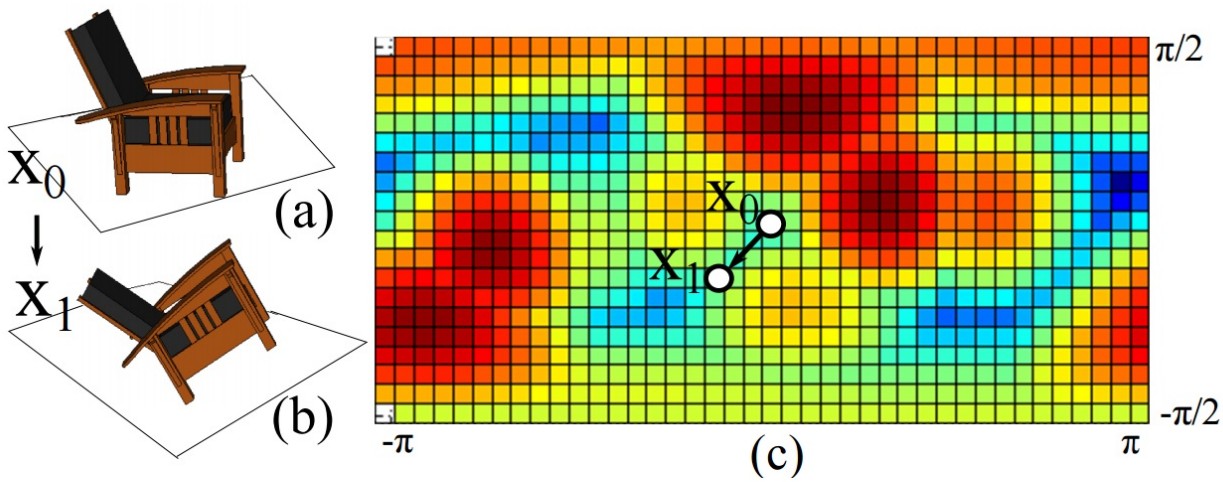


Figure 3.3: (a) A chair in a stable state x_0 is moved to (b) an unstable state x_1. (c) The landscape of potential energy is calculated by Eq. (3.3) over two rotation angles where x_0 is a local minimum and x_1 is a saddle point passing which, the chair will fall to a deeper energy basin (blue).

on the sphere $(\phi, \theta) : S^2 \rightarrow R$ in Fig. 3.3.(c) using the two pose angles ($\theta \in [-\pi, \pi], \theta \in [-\pi/2, \pi/2]$). 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 [Kri97] for further details.

Case II: position change. Imaging a cup on a desk at stable equilibrium state x_0, one can push it to the edge of the table. Then it falls to the ground and releases energy to reach a deeper minimum state x_1. The energy change needed to move the cup is

$$\varepsilon_r(x0 \rightarrow x_1) = (c - t) \cdot mg - f, \tag{3.4}$$

where $t \in R^3$ is the translation parameter (shortest distance to the edge of the desk), and f is friction defined as $f = f_c\sqrt{(t_1 - c_1)^2 + (t_2 - c_2)^2}$ given the friction coefficient f_c. Note for common indoor scenes, we choose fc as 0.3 as common material such as wood. Therefore the energy landscape can be viewed as a map from 3D space $R^3 \rightarrow R$.

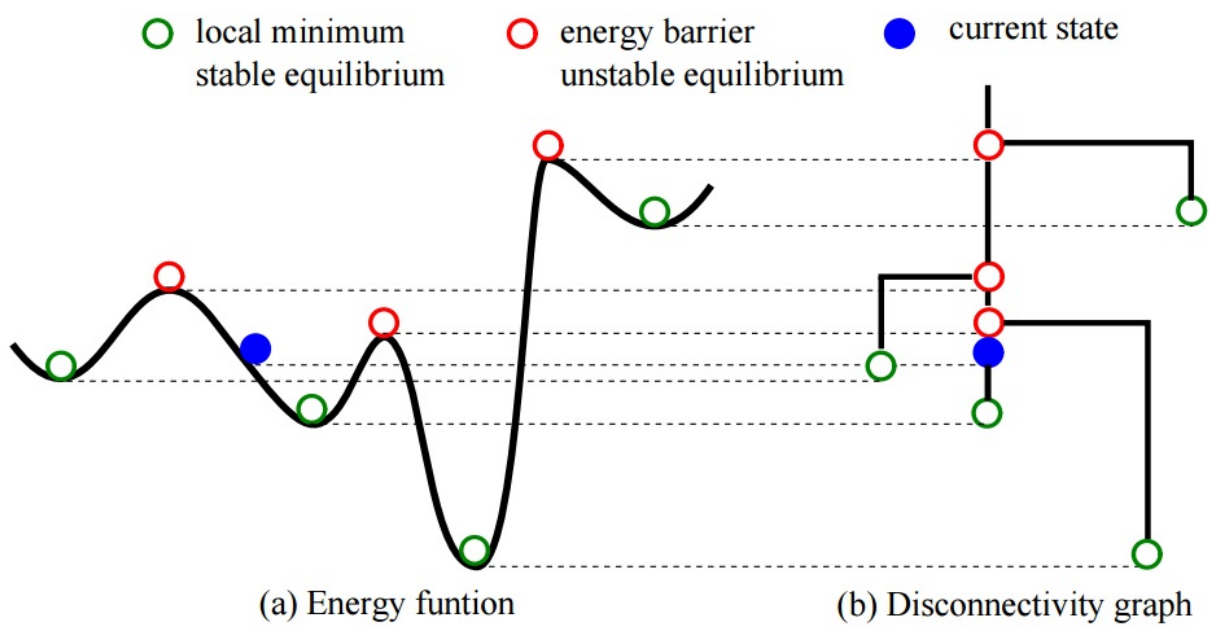


Figure 3.4: (a) Energy landscapes and its corresponding disconnectivity graph (b).

In both cases, we observe that object stability is only local and relative, and can be changed subject to disturbance (gravity, wind, mild earthquake, and human activity).

3.3.2 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 current state (according to the current interpretation of the scene). Therefore, we represent the energy landscape by a so-called disconnectivity graph (DG) which has been used in studying the spinglass models in physics [Wal03]. 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 [Wal03]. From the DG, we can conveniently calculate two quantities: Energy absorption and Energy release during the state changes.

Definition 2. The energy absorption $\Delta E(x_0 \to \tilde{x})$ is the energy absorbed from the

perturbations, which moves the object from the current state x_0 to an unstable equilibrium $\tilde{x}$ (say a local maximum or energy barrier).

For the chair in Fig.3.3, its energy absorption is the work needed to push it in one direction to an unstable state x_1. For the cup example, its energy barrier is the work needed (to overcome friction) to push it to the edge. In both cases, the energy depends on the direction and path of movement.

Definition 3. Energy release $\Delta E(x_0 \to x_0')$ is the potential energy released when an object moves from its unstable equilibrium $\tilde{x}$ to a minimum x_0' which is lower but connected by the energy barrier.

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

With DG, we define object stability in 3D space.

Definition 4. The stability $S(a, x_0, W)$ of an object a at state x_0 in the presence of a disturbance work W is the maximum energy that it can release when it moves out the energy barrier by the work W.

$$S(a, x_0, W) = \max_{x_0'} \Delta E(\tilde{x} \to x_0')\delta([\min_{\tilde{x}} \Delta E(x_0 \to \tilde{x})] \leq W), \tag{3.5}$$

where $\delta()$ is an indicator function and $\delta(z) = 1$ if condition z is satisfied otherwise $\delta(z) = 0$. $\Delta E(x_0 \to \tilde{x})$ is the energy absorbed, if it is overcome by W, then $\delta() = 1$, and thus the energy $\Delta E(\tilde{x} \to x_0')$ is released. We find the easiest direction $\tilde{x}$ to minimize the energy barrier and the worst direction x_0' to maximize the energy release.

3.4 Physical reasoning

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

3.4.1 Contact graph and group labelling

The contact graph is an adjacency graph $G =< V, E >$, where $V = \{v_1, v_2, ..., v_k\}$ is the 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.1.3.(e) where each node corresponds to a primitive in Fig. 1.3.(c). 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 the stability is re-calculated according to O_i. The optimal labelling L^* can be determined by the optimization of a global energy function, for a work level W

$$E(L|G;W) = \sum_{O_i \in L} (S(O_i, x(O_i), W) + F(O_i)) \tag{3.6}$$

where $x(O_i)$ is the current state of grouped object O_i . The new term F represents a penalty function expressing the scene prior and can be decomposed into parts.

$$F(O_i) = \lambda_1 f_1(O_i) + \lambda_2 f_2(O_2) + \lambda_3 f_3(O_i), \tag{3.7}$$

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 designed by 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 $\{v_j, v_k \in O_i\}$, 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 arise if the mouse is assigned by same label to the table.

3.4.2 Inference of Maximum stability

As the label of primitives are coupled with each other, we adopt the graph partition algorithm Swendsen-Wang Cut (SWC) [BZ05] for efficient MCMC inference. To obtain globally optimal L^* by the SWC, the next 3 main steps works iteratively until convergence.

(i) Edge turn-on probability. Each edge $e \in E$ is associated with a Bernoulli random variable $\mu_e \in \{on, off\}$ 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 probability as:

$$q_e = p(\mu_e = on|v_i, v_j) = exp(-(F(v_i, v_j)/T)), \tag{3.8}$$

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 of different categories. Then all the remaining edges are turned on independently with the 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_i from the set formed in step (ii) with a uniform probability, and then flips the labels of all nodes in Π_i to a category $c \in 1, 2, ..., C$. The flip is accepted with probability [BZ05]:

$$\alpha(L \to L') = min(1, \frac{Q(L' \to L)E(L'|G; W)}{Q(L \to L')E(L|G; W)}) \tag{3.9}$$

Fig. 3.5 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 (3.9) shown in (b) and (c), resulted the energy v.s. iterations in (d). It is worth noticing that 1) in case of 3.5 (b), the little chair is not grouped to floor, since the penalty term A3 penalize the legs to fix to floor. 2) On the other hand, we increase the disturbance W in (3.5), the chair is fixed to floor.

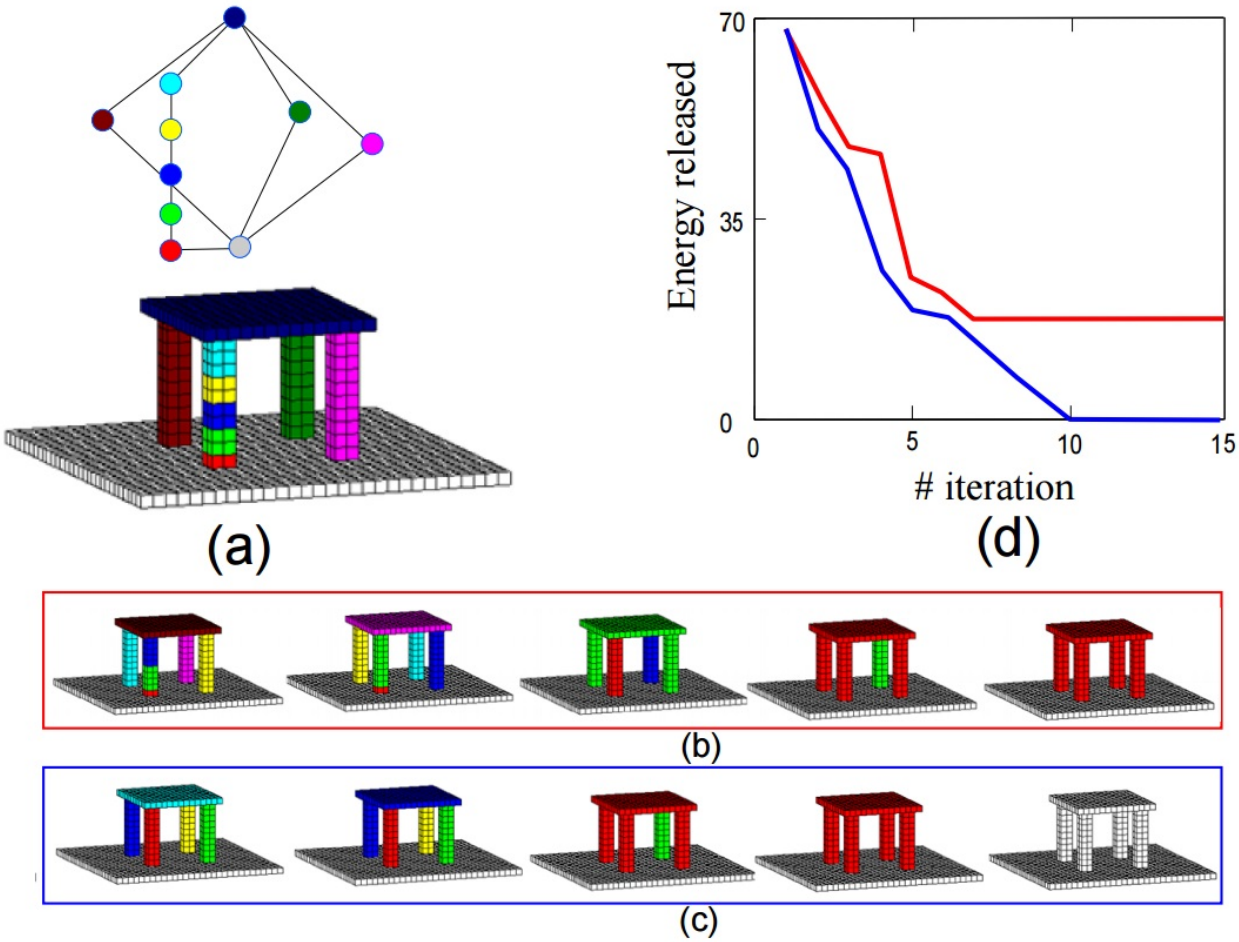


Figure 3.5: Example of illustrating the Swendsen-Wang sampling process. (a) Initial state with corresponding contact graph. (b) shows the grouping proposals accepted by SWC at different iterations. (c) convergence under larger disturbance W and consequently the table is fixed to the ground. (d) shows two curves of Energy released v.s. number of iteration in SWC sampling corresponding to (b) and (c)

3.5 Experimental result

We quantitatively evaluate our method in terms of

- 1) single depth image segmentation.
- 2) volumetric completion evaluation.
- 3) physical inference accuracy evaluation.
- 4) intuitive physical reality (by videos in supplementary).

All these evaluations are based on three datasets:

- i) NYU depth dataset V2 [SHK12] including 1449 RGBD images with manually labeled ground truth.
- ii) A set synthesized depth map and volumetric images simulated from CAD scene data.
- iii) 13 reconstructed 3D scene data captured by Kinect Fusion [NIH11] gathered from office and residential rooms with ground truth labeled by a dense mesh coloring.

3.5.1 Evaluating single depth image segmentation

Two evaluation criterion: "Cut Discrepancy" and "Hamming Distance" mentioned in [CGF09] 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 result shown in Fig. 3.6, our segmentation by physical reasoning is with lower error rate than the another two: region growing segmentation [PVB08], and our geometric reasoning.

Fig. 3.6 shows some examples for comparing point cloud segmentation result [PVB08] 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, and stabilities etc.

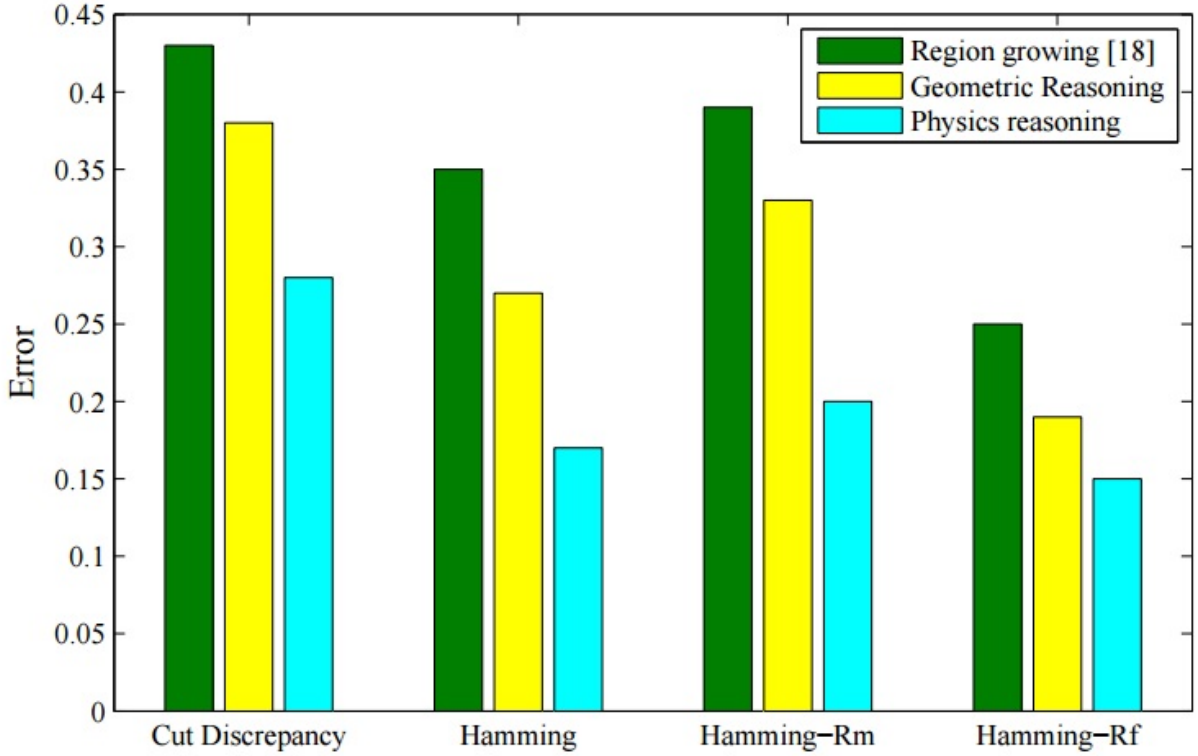


Figure 3.6: Segmentation accuracy comparison of three methods: Region growing method [PVB08], result of our geometric reasoning and physical reasoning by one Cut Discrepancy and three Hamming Distance

3.5.2 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

	Octree [SNI05]	Invisible space	Vol. com
Precision	98.5%	47.7%	**94.1%**
Recall	7.8%	95.1%	**87.4%**

Table 3.1: Precision and recall of Volumetric completion. Comparison of three method: 1) voxel-based representation generated by Octree algorithm [SNI05], 2) voxels in surface and invisible space (sec. 2.2), and 3) our volumetric completion.

relations	Discriminative	Greedy	SWC
Precision	20.5%	66%	**81.8%**
Recall	42.2%	60.3%	**78.1%**

Table 3.2: Results of inferring the fixed joints and support relations between primitives. Accuracy is measured by nodes of contact graph whose label is correctly inferred divided by the total number of labeled nodes.

testing data. Tab. 3.5.2 shows the result that our method has much better accuracy than traditional Octree method such as [SNI05]. 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 of using 3D feature priors as similar as one in [SHK12], and greedy inference method such as marching pursuit algorithm for physical inference. The result shown in Tab. 3.5.2 is evaluated by the average over 13 scene data captured by Kinect Fusion.

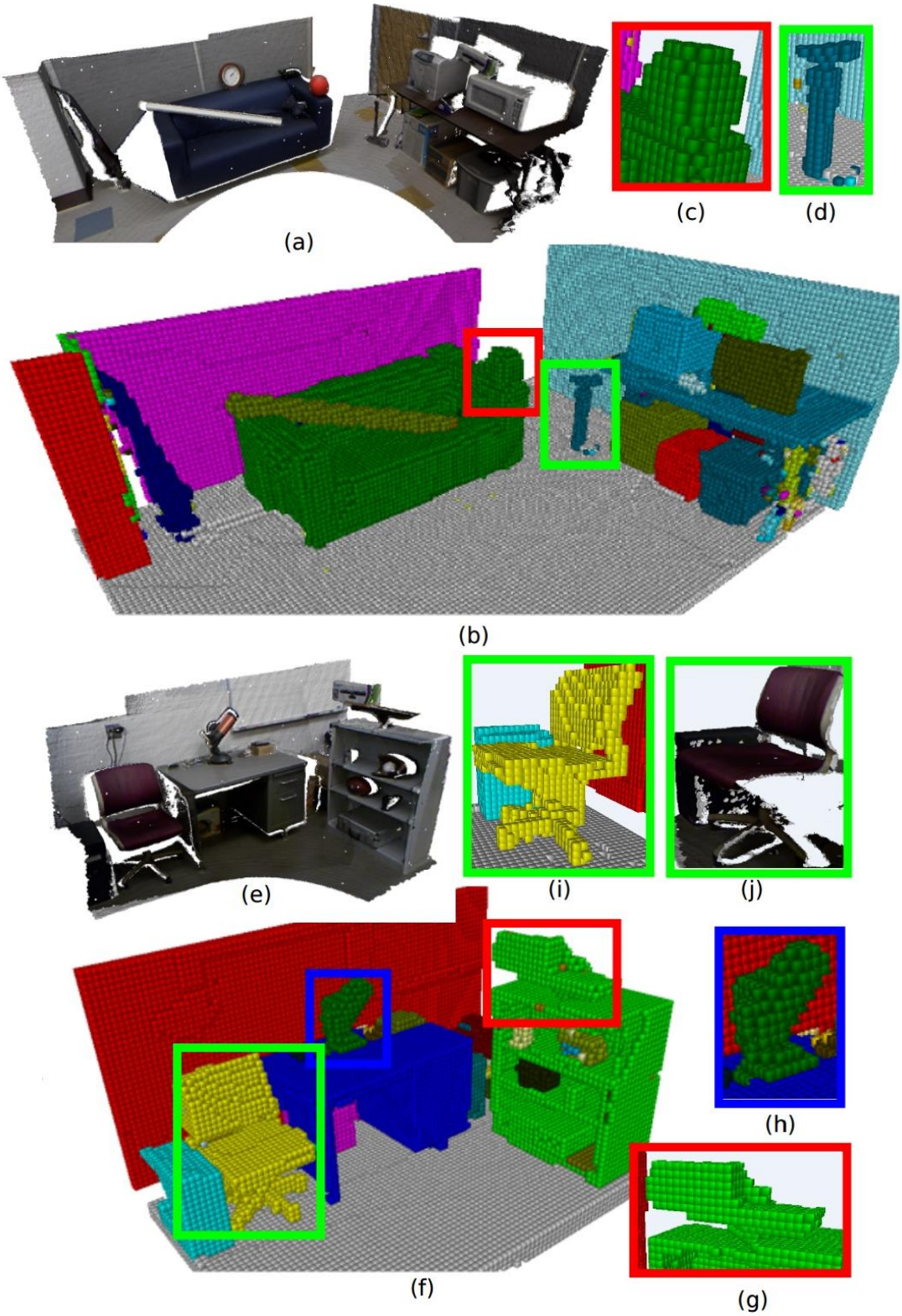


Figure 3.7: Example result. (a) and (e): data input. (b) and (f): volumetric representation of stable objects. (c): the ball is fixed onto the handle of sofa. (d): the pump is unstable (see text). (i): a irregular case of (g). (j): hidden voxels under chair compared to (h).

Figure 3.7 (a)-(d) and (e)-(j) show two examples from the results. Here we discuss some irregular cases by close-ups in the figures.

Case I: Figure 3.7 (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. (3.5). 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 3.7 (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.(3.7) penalized it to be fixed onto floor. So is the lamp not fixed to table in Figure 3.7 (h).

Case III: Figure 3.7 (g) the "empty Kinect box" with its base is fixed together onto the shelf, because of the miss segmentation of base, i.e., the lower part of base is miss merged to top of shelf.

Case IV: Figure 3.7 (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 hidden part is not accurate geometrically, but it helps to form a stable object physically. In contrast, original point cloud shown in Figure 3.7 (j) misses more data.

CHAPTER 4

Detecting Potential Falling Objects by Inferring Human Action and Natural

4.1 Introduction

The recent development of consumer-grade range cameras,6 such as the Kinect camera, has attracted increasing studies in the field of 3D scene understanding [AKJ13, JKJ13, SNI05, XF14]. However, most of existing work is focused on locating and naming the object in the scene, and leaves a big gap to answer human-level scene understanding questions, such as: how does a human interact with a scene? how does the scene response to the action? What and where are potential dangers in the environment?

4.1.1 Motivation

Traditional approaches, e.g., Shi and Ks[SF83] and Tu et al.[TC03], for scene understanding have mostly focused on segmentation and object recognition from 2D/3D images. Such representations lack important physical knowledge, 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 scenarios illustrate the importance of physical knowledge.

i) Safety surveillance robots. 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 poses useful constraints for the plausible interpretations (parses) in scene understanding.

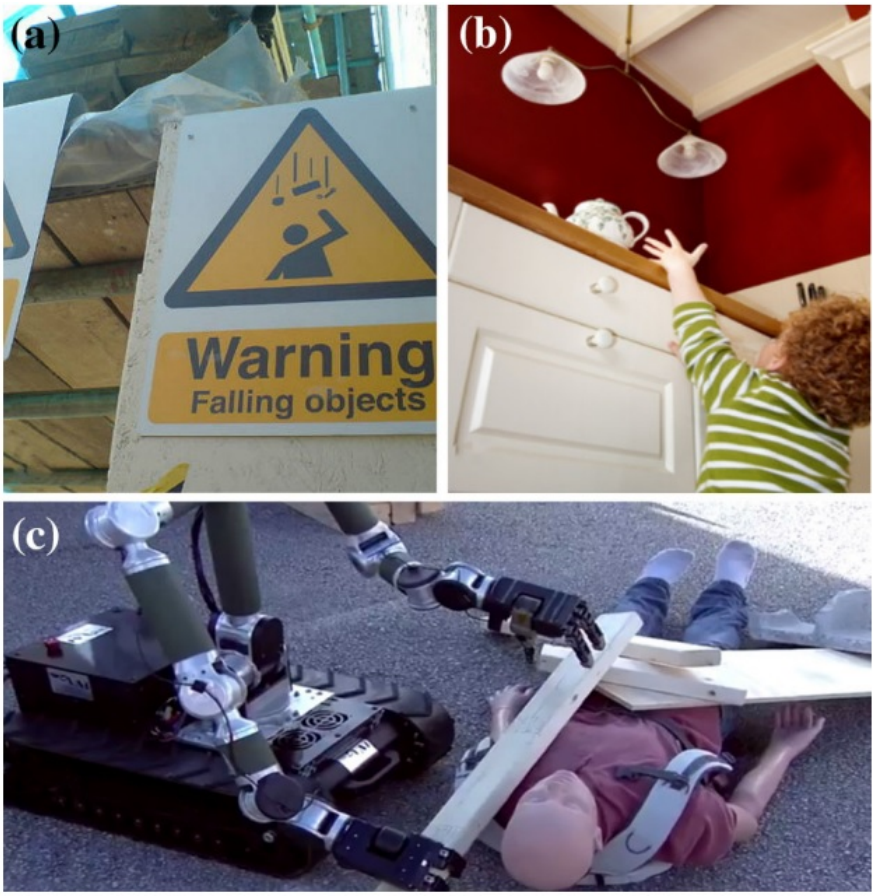


Figure 4.1: A safety-aware robot can be used to detect potentially physically unstable objects in a variety of situations: **a** falling objects at a constructions site, **b** the human assistant for baby proofing, and **c** the disaster rescue (from the recent DARPA Robotics Challenge), where the Multi-Arm robot needs to understand the physical relationships among the debris

ii) Human assistant robots Objects have the risk to fall onto/hit people at workplaces, such as the construction site in Fig. 4.1a. To prevent objects from falling freely from one level to another, safety surveillance ensures that people will not get hurt by their environment, especially for children, elders and people with disabilities. As the example shows in Fig. 4.1b, we see a child is reaching for a teapot, and we can predict possible consequences of his actionthe teapot may fall, and the child may get hurt by the falling teapot.

iii) Disaster rescue robots. The Fig.4.1 (c) shows a HDR-IAI Multi-Arm robot rescuing people during a mock disaster of 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.

4.1.2 Overview of our approach

We address the problem of detecting potential falling objects by inferring hidden "causes" (disturbance) and reasoning possible "effect" (falling) using intuitive mechanics. Taking a 3D point cloud as the input as shown in Fig.4.2 (a), our method first segments the point cloud and recovers volumetric 3D objects in the scene following a recent approach by Zheng et al. [ZZY13], and predicts the walkable area by hallucinating the human actions [GGV11, GSE11, JKS13]. 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.4.2 (b).

i) We infer the disturbance field caused by earthquake or wind, as well as the human activities. A disturbance field representing the possible physical work applied to each position in the 3D space. We use the motion capture data of human actions, as the red stick figures in Fig.4.2 (b), and situate 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 a primary motions on the 2D ground plan which recodes the visiting frequency and walking direction for each walkable position, and add detailed secondary body part motions in 3D space on top. We estimate the distribution of primary motions by synthesizing human walking trajectories following two simple observations: (a) A rational agent mostly

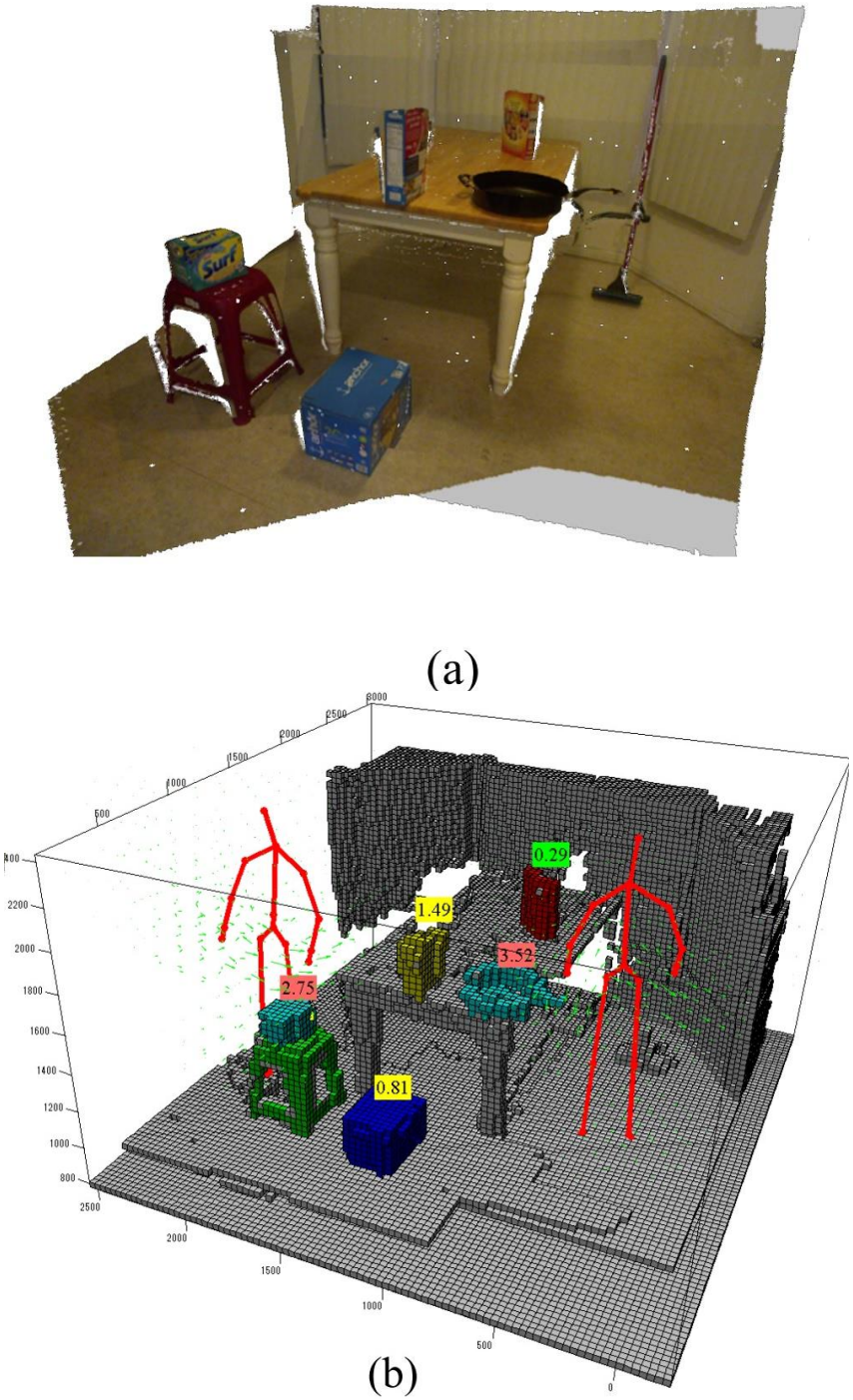


Figure 4.2: (a) The input point cloud; (b) Imagined human action field and detected potential falling objects with red tags

walks along a shortest path with minimal effort; (b) A agent has a basic need to travel between any two walkable positions in the scene. As a result, a convex corner, like the table corners in Fig.4.2 (b), has a high probability to be visited, and the pan on the corner of the table are less safe than others. Similarly, the box on the chair is easy to be knocked off the stool by a swinging hand as well.

ii) We then reason "effects" (falling) of each possible disturbance (an accidental collision) by intuitive mechanism. We first decompose the velocity of input disturbance according to the directions of rotational movement (rolling) and translational movement (sliding) by a parallelogram rule. And we calculate the initial kinetic energy of object after a collision as an input work to the system. According to two principles: conservation of kinetic energy and conservation of momentum, we can infer that the velocity of the object after the collision. We then 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 calculate the risk of releasing the energy in reaching a deeper minimum.

In experiments, we quantitatively evaluated the accuracy of potential falling object detection, as well as the ranking of falling risk w.r.t. human judgments on a challenging dataset.

4.2 Definition of the falling risk

We measure the risk of a potential falling object as illustrated in Fig.4.3. The curve represents the change of potential energy in terms of different positions. At the beginning, an object a stays in the position x_0 which is a stable equilibrium. When a work W applies to the object, it start to move upward towards the position of unstable equilibrium $\tilde{x}$. The total energy needed to go over the unstable equilibrium $\Delta E(x_0 \rightarrow \tilde{x})$ is called "energy barrier". If the work is larger than the energy barrier $W \geq \Delta E(x_0 \rightarrow \tilde{x})$, then the object will fall over the unstable equilibrium. In this way, we define the falling risk as:

Definition 1. The falling risk $R(a, x_0, W)$ of an entity a at x_0 in the presence of a disturbance work W is the maximum energy that it can release when it moves out the

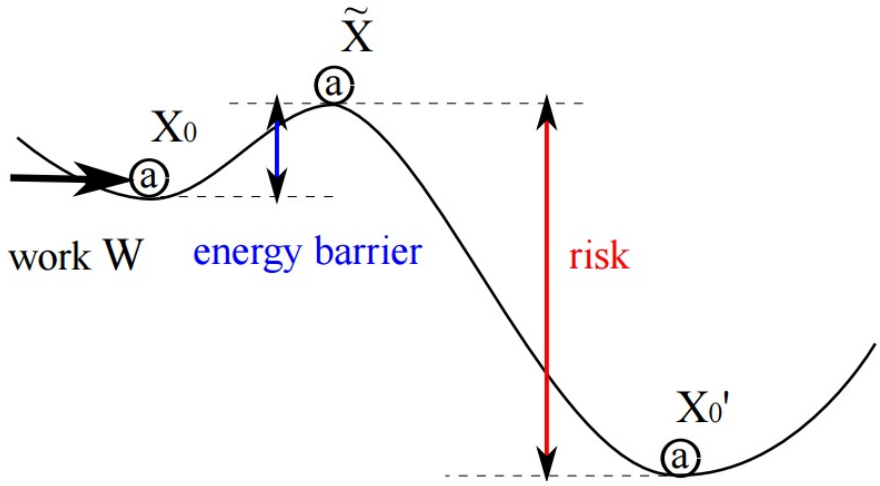


Figure 4.3: An illustration of falling risk definition and other basic concepts on the potential energy curve;

energy barrier by the work W.

$$R(a, x_0, W) = \delta[W \geq \Delta E(x_0 \to \tilde{x})]\Delta E(\tilde{x} \to x_0') \tag{4.1}$$

$\delta()$ is an indicator function and $\delta(z) = 1$ if condition z is satisfied otherwise $\delta(z) = 0$.

Definition 2. The falling risk $R(a, x_0)$ of an entity a at position x_0 in the presence of a disturbance field $p(W, x)$ is the expected risk with respect to the disturbance distribution.

$$R(a, x_0) = \int p(W, x_0) R(a, x_0, W) \Delta W, \tag{4.2}$$

The energy barrier $\Delta E(x_0 \to \tilde{x})$ is the minimum energy needed to move from the current state (say a local minimum) x_0 to an unstable equilibrium $\tilde{x}$. For example, as shown in Fig. 4.4, when a cone is currently in stable state B, its energy barrier is the minimum work needed to push it out of the current energy basin. Passing that point B', the cone will fall to a new stable state at lower position. Also, for example, when a cup is at the center of the table, its energy barrier is the minimum work needed (to overcome friction) to push it to the edge.

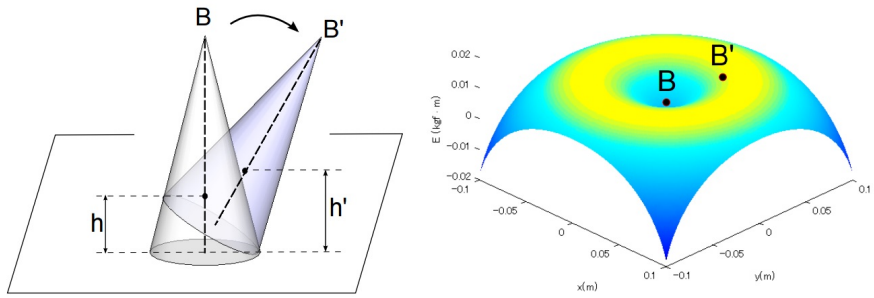


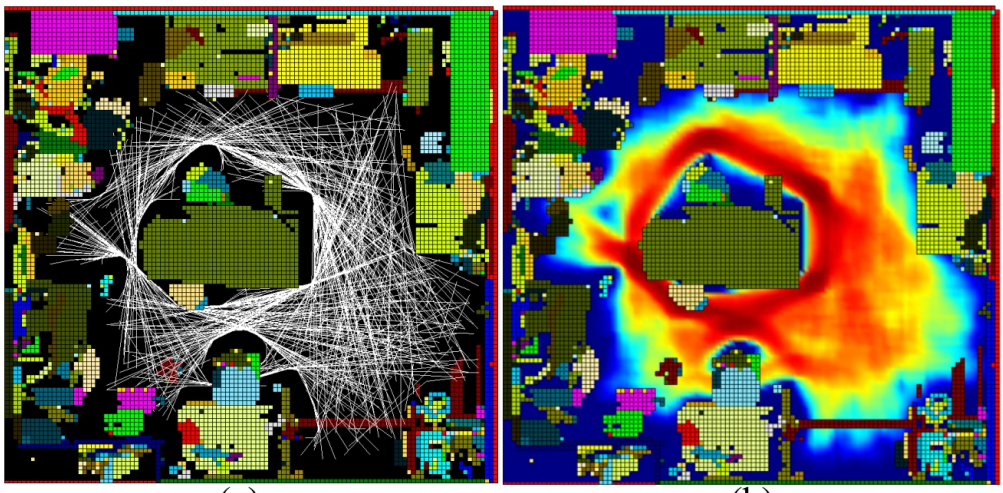
Figure 4.4: An simple example that a cone is being knocked down. It is pushed up from the stable equilibrium B, and about to go over the energy barrier B'. The correspondent potential energy map is on the right.

The potential falling risk $\Delta E(\tilde{x} \rightarrow x_0')$ is the energy released when an entity moves from its unstable equilibrium $\tilde{x}$ to a lower minimum x_0. For example, when the cup falls of from the edge of the table to the ground. The higher the table, the larger the energy risk $\Delta E(\tilde{x} \rightarrow x_0')$.

With the definition of the potential falling object, we introduce the inference of the disturbance field in Sect.III and the calculation of potential energy and initial kinetic energy given a disturbance in Sect.IV.

4.3 Inferring the disturbance field

Taking a 3D point cloud as the input as shown in Fig.4.2 (a), our method first segments the point cloud and recovers volumetric 3d objects the scene following a recent approach by Zheng et al. [ZZY13], and predict the walkable area and sittable area by hallucinating the human actions [GGV11, GSE11]. The result is shown in Fig.4.2 (b). In order to approximate arbitrary shape of 3D objects, we discretize the 3D space to voxels, which are the smallest units in the space. So that all the 3D entities are represented by a group of voxels. In such recovered 3D environment, we then estimate disturbance field caused by natural forces and human actions.



(a) (b)

Figure 4.5: 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.

4.3.1 Natural disturbance field

Despite the gravity applies a constant downward force to all the voxels, other natural disturbances such as earthquakes and winds are also present in a natural scene.

1) Earthquake transmits energy by forces of interactions between contacting faces, 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 need to simulate the fluid flow in the space. Here, we simplify it as an uniformly distributed field over the space.

4.3.2 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.4.5 and

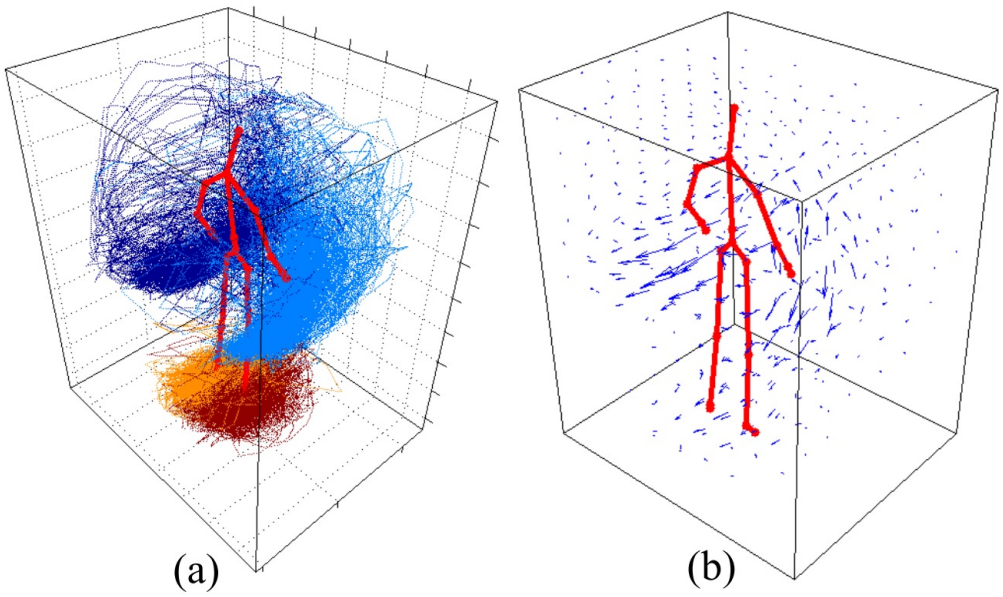


Figure 4.6: Secondary motion field: (a) Secondary motion trajectories of hands and feet from motion capture data; (b) Distribution of the secondary motion field. Long vectors represent large velocity of body movement.

the secondary motions i.e. the body parts movements in Fig.4.6. 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.

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) A 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 that connecting these two positions as shown in Fig.4.5 (a). And we calculate the walking frequency as well as walking directions based on the synthesized trajectories. Fig.4.5 (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 the Fig.4.5 (b), we can see some more details 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 a concave corners. A hallway connecting two walkable area is also frequently visited, and objects in the hallway are less safe too. It is worth noting that the distribution of moving direction is also very distinctive, it help us to locate human body move in the right direction to generate the human disturbance field.

Secondary motion field is the movement thats not part of the main action e.g. arms swinging while walking. But 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 use the Kinect camera to collect human motion capture data Fig.4.6 (a), and then calculate the distribution of moving velocity as shown in Fig.4.6 (b).

The primary motion field further convolves with secondary motion field, thus generate a dense disturbance field that capturing 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 cup 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 demonstrate in Fig.4.7.

4.4 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 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 cup falling off from a table, huge amount of action need to be simulated until meeting the case that human body acting on the cup. The relation between intuitive physical model and human psychology was discussed by recent

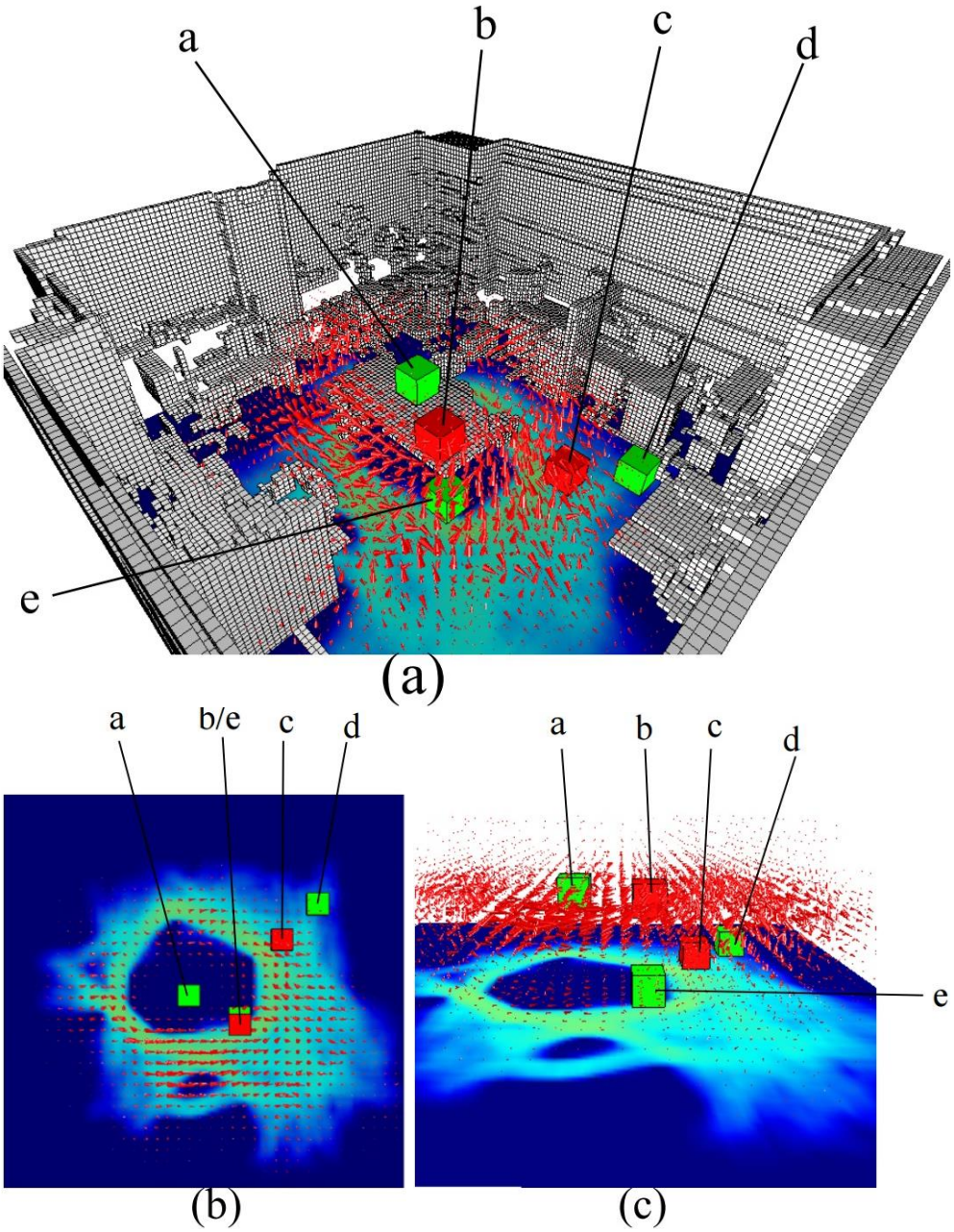


Figure 4.7: The integrated human action field by convolving primary motions with secondary motions. The objects **a-e** are five typical cases in the disturbance field: the object **b** on edge of table and the object **c** along the passway exhibit more disturbances (accidental collisions) than other objects such as **a** in the center of the table, **e** below the table and **d** on a concave corner of space.

cognitive study [HBT11].

In this chapter, to obtain a simple intuitive physical model 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: 0.6, uniform density: $700kg/m^3$).
- 3. A scene is a dissipative mechanical system that total mechanical energy along any trajectory is always decreasing caused by friction, while kinetic and potential energy may be traded off at different states due to elastic collision.

4.4.1 Initial kinetic energy after an elastic collision

We now calculate the initial kinetic energy, which is considered as the input work in Fig. 4.3 after an elastic collision. Here, we simplify objects as mass points to illustrate the simple idea, we will extend the model to more general rigid bodies with arbitrary shapes and arbitrary collision points in the next sub-section.

A head-on elastic collision between two bodies can be represented by velocities in one dimension along a line passing through the bodies. If the velocities are u_1 and u_2 before the collision and v_1 and v_2 after, the equations expressing conservation of momentum and kinetic energy are:

$$m_1u_1 + m_2u_2 = m_1v_1 + m_2v_2 \tag{4.3}$$

$$\frac{1}{2}m_1u_1^2 + \frac{1}{2}m_2u_2^2 = \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2 \tag{4.4}$$

Considering the case that one hand with m_1 knocked off a cup with m_2, we set the initial velocities of hand as u_1 and the cup is still $u_2 = 0$. The final velocity of the cup is given by

$$v_2 = (\frac{2m_1}{m_1 + m_2})v_1. \tag{4.5}$$

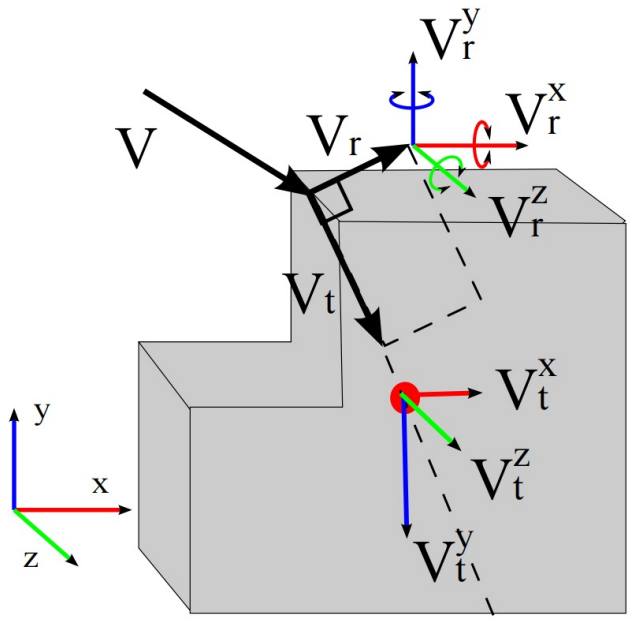


Figure 4.8: The decomposition of action velocity. The gray polygon represents an object with its center of mass on the red dot. The action velocity V first decompose as a rotational velocity V_r and translational velocity V_t, and each velocity is further decomposed as three components along three dimensions.

If the cup has the same mass as the hand, then the hand that was moving is now stopped and the cup is moving away at speed u1. However, if the hand collide with a table with much greater mass, then the table will be little affected by a collision while the hand will be rebounded back. Given the initial velocity of the object, we can easily calculate the initial kinetic energy, which is also the input work in Fig.4.3:

$$W = E_k = \frac{1}{2} m_2 v_2^2 = \frac{2m_1^2 m_2}{(m_1 + m_2)^2} u_1^2 \tag{4.6}$$

4.4.2 Decomposition of the force, the velocity and the momentum

Here, we treat the object as a rigid body with arbitrary shape. As shown in Fig.4.8, the input force V can be decomposed to a force V_t along a line passing through the center of mass and another force V_r perpendicular to V_t. The former force V_t generates an translational movement, while the latter force V_r generates an rotational movement. V_t can further be

decomposed as three velocities V_t^x, V_t^y, V_t^z along three axes, and V_r is decomposed as three rotational velocity V_r^x, V_r^y, V_r^z around three axes. The input force or momentum can be decomposed in the same way.

Consider the object supported by a flat surface from the bottom, we can ignore V_t^y because it will be rebounded back along the y axis as we discussed before. We can also ignore the V_r^y because the rotation around the y axis will not change potential energy, and it also suffer a large friction at the time.

4.4.3 Potential energy

As we discussed in the Sect.II, we calculate an energy map of potential energy. By comparing the input work with the energy landscapes on potential energy map, we calculate the falling risk according to Eq.4.1 and Eq.4.2. In a same spirit of decomposition above, fortunately we can decompose the change of potential energy according to rotation (rolling by itself) movement and translation (position change) movement. By ignoring the translation and rotation along y axis, we calculate the rotational energy map according to two vectors V_r^x, V_r^z, which can be also projected onto spherical coordinate system see [Kri97]; and calculate the translational energy map according to the V_t^x, V_t^z.

Fig.4.9 shows a simple example, giving energy a book is falling off table. We roughly decompose this process into two sub-steps: 1) it rolls from stable state (in black) to unstable state (in blue); and 2) it falls off to the position (in yellow) as a mass point. Therefore we can draw the state change (along the blue and yellow arrows) on the energy maps shown in Fig. 4.9 (b) and (c) respectively. In each energy map, red means high potential energy, whereas blue means low potential energy. We can see that the object is initially lying at the energy minimum (stable equilibrium) on both maps, and it need some work to push out of the unstable equilibrium. Once it is pushed into the unstable states, the case in Fig.4.9 (c) releases much more energy than that in Fig.4.9 (b).

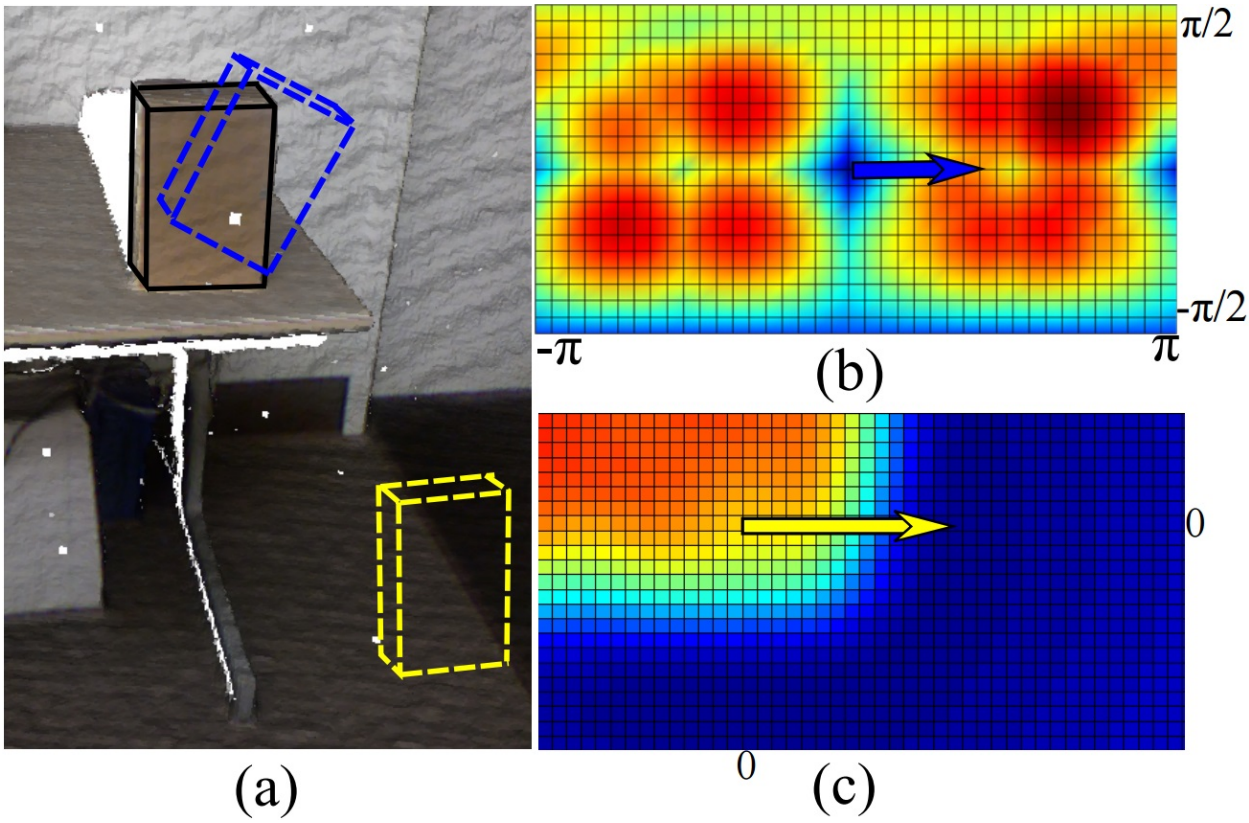


Figure 4.9: Potential energy map for (b) the rotational movement and (c) the displacement movement of the box on a table in (a).

4.5 Experiments

In our experiments, we evaluate our approach by two datasets of large-scale point clouds. The first dataset captured by Microsoft Kinect sensor contains 100 scenes, and each scene is composed by 20-30 rgb-depth images with a powerful SLAM algorithm [NIH11]. Another dataset is captured by a high-end 3D sensor Leica ScanStation C10. It contains 20 large scenes, and each snapshot of the sensor scans 260 rgb-depth images covers a panorama of the scene.

4.5.1 Qualitative evaluation

As shown in Fig. 4.10, we compare the potential falling objects under three different disturbance fields: 1) The human action field in Fig. 4.10 (b,e); 2) The wind field (an uniform directional field) in Fig. 4.10 (c,f) and 3) earthquake (random forces on contacting object surface) in Fig. 4.10 (d,g). As we can see the cups with red tags are detected potential falling objects, which are very close to human judgments.

In Fig.4.11, we show four large-scale point clouds in each row, where (a) shows input 3D point clouds with rgb color for reference; (b) illustrates inferred human action fields, the larger and more complex environment like the last scene on the bottom exhibits more sophisticated motion patterns, which beautifully matches with human motion patterns; (c) shows a overview of potential falling objects with their risk scores on yellow tags; and (d) shows the zoom-in details of some typical successful and failure detection examples. Some false positives may caused by highly occlusions.

4.5.2 Quantitative evaluation

We conduct two quantitative evaluations:

Accuracy of potential falling object detection. In this experiment, we first manually labeled 83 potential falling objects from 20 large scale point clouds, some of them are shown in 2 5 7 9 10 Fig.11. The groundtruth come from majority vote ($> 50\%$) of 10 partic-

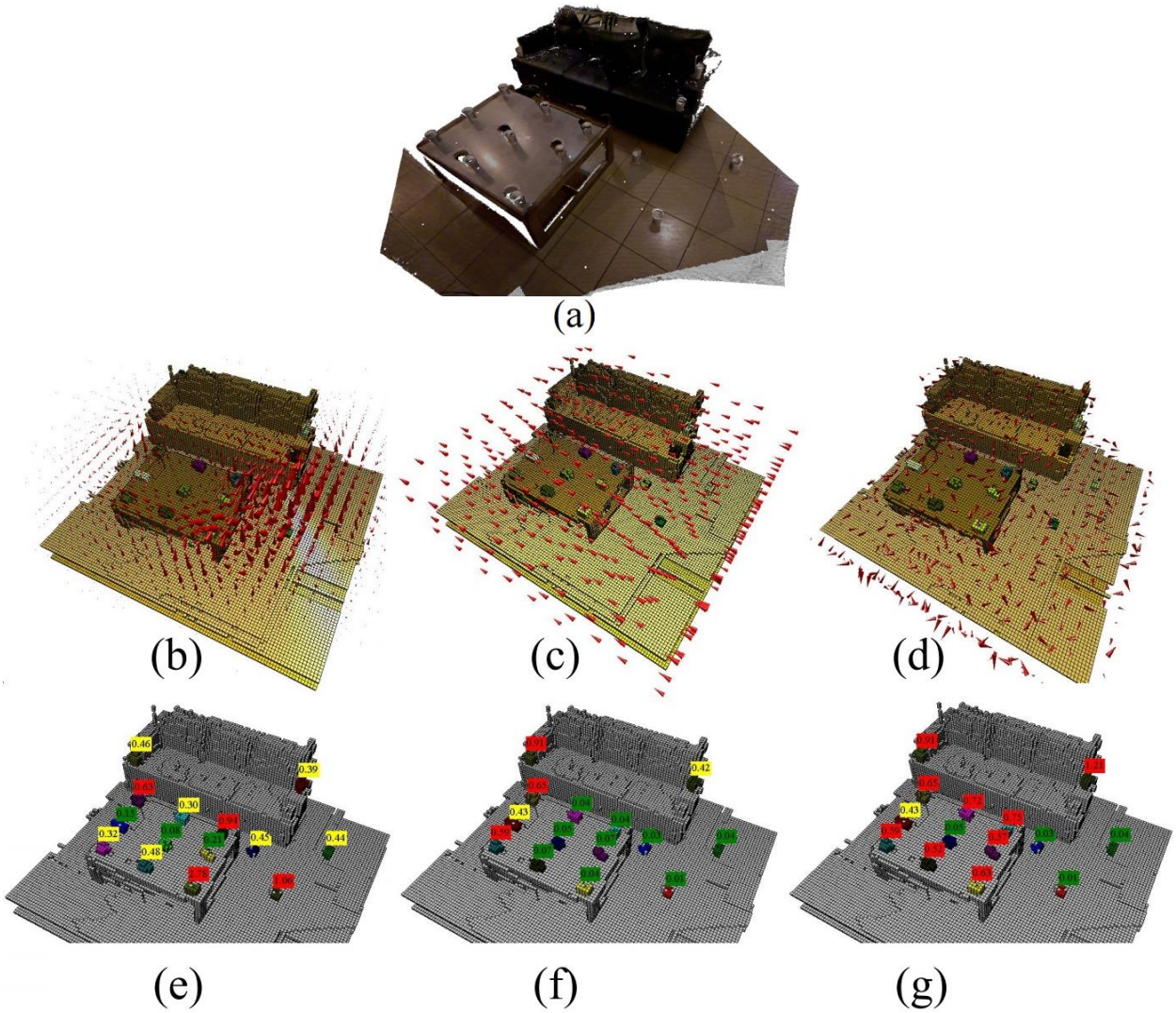


Figure 4.10: The potential falling objects (with red tags) under the human action field (b,e), the wind field (c,f) and the earthquake field (d,g) respectively. The results match with human perception: (i) objects around table corner are not safe w.r.t human walking action; (ii) object along the edge 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.

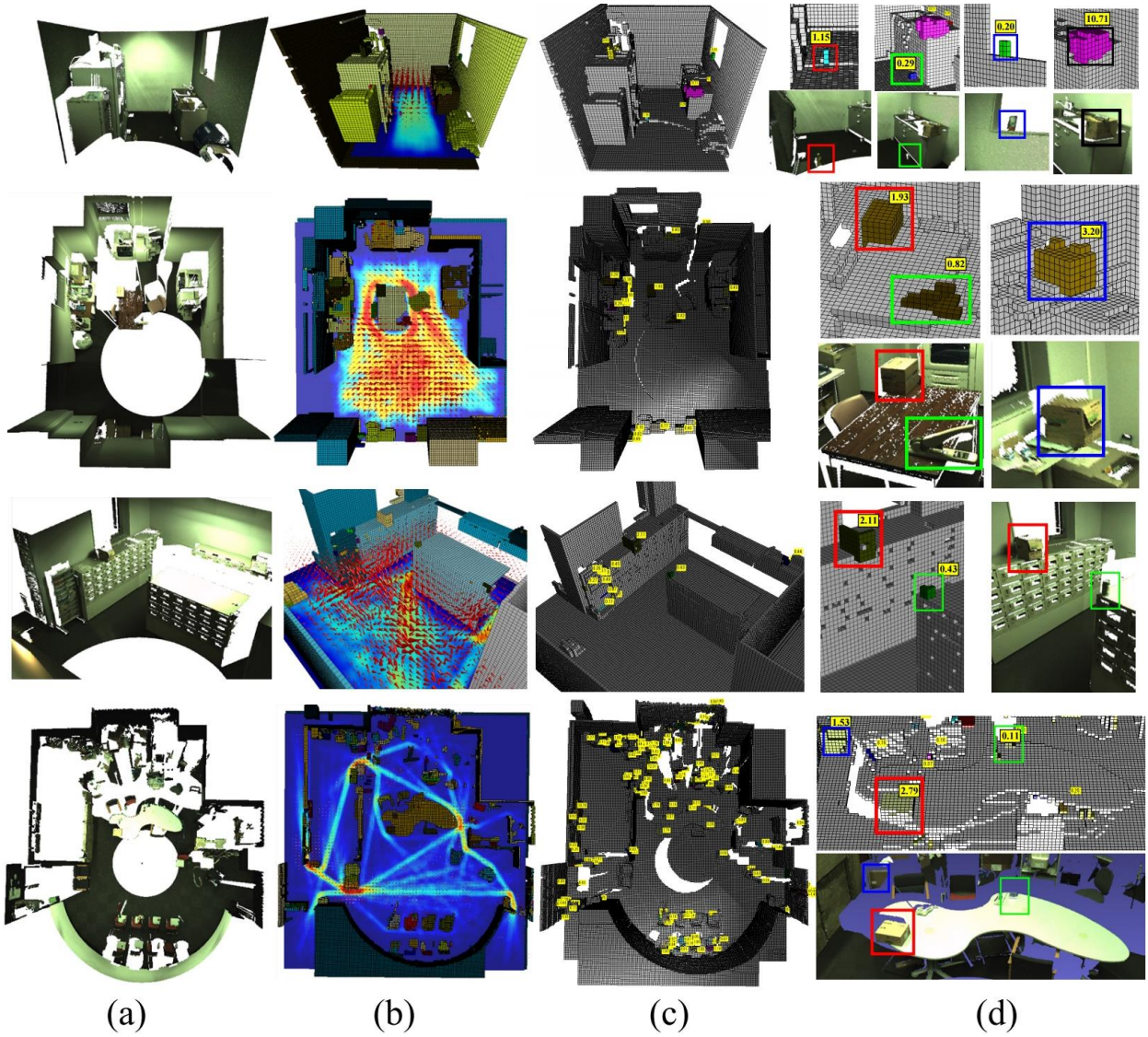


Figure 4.11: (a) Input 3D scene point clouds; (b) Inferred human action fields and segmented objects shown in different colors; (c) Detected potential falling objects with their risk scores on the yellow labels; (d) Some zoom-in details of detected potential falling objects. See text for more explanation.

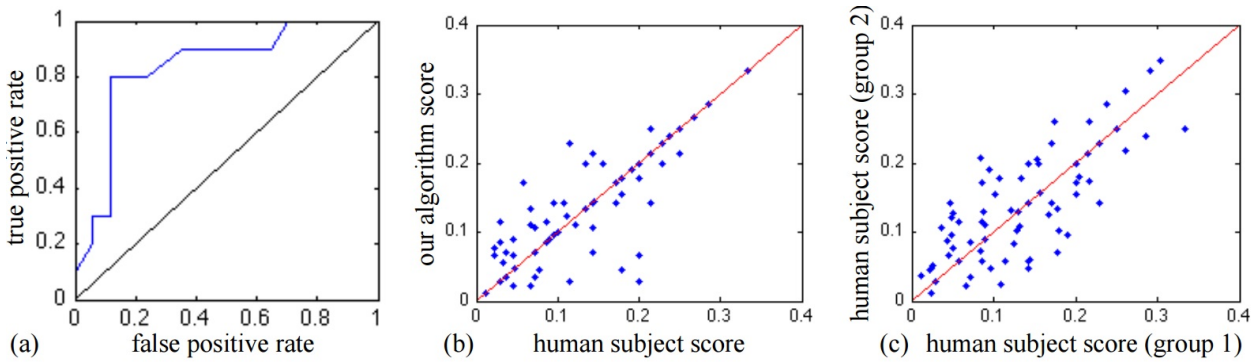


Figure 4.12: (a) The ROC curve of the potential falling object detection (b) The correlation between the falling risk ranking by our algorithm and the ranking by human subjects. (c) The correlation between the falling risk ranking by two different random split groups of human subjects.

ipants. We calculated the ROC curve of potential falling object detection by our proposed approaches in Fig.12 (a). It is shown that our algorithm can reliably detect potential falling objects with 80% true positive rate and keep a 20% false positive rate at the same time.

Ranking of falling risk. The human judgments of potential falling objects can be very subjective, and they may not be reliable ground truths. Instead of calculating the error rate, we compare the ranking of several potential falling objects in a scene with the ranking of human judgment in this experiment. We asked 10 participants to choose a reasonable order of the object according to their falling risk. The results are shown in Fig.4.12 (b) where the model output fit well with the human judgment, but still keep a certain variance. Then we conducted a similar experiment. We random split the participants into two groups, and evaluate the correlation between these two groups. As shown in Fig.4.12 (c), the correlation between human judgments keep the same amount of variance as the correlation between model and human. It is also interesting to note that the variance is larger when the risk score is low (lower left corner of Fig.4.12 (b,c)), or say the falling risk judgment will become less ambiguous when the risk is higher.

The similar judgment correlation between machine and human in Fig.4.12(b,c) implies the algorithm may pass the Turing test because the judge cannot reliably tell the machine

from the actual human according to the answers.

CHAPTER 5

Conclusion

In this dissertation, we propose several novel approaches for the three tasks and the extensive experiment results on public datasets demonstrate the superiority and defectiveness of our methods.

Firstly we presented a probabilistic method for single-view 3D scene reconstruction using geometric commonsense and specify a generic probabilistic formula to solve multiple 3d parsing problems simultaneously. We developed a stochastic optimization algorithm to search the optimal parse graph with both bottom-to-up and top-down computations. In evaluations, we collected a new image dataset to include a variety of scene categories and annotated their 3D scene models. Results with comparisons demonstrated that our method is capable of accurately reconstructing a wide variety of scene categories. We also demonstrate that our method can be used to disclose the valid commonsense knowledge used to explained an image. As the first piece of works in its catalog, our studies are able to enhance our understandings of geometric commonsense knowledge and their critical role in computer vision. The developed techniques can also be easily extended to the broader scope of commonsense reasoning.

Secondly 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.

Thirdly also presents a novel approach for detecting potential unsafe objects. We demon-

strated 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-ofthe-art algorithms. Further studies along this way, including friction, material properties, causal reasoning, can be very interesting dimensions of vision research.

REFERENCES

[AFS06] Marco Attene, Bianca Falcidieno, and Michela Spagnuolo. "Hierarchical mesh segmentation based on fitting primitives." *The Visual Computer*, **22**(3):181–193, 2006.

[AKJ13] Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. "Contextually guided semantic labeling and search for three-dimensional point clouds." *The International Journal of Robotics Research*, **32**(1):19–34, 2013.

[BLC00] Michael M Blane, Zhibin Lei, Hakan Çivi, and David B Cooper. "The 3L algorithm for fitting implicit polynomial curves and surfaces to data." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(3):298–313, 2000.

[BMR82] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. "Scene perception: Detecting and judging objects undergoing relational violations." *Cognitive psychology*, **14**(2):143–177, 1982.

[BZ05] Adrian Barbu and Song-Chun Zhu. "Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(8):1239–1253, 2005.

[CF14] Ricardo Cabral and Yasutaka Furukawa. "Piecewise planar and compact floorplan reconstruction from images." In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 628–635. IEEE, 2014.

[CGF09] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. "A benchmark for 3D mesh segmentation." In *Acm transactions on graphics (tog)*, volume 28, p. 73. ACM, 2009.

[CRZ00] Antonio Criminisi, Ian Reid, and Andrew Zisserman. "Single view metrology." *International Journal of Computer Vision*, **40**(2):123–148, 2000.

[CTW12] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. "A tree-based context model for object recognition." *IEEE transactions on pattern analysis and machine intelligence*, **34**(2):240–252, 2012.

[CY03] James M Coughlan and Alan L Yuille. "Manhattan world: Orientation and outlier detection by bayesian inference." *Neural Computation*, **15**(5):1063–1088, 2003.

[CZ00] Antonio Criminisi and Andrew Zisserman. "Shape from Texture: Homogeneity Revisited." In *BMVC*, pp. 1–10, 2000.

[DBK13] Luca Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. "Understanding bayesian rooms using composite 3d object models." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 153–160, 2013.

[DFL12] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. "Scene semantics from long-term observation of people." In *European conference on computer vision*, pp. 284–298. Springer, 2012.

[DGB11] Luca Del Pero, Jinyan Guan, Ernesto Brau, Joseph Schlecht, and Kobus Barnard. "Sampling bedrooms." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 2009–2016. IEEE, 2011.

[DM15] Ernest Davis and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence." *Communications of the ACM*, **58**(9):92–103, 2015.

[DSS93] Randall Davis, Howard Shrobe, and Peter Szolovits. "What is a knowledge representation?" *AI magazine*, **14**(1):17, 1993.

[DST00] Frank Dellaert, Steven M Seitz, Charles E Thorpe, and Sebastian Thrun. "Structure from motion without correspondence." In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pp. 557–564. IEEE, 2000.

[EEV15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes challenge: A retrospective." *International Journal of Computer Vision*, **111**(1):98–136, 2015.

[FBB10] RW Fleming, M Barnett-Cowan, and HH Bülthoff. "Perceived object stability is affected by the internal representation of gravity." In *33rd European Conference on Visual Perception*, 2010.

[FCS09] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. "Manhattan-world stereo." In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1422–1429. IEEE, 2009.

[FDG14] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. "People watching: Human actions as a cue for single view geometry." *International journal of computer vision*, **110**(3):259–274, 2014.

[FGM10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. "Object detection with discriminatively trained part-based models." *IEEE transactions on pattern analysis and machine intelligence*, **32**(9):1627–1645, 2010.

[FV10] Pedro F Felzenszwalb and Olga Veksler. "Tiered scene labeling with dynamic programming." In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3097–3104. IEEE, 2010.

[GEH10] Abhinav Gupta, Alexei Efros, and Martial Hebert. "Blocks world revisited: Image understanding using qualitative geometry and mechanics." *Computer Vision–ECCV 2010*, pp. 482–496, 2010.

[GGV11] Helmut Grabner, Juergen Gall, and Luc Van Gool. "What makes a chair a chair?" In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1529–1536. IEEE, 2011.

[GHK10] Abhinav Gupta, Martial Hebert, Takeo Kanade, and David M Blei. "Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces." In *Advances in neural information processing systems*, pp. 1288–1296, 2010.

[GSE11] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. "From 3d scene geometry to human workspace." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1961–1968. IEEE, 2011.

[HBT11] Jessica Hamrick, Peter Battaglia, and Joshua B Tenenbaum. "Internal physics models guide probabilistic judgments about object dynamics." In *Proceedings of the 33rd annual conference of the cognitive science society*, pp. 1545–1550. Cognitive Science Society Austin, TX, 2011.

[HEH05] Derek Hoiem, Alexei A Efros, and Martial Hebert. "Geometric context from a single image." In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pp. 654–661. IEEE, 2005.

[HEH08a] Derek Hoiem, Alexei A Efros, and Martial Hebert. "Closing the loop in scene interpretation." In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

[HEH08b] Derek Hoiem, Alexei A Efros, and Martial Hebert. "Putting objects in perspective." *International Journal of Computer Vision*, **80**(1):3–15, 2008.

[HGS09] Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. "Cascaded classification models: Combining models for holistic scene understanding." In *Advances in Neural Information Processing Systems*, pp. 641–648, 2009.

[HHF10] Varsha Hedau, Derek Hoiem, and David Forsyth. "Thinking inside the box: Using appearance models and context based on room geometry." *Computer Vision–ECCV 2010*, pp. 224–237, 2010.

[HR14] Mohsen Hejrati and Deva Ramanan. "Analysis by synthesis: 3d object recognition by object reconstruction." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2449–2456, 2014.

[HSP14] Christian Hane, Nikolay Savinov, and Marc Pollefeys. "Class specific 3d object shape priors using surface normals." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–659, 2014.

[HZ03] Feng Han and Song-Chun Zhu. "Bayesian reconstruction of 3d shapes and scenes from a single image." In *Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003. HLK 2003. First IEEE International Workshop on*, pp. 12–20. IEEE, 2003.

[HZ09] Feng Han and Song-Chun Zhu. "Bottom-up/top-down image parsing with attribute grammar." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(1):59–73, 2009.

[HZC13] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. "Joint 3D scene reconstruction and class segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 97–104, 2013.

[JGS13] Zhaoyin Jia, Andrew Gallagher, Ashutosh Saxena, and Tsuhan Chen. "3d-based reasoning with blocks, support, and stability." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2013.

[JKJ13] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. "A category-level 3d object dataset: Putting the kinect to work." In *Consumer Depth Cameras for Computer Vision*, pp. 141–165. Springer, 2013.

[JKS13] Yun Jiang, Hema Koppula, and Ashutosh Saxena. "Hallucinated humans as the hidden context for labeling 3d scenes." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2993–3000, 2013.

[JS13] Yun Jiang and Ashutosh Saxena. "Infinite Latent Conditional Random Fields for Modeling Environments through Humans." In *Robotics: Science and Systems*, pp. 1–8, 2013.

[Kri97] David J Kriegman. "Let them fall where they may: Capture regions of curved objects and polyhedra." *The International Journal of Robotics Research*, **16**(4):448–472, 1997.

[KST09] Panagiotis Koutsourakis, Loic Simon, Olivier Teboul, Georgios Tziritas, and Nikos Paragios. "Single view reconstruction using shape grammars for urban environments." In *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1795–1802. IEEE, 2009.

[LHK09] David C Lee, Martial Hebert, and Takeo Kanade. "Geometric reasoning for single image structure recovery." In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2136–2143. IEEE, 2009.

[LK81] Bruce D Lucas, Takeo Kanade, et al. "An iterative image registration technique with an application to stereo vision." 1981.

[LML16] Xiaobai Liu, Yadong Mu, and Liang Lin. "A Stochastic Grammar for Fine-grained 3D Scene Reconstruction." In *IJCAI*, 2016.

[LMP01] John Lafferty, Andrew McCallum, Fernando Pereira, et al. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pp. 282–289, 2001.

[LPY12] Bo Li, Kun Peng, Xianghua Ying, and Hongbin Zha. "Vanishing point detection using cascaded 1D Hough Transform from single images." *Pattern Recognition Letters*, **33**(1):1–8, 2012.

[LZZ14] Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. "Single-view 3d scene parsing by attributed grammar." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 684–691, 2014.

[Mar82] David Marr. "Vision." 1982.

[MFY16] Roozbeh Mottaghi, Sanja Fidler, Alan Yuille, Raquel Urtasun, and Devi Parikh. "Human-machine CRFs for identifying bottlenecks in scene understanding." *IEEE transactions on pattern analysis and machine intelligence*, **38**(1):74–87, 2016.

[MZY11] Hossein Mobahi, Zihan Zhou, Allen Y Yang, and Yi Ma. "Holistic 3d reconstruction of urban structures from low-rank textures." In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 593–600. IEEE, 2011.

[NIH11] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. "KinectFusion: Real-time dense surface mapping and tracking." In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pp. 127–136. IEEE, 2011.

[NN90] Shree K Nayar and Yasuo Nakagawa. "Shape from focus: An effective approach for rough surfaces." In *Robotics and Automation, 1990. Proceedings., 1990 IEEE International Conference on*, pp. 218–225. IEEE, 1990.

[PF05] Stéphane Petti and Thierry Fraichard. "Safe motion planning in dynamic environments." In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pp. 2210–2215. IEEE, 2005.

[PL11] Mike Phillips and Maxim Likhachev. "Sipp: Safe interval path planning for dynamic environments." In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 5628–5635. IEEE, 2011.

[PM11] Jan Prokaj and Gérard Medioni. "Using 3d scene structure to improve tracking." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1337–1344. IEEE, 2011.

[PT11] Nadia Payet and Sinisa Todorovic. "Scene shape from texture of objects." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 2017–2024. IEEE, 2011.

[PVB08] Jann Poppinga, Narunas Vaskevicius, Andreas Birk, and Kaustubh Pathak. "Fast plane detection and polygonalization in noisy 3D range images." In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pp. 3378–3383. IEEE, 2008.

[RM03] Xiaofeng Ren and Jitendra Malik. "Learning a Classification Model for Segmentation." In *ICCV*, volume 1, pp. 10–17, 2003.

[RSN13] Jesús Martínez del Rincón, Maria J Santofimia, and Jean-Christophe Nebel. "Common-sense reasoning for human action recognition." *Pattern Recognition Letters*, **34**(15):1849–1860, 2013.

[SF83] Qing-Yun Shi and King-Sun Fu. "Parsing and translation of (attributed) expansive graph languages for scene analysis." *IEEE transactions on pattern analysis and machine intelligence*, (5):472–485, 1983.

[SHK12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. "Indoor segmentation and support inference from rgbd images." *Computer Vision–ECCV 2012*, pp. 746–760, 2012.

[SNI05] Ryusuke Sagawa, Ko Nishino, and Katsushi Ikeuchi. "Adaptively merging large-scale range data with reflectance properties." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(3):392–405, 2005.

[SRF14] Julian Straub, Guy Rosman, Oren Freifeld, John J Leonard, and John W Fisher. "A mixture of manhattan frames: Beyond the manhattan world." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3770–3777, 2014.

[SSN09] Ashutosh Saxena, Min Sun, and Andrew Y Ng. "Make3d: Learning 3d scene structure from a single still image." *IEEE transactions on pattern analysis and machine intelligence*, **31**(5):824–840, 2009.

[SU12] Alexander Schwing and Raquel Urtasun. "Efficient exact inference for 3d indoor scene understanding." *Computer Vision–ECCV 2012*, pp. 299–313, 2012.

[TC03] Zhuowen Tu, Xiangrong Chen, et al. "Image parsing: Unifying segmentation, detection, and recognition." In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 18–25. IEEE, 2003.

[TNC13] Eno Toppe, Claudia Nieuwenhuis, and Daniel Cremers. "Relative volume constraints for single view 3D reconstruction." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 177–184, 2013.

[TZ02] Zhuowen Tu and Song-Chun Zhu. "Image segmentation by data-driven Markov chain Monte Carlo." *IEEE Transactions on pattern analysis and machine intelligence*, **24**(5):657–673, 2002.

[Wal03] David Wales. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

[WFU15a] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. "Holistic 3d scene understanding from a single geo-tagged image." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3964–3972, 2015.

[WFU15b] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. "Lost shopping! monocular localization in large indoor spaces." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2695–2703, 2015.

[WJW13] Shuo Wang, Jungseock Joo, Yizhou Wang, and Song-Chun Zhu. "Weakly supervised learning for attribute localization in outdoor scenes." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3111–3118, 2013.

[WPC05] Danny Wyatt, Matthai Philipose, and Tanzeem Choudhury. "Unsupervised activity recognition using automatically mined common sense." In *AAAI*, volume 5, pp. 21–27, 2005.

[WPP07] Shiaokai Wang, William Pentney, Ana-Maria Popescu, Tanzeem Choudhury, and Matthai Philipose. "Common Sense Based Joint Training of Human Activity Recognizers." In *IJCAI*, volume 7, pp. 2237–2242, 2007.

[WZZ13] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. "Modeling 4d human-object interactions for event and object recognition." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3272–3279, 2013.

[XF14] Jianxiong Xiao and Yasutaka Furukawa. "Reconstructing the world's museums." *International journal of computer vision*, **110**(3):243–258, 2014.

[XOT13] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. "Sun3d: A database of big spaces reconstructed using sfm and object labels." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1625–1632, 2013.

[XRT12] Jianxiong Xiao, Bryan Russell, and Antonio Torralba. "Localizing 3D cuboids in single-view images." In *Advances in neural information processing systems*, pp. 746–754, 2012.

[ZFF14] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. "Reasoning about object affordances in a knowledge base representation." In *European conference on computer vision*, pp. 408–424. Springer, 2014.

[Zha00] Zhengyou Zhang. "A flexible new technique for camera calibration." *IEEE Transactions on pattern analysis and machine intelligence*, **22**(11):1330–1334, 2000.

[ZL05] Myrka Zago and Francesco Lacquaniti. "Visual perception and interception of falling objects: a review of evidence for an internal model of gravity." *Journal of Neural Engineering*, **2**(3):S198, 2005.

[ZP13] C Lawrence Zitnick and Devi Parikh. "Bringing semantics into focus using visual abstraction." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3009–3016, 2013.

[ZP14] Bernhard Zeisl, Marc Pollefeys, et al. "Discriminatively trained dense surface normal estimation." In *European conference on computer vision*, pp. 468–484. Springer, 2014.

[ZST14] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. "Panocontext: A whole-room 3d context model for panoramic scene understanding." In *European Conference on Computer Vision*, pp. 668–686. Springer, 2014.

[ZTC99] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. "Shape-from-shading: a survey." *IEEE transactions on pattern analysis and machine intelligence*, **21**(8):690–706, 1999.

[ZTI10] Bo Zheng, Jun Takamatsu, and Katsushi Ikeuchi. "An adaptive and stable method for fitting implicit polynomial curves and surfaces." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(3):561–568, 2010.

[ZZ11] Yibiao Zhao and Song-Chun Zhu. "Image parsing with stochastic scene grammar." In *Advances in Neural Information Processing Systems*, pp. 73–81, 2011.

[ZZC15] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. "Understanding tools: Task-oriented object modeling, learning and recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2855–2864, 2015.

[ZZY13] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. "Beyond point clouds: Scene understanding by reasoning geometry and physics." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3127–3134, 2013.