Information Sciences 320 (2015) 320–332

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Towards efficient support relation extraction from RGBD images

Feng Xue, Shan Xu, Chuan He, Meng Wang, Richang Hong [*]

School of Computer and Information, Hefei University of Technology, Hefei, China

ARTICLE INFO

Article history:
Received 15 June 2014
Received in revised form 5 December 2014
Accepted 13 December 2014
Available online 7 January 2015

Keywords:
RGBD image
Support relation
Scene understanding
Image segmentation

ABSTRACT

To extract reasonable support relations from "RGB + depth" (RGBD) images, it is very important to achieve good scene understanding. This paper proposes a novel approach to extracting accurate support relationships by analyzing the RGBD images of indoor scenes. Noting that the support relations and structure classes of indoor images are inherently related to physical stability, we construct an improved energy function that embodies this stability. We then infer the support relations and structure classes from indoor RGBD images by minimizing this energy function. Moreover, the authors succeed in improving the segmentation quality of RGBD images using the inferred results as input. Compared with previous methods, our approach produces more reasonable support relations and structure classes, where physical stability function is taken into account for resolving the optimization problem. We use the NYU-Depth2 dataset as the training data, and experimental results show that the proposed RGBD image segmentation method based on support relation abstraction produces more accurate results than segmentation methods based on ground-truth support relations.

1. Introduction

Scene understanding aims to interpret and describe a given scene using image processing and analysis techniques. Traditional scene understanding only answers the question of "what is in a scene" and provides descriptive tags for the objects found therein. With the rapid development of computer techniques, these tags have become insufficient for computer vision applications. Therefore, excavating more image cues, such as three-dimensional (3D) structure and physical relations of indoor objects has become a research topic of great interest in computer vision. For some robotic and computer vision applications, precise, automatic support relation extraction of different objects from the scenes are needed for robots to make reasonable decisions. For example, when a robot wants to pick up a cup in a room, it needs to identify not only the cup, but also enough support relations among the objects surrounding the cup so that the adjacent objects remain undisturbed when the cup is removed.

As many researchers have found, physical relations play an important role in computer vision and human interactive applications. Physical support is the most important aspect of these relationships for three reasons. First, it intrinsically reflects the support relations between adjacent objects in the given scene, and is helpful for object recognition. For instance, an object on a desk may be a cup, bowl, or other decorative objects, but cannot be a kitchen cabinet. Second, it may also function as supplementary knowledge for machine vision and scene understanding. When a robot seeks to extract a

[*] Corresponding author. Tel.: +86 13866188972.
E-mail address: hongrc.hfut@gmail.com (R. Hong).

http://dx.doi.org/10.1016/j.ins.2014.12.035
0020-0255/© 2015 Elsevier Inc. All rights reserved.

cookbook that is under a cellphone, it should remove the phone and grab the cookbook in the proper sequence. Finally, support relation priors can be employed to make image segmentation results more accurate.

Recently, the extraction of support relations between objects in "RGB + depth" (RGBD) images has become a new and popular research topic in computer vision. In [22], Silberman et al. initially provided a classic framework for physical support relation extraction from an indoor scene. In their work, standard structure classes and support relation models were formalized for the first time. Typical constraints were also defined and used as the theoretical basis for subsequent support relation inference. However, because of insufficient explorations of the structure classes and support relations of objects, the inferred results in [22] were not accurate enough, and hence the segmentation was unsatisfactory.

Based on the previous work in [22], this paper presents a reliable approach to extracting support relations from RGBD images to obtain accurate relations among indoor objects. The proposed algorithm in this paper segments the input RGBD images into regions that are in accord with individual objects within the scene, and then set up a series of constraints to infer accurate support relations between the segmented regions.

The main contributions of our work are as follows:

(1) We design a new support relation inference method that achieves more accurate inferred results than [22]. The inferred support relations assist scene understanding and other computer vision tasks.
(2) Based on the proposed support relation inferences, segmentation of RGBD images is improved.

2. Related works

There have been many research efforts in the area of scene understanding and support relation extraction. Lee et al. [16] presented an automatic algorithm that could recognize 3D structures inside buildings using an image line detection technique. This algorithm works well, even for hidden objects. In [17], Lee et al. proposed a method that reconstructs the spatial layout of a given scene by inferring the volume of indoor objects. Nempont et al. [21] proposed a new structural interpretation method for complex scenes based on a constraint propagation algorithm. Their method has been illustrated through the example of brain structure recognition in magnetic resonance images. Guo et al. [4] presented a highly distinctive local surface feature called the TriSI feature, and then proposed an effective 3D object recognition algorithm.

Gupta et al. [5] attempted to combine 3D scene geometry and human actions. They proposed a human-centric scene representation that contained a set of reachable human pose states in the scene. However, their work was preliminary and the geometric estimation and human pose analysis could be further improved. Hedau et al. [9] proposed a method to extract free spaces of an indoor scene from a single image and furthermore evaluated their free space estimates in a 3D scene. The authors of [6] built a scene understanding system based on RGBD images that could be used for contour detection, bottom-up grouping, and semantic segmentation. The method in [6] works well on most scene surfaces, but may run into difficulties with some small objects, as discussed in [6]. However, the methods mentioned above mainly focused on geometric inference and object recognition in different scenes, where physical relations among objects are neglected in most cases. Unlike these methods, our goal is to exactly extract the physical relations between adjacent objects using some pre-defined constraints.

As discussed in Section 1, physical relations are indispensable facts in many applications. Because of this, many researchers have switched their focus from traditional scene understanding to physical relation reasoning. Grabner et al. [7] provided a framework for computing support relations between human and objects based on RGBD images. The framework could filter out places where a person could sit or lie down. For example, the algorithm could find the location of a chair upon which a human could sit. However, Grabner's method did not involve support relations between objects. By integrating geometric and physical features, Cupta el al. [8] presented a physical relation description between each pair of objects in a depth image that embodied force, volume, depth, and other factors. Jiang et al. [11] proposed a graphical model to infer locations where some objects could be placed in the scene. The robotic experiments in [11] showed that the algorithm could automatically and stably place known and new objects with high precision. A model based on an "intuitive physics engine" was proposed in [1] to simulate a human's rapid physical inferences to explain how people communicate with others.

Other researchers have shown that stability is useful for scene understanding. In [24], Zheng et al. used a set of point cloud data to infer the geometric features and understand scene structures. In their method, gravity was used as a support stability constraint to analyze the depth image. Their primary goal was to extract the physical stabilities of objects from the point cloud, while we attempt to infer support relations according to a physical stability constraint. Jia et al. [12] produced a 3D representation of the image scene based on a joint optimization over image segments, block fitting, supporting relations, and object stability. In our approach, stability has been used as a useful constraint to infer support constraint. Silberman et al. [22] first parsed the RGBD images into floor, walls, supporting surfaces, and object regions to recover their support relationships. They then used labeled support relations to improve segmentation, but the inferred accuracy rate (around 50–60%) was not desirable.

As many researchers have found, depth information is an efficient supplemental cue for object recognition [13,14,20] and motion detection [15], and can remarkably improve classification accuracy. The work in this paper is mostly related to [12,22,23], however, we propose the novel idea of using RGBD cues to recover support relations between objects in images. In addition, we build a new "closed-loop" framework to support relation extraction and image segmentation. Experimental results in Section 4 show the feasibility and accuracy of the proposed algorithm.

3. Support relation extraction

3.1. Overview

An overview of the proposed physical relation extraction method is shown in Fig. 1. Our algorithm starts from a set of RGBD images and proceeds as follows:

(1) The input depth data are preprocessed. We first retrieve space coordinates from depth images as in [22], extract 3D planes using Random Sample Consensus (RANSAC) [3], and classify each image pixel into corresponding planes using a graph cuts algorithm.
(2) The hierarchical segmentation algorithm is then used to segment the input image into regions that correspond to objects.
(3) We extract SIFT (Scale-invariant feature transform) feature descriptors from segmented regions.
(4) A logistic regression classifier is employed to train the structure classes and support relations on the given dataset.
(5) The new energy function proposed in this paper that integrates physical stability is finally used to infer the structure classes and support relations.

There are two innovations in the proposed support relation extraction framework:

(1) A physical stability constraint is used in the inference procedure, making the extracted relations more accurate and rational. We formalize physical stability as an additional factor and integrate it into the energy function to be minimized by an integer program, as detailed in Section 3.4.3.
(2) Image segmentation is the middle step of the inference algorithm, as shown in Fig. 1. We have found that well segmented results promote the accuracy of the inferred support relations, while accurately extracted support relations improve the segmentation quality in return. Based on this observation, we build a new "closed-loop" framework of support relation extraction and image segmentation, as shown in Fig. 2.

In Fig. 2, the two roles of support relation and image segmentation are mutually called by each other. Initially, we use hierarchical segmentation results [10] to infer support relations. The output support relations and structure classes are then fed back as new features to the second round of image segmentation, which helps to improve the segmentation quality. Finally, we repeat the inference phase using the improved segmentation. Theoretically, this kind of recursive calling between the support relationship inference and image segmentation can further improve the accuracy of the inferred relations. However, the computation cost can become very high. The experiments in Section 4 show the feasibility and effectiveness of our proposed method.

3.2. RGBD image preprocessing

The depth data in an RGBD image includes abundant space and geometric information. Some preprocessing should be performed before further relation inference. In our implementation, the detailed preprocessing phase is as follows:

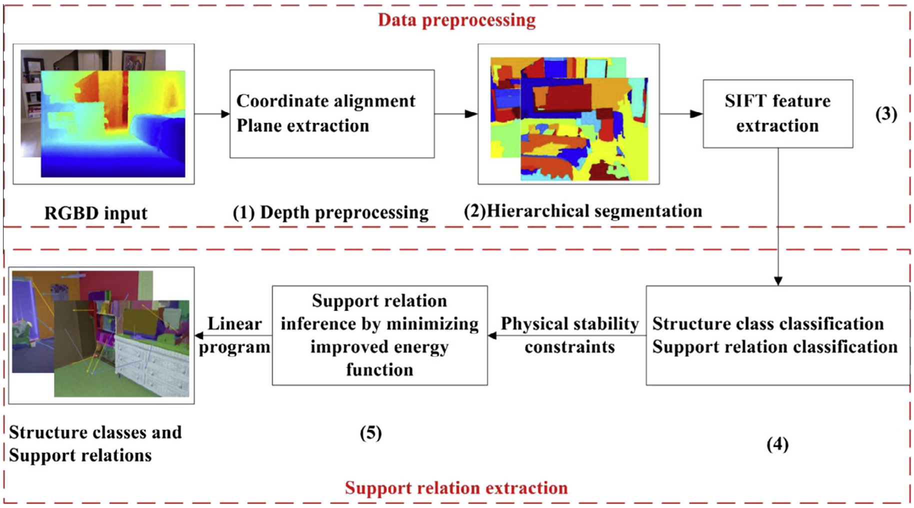

Data preprocessing
Coordinate alignment
Plane extraction
SIFT feature extraction
(3)
RGBD input
(1) Depth preprocessing
(2)Hierarchical segmentation
Linear program
Support relation inference by minimizing improved energy function
Physical stability constraints
Structure class classification
Support relation classification
Structure classes and Support relations
(5)
(4)
Support relation extraction

Fig. 1. Overview of proposed support relation extraction.

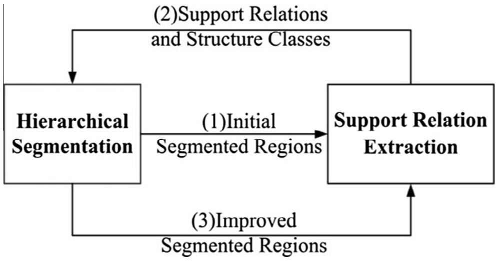


Fig. 2. "Closed-loop" architecture of support relation extraction and image segmentation.

Coordinate registration. Coordinate registration attempts to align the image coordinate with the 3D world coordinate. For each pixel, we have image coordinate (u,v), 3D coordinate (X,Y,Z), and the local surface normal (N_X,N_Y,N_Z). As discussed in [22], the coordinate alignment is based on the Manhattan world assumption [2] that most of the visible surfaces are located along one of three orthogonal directions. We can then compute each surface normal at each pixel by sampling the surrounding pixels within a depth threshold and fitting a least squares plane. The main directions of a given pixel can then be determined by extracting lines from images and computing the mean-shift modes of the surface normal. For instance, if a surface is perpendicular to the Y direction, we sample the other orthogonal candidates and compute the score of the triple as follows:

$$S(v_1,v_2,v_3)=\sum_{j=1}^{3}(N_j+L_j) \tag{1}$$

$$SN_j=\frac{k_N}{N_N}\sum_{i}^{Num_N}\exp\left(\frac{-(N_i\cdot V_j)^2}{\sigma^2}\right)\quad j=1,2,3$$

$$SL_j=\frac{k_L}{N_L}\sum_{i}^{Num_L}\exp\left(\frac{-(L_i\cdot V_j)^2}{\sigma^2}\right)\quad j=1,2,3$$

Here, v_1, v_2, and v_3 are the three principal directions, N_i denotes the normal of a pixel, L_i denotes the direction of a straight line, Num_N and Num_L are the number of points and lines on each surface, respectively, and k_N and k_L represent the weights of the 3D normals and line scores, respectively.

When the score of each candidate orthogonal direction has been calculated, we choose the candidate coordinate with the maximal score as the image coordinate and align the three orthogonal directions of image to the 3D world coordinates.

Plane labeling. When the image coordinates have been fixed, we need to classify each pixel into a labeled plane. Here, RANSAC [3] is employed to cluster image pixels into planes. Pixels along the coordinate axes are sampled, together with neighboring points at a fixed distance (e.g., 20 pixels) in both the horizontal and vertical directions. While thousands of planes are proposed, the only planes that are retained are those for which the number of pixels is greater than a threshold (e.g., 2500) after RANSAC and non-maximal suppression. For each plane, as in [22], a segmentation method based on a graph cuts technique is used to determine which image pixels correspond to each plane. Interested readers are referred to [22] for details.

The preprocessing phase described above has three outcomes: (1) Image depth coordinates are completely aligned with the 3D world coordinates. This is essential for computing the height of each pixel above the ground. (2) We can now easily find the ground by filtering the surfaces by height. Ground inference is very important, as it is the first step for support relation inference. (3) The classified surfaces and their corresponding pixels are useful for the resulting inference steps.

3.3. Depth image segmentation

Before extracting the support relations, we need to segment the input image into different regions that are in accord with individual objects or surface instances within the scene. In this section, we introduce an improved hierarchical segmentation algorithm to achieve better segmentation quality. At the first stage, hierarchical segmentation proceeds as a common segmentation procedure, and then the segments are merged into bigger regions based on learned similarities. Our method starts from the watershed algorithm [23] that generally segments an image into more than 1000 subregions. This over-segmentation ensures that very few regions overlap more than one object.

Given this over-segmentation, the next step of the hierarchical segmentation is to iteratively merge the subregions with minimum boundary strengths until a given threshold is reached. Boundary strengths are predicted by a decision tree classifier that is trained on the RGBD features. The segmentation result of this step is used in our support inference algorithm.

However, the result of hierarchical segmentation [10] is not desirable, hence we interactively use the inferred support relations to improve the hierarchical segmentation results, as shown in Fig. 2.

The prediction accuracy of the boundary strengths determines the quality of the segmentation results. Therefore, the key problem is the selection of features used to train the classifier. Silberman et al. [22] proposed five different features: RGB, Depth, RGBD, RGBD combined with support relations, and RGBD combined with structure classes and support relations. Silberman et al.'s experiments showed that the segmentation algorithm using the combined feature of RGBD, structure classes, and support relations produced the best segmentation results. In fact, the support relationships and structure classes used for segmentation in [22] were manually annotated, and small objects were often ignored. As can be seen in the first row of Fig. 3, the decorative objects in the red rectangles in (a) are ignored in the ground-truth segmentation of (b). In the second row, the quilts in (c) are also ignored on the right of the segmentation of (d). These small missing objects in the segmentation cause insufficient accuracy in the inferred support relations.

There is a one-to-one correspondence between the inferred relations and segmented regions in this paper, namely, each region has its support. The outputs of the support inference algorithm are fed back into the hierarchical segmentation algorithm shown in Fig. 2. In our implementation, RGBD, the inferred support relations and structure classes are used as the input training features of the decision tree classifier in the segmentation stage.

3.4. The proposed inference method

3.4.1. Model definition

When the image segmentation is completed, we can classify individual objects or surface instances within the image into different segmented regions. However, this information is often not enough to infer the support relations. The first step to infer support relations between objects is to model the support relationships. A support relationship model usually has two aspects: support objects and support types. The classic support model for indoor objects of [22] is used in this paper. In this model, regions are categorized into those: (1) supported by a visible region, (2) supported by an invisible region, and (3) requiring no support. Support types involve two categories: those supported from below and those supported from behind, as shown in Fig. 4.

Given an image split into N regions, we model the support relationship as follows:

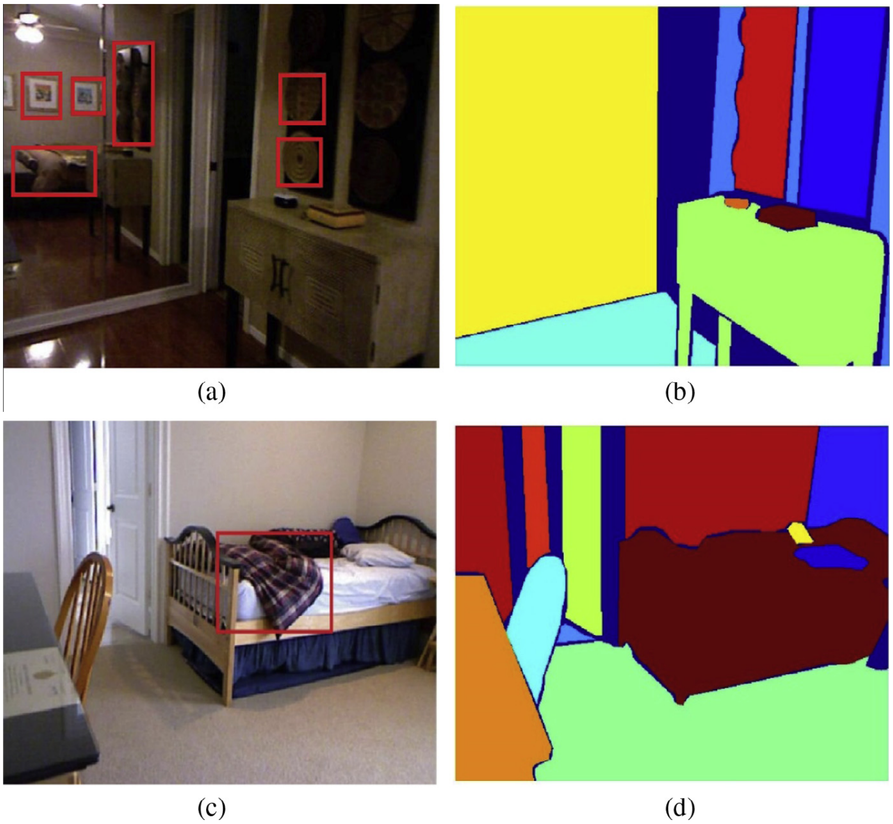


Fig. 3. Analysis of the ground-truth segmentation: (a) and (c) are RGB images of indoor scenes, and (b) and (d) are the ground-truth segmentations of (a) and (c), respectively.

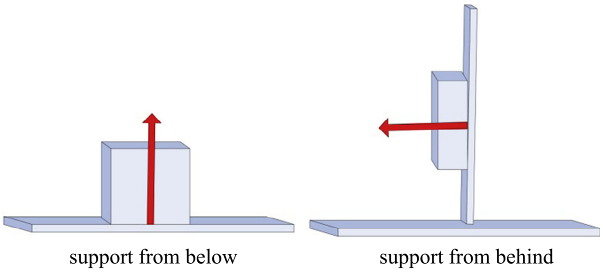


Fig. 4. Support types.

(1) Let *Si* encode the supporter of region *Ri*. Here, (a) $Si \in \{1, 2, \ldots, r\}$ indicates *Ri* is supported by a visible region, (b) $Si = r + 1$ indicates that *Ri* is supported by an invisible object, and (c) $Si = ground$ indicates that the *Ri* is the ground and needs no support.

(2) Let *STi* denote the support type of region *Ri*, where $STi = 0$ indicates that *Ri* is supported from below and (b) $STi = 1$ indicates that *Ri* is supported from behind.

When inferring support relationships, prior knowledge about object types can help predict support relations. For example, it is unlikely that a small object supports a very large object. Therefore, we model structure classes as important auxiliary information regarding support inference. The structure classes model proposed here divide objects in the scene into four classes: (1) ground, (2) furniture (i.e., large objects that cannot be moved easily) (3) small objects (that can be moved easily), and (4) structure (e.g., a ceiling or wall). This model sufficiently defines most objects in indoor scenes and can classify the structure classes well. The structure class of region *Ri* is encoded by *SCi*, which may be ground ($SCi = 1$), furniture ($SCi = 2$), small objects ($SCi = 3$), or structure ($SCi = 4$).

The model is illustrated in Fig. 5, where each region has three features: structure class, support region, and support type. In Fig. 5, region 1 is an object supported by a visible region from behind ($STi = 1, SCi = 3$), regions 2 and 3 are structures supported by the ground from below ($STi = 0, SCi = 4$), region 4 is the ground without any support ($Si = ground, SCi = 1$), and region 5 is supported by a visible region from below ($STi = 0, SCi = 3$).

3.4.2. Basic feature selection and classification

In this section, we first implement the extraction of support relations in the RGBD images using some basic features in the regions, without considering any physical constraints, as in [22]. In reality, the structure classes and support relations are subject to some physical constraints. Therefore, the inferred results are undesirable and are mainly used as the baseline of the improved relations inference method. In Section 3.4.3, we systematically describe our method that integrates some additional physical constraints into the energy minimization function.

Selection of classification features: during classification, features of different regions in the scene are extracted from training samples that are used to describe the characteristics of these regions.

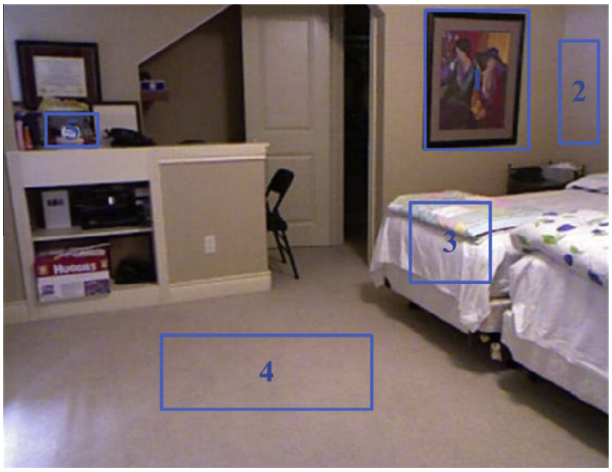


Fig. 5. Example of the support inference model.

In the conventional method, feature descriptors are selected to represent the character of image regions. Better feature descriptions lead to more distinct regions and better classification results can be obtained. Feature descriptor generation normally extracts interest points from the image and then uses multi-dimensional vectors to represent different features. Thus, an image region can be described by combining all the descriptors therein.

SIFT is a well-known feature descriptor; it is robust and has a high identifying ability [18,19]. SIFT basically includes a feature detector and feature descriptor. The detector extracts a number of attributed regions from an image in a way that is consistent with variations in illumination, viewpoint, and other viewing conditions. The descriptor associates with the regions a signature that identifies their appearance compactly and robustly. Consequently, we use SIFT descriptors as the training feature when classifying structure classes and support relations.

Classification of structure classes: To predict probabilistic label structures with different classes, we need to first build a classifier. In this paper, we chose the logistic regression classifier, which is widely used for probabilistic predictions. We denote the structure class features as D_{SC}, and term the structure class features (SIFT) used in D_{SC} as F_i^{sc}. The feature D_{SC} outputs the probability that a given region should be labeled with each structure, and this probability is used to determine the final structure class of a region.

Classification of support relations: To classify structure classes, we use SIFT descriptors and logistic regression classifier D_{SP} to get the support relations between regions. Training features should be asymmetric, i.e., feature vector $F_{i,j}^{sp}$ indicates whether Ri supports Rj, but not vice versa. To train classifier D_{SP}, each $F_{i,j}^{sp}$ is marked with a label $L^s \in \{1,\ldots,4\}$ to indicate the support relations between Ri and Rj. A label of "1" indicates that Ri is supported by Rj from the bottom, "2" indicates Ri is supported by Rj from behind, and "3" indicates that Rj is the ground. The label "4" indicates that there is no support relationship between the two regions.

3.4.3. Support relation inference using improved energy function

The structure classes and support relations in Section 3.4.2 are inferred under the assumption that they are not affected by physical constraints. In practice, the structure classes and support relations of indoor scenes are bound to be constrained by various physical factors, such as physical rationality and common sense. To make the inferred results more accurate, we add some new constraints to embody these practical factors.

Suppose an image is segmented into N regions $R = \{R_1, \ldots, R_N\}$. We aim to infer the most probable joint assignment of support regions $S = \{S_1, \ldots, S_N\}$, support types $ST = \{0,1\}^N$, and structure classes $SC = \{0,4\}^N$. We solve this problem by minimizing an energy function, where $E(S,ST,SC)$ is the energy of the model $\{S,ST,SC\}$, and $\{S^*,ST^*,SC^*\}$ is the most probable joint assignment, given by:

$$\{S^*, ST^*, SC^*\} = \arg\min_{S,ST,SC} E(S,ST,SC) \tag{2}$$

The energy of the model in Eq. (2) is divided into three parts: the energy of support E_s, the energy of structure classes E_{SC}, and the energy of the physical stability constraint E_{PS}. Consequently, the energy model is formally defined as:

$$E(S,ST,SC) = E_s(S,ST) + E_{SC}(SC) + E_{PS}(S,ST,SC) \tag{3}$$

$$E_s(S,ST) = -\sum_i^R \log(D_{SP}(F_{i,j}^{sp}|Si, STi))$$

$$E_{SC}(SC) = -\sum_i^R \log(D_{sc}(F_i^{sc}|SCi)$$

where D_{SP} is the support classifier trained (as detailed in Section 3.4.2), $F_{i,j}^{sp}$ are the support features for Ri and Rj, and F_i^{sc} are the structure features for Ri.

Physical stability constraint: physical stability constraint E_{PS} consists of three items:

(1) **Structure class constraint C_{sc}**: This structure class constraint is imposed onto the support relations of a given region. In general, in an indoor scene, the structure classes of a region have a significant impact on its support relations. For example, furniture ($SCi = 2$) or structures ($SCi = 4$) can only be supported by the ground, and their support regions are lower than any other regions in the scene. The ground ($SCi = 1$) requires no support and is the lowest region in the scene coordinate system. Small objects ($SCi = 3$) have a wide choice of support regions because they can be moved easily. For these objects, the only rule is that the support region Si should not be higher than the supported region Ri. Generally, the structure class constraint can be formulated as:

$$C_{sc}(S_iSC_i) = \begin{cases} k1 & \text{if } SCi = 1 \ AND \ p_j^L > p_i^L, \forall R_j \\ k2 & \text{if } SCi = 3 \ AND \ p_{Si}^L \leqslant p_i^H \\ k3 & \text{if } SCi = 2 \text{ or } 4 \ AND \ P_{Si}^L \leqslant p_i^H \ AND \ p_j^L > P_{Si}^h, \forall R_j \\ \infty & \text{otherwise} \end{cases} \tag{4}$$

where p_i^H and p_j^L encode the highest and lowest points in 3D regions *Ri* and *Rj*, respectively, and *k*1, *k*2, and *k*3 are integers.

(2) **Adjacency constraint C_{adj}**: In a real indoor scene, a region and its support must be adjacent to each other to ensure physical stability. Formally, the adjacency constraint is defined as:

$$C_{adj}(S_i, ST_i) = \begin{cases} (p_i^L - p_{si}^H)^2 & \text{if } ST_i = 0 \\ D^2(R_i, S_i) & \text{if } ST_i = 1 \end{cases} \tag{5}$$

where *D*(*Ri*, *Si*) is the minimum horizontal distance between region *Ri* and its support region *Si*.

(3) **Support constraint C_s** : In the scene coordinate system, all the regions are higher than the ground and usually need a support. This constraint is defined as:

$$C_S(S_i, SC_i) = \begin{cases} \infty & \text{if } S_i = ground \ AND \ SCi \neq 1 \\ \infty & \text{if } SC_i = 1 \ AND \ p_j^L < p_i^L, \exists R_j \\ k & \text{otherwise} \end{cases} \tag{6}$$

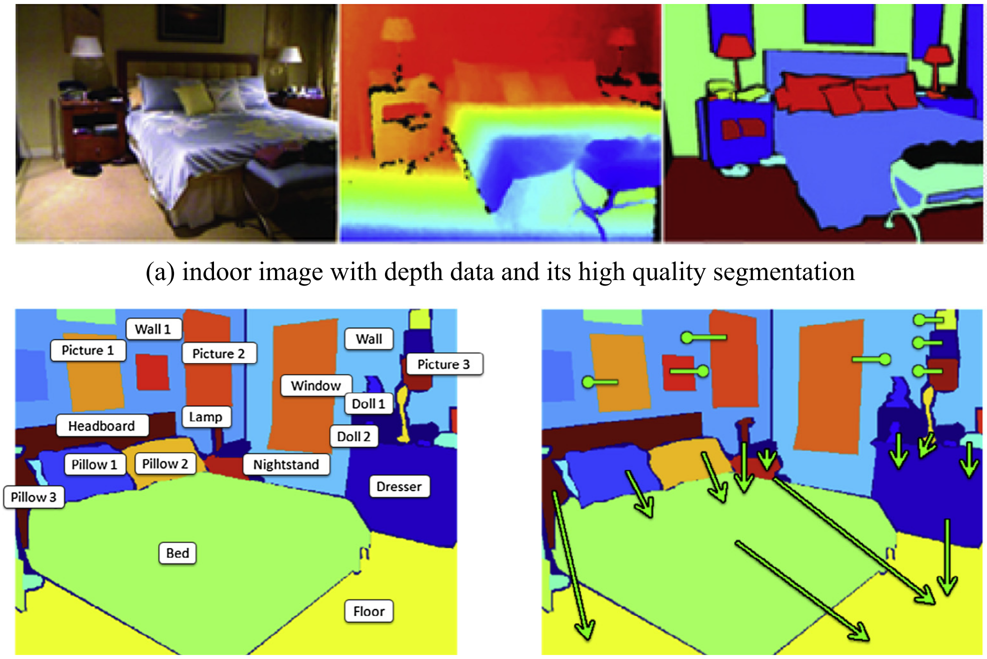
where *k* is an integer.

The three constraints C_{sc}, C_{adj}, and C_s have different influences on the physical stability energy E_{ps}. In the implementation, we combine these three constraint items with different weights and determine the optimal support relations by tuning them. At this point, the energy of stability can be formulated as:

$$E_{PS}(S, ST, SC) = \alpha_1 C_{sc}(S_i, SC_i) + \alpha_2 C_{adj}(S_i, ST_i) + \alpha_3 C_S(S_i, SC_i) \tag{7}$$

The energy minimization in Eq. (2) can be formulated in terms of an integer programming problem. The total number of regions in the scene equals the visible regions plus a hidden region. Let $N^* = N + 1$ be the total number of regions. First, Boolean indicator variable $Bsr_{i,j} : 1 \leqslant i \leqslant N, \ 1 \leqslant j \leqslant 2N^* + 1$ is introduced to represent support region *S* and support type *ST*. If $Bsr_{i,j} = 1, 1 \leqslant j \leqslant N^*$, then region *Ri* is supported by region *Rj* from below. Similarly, $Bsr_{i,j} = 1, N^* + 1 \leqslant j \leqslant 2N^*$ indicates that region *Ri* is supported by region R_{j-N*} from behind, and $Bsr_{i,2N^*+1} = 1$ indicates that region *Ri* is the ground. Boolean variable $Bsc_{i,\lambda}$ indicates the structure classes of the regions, where $Bsc_{i,\lambda} = 1$ indicates that the structure class of region *Ri* has a value of λ. Furthermore, $\chi_{i,j}^{\lambda,\gamma}$ represents the case $Bsr_{i,j} = 1, Bsc_{i,\lambda} = 1$, and $Bsc_{j,\lambda} = 1$. Using this over-complete representation, we can formulate the minimum energy inference problem as an integer program:

$$\arg \min_{Bsr, Bsc, \chi} \sum_{i,j} \varphi_{i,j}^{sr} Bsr_{i,j} + \sum_{i,\lambda} \varphi_{i,\lambda}^{sc} Bsc_{i,\lambda} + \sum_{i,j,\lambda,\gamma} \varphi_{i,j,\lambda,\gamma}^{\chi} \chi_{i,j}^{\lambda,\gamma} \tag{8}$$



(a) indoor image with depth data and its high quality segmentation

(b) labeled objects of the indoor scene

(c) standard of support relations

Fig. 6. RGBD images and their segmentations, labeled objects and support relations in NYU-Depth2.

In Formulation (8), the support energy E_s (Eq. (3)) and the adjacency constraint C_{adj} of the energy (Eq. (5)) are encoded in the integer program objective by coefficients $\varphi^{sr}_{i,j}$. We also encode the structure class energy E_{sc} (e.g., Eq. (3)) and support constraint C_s of the energy (Eq. (6)) in the integer program objective with coefficients $\varphi^{sc}_{i,\lambda}$. Structure constraint C_{sc} (Eq. (4)) is encoded in the integer program objective by coefficients $\varphi^{\chi}_{i,j,\lambda,\gamma}$.

Eqs. (9)–(11) satisfy support constraint C_s (Eq. (6)).

$$\sum_j Bsr_{i,j} = 1, \sum_\lambda Bsc_{i,\lambda} = 1, \quad \forall i \tag{9}$$

$$\sum_{j,\lambda,\gamma} \chi^{\lambda,\gamma}_{i,j} = 1, \quad \forall i \tag{10}$$

$$Bsr_{i,2N^*+1} = Bsc_{i,1} \quad \forall i \tag{11}$$

Eqs. (12) and (13) ensure that the value of $\chi^{\lambda,\gamma}_{i,j}$ is right for its definition.

$$\sum_{\lambda,\gamma} \chi^{\lambda,\gamma}_{i,j} = Bsr_{i,j}, \quad \forall i,j \tag{12}$$

$$\sum_{j,\gamma} \chi^{\lambda,\gamma}_{i,j} \leqslant Bsc_{i,\lambda}, \quad \forall i,\lambda \tag{13}$$

Eq. (14) satisfies structure constraints item C_{sc}, and Eq. (15) ensures that all indicator variables take integral values.

$$\chi^{4,1}_{i,j} = Bsr_{i,j}, \ \chi^{2,1}_{i,j} = Bsr_{i,j}, \quad \forall i,j \tag{14}$$

$$Bsr_{i,j}, \ Bsc_{i,\lambda}, \ \chi^{\lambda,\gamma}_{i,j} \in \{0,1\}, \quad \forall i,j,\lambda,\gamma \tag{15}$$

The integer program problem defined in Eq. (15) is an NP hard problem. We can convert the integer program into a linear program by revising Eq. (15) as follows:

$$Bsr_{i,j}, \ Bsc_{i,\lambda}, \ \chi^{\lambda,\gamma}_{i,j} \in [0,1], \quad \forall i,j,\lambda,\gamma \tag{16}$$

4. Experimental results

4.1. Dataset

Our experiments were carried out on the NYU-Depth dataset [22]. The NYU-Depth2 dataset is a set of images manually collected by a Kinect. The RGBD images in the earlier NYU-Depth1 dataset were mostly sampled from indoor videos. The NYU-Depth2 extended NYU-Depth1 by the addition of 464 images of three cities. There are 1449 labeled indoor images, including standards for scene segmentation and support relations. The dataset has the following advantages: (1) there are sufficient image samples, (2) all RGBD images have been labeled carefully, and (3) all images are well segmented and the support relations have good manual labels.

Generally speaking, NYU-Depth2 is widely used by researchers for depth image processing and analysis, and is currently used as a benchmark for RGBD image research. An example image is shown in Fig. 6.

4.2. Experimental results

4.2.1. Support relations evaluation

Usually, support relation inference is based on image segmentation. Ground-truth support relations in NYU-Depth2 are mainly extracted from manually segmented images, while our method is based on hierarchical segmentation. For the sake of convenience, we term the result regions of manually segmented method as *Rg*, and the result regions of the hierarchical segmentation method as *Rh*. Normally, there are more small fragmentized regions output by the hierarchical segmentation than the manual segmentation. These fragmentized regions have no supporters or support types, which may introduce a

Table 1
Predicting accuracy of support relations without support type.

	Ground truth segmentation	Initial segmentation	Improved segmentation
Image plane baseline	63.9	24.5	25.9
Structure class baseline	72.0	46.6	47.8
Energy minimized method	75.9	54.1	55.4
Improved energy minimized method	77.4	56.2	58.6

Table 2
Predicting accuracy of support relations with support type.

	Ground truth segmentation	Initial segmentation	Improved segmentation
Image plane baseline	50.7	22.9	23.4
Structure class baseline	57.7	43.0	44.0
Energy minimized method	72.6	53.5	54.9
Improved energy minimized method	74.5	55.3	56.0

penalized evaluation. To avoid this issue, a relation mapping from *Rg* to *Rh* is performed as follows: first, the support relations of *Rg* are replaced with those of *Rh*. We then map the relations of the ground-truth to *Rh* and reassign new support relations of old regions without changing the values.

Once the mapping between the relations and segmented regions has been established, we can evaluate the prediction accuracy of the support relations. There are two accuracy criteria:

(1) Accuracy with respect to support type *P1* = *PSR*/*TSR*, where *PSR* is the number of regions whose support relations are correctly predicted, and *TSR* is the total number of segmented regions with labeled relations.
(2) Accuracy with respect to support type *P2* = *PSTR*/*TSR*, where *PSTR* is the number of regions in which both the support regions and support type are correctly predicted, and *TSR* is the total number of segmented regions with labeled relations.

Silberman et al. evaluated their results against three baselines, namely, image plane rules, structure class rules, and support classifiers. We present some comparative results between the proposed algorithm and these baselines. The prediction support accuracies of our method and [22] are listed in Tables 1 and 2 for comparison. In the five listed methods, the inference accuracies of the last two methods based on the minimized energy function are significantly higher than those of the other three. Compared with the original minimized energy inference method, the improved method outperformed the first two methods by 10% and outperformed the original minimized energy inference method by 2–3%. We also note that in each baseline, there is an improvement in accuracy of around 2% when the proposed segmentation algorithm (described in Section 3) is adopted instead of the segmentation method used in [22].

As shown in Tables 1 and 2, the prediction accuracies are still not very high, which may currently limit the application of this technique. However, with the further promotion of data precision and the appearance of more advanced algorithms, the technique of support relation extraction will have increasing applications, such as scene understanding and robotics.

As a comparison, we present some support relation maps of indoor RGBD images generated by different methods. In Fig. 7, the first row are input images with depth data, and the second through fifth rows are inferred support relations using image rule baseline, structure class baseline, minimized energy method, and the improved method, respectively. Blue arrows indicate correctly inferred regions, and orange arrows indicate incorrectly inferred regions. Green crosses indicate correctly identified support from behind, while red crosses indicate incorrectly identified support from behind. The circled cross denotes the ground. As can be seen in Fig. 7, the proposed method is superior to the other three methods, with more accurate results and fewer errors. However, in the third image, it is difficult to obtain good inference in both our method and the method in [22]. The main cause of this poor inference is that both methods are highly dependent on the accuracy of the ground inference. A wrongly inferred ground will introduce many erroneous inferences.

4.2.2. Evaluation of the improved segmentation method

To evaluate the improved segmentation method, we present a comparison of evaluation scores of the five baselines and the proposed method in this paper according to the overlap criteria in [10]. In our implementation, the proposed method (the last row) outperforms all the above five methods by 5–15%. Moreover, we note that in Table 3, for each feature, weighted scores are generally higher than unweighted ones. The main cause of this phenomenon is that the bigger regions, whose accuracies are on average higher than those of small regions in the segmentation phase, are often assigned larger weights during the weighted evaluation phase.

In this paper, we use different feature combinations to train the boundary strength classifier. In addition to presenting the comparisons of the overlapping criteria in Table 3, we also compared the receiver operating characteristic curves (ROCs) of the improved classifier and the classifier in [22]. As shown in Fig. 8, the area under curve values of our method (train AUC: 0.868276, test AUC: 0.677272) are higher than that of Silberman et al.'s method (train AUC: 0.864478, test AUC: 676329), indicating that the segmentation classifier has better performance and stronger classification ability.

4.3. Discussion

In this paper, we use the support relations as an auxiliary cue to help to segment RGBD images. As illustrated in Tables 1 and 2, in our framework, the closed-loop feedback formed by support inference and improved segmentation makes our algorithm surpass the algorithm of [22] with respect to support accuracy and segmentation precision. The main differences

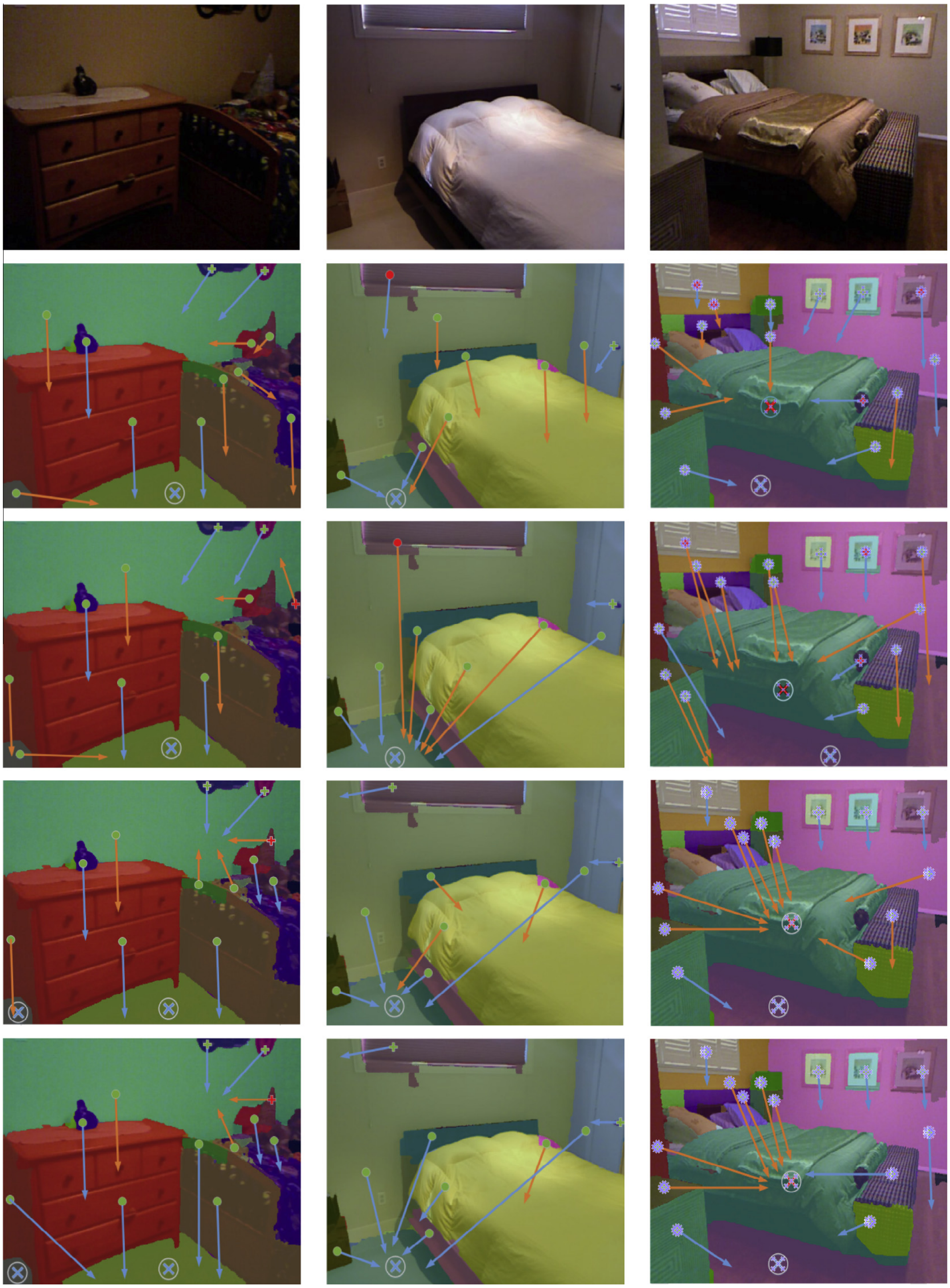


Fig. 7. Comparison of inferred support relations of different methods.

between the method in this paper and [22] are that: (1) we set up a series of new constraints for physical support relations and structure classes in which the physical stability constraint is integrated into the target optimization function and (2) we use the inferred relations for image segmentation instead of the ground-truth support relations used in [22]. Because the ground-truth support relations often ignore the relations of small objects, our proposed method produces better segmentation.

Table 3
Evaluation scores using different features based on the overlap criteria. The unweighted score is the ratio of correctly segmented pixels to total number of image pixels. For the weighted scores, bigger regions are assigned larger weights for scoring.

Features	Weighted score	Unweighted score
RGB features	52.5	49
Depth feature	56.5	46.8
RGBD features	64.7	52.7
RGBD + Structure classes	62.9	54.5
RGBD + Structure classes + Support relations in [21]	63.4	55.3
RGBD + Structure classes + Our support relations	68.0	56.3

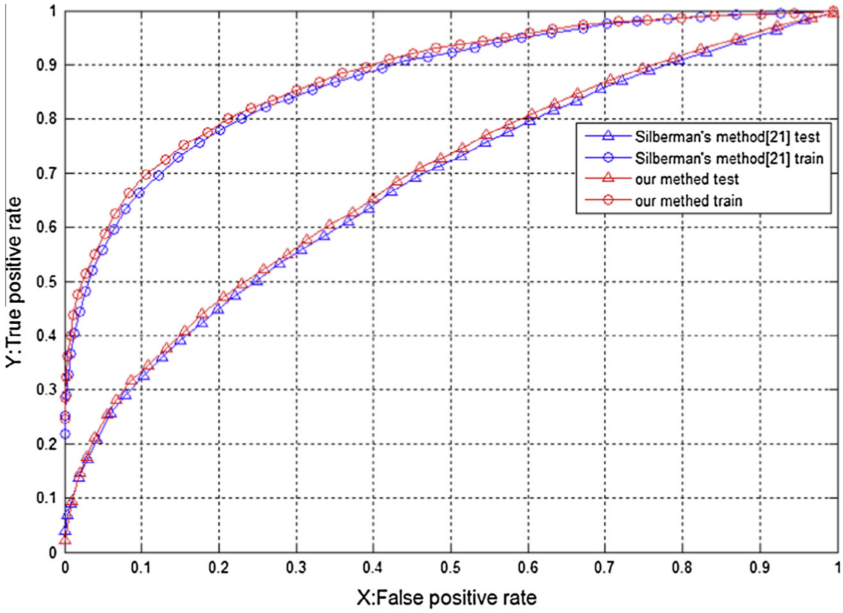


Fig. 8. ROC map comparison of accuracies for the proposed method and [22] on training and test datasets.

5. Conclusions

This paper presents a new approach for physical relation inference between objects from RGBD images. The input depth image is first segmented by a hierarchical segmentation algorithm, and then features are extracted to train the support relations and structure classes classifiers. Next, an improved constraint energy function is formulated, and then the support relations are inferred by resolving an integer programming problem. Moreover, the proposed support relation extraction method can be used to iteratively improve the image segmentation. Experimental results show that there is an improvement in accuracy of around 5–6% in support relation inference compared with the method in [22].

Acknowledgments

This research is support by the National Natural Science Foundation of China (Nos. 61202283, 61472115), and the ChunHua Research Project of Hefei University of Technology (No. 2013HGCH0019). We are very grateful to the datasets and code shared by Silberman et al. that we used in this paper.

References

[1] P.W. Battaglia, J.B. Hamrick, J.B. Tenenbaum, Simulation as an engine of physical scene understanding, in: Proceedings of the National Academy of the Sciences of the United States of America, vol. 110, 2013, pp. 18327–18332.
[2] J. Coughlan, A. Yuille, Manhattan world: orientation and outlier detection by Bayesian inference, Neural Comput. 15 (5) (2003) 1063–1088.
[3] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[4] Y.L. Guo, F. Sohel, M. Bennamoun, J.W. Wan, M. Lu, A novel local surface feature for 3D object recognition under clutter and occlusion, Inf. Sci. 293 (2015) 196–213.

[5] A. Gupta, S. Satkin, A. Efros, M. Hebert, From 3D scene geometry to human workspace, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1961–1968.

[6] S. Gupta, P. Arbelaez, J. Malik, Perceptual organization and recognition of indoor scenes from RGBD images, In: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 564–571.

[7] H. Grabner, J. Gall, L.J.V. Gool, What makes a chair a chair?, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1529–1536.

[8] A. Gupta, A.A. Efros, M. Hebert, Blocks world revisited: image understanding using qualitative geometry and mechanics, in: European Conference on Computer Vision, 2010, pp. 482–496.

[9] V. Hedau, D. Hoiem, D. Forsyth, Recovering free space of indoor scenes from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2807–2814.

[10] D. Hoiem, A.A. Efros, M. Hebert, Recovering occlusion boundaries from an image, Int. J. Comput. Vis. 91 (3) (2011) 328–346.

[11] Y. Jiang, M. Lim, C. Zheng, A. Saxena, Learning to place new objects in a scene, Int. J. Robotics Res. 31 (9) (2012) 1021–1043.

[12] Z. Jia, A.C. Gallagher, A. Saxena, T. Chen, 3D-based reasoning with blocks, support, and stability, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1–8.

[13] A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko, T. Darrell, A category-level 3-D object dataset: putting the Kinect to work, in: IEEE International Conference on Computer Vision Workshops, 2011, pp. 1168–1174.

[14] Y. Jiang, H. Koppula, A. Saxena, Hallucinated humans as the hidden context for labeling 3D scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2993–3000.

[15] H. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from RGB-D videos, Int. J. Robotics Res. 32 (8) (2013) 951–968.

[16] D. Lee, M. Hebert, T. Kanade, Geometric reasoning for single image structure recovery, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2136–2143.

[17] D. Lee, A. Gupta, M. Hebert, T. Kanade, Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces, in: Advances in Neural Information Processing Systems, 2010, pp. 1288–1296.

[18] D.G. Lowe, Object Recognition from local scale-invariant features, in: IEEE International Conference on Computer Vision Workshops, 1999, pp. 1150–1157.

[19] W. Liu, X.T. Li, Q.H. Huang, X.L. Li, GA-SIFT: a new scale invariant feature transform for multispectral image using geometric algebra, Inf. Sci. 281 (2014) 559–572.

[20] K. Lu, Q. Wang, J. Xue, W.G. Pan, 3D model retrieval and classification by semi-supervised learning with content-based similarity, Inf. Sci. 281 (2014) 703–713.

[21] O. Nempont, J. Atif, I. Bloch, A constraint propagation approach to structural model based image segmentation and recognition, Inf. Sci. 246 (2013) 1–27.

[22] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, 2012, pp. 746–760.

[23] L. Vincent, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, IEEE Trans. Pattern Anal. Mach. Intell. 13 (6) (1991) 583–598.

[24] B. Zheng, Y. Zhao, J.C. Yu, Beyond point clouds: scene understanding by reasoning geometry and physics, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3127–3134.