Semantic Segmentation of RGBD Images with Mutex Constraints

Zhuo Deng
Temple University
Philadelphia, USA
zhuo.deng@temple.edu

Sinisa Todorovic
Oregon State University
Corvallis, USA
sinisa@eecs.oregonstate.edu

Longin Jan Latecki
Temple University
Philadelphia, USA
latecki@temple.edu

Abstract

In this paper, we address the problem of semantic scene segmentation of RGB-D images of indoor scenes. We propose a novel image region labeling method which augments CRF formulation with hard mutual exclusion (mutex) constraints. This way our approach can make use of rich and accurate 3D geometric structure coming from Kinect in a principled manner. The final labeling result must satisfy all mutex constraints, which allows us to eliminate configurations that violate common sense physics laws like placing a floor above a night stand. Three classes of mutex constraints are proposed: global object co-occurrence constraint, relative height relationship constraint, and local support relationship constraint. We evaluate our approach on the NYU-Depth V2 dataset, which consists of 1449 cluttered indoor scenes, and also test generalization of our model trained on NYU-Depth V2 dataset directly on a recent SUN3D dataset without any new training. The experimental results show that we significantly outperform the state-of-the-art methods in scene labeling on both datasets.

1. Introduction

This paper addresses the fundamental problem of semantic scene segmentation of indoor scenes. Assigning class labels to every pixel in real-world images is challenging, as objects may be heavily occluded, appear in a wide range of configurations, and viewed from different camera viewpoints and distances. In addition, indoor scenes typically consist of a relatively large number of alike objects that are often cluttered and in disorder, reflecting various lifestyles. Our goal is to partition the image by identifying subimage ownership among occurrences of distinct object classes.

The recent advent of Microsoft Kinect alleviated some of these challenges, and thus enabled an exciting new direction of approaches to semantic scene segmentation [1, 2, 3, 4, 9, 11, 12, 17]. Equipped with an active infrared structured light sensor, Kinect is able to provide the depth information of objects in the scene which is aligned synchronously with their color images. Since indoor scenes are typically characterized by large planar surfaces (e.g., floor, walls, table tops), and objects can often be interpreted in relation to those surfaces, semantic scene segmentation can be largely facilitated by properly integrating visual cues with detailed and accurate geometric structure of the scene surfaces provided by Kinect.

Recent work has demonstrated that the depth information can be readily used to leverage rich geometric structure of indoor scenes toward their robust semantic segmentation. The SLAM technology was used to merge multiple RGBD images into a single point cloud and densely label it with Markov Random Field (MRF) [17]. Scenes were labeled by incorporating SIFT features and 3D location priors into a Conditional Random Field (CRF) [4]. A CRF with higher order cliques was used to encourage all regions in them to take the dominant label [11]. [1] extended the Kernel Descriptors (KDES) [18] by introducing depth gradient and spin normal descriptors, and labeled scenes by combining MRF with segmentation tree. In [3], geometric features were integrated with traditional visual features through support vector machines, or with high level features from object detection [12]. Instead of designing hand crafted features, a multiscale convolutional network was used to learn features directly from RGBD images [9].

Although designing distinct features from RGBD images has achieved much progress for indoor semantic segmentation, how to jointly model local and long range object spatial configurations by taking advantage of available geometric structure of indoor scenes is not fully explored. We find that there is still room for improvement.

In this paper, we propose a holistic framework for reasoning about object classes and their co-occurrences, and spatial layouts based on geometric structure of indoor scenes as well as on common sense knowledge. We model the scene by a CRF grounded to regions of the low-level generalized gPb-UCM segmenter [3]. Geometric and visual information of objects are integrated into unary potentials. The pairwise potentials encode local object configurations based on several typical geometric patterns. In this way, we pose semantic scene segmentation as the problem of assigning class labels to image regions in the CRF inference. As common, we cast CRF inference as a quadratic programming (QP) problem.

As our key contribution, we incorporate in our QP *qualitative* common-sense constraints from domain knowledge in a principled manner. We focus on mutual exclusion (mutex) constraints that specify *negation* (inconsistency) rules about object configurations in the real physical world. For example, a chair should *not* be on top of a TV, and a floor should *not* occur above a dishwasher. In scene labeling, mutex constraints are binary relations specifying inconsistent class label assignments to pairs of image regions, and can be expressed without any higher-order potentials. Also, model expressiveness is significantly increased as they can enforce *long-range* consistency constraints on the solution. With mutex constraints, our approach can make use of rich and accurate geometric structure coming from Kinect in a principled manner.

Prior works have already demonstrated the importance of domain constraints, in general, as they help resolve competing hypotheses when visual cues are not sufficient for scene interpretation. Constraints are typically incorporated in CRFs as features of the pairwise potentials [5, 6, 2]. More sophisticated methods use higher-order constraints, beyond pairwise [7, 8, 11]. Instead of working our way through higher-order constraints, we focus on exclusion commonsense rules, i.e., hard rules that exclude nonsense configurations.

In this paper, we show that mutex constraints can be compactly expressed in a quadratic equality form, and rigorously enforced in a principled manner. As smoothness and constraints are typically combined in the pairwise potential, traditional formulations of CRF inference may not guarantee that hard constraints are all satisfied. This could yield non-sensical results. We address this problem by expressing the mutex constraints as quadratic constraints of our QP. The most closely related works are [10] and [14], both of which utilize mutex relations to constrain the CRF inference. However, both [10] and [14] work with 2D images only. The goal of [10] is foreground object segmentation in videos, while [14] is focused on scene labeling. In contrast, the focus of our work is on 3D mutex relations representing common sense knowledge. Since understanding RGB-D indoor scenes is an arguably more complex task [1, 19], in addition, we utilize 3D geometric patterns and spatial object correlation for edge potential estimation, instead of the standard Potts model in [10]. Moreover, we are using a sparsely connected CRF model.

In this paper, we empirically demonstrate that enforcing *qualitative* mutex constraints can significantly improve quantitative measures of performance. The effectiveness of our approach is evaluated on the indoor scene NYU dataset V2 [2] and a recent SUN3D dataset [16]. Our labeling accuracy significantly outperforms the state of the art [3, 12].

In the rest of this paper: Sec. 2 formulates our CRF model and CRF inference as QP for semantic segmentation; Sec. 3 specifies unary and pairwise potentials that are used to compute the affinity matrix for our QP; Sec. 4 describes how to estimate mutex constraints from training data; and Sec. 5 presents experimental results and related discussion.

2. CRF for Semantic Segmentation

This section formulates our CRF model of a scene grounded on low-level segments (also called superpixels), and casts semantic segmentation as the MAP assignment of class labels to superpixels. We begin by specifying the quadratic objective of the MAP assignment problem, and then extend that formulation to include mutex constraints, resulting in our integer QP with quadratic constraints.

2.1. CRF and the MAP Assignment as QP

As in [2, 3, 12], we partition an image, $I(x, y)$, into a set of segments $\mathbb{S} = \{s_i : i = 1, \dots, N\}$, $|\mathbb{S}| = N$, using variants of the gPb-UCM hierarchical segmentation algorithm [13]. Each segment, $s_i \in \mathbb{S}$, can take one object class label, l_i, from the set of labels $l_i \in \mathbb{L}$, $|\mathbb{L}| = L$. Each label assignment to a superpixel, (s_i, l_i), can be represented as a node of the association graph $\mathbb{G} = (\mathbb{V}, \mathbb{E}, A)$, where $\mathbb{V} = \mathbb{S} \times \mathbb{L}$ is the set of nodes, $|\mathbb{V}| = N \cdot L$, and $\mathbb{E} \subset \mathbb{V} \times \mathbb{V}$ is the set of graph edges. We define $((s_i, l_i), (s_j, l_j)) \in \mathbb{E}$ if s_i and s_j are *spatially adjacent*, which means that their shared boundary in $I(x, y)$ contains at least one pixel and the minimal 3D distance between point clouds projecting to s_i and s_j is very close. A is the adjacency matrix (or the affinity matrix) of $\mathbb{G}$, with size $(N \cdot L) \times (N \cdot L)$.

We define a CRF over $\mathbb{G}$. To this end, we associate a latent binary random variable $X_{s_i,l_i} \in \{0, 1\}$ with every node $(s_i, l_i) \in \mathbb{V}$. When X_{s_i,l_i} is instantiated to value $x_{s_i,l_i} = 1$ then the CRF assigns class label $l_i \in \mathbb{L}$ to superpixel $s_i \in \mathbb{S}$. The column vector of all instantiations of the assignment random variables is denoted as $\mathbf{x} = [\dots, x_{s_i,l_i}, \dots]^\top \in \{0, 1\}^{N \cdot L}$.

We use the affinity matrix A to specify the unary and pairwise potentials of the conditional log-likelihood of the CRF. In particular, the diagonal elements $A((s_i, l_i), (s_i, l_i))$ encode the unary potentials corresponding to log-likelihoods of label assignments $x_{s_i,l_i} = 1$. The off-diagonal elements $A((s_i, l_i), (s_j, l_j))$ encode the pairwise potentials corresponding to joint log-likelihoods of label assignments $x_{s_i,l_i} = 1$ and $x_{s_j,l_j} = 1$.

From above, the conditional log-likelihood of the CRF is specified as

$$\log P(\mathbf{x}|\mathbb{G}) = \sum_{(s_i,l_i)\in V} A((s_i,l_i),(s_i,l_i))x_{s_i,l_i} + \sum_{((s_i,l_i),(s_j,l_j))\in E} A((s_i,l_i),(s_j,l_j))x_{s_i,l_i}x_{s_j,l_j} - Z(\mathbb{G}), \quad (1)$$

where $Z(\mathbb{G})$ is the partition function.

From (1), it follows that the semantic scene segmentation problem can be formulated as finding the MAP assignment $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\Omega} P(\mathbf{x}|\mathbb{G})$, where Ω is the space of allowed solutions. Note that the MAP assignment is independent of $Z(\mathbb{G})$. Thus, we can compactly express the MAP assignment problem as the following integer QP with linear constraints:

$$\begin{aligned} &\text{QP-L}: \quad \text{maximize} \quad \mathbf{x}^\top A\mathbf{x} \\ &\text{s.t.} \quad \text{for all } s_i \in \mathbb{S}, \sum_{l_i\in\mathbb{L}} x_{s_i,l_i} = 1, \ \ \mathbf{x} \in \{0,1\}^{N\cdot L}. \end{aligned} \quad (2)$$

The linear constraints in the QP-L, given by (2), ensure that every superpixel in the image gets assigned a unique class label. In the following, we extend QP-L such that the resulting QP encodes mutex constraints.

2.2. QP with Mutex Constraints

This section formulates mutex constraints in a quadratic equality form, combines them with the linear constraints of QP-L, and thus expresses the MAP assignment problem as an integer QP with quadratic equality constraints.

Mutex constraints are aimed at prohibiting certain nonsensical label assignments to superpixels in the image. We eliminate this hypothesis by enforcing $x_{s_i,l_i}\cdot x_{s_j,l_j}$=0. That is, only one of the two label assignments is allowed. If one is accepted as a solution then it automatically prevents the other one. Using the notation introduced in Sec. 2.1, it follows that all mutex constraints can be compactly represented as

$$\text{Quadratic mutex constraints (QMC)}: \quad \mathbf{x}^\top M\mathbf{x} = 0, \quad (3)$$

where M is a $(N \cdot L) \times (N \cdot L)$ binary mutex matrix. Note that when matrix elements are set to one, $M((s_i,l_i),(s_j,l_j)) = 1$, then the corresponding assignments are prohibited and hence $x_{s_i,l_i} = 0$ and/or $x_{s_j,l_j} = 0$ in order to enforce $x_{s_i,l_i}\cdot 1\cdot x_{s_j,l_j}$=0. Conversely, when $M((s_i,l_i),(s_j,l_j)) = 0$ then superpixels s_i and s_j may be assigned any class labels, because $x_{s_i,l_i}\cdot 0 \cdot x_{s_j,l_j} = 0$. If the sum of each row of M is at least one, then M represents global mutex constraints. This means that at least one constraint applies to each variable.

Further, it is convenient to merge the set of linear constraints of QP-L — namely that for all $s_i \in \mathbb{S}$, $\sum_{l_i\in\mathbb{L}} x_{s_i,l_i} = 1$ — with the quadratic mutex constraints (QMC) in (3). For every superpixel s_i, we set all matrix elements $M((s_i,l_i),(s_i,l_i')) = 1$, if $l_i \neq l_i'$. This prohibits illegal assignments of two (or more) distinct labels to a single superpixel.

From (2) and (3), we finally derive the MAP assignment problem as the integer QP with quadratic constraints:

$$\begin{aligned} \text{QP-Q}: \quad &\text{maximize} \quad \mathbf{x}^T A\mathbf{x} \\ &\text{subject to} \quad \mathbf{x}^\mathbf{T} M\mathbf{x} = 0 \ , \ \ \mathbf{x} \in \{0,1\}^{N\cdot L}. \end{aligned} \quad (4)$$

For solving QP-Q in (4), we follow the line search algorithm of [14] by relaxing QP-Q to the continuous domain

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \ \mathbf{x}^\top(\mathbf{A} - \lambda\mathbf{M})\mathbf{x} \ \text{ subject to } \ \mathbf{x} \in [0,1]^{N\cdot L} \quad (5)$$

where $\lambda > 0$ is a sufficiently large regularization parameter.

Let $f(\mathbf{x}) = \mathbf{x}^\top(\mathbf{A} - \lambda\mathbf{M})\mathbf{x}$ denotes the target function. The algorithm in [14] seeks binary solutions in each step. For a given initial vector $\mathbf{x}_0$ with $f(\mathbf{x}_0) > 0$, it increases f in each iteration until it converges to a MAP assignment $\mathbf{x}^*$. Although the formulation is relaxed the returned solutions $\mathbf{x}^*$ are binary in all experiments in [14] and in all our experiments.

Now we show that a binary solution $\mathbf{x}^*$ implies that all mutex constraints are satisfied, i.e., $(\mathbf{x}^*)^\top(\mathbf{M})\mathbf{x}^* = 0$. Suppose that this fact is not true, i.e., there exists i with $\mathbf{x}_i^* = 1$ that violates a mutex constraint. Then $(\mathbf{x}^*)^\top\mathbf{M}\mathbf{x}^* \geq 1$. Let λ be equal to the sum of all elements of $\mathbf{A}$. Because then $(\mathbf{x}^*)^\top\mathbf{A}\mathbf{x}^* \leq \lambda$, we obtain

$$f(\mathbf{x}^*) = (\mathbf{x}^*)^\top\mathbf{A}\mathbf{x}^* - \lambda(\mathbf{x}^*)^\top\mathbf{M}\mathbf{x}^* \leq 0.$$

A contradiction, since $f(\mathbf{x}^*) \geq f(\mathbf{x}_{(0)}) > 0$.

In the following two sections, we explain how to compute the affinity matrix A, and estimate the mutex matrix M from training data. In the experimental section we discuss our initialization strategy of selecting initial vectors $\mathbf{x}_0$.

3. The affinity matrix A

This section explains how to compute the unary and pairwise potentials organized in the affinity matrix A.

3.1. The Unary Potential

Recall that elements of the affinity matrix A encode the unary and pairwise potentials of our CRF (see Sec. 2.1).

We specify the unary potential of each label assignment (s_i,l_i) as follows:

$$A((s_i,l_i),(s_i,l_i)) = \begin{cases} P(l_i|F,m), & \text{if m = 1} \\ P(l_i|F,a,h,pt), & \text{otherwise} \end{cases} \quad (6)$$

where F are appearance and geometric features of region s_i used in [3], a is the angle between normal vector of s_i and gravity direction ($[0, \pi]$), h is the estimated absolute height above ground, pt is detected plane type $P(pt|l_i)$ [2] (vertical boundary, horizontal boundary, vertical plane, horizontal plane, plane, non-plane), and the binary variable m indicates if a majority of depth information is missing in s_i. For simplicity, we ignore denotation s_i in the following formulas. Assume these observations are independent from each other, then (6) can be further decomposed based on Chain Rule:

$$A((s_i, l_i), (s_i, l_i)) = \begin{cases} P(l_i|F)P(m|l_i), & \text{if m = 1} \\ P(l_i|F)P(a|l_i)P(h|l_i)P(pt|l_i), & \text{o.w.} \end{cases} \tag{7}$$

Probability Estimation: The posterior probability $P(l_i|F)$ is the output of Multi-Class Logistic Regression in [3]. The likelihoods of $P(pt|l_i)$ and $P(m|l_i)$ are estimated directly as corresponding histograms on training dataset. For the estimation of likelihood $P(h|l_i)$, it is worth noting that the absolute height h is different from the relative height in previous works such as [2, 3], where it is defined as the height above the lowest point in the image. Typically, the relative height information becomes misleading when the floor doesn't show up in the image. As shown in the left image of Fig.1, the horizontal plane is very close to the lowest point of the 3D scene, but actually it is a counter instead of a floor. To solve this problem, we assume that indoor images are captured by human in a natural way. We firstly extract statistical distribution of absolute camera height h_{cam} and for each object class from a training set. We plot the normalized histogram of absolute camera height of training set in the right image of Fig.1. It is observed that it roughly obeys a Gaussian distribution. Since height is continuous, the probability density of object l_i, $f_{l_i}(h)$, is derived by Kernel Density Estimation:

$$f_{l_i}(h) = \frac{1}{nb} \sum_{i=1}^{n} K(\frac{h - h_i}{b}) \tag{8}$$

where K is a Gaussian kernel smoother and b is bandwidth. Then the likelihood $P(h|l_i)$ is computed as follows:

$$P(h|l_i) = \int_{\mu_c - h' - 3\sigma_c}^{\mu_c - h' + 3\sigma_c} f_{l_i}(h) \, \mathrm{d}h \tag{9}$$

where μ_c and σ_c are mean and variance of absolute camera height respectively, and h' is a relative height difference between object and camera. The likelihood $P(a|l_i)$ is estimated in a similar way.

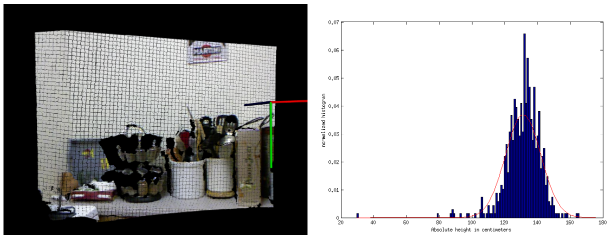

Figure 1. Left image: an example of indoor scene (point cloud attached with colors). Camera position and orientation are represented by three orthogonal color sticks. Right image: the normalized histogram of absolute camera height on training set of NYU-V2. The mean value of camera height is around 131 cm.

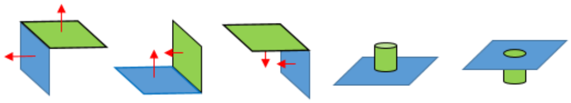

Figure 2. Geometric pairwise patterns. Red arrow represents normal vector direction. Blue or green planes indicate that the superpixel is covered by one detected plane structure.

3.2. The Pairwise Potential

Further, for all edges in the association graph $\mathbb{G}$, $((s_i, l_i), (s_j, l_j)) \in \mathbb{E}$, we encode the pairwise potentials as the off-diagonal elements of the affinity matrix A. Consider the available 3D geometric information, we define five special pairwise patterns, as is shown in Fig 2. While detected edges in 2D image often indicate object boundaries, pairwise patterns imply certain local configurations in 3D space. For example, ”cabinet” and ”counter” usually satisfy the first pattern, while the fourth pattern implies ”table” or ”counter” supports other ”props”.

Co-occurrence Probability Estimation: Except for the five defined patterns above, the other pairwise patterns are considered as one category. We compute adjacency co-occurrence probabilities of the two classes $\Psi^{(k)}(l_i, l_j), k = 1, 2, ...6$ from training data as

$$\Psi^{(k)}(l_i, l_j) = \frac{N^{(k)}(l_i, l_j)}{N^{(k)}(l_i) + N^{(k)}(l_j) - N^{(k)}(l_i, l_j)} \tag{10}$$

where $N^{(k)}$ is a function that counts the total number of training images where the event shows up in pattern k. It is worth noting that the first five adjacency co-occurrence probabilities are asymmetric. They also differ from mutex constraints in that the latter captures long-range inconsistency constraints, whereas the former are treated as “soft” constraints that only favors certain pairs of labels at spatially adjacent locations, but in no way strictly prohibit any particular pair.

4. Mutex Constraints for Scene Labeling

This section defines the mutex constraints and describes how to estimate them. We use three types of mutex constrains.

Global object co-occurrence constraints encode which objects cannot occur together in a scene. They are called global, because these constraints do not account for a particular spatial layout of co-occurrence. For example, under normal conditions, it is impossible to see both toilet and white board in the same room. In [8], similar co-occurrence constraints are incorporated into the energy function as negative logarithmic potential. Instead, we formulate them as hard constraints using the $(NL) \times (NL)$ binary matrix mutex M_{co} for each pair $v_i = (s_i, l_i)$ and $v_j = (s_j, l_j)$:

$$M_{co}(v_i, v_j) = \begin{cases} 1, & \text{if regions } s_i \text{ and } s_j \text{ with labels} \\ & l_i \text{ and } l_j \text{ never co-exist in a scene} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Relative height relationship constraints: We observe that relative height relationships typically hold in most indoor scenes. For example, the floor should be lower than chairs, and the ceiling should be higher than pictures. Thanks to Kinect technology, we can easily access depth data for each pixel. Given raw depth data, we align 3D points with gravity direction so that the floor plane lies in $X - Z$ plane, and Y axis represents the height information. The relative height relationship is represented as the $(NL) \times (NL)$ binary matrix M_{rh}:

$$M_{rh}(v_i, v_j) = \begin{cases} 1, & \text{if estimated height relation between} \\ & \text{regions } s_i \text{ and } s_j \text{ contradicts true} \\ & \text{relative locations of objects } l_i \text{ and } l_j. \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Object local support relationship constraints encode basic physical configuration rules of indoor scenes. For instance, counters are usually supported by cabinets, and televisions are supported by dressers. The inverse of these support relations would contradict common-sense knowledge about the real world. We call these constraints local, since they only regulate support relationship between two spatially adjacent regions. In order to evaluate the support relationship of two neighboring regions, we first project 3D points of both regions onto the X-Z plane. If these two projected regions have overlapping area, a support relationship does exist between them. We use a variant of Jaccard Index to measure a ratio of the overlapping area. Let $\alpha(s'_i)$ denote the area of the projected region s_i onto the ground plane. Then, we define the variant of Jaccard Index as

$$\alpha_{ratio}(s'_i, s'_j) = \frac{\alpha(s'_i \bigcap s'_j)}{\min(\alpha(s'_i), \alpha(s'_j))} \tag{13}$$

In practice, considering errors from Kinect depth measurement [15] and low level segmentation, we relax the condition to tolerate small overlaps that α_{ratio} is below certain threshold θ. We set $\theta = 0.1$ in all experiments. The support relation constraints are then encoded into the $(NL) \times (NL)$ binary matrix M_{sup}:

$$M_{sup}(v_i, v_j) = \begin{cases} 1, & \text{if } s_i \text{ cannot support } s_j \text{ w.r.t. real} \\ & \text{support relation of objects } l_i \text{ and } l_j \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

Generally, we say region s_i can support s_j when the corresponding $\alpha_{ratio} > \theta$, and the centroid height of s_i is lower than that of s_j, given object l_i can support object l_j in the real world.

Finally, the aforementioned three mutex matrices are merged into the unique mutex matrix M as

$$M(v_i, v_j) = M_{co}(v_i, v_j) \vee M_{rh}(v_i, v_j) \vee M_{sup}(v_i, v_j) \tag{15}$$

To merge the set of linear constraints of QP-L in (2), we set all matrix elements $M((s_i, l_i), (s_i, l'_i)) = 1$, if $l_i \neq l'_i$.

Mutex constrains learning: Denote a pair of nodes as $v_i = (si, li)$ and $v_j = (sj, lj)$. We make the assumption that the training set is sufficiently large. For global object co-occurrence constraints, if object class l_i and l_j have been observed present together in at least one training image, then $M_{co}(v_i, v_j) = 0$, otherwise $M_{co}(v_i, v_j) = 1$.

For relative height constraints, we use two auxiliary matrices M_{auxH} and M_{auxL} obtained from training images to encode height relationship rules w.r.t highest point and lowest point respectively. For example, $M_{auxH}(l_i, l_j) = 1$ means the highest point of class l_i always is higher than that of class l_j, while $M_{auxL}(l_i, l_j) = 1$ indicates the lowest point of class l_i always is lower than that of class l_j. Otherwise, no height relative constraint applies to class pair (l_i, l_j). Therefore, $M_{rh}(v_i, v_j) = 0$ when observed height relationship between node v_i and v_j does not violate any one of rules encoded in auxiliary matrices, otherwise $M_{rh}(v_i, v_j) = 1$.

For local support constraints, we compute the probability of class l_i support class l_j, $P_s(l_i, l_j)$, as the number of positive instances divided by the total number of spatially adjacent regions assigned with labels l_i and l_j. Here,two regions are spatially adjacent if their shared boundary contains at least one pixel and the minimal 3D distance between point clouds is less than $5cm$. Class l_i can not support class l_j if $P_s(l_i, l_j) < 5\%$.

5. Experiments

We evaluate our framework on the New York Univeristy (NYU) Depth dataset (v2) and Princeton University SUN3D dataset [16]. The NYU dataset contains 1449 pairs of aligned RGB and depth images which are captured from 27 different indoor scene categories, such as bedrooms, classrooms, kitchens, furniture stores and so forth. In [2]

894 subclasses were grouped into four super-categories: ground, furniture, props and structure for sematic segmentation. [3] extended the total number of object classes for sematic segmentation task from four to 40 classes. In our experiments we follow the settings in [3]. Since only a small portion of images has been labeled in the SUN3D dataset, we use the officially released eight annotated sequences and extract 65 keyframes that cover the content of sequences as much as possible.

Inference settings: As is described in section 3.1 , the number of nodes in the weighted graph is relevant to both over-segmentation and class labels. For some extremely complex scenes, the number of regions in the over-segmentation is around 600. But typically the number is around 140. We sort the unary potentials in decreasing order and choose the first k labels as candidates for each superpixel in graph construction stage. If k is too large, it will increase the computational cost and reduce the chance of selecting a correct label. If k is too small, it has a high probability that correct label is not in the candidate list. In the experiment, we set $k = 5$.

As the solver for finding maximum weight subgraph [14] usually converges to a local optimum, multiple initializations are needed to obtain a better performance. We train a SVM classifier by taking unary potentials as features for predicting confidence of each region and rank regions in decreasing order according to it. Then a weighted sampling mechanism is adopted to select a triple of regions as initializations each time. In other words, we set $\mathbf{x}^{(0)}(i) = 1$ if region v_i is selected as one of the three initialization regions. Otherwise, $\mathbf{x}^{(0)}(i) = 0$. Start from $\mathbf{x}^{(0)}$, we obtain a subgraph denoted by the indicator vector $\mathbf{x}^*$. In order to enforce the final solution always satisfies the mutex constraints $\mathbf{x}^\top M\mathbf{x} = 0$, the parameter λ is set to 1000. We compute x^* in (5) t times and select the one with highest energy score as the best solution according to $f(\mathbf{x}^*) = \mathbf{x}^{*\top}\mathbf{A}\mathbf{x}^*$. In our experiment, t is set to 1000.

Performance on NYU dataset: We present both qualitative and quantitative evaluation of our semantic segmentation algorithm. In order to compare our performance directly with the state of the art results in [3, 20, 12], we use the same three metrics: pixel frequency weighted average Jaccard Index, average Jaccard Index and pixel accuracy. We present the quantitative evaluation results in Table. 1. We list the best labeling result from [3, 20, 12] in the first three rows of the table respectively. [20] is a journal version of [3]. [12] improved the performance of [3] by using object detections to compute additional features for superpixels. The last row contains labeling results of our inference with mutex constraints. We achieve the best performance in the 40-class segmentation task. In particular, we outperform [3] by 3.4% (fwavacc), 5.4% (avacc) and 5.9% (pixacc), and outperform [12] by 1.5% (fwavacc), 3.1% (avacc) and 3.5% (pixacc).

	NYUV2			SUN3D		
	fwavacc	avacc	pixacc	fwavacc	avacc	pixacc
no co-occur	47.2	28.9	61.9	49.4	27.1	64.1
no rel-h	48.0	30.6	63.2	49.8	27.3	64.5
no support	48.4	31.0	63.6	50.9	28.0	65.3
full	48.5	31.5	63.8	51.0	28.2	65.7

Table 3. Ablation Study: We remove the different mutex constraints from the full system and study how the performance degrades.

In order to demonstrate the effectiveness of mutex constraints, we list the corresponding labeling results obtained by removing mutex constraints from our CRF model in the forth row. In addition, we replace our unary potential in (6) with the output of multi-class logistic regression from [3] while keeping the rest of our model unchanged. As shown in fifth row, the performance is slightly worse than our best performance. It indicates that the proposed unary potential formulation in Sec. 3.1 is useful for the CRF inference.

Performance on SUN3D dataset: It is worth noting that all 65 images are only used as test set, since we used the system trained on the NYU dataset. In other words, all the parameters and classifiers are exactly the same as those used in the NYU dataset. As only 33 classes are present in the labeled images based on the definition of 40 classes task above, after we obtain the semantic segmentation results for original 40 classes, we project unseen 7 labels into 33 classes. ”floor mat” merges to ”floor” class, ”dresser” merges to ”other furniture” and the other five merge to ”other props”. As is shown in Table.2, our model outperforms [3] by 2.8% (fwavacc), 3.4% (avacc) and 5.6% (pixacc). This results clearly demonstrate the generalization power of the proposed model with mutex constraints. We can observe that there are several zero terms in Table.2. This might because of the difference in variance of object instance appearances between training set in NYU dataset and SUN3D dataset.

We study the impact of each of our three classes of mutex constraints on the performance of our proposed system in Table 3. As can be seen all the constraints contribute to the performance. The most significant mutex constraints are co-occurrence followed by relative height.

Finally we provide some qualitative examples to demonstrate the effectiveness of our CRF inference model with mutex constraints in Fig. 3. The region labelings shown in the second column are directly from [3]. It can be observed that some common sense object configuration rules are violated. For example, the counter (row 2, col 2) is fully supported by a door, and the sofa region (row 4, col 2) has been divided into sofa and bed. The labeling of the same scene turns out to be much more reasonable after enforcing

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror	floormat	clothes	ceiling
[3]	**68.2**	81.3	46.2	57.1	36.9	41.2	25.9	14.4	33.5	18.5	42.1	51.5	41.9	5.8	4.4	28.5	19.6	30.2	21.5	23.4	7.4	**61.2**
[20]	67.9	**81.5**	45.0	60.1	41.3	47.6	29.5	12.9	34.8	18.1	40.7	51.7	41.2	6.7	5.2	26.9	25.0	32.8	21.2	30.7	7.7	**61.2**
[12] (R-CNN)	68.0	81.3	44.9	65.0	**47.9**	47.9	29.9	**20.3**	32.6	18.1	40.3	51.3	42.0	11.3	3.5	29.1	**34.8**	34.4	16.4	28.0	4.7	60.5
Ours (noMutex)	66.9	81.0	42.9	55.7	33.5	41.2	28.2	14.0	32.9	20.3	41.2	51.2	41.6	6.6	6.2	29.5	20.0	30.4	21.6	23.4	8.8	61.1
Ours ([3]+mutex)	65.1	80.4	48.5	65.2	41.9	51.8	35.3	18.8	**35.1**	**33.9**	**49.1**	49.0	**49.6**	**11.5**	**9.6**	44.8	17.1	34.1	**34.8**	31.8	**14.8**	56.9
Ours (mutex)	65.6	79.2	**51.9**	**66.7**	41.0	**55.7**	**36.5**	**20.3**	33.2	32.6	44.6	**53.6**	49.1	10.8	9.1	**47.6**	27.6	**42.5**	30.2	**32.7**	12.6	56.7

	books	fridge	tv	paper	towel	showercurtain	box	whiteboard	person	nightstand	toilet	sink	lamp	bathtub	bag	o-struct	o-furni	o-props	**fwavacc**	**avacc**	**pixacc**
[3]	7.0	16.1	7.6	15.7	25.8	7.1	**2.1**	11.7	1.4	21.5	45.4	32.5	23.3	32.6	0	8.0	3.9	21.6	45.1	26.1	57.9
[20]	7.5	11.8	15.8	14.7	20.0	4.2	1.1	10.9	1.4	17.9	48.1	**45.1**	31.1	19.1	0.0	7.6	3.8	22.6	45.9	26.8	58.3
[12] (R-CNN)	6.4	14.5	**31.0**	14.3	16.3	4.2	**2.1**	**14.2**	0.2	27.2	**55.1**	37.5	**34.8**	**38.2**	**0.2**	7.1	6.1	23.1	47.0	28.4	60.3
Ours (noMutex)	8.1	16.2	9.8	16.7	27.0	9.1	**2.1**	11.2	5.7	21.7	47.1	36.5	23.3	32.6	0	7.8	5.4	23.3	44.8	26.6	60.5
Ours ([3]+mutex)	**13.2**	20.9	9.5	25.7	**32.3**	22.8	**2.1**	1.0	6.0	18.1	50.4	35.0	29.2	28.9	0	9.4	**8.6**	24.9	47.9	30.4	63.1
Ours (mutex)	8.9	**21.6**	19.2	**28.0**	28.6	**22.9**	1.6	1.0	**9.6**	**30.6**	48.4	41.8	28.1	27.6	0	**9.8**	7.6	**24.5**	**48.5**	**31.5**	**63.8**

Table 1. Performance on 40-class semantic segmentation on the NYU-Depth V2 data set. We compare directly with the best results obtained in [3, 12, 20]. The fourth row shows results of our model without mutex constraints. The fifth row shows results of our model with mutex constraints where our unary potential is replaced with the output of multi-class logistic regression in [3]. The last row contains labeling results of our full model.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookslf	picture	counter	blinds	desk	nightstd	curtain	toilet	pillow
[3]	**64.8**	89.9	2.8	29.5	45.1	38.3	42.9	**16.5**	21.8	**15.7**	13.9	40.9	**21.3**	0.05	56.6	18.5	**56.4**	6.1
Ours (noMutex)	63.7	**90.1**	5.6	42.9	45.8	38.7	50.6	3.5	26.2	12.3	10.5	43.8	19.7	0	58.0	12.8	51.3	13.4
Ours ([3]+mutex)	60.9	89.3	14.5	45.1	46.6	**42.3**	64.8	5.7	**36.4**	0.5	11.3	47.7	6.8	**0.08**	59.4	20.1	55.4	14.1
Ours	61.1	88.8	**19.8**	**46.3**	**51.1**	41.9	**69.7**	9.3	34.9	2.0	**21.8**	**49.4**	5.2	0	**62.3**	**20.8**	56.0	**16.7**

	mirror	sink	clothes	ceiling	lamp	fridge	tv	bathtub	towel	bag	box	whtbrd	ostruct	ofurn	oprops	**fwavacc**	**avacc**	**pixacc**
[3]	0	**81.4**	0	61.0	0.57	21.8	0	23.7	**13.8**	0	0	**21.2**	0	1.6	10.7	48.2	24.8	60.1
Ours (noMutex)	0.16	80.1	0	70.5	4.9	27.3	0	**24.3**	11.1	0	0	17.8	0	**3.4**	11.8	48.8	25.5	61.2
Ours ([3]+mutex)	0	70.6	0	93.2	9.9	65.7	0	20.3	1.0	0	0	15.5	0	2.8	**12.7**	50.1	27.6	64.7
Ours	**0.2**	67.2	0	**93.6**	**11.5**	**65.7**	0	15.3	2.2	0	0	5.6	0	2.3	11.0	**51.0**	**28.2**	**65.7**

Table 2. Performance of 33 classes semantic segmentation task on the SUN3D dataset. All 64 images are used as the test set. Note: since [20] and [12] did not report any results on SUN3D, we cannot include them here.

mutex constraints during inference. As is shown in the row 2 and column 3 image, the door area is labeled correctly as cabinet and the labelings of other regions are improved too. Also the big sofa region (row 4, col 3) has been correctly recognized after our inference. The last row shows one labeling example from SUN3D dataset.

6. Conclusion

We present a novel method for indoor scene semantic segmentation from RGB-Depth images. We effectively utilize available 3D geometric structures of indoor scenes and learn object relationships directly from training set. Our experimental results demonstrate incorporating hard mutex constraints into a soft CRF model can significantly increase the labeling accuracy. The proposed approach outperforms the state of the art methods on very challenging NYU-v2 RGBD dataset and SUN3D dataset for indoor scene semantic segmentation.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1302700, IIS-1302164, and OIA-1027897.

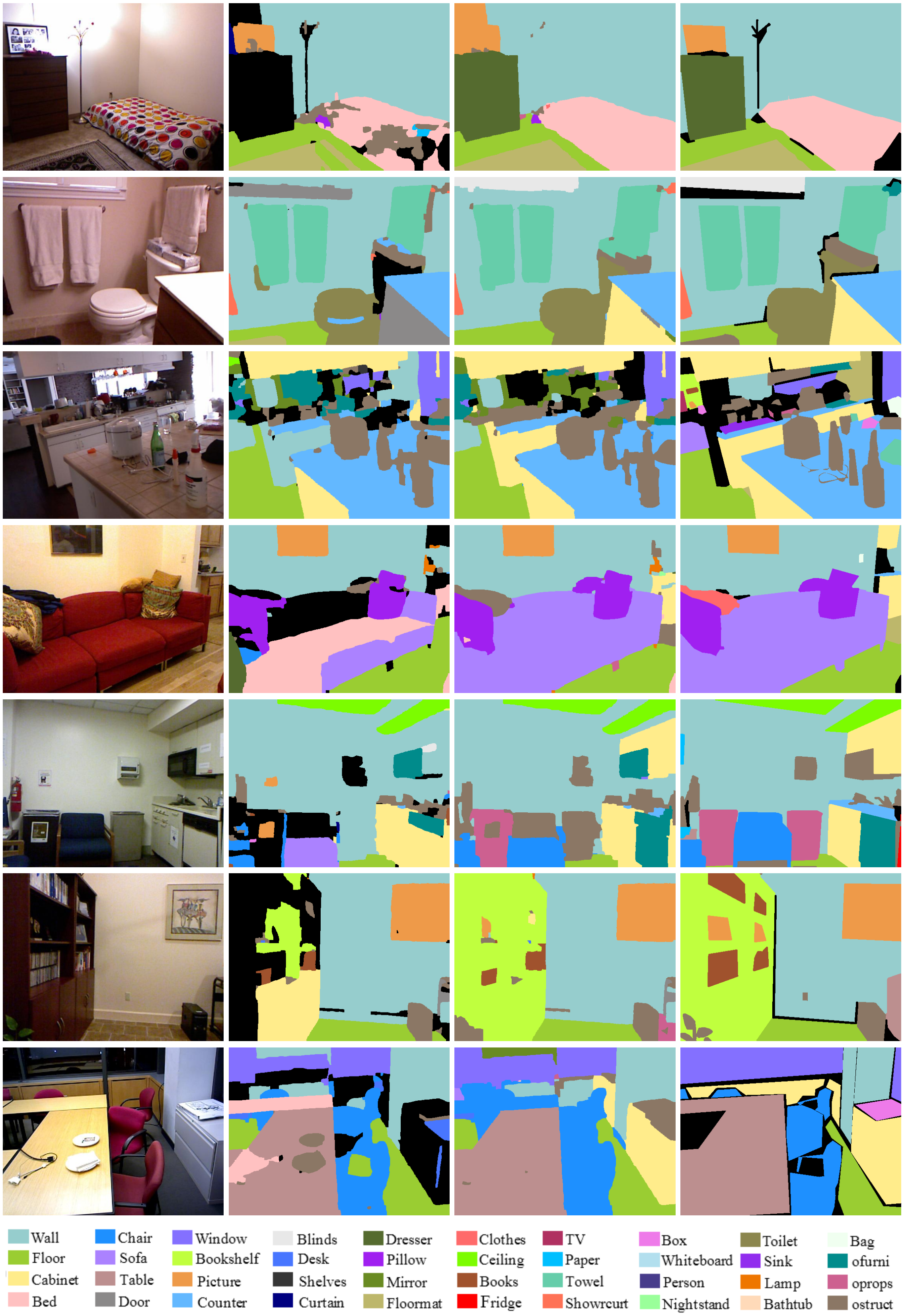

Figure 3. Examples of indoor scene semantic segmentation obtained by our system. Column 1 shows the original RGB images, column 2 shows the results from [3], column 3 shows our results after inferring with hard mutex constraints and column 4 shows the ground truth (black areas are unlabeled). Recommend to view in color.

References

[1] Ren, X., Bo, L., Fox, D.: RGB-(D) scene labeling: Features and algorithms. CVPR (2012)

[2] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.:Indoor segmentation and support inference from rgbd images. ECCV (2012)

[3] Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. CVPR (2013)

[4] Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. ICCVW (2011)

[5] Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. IJCV (2008)

[6] Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. CVPR (2008)

[7] Kohli, P., Torr, P.H.: Robust higher order potentials for enforcing label consistency. IJCV (2009)

[8] Ladicky, L., Russell, C., Kohli, P., Torr, P.H.: Graph cut based inference with co-occurrence statistics. ECCV (2010)

[9] Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. ICLR (2013)

[10] Roy, A., Todorovic, S.: Scene labeling using beam search under mutex constraints. CVPR (2014)

[11] Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Geometry driven semantic labeling of indoor scenes. ECCV (2014)

[12] Gupta, S., Girshick, R., Arbelez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. ECCV (2014)

[13] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI (2011)

[14] Ma, T., Latecki, L J.: Maximum weight cliques with mutex constraints for video object segmentation. CVPR (2012)

[15] Khoshelham, K.: Accuracy analysis of kinect depth data. ISPRSW (2011)

[16] Xiao, J., Owens, A., Torralba, A.: SUN3D: A database of big spaces reconstructed using sfm and object labels. ICCV (2013)

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

[18] Bo, L., Ren, X., Fox, D.: Kernel descriptors for visual recognition. NIPS (2010)

[19] Quattoni, A., Torralba, A.: Recognizing indoor scenes. CVPR (2009)

[20] Gupta, S., Arbelaez, P., Girshick, R., Malik, J.: Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation. IJCV (2014)