Unsupervised Segmentation of RGB-D Images

Zhuo Deng, Longin Jan Latecki

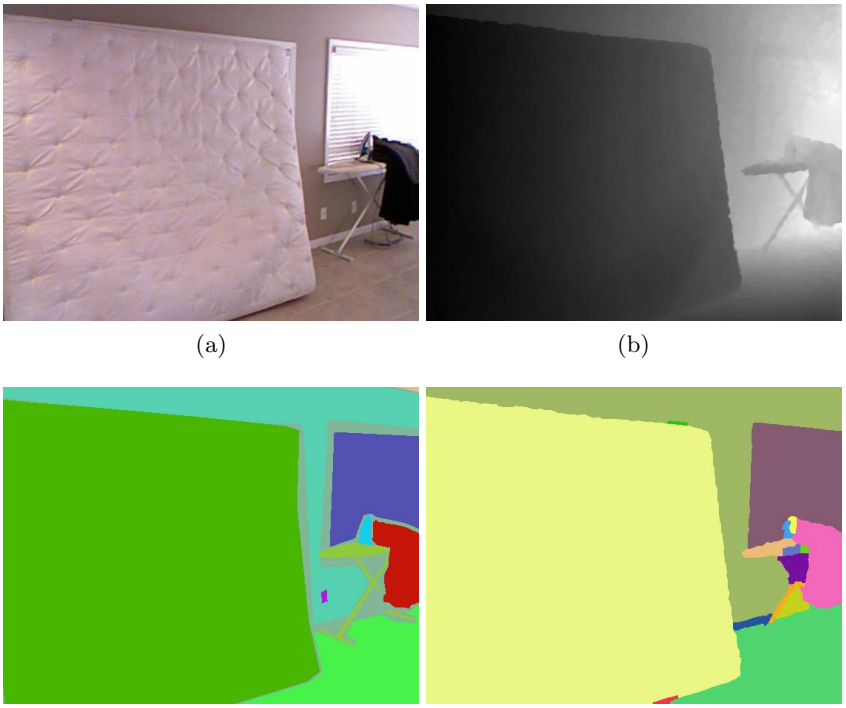
Dept. of Computer and Information Sciences,Temple University, Philadelphia, USA,
{zhuo.deng, latecki}@temple.edu

Abstract. While unsupervised segmentation of RGB images has never led to results comparable to supervised segmentation methods, a surprising message of this paper is that unsupervised image segmentation of RGB-D images yields comparable results to supervised segmentation. We propose an unsupervised segmentation algorithm that is carefully crafted to balance the contribution of color and depth features in RGB-D images. The segmentation problem is then formulated as solving the Maximum Weight Independence Set (MWIS) problem. Given superpixels obtained from different layers of a hierarchical segmentation, the saliency of each superpixel is estimated based on balanced combination of features originating from depth, gray level intensity, and texture information. We want to stress four advantages of our method: 1) Its output is a single scale segmentation into meaningful segments of a RGB-D image; 2) The output segmentation contains large as well as small segments correctly representing the objects located in a given scene; 3) Our method does not need any prior knowledge from ground truth images, as is the case for every supervised image segmentation; 4) The computational time is much less than supervised methods. The experimental results show that our unsupervised segmentation method yields comparable results to the recently proposed, supervised segmentation methods [1, 2] on challenging NYU Depth dataset v2.

1 Introduction

Unsupervised Image Segmentation (UIS) is one of the oldest and most widely researched topics in the area of computer vision, of which the goal is to partition an image into several groups of pixels that are visually meaningful using only the information provided by the single image.

In the past few decades, many great accomplishments have been made in this field from the early techniques [4, 5], which usually are based on the region splitting or merging framework to more recent works which tend to either integrate global constraints into grouping task, such as intra-region consistency and inter-region dissimilarity [6–9], or formulate segmentation problem under clustering framework [10]. However, unsupervised image segmentation has remained an unsolved problem of computer vision, since RGB color information alone of a single image often does not provide sufficient information to successfully complete this task. There are many reasons for this, e.g., lack of distinctive features



(a) (b)

(c) (d)

Fig. 1: A typical indoor scene and our segmentation results. (a) Original RGB image obtained from Kinect camera. (b) Depth image, the missing values of which has been filled by the approach in [3]. (c) Ground truth segmentation. (d) Final segmentation result based on the proposed method.

and instability of features due their sensitivity to illumination variation. Generally speaking, UIS is extremely difficult since incorrect segmentations (either too fine or too coarse) can be easily derived, even when employing algorithms that require the user to guess the number of segments.

Recently, with the advent of Microsoft Kinect, the landscape of various vision-related tasks has been changed. Firstly, using an active infrared structured light sensor, the Kinect can provide directly the depth information that is hard to infer from traditional RGB images. Secondly, RGB and depth information are generated synchronously and can be easily aligned, which makes their direct integration possible. A wide range of research works have demonstrated that RGB-D information is useful for improving the performance of vision tasks such as object recognition [11], scene labeling [1], body pose estimation [12], saliency detection [35] etc. The depth information itself is also very helpful for scene geometric structure estimation.

The main goal of this paper is to explore the impact of RGB-D information on improving the unsupervised image segmentation. As we will demonstrate, the improvement is dramatic to the point that for many scenes the segmentation results are comparable to the results of supervised segmentation. Both supervised and unsupervised image segmentation that return a single scale complete image segmentation face the same problem of obtaining image segments correctly representing the scene objects of varying sizes. In particular, segments belonging to a single segmentation result may differ dramatically, some segments may fill nearly the whole image, representing objects like sofas in close view, and some may have area smaller that 1/100 of the image area. To solve this problem, we formulate the single scale segmentation as finding a maximum weight independent set (MWIS). This way we can automatically partition an RGB-D image into several salient regions with no need to specify either the number or sizes of regions in advance. A representative example is shown in Fig. 1.

The MWIS segmentation has been proposed for RGB images in [9]. It yields good segmentation results when foreground objects are very different from the background, since only then the region saliency measure is able to provide useful segment weights. Due to specific of RGB-D images, our saliency measure is very different and more informative. The main contribution of the proposed approach is a definition of region saliency measure that incorporates both RGB and depth information. As stated above such measure needs to properly balance the color and depth information, since for many objects only one of them is informative.

We test our method on the NYU depth dataset [1] and compare it to supervised hierarchical segmentation approaches in [1, 2]. [1] starts from an over-segmentation, and adapts the algorithm in [13] to iteratively merge regions based on boundary strength. This approach is supervised, since the boundary strength needs to be learned from labeled instances. Similarly, [2] trains oriented contour detectors based on features extracted from watershed over-segmentation contours. Finally, initial over-segmentation regions are merged based on the average strength of oriented contour detectors. Although our method is unsupervised, it obtains comparable results to [1, 2]. Moreover, we also compare our approach to an unsupervised segmentation method in [14]. It extends the work of [6] by creating an extra edge on the original graph, of which the weight is measured based on the angle difference of surface normals obtained from depth information. In addition, we also use gpb-owt-ucm as a baseline where depth information is not used. We evaluate the segmentation quality based on five standard measures: Probabilistic Rand Index (PRI) [15], Variation of Information (VI) [16], Global Consistency Error (GCE) [17], Boundary Displacement Error (BDE) [18] and Jaccard Index (JI)[8]. Our approach significantly outperforms [14] in all five measures, which clearly demonstrates the superiority of the proposed combination of color and depth information.

2 Related Works

Image segmentation is a fundamental problem and has been studied extensively. Classic image segmentation approaches include normalized cuts [7], minimum spanning tree [6], meanshift [10], and gPb-OWT-UCM[8]. However, these approaches can only obtain segmentation results comparable to humans if their parameters are known in advance or in other words manually tuned. For example, the normalized cuts requires assigning a specific number of regions at the beginning. Therefore, these algorithms are usually run with different parameter settings, which yields multi-scale image segmentation results. While multi-scale results are very useful for many supervised methods for object detection, scene labeling or image segmentation, it is hard to utilize them to obtain a single segmentation result of an RGB image in unsupervised setting.

One common drawback of these unsupervised segmentation techniques is that they have no prior knowledge about the geometric structure of the scene, which leads to the segmentation to be either too coarse if two spatially separated regions have similar appearance or too fine when one planar region contains subregions with different textures. Although recent approaches that try to infer the 3D structure of the scene given only a single RGB image, e.g., [19–23], they are limited to very simple structures.

The emergence of the RGB-D technology provides a great opportunity to take advantages of merits from both RGB and depth information. Some of the recent works on unsupervised RGB-D segmentation integrate the image segmentation with plane fitting [24, 25]. In [24], the RGB-D segmentation is formulated as iterative refinement of the pixel-to-plane assignment and optimized as discrete labeling in a Markov Random Field (MRF), with plane merging controlled by a threshold. [25] formulates the plane fitting as a linear least-squares problem and infers the segmentation of the scene in a Bayesian framework. The other unsupervised segmentation works are trying to adapt the classic segmentation algorithms into the RGB-D field. [26] first detects edges on RGB images and computes triangular tessellation of images based on edge information by the Delaunay Triangulation algorithm. Then a variant of N-cut is applied to the graph constructed from the triangular regions. Finally the segments from N-cuts are used to suggest groupings of depth samples from depth image. [27] extends the work in [26] to segment the Manhattan structure of an indoor scene from a single RGB-D frame into floor plane and walls. In contrast to these approaches, our method is not limited to planar structures in the scene. Similar to our work, in [28], image segmentation is formulated as finding high-scoring maximal weighted cliques in a graph connecting non-overlapping putative figure-ground segment hypothesis. In [36], the pylon model is proposed to find a globally optimal subset of segment pool and their labels through graph-cuts and max-margin learning. But both [28] and [36] are supervised whereas ours is an unsupervised method. Except for unsupervised segmentation, supervised segmentation also benefits from the RGB-D technology. One of the most recent works is [1], where regions with minimum boundary strength are iteratively merged in a hierarchical framework. The boundary is predicted by a trained boosted decision tree clas-

sifier based on labeled instances. The other one proposed in [2] utilize depth information to train several oriented contour detectors. Hierarchical segmentation is constructed by merging regions of initial over-segmentation based on the average strength of those oriented contour detectors. Unlike the above works, the proposed approach is completely unsupervised, since it does not require any parameter learning from labeled instances, nor we make any assumptions about the number of regions to be segmented.

3 General Framework

3.1 Hierarchical image segmentation

To partition one image into superpixels, there are several excellent algorithms such as the gPb-OWT-UCM method of [8], the minimum spanning tree segmentation [6], the multi-scale normalized cuts [29], mean shift segmentation [10], and watershed based segmentation [30]. In this paper, we adapt the method introduced in [8] to integrate both RGB and depth information for hierarchical segmentation. In [8], firstly an over-segmentation is derived based on the watershed transformation of the gradient map, which is a linear combination of brightness, color, texture gradients and spectral signal. Following the multiple cues combination framework, we integrate depth and normal gradients directly into the final gradient map. Suppose we denote an image as $I(x, y)$, the gradient map $G(x, y)$ is represented as

$$G(x, y) = w_b G_b + w_c G_c + w_d G_d + w_n G_n + w_s G_s, \tag{1}$$

where G_b and G_c are brightness and color gradient signals respectively, which are computed in the CIE-LAB color space. G_d is the gradient signal estimated based on depth image. G_n represents the normal signal where the difference of two normal vectors $\mathbf{n}_i$ and $\mathbf{n}_j$ is measured as

$$Dist(\mathbf{n}_i, \mathbf{n}_j) = sin(acos(\frac{\mathbf{n}_i \bullet \mathbf{n}_j}{|\mathbf{n}_i||\mathbf{n}_j|}), \tag{2}$$

and G_s is the spectral signal. All the gradient signals except for the spectral signal are estimated by convolving a 3×3 sobel kernel with signals themselves. Then an over-segmentation is obtained by applying the watershed transformation to $G(x, y)$. In order to present the hierarchical segmentation, Ultrametric Contour Map (UCM) is used to capture the average strength of shared boundary between two adjacent regions based on $G(x, y)$. For an input RGB-D image, we obtain 7 scales of hierarchical image segmentation by adjusting the strength threshold θ_g on the UCM. We denote with V the set of all superpixels from all scales and from both RGB and D images.

3.2 Saliency measure of superpixels

The goal of this section is to compute the saliency measure for each superpixel in V. For RGB-D segmentation, a critical issue is how to integrate depth information with RGB information in order to obtain a weight of each superpixel.

Previous works such as [31] and [24] assign a fixed importance weight to RGB and depth information respectively based on parameter training or empirical setting. However, it is not the case that depth information is more important than RGB information nor vice versa. In reality, when we are trying to identify a salient object from its background, the criteria used always change. For example, based on depth it is easy to separate the surface of a desk from the floor. Whereas, to distinguish a bedsheet from a bed frame, color or texture properties are more helpful. Based on this intuition, we propose a novel weighting scheme to estimate the saliency of superpixels in RGB-D images.

We estimate the saliency by combining three kinds of information: depth, gray level intensity, and textures. Suppose we denote a superpixel as $S_i \in V$ and given depth image $I_d(x, y)$, and RGB image $I_c(x, y)$. We extract gray scale image $I_g(x, y)$ from $I_c(x, y)$. The corresponding saliency measures $C_d(S_i)$, $C_g(S_i)$, $C_t(S_i)$ are defined below. The higher their values, the more uniform is superpixel S_i. We define the saliency of superpixel S_i as their weighted average

$$w(S_i) = W_{area}(w_1 C_d(S_i) + w_2 C_g(S_i) + w_3 C_t(S_i)), \tag{3}$$

where $w_1, w_2, w_3 \geq 0$, $w_1 + w_2 + w_3 = 1$,

$$W_{area} = (1 - \exp(-\eta \frac{|S_i|}{|I(x, y)|}))$$

is used to slightly favor larger regions. The weights w_1, w_2, w_3 are dynamically assigned so that the value of most informative of the three saliency measures $C_d(S_i)$, $C_g(S_i)$, $C_t(S_i)$ has the higher weight. We have three constant values $\alpha > \beta > \gamma > 0$ for the weights and assign the largest value to the largest feature, e.g., if $C_d(S_i) > C_g(S_i) > C_t(S_i)$, then $w_1 = \alpha, w_2 = \beta, w_3 = \gamma$.

Unlike [35] where the relationship between saliency and depth is trained by fitting a GMM, we directly define the confidence from depth information $C_d(S_i)$ as

$$C_d(S_i) = \exp(\frac{-std(\{G_d(p)|p \in S_i\})}{|\underset{p \in S_i}{avg}(\{I_d(p)\}) - \underset{p \in S^i_{ext}}{avg}(\{I_d(p)\})|}) \tag{4}$$

where $p = (x, y)$ represents a pixel at position (x, y), S^i_{ext} denotes the neighboring area of S_i, and $G_d(x, y)$ represents the gradient map of $I_d(x, y)$. This term encourages the planar region that has high contrast to its surrounding area on the depth value.

The gray scale confidence is defined as

$$C_g(S_i) = \exp(\frac{-\underset{p \in S_i}{std}(\{I_g(p)\})}{\underset{p \in S^i_{ext}}{std}(\{I_g(p)\})}). \tag{5}$$

The region where pixels have similar intensity value within it and dissimilarity is high with respect to its neighbor area should be assigned a heavier weight.

In order to estimate the weight from the texture perspective, we firstly apply the Maximum Response (MR8) filter bank [32] to the gray scale image $I_g(x, y)$.

MR8 filter bank consists of 38 filters (6 orientations at 3 scales for 2 oriented filters and 2 isotropic filters) and the number of filter responses is reduced to eight. Each pixel of $I_g(x, y)$ is attached with a filter response vector $\mathbf{f}_r$. Then K-means clustering are used to extract k "vector words". Each vector $\mathbf{f}_r$ is assigned an integer label of the "vector word" which is closest. In order to measure the texture saliency, we use the J-measure proposed in [33] that is based on spatial distributions of pixels of similar properties. Suppose there are n_c different labels in S_i, C_i denotes all pixels in S_i with the same quantized label, and N_i is the number of pixels in C_i. The center of C_i is denoted as $m_i = \frac{1}{N_i}\sum_{p\in C_i} p$. We define

$$S_W = \sum_{i=1}^{n_c}\sum_{p\in C_i} ||p - m_i||^2 \tag{6}$$

and observe that S_W is small if there are compact clusters of labels in S_i while it is large if pixels with different labels are uniformly distributed in S_i. We also define the spread of all pixels in S_i as

$$S_T = \sum_{p\in S_i} ||p - m||^2 \tag{7}$$

where m is the central point of S_i. The texture salience is then defined as

$$C_t(S_i) = \exp(\frac{S_W - S_T}{S_W}) \tag{8}$$

If all the pixel labels are distributed uniformly over the entire superpixel area, the value of $C_t(S_i)$ is large. In contrast, it is small if there are compact clusters of labels in S_i.

3.3 Final Segmentation as MWIS

We first construct a graph composed of superpixels $S_i \in V$ as its nodes, where $|V| = n$ We assign to each node $S_i \in V$ a weight $w_i = w(S_i)$ defined in formula (3). We observe that all weights are nonnegative and denote with $\mathbf{w} = [w_1, w_2, ..., w_n]^\top$ the weight vector.

The adjacency matrix M is defined as follows. An edge exists between two superpixels S_i and S_j if they overlap, i.e., $M_{ij} = 0$ if $S_i \cap S_j = \emptyset$ and $M_{ij} = 1$ otherwise. We obtain an undirected graph $G = (V, M, \mathbf{w})$.

In graph theory, an *independent* set is a set of vertices in a graph where no two vertices are adjacent. The *maximal independent set* is an independent set which has the largest number of vertices. In the case we have a weight attached to each vertex, the *maximum weight independent set (MWIS)* is an independent set with the largest sum of the node weights.

An indicator vector, $\mathbf{x} = [x_1, x_2, ..., x_n]^\top \in \{0, 1\}^n$, is used to denote any subset B of the graph nodes, where $x_i = 1$ means node $S_i \in B$ and $x_i = 0$ means node $S_i \notin B$. When B is an independent set and $\mathbf{x}$ its indicator vector, we have $\forall(i, j), x_i \cdot x_j = 0$ if $M_{ij} = 1$. Hence it holds that $\mathbf{x}^\top M\mathbf{x} = 0$. Therefore,

$\mathbf{x}^*$ representing the MWIS can be obtained as the solution of the following quadratically constrained integer linear program

$$\begin{aligned} &\mathbf{x}^* = \underset{\mathbf{x}}{argmax} \;\; \mathbf{w}^\top \mathbf{x} \\ &s.t.\; \forall i \in V : x_i \in \{0,1\}, and \;\; \mathbf{x}^\top M \mathbf{x} = 0 \end{aligned} \tag{9}$$

We solve the program (9) with the algorithm introduced in [9]. The solution vector $\mathbf{x}^*$ selects superpixels that compose our final single scale segmentation of a given image.

4 Experiments

This section presents both qualitative and quantitative evaluation of our unsupervised segmentation algorithm on 1449 pairs of aligned RGB and depth images from the NYU Depth Dataset V2 [1]. Detailed ground truth segmentation is provided for each image. This data set is very challenging for segmentation, even with RGB-D information, because of poor illumination, often rendering RGB information useless, cluttered non-planar stuff (eg. bedsheets, sofa, clothes etc), which strongly limits the depth cues, large variation of scene types, and non-perfect depth measurement. In particular, depth images contain "black holes" due to missing data, and random error of depth measurements increase quadratically with the increasing distance from the sensor [34]. Also the average density of depth measurements decreases when the distance to the objects increases, since the resolution of Kinect is fixed at $480 * 640$.

In order to evaluate our algorithm quantitatively, five standard evaluation measures are employed. The first one is Probabilistic Rand Index (PRI), which estimates the ratio between pairs of pixels, whose labelings are consistent in both ground truth and estimated segmentation, and the total number of pixel pairs. Variation of Information (VI) measures the distance between two segmentations by the average conditional entropy of one segmentation given the other. Global Consistency Error (GCE) measures the extent to which one segmentation can be viewed as a refinement of the other. The Boundary Displacement Error (BDE) measures the average displacement error of boundary pixels between two segmented images. Particularly, it defines the error of one boundary pixel as the distance between the pixel and the closest pixel in the other boundary image. The Jaccard Index (JI) measures similarity between two segmentations, and is defined as the size of the intersection divided by the size of the union of the two segmentations.

We first compare our method to the two baseline UIS methods: in [8], depth information is not used and in [14], normal vector information is applied. For [8], we select the best layer from the hierarchical segmentation based on the five evaluations. As can be seen in Table 1, our method significantly outperforms both of the baseline methods on all five evaluation measures. Surprisingly, the result of [8] is slightly better than [14]. We also compare our approach to two recent RGB-D supervised segmentation methods proposed in [1, 2]. Therefore, following

Fig. 2: Two examples to illustrate the benefits of using depth information. The first column contains two original RGB images from Kinect. The second column is the segmentations only based on RGB information. The third column contains the corresponding segmentations based on both RGB and depth information.

the same dataset split setting, training set contains 795 images, and performance is evaluated on 654 test images. Since the algorithm in [1] outputs a hierarchical segmentation composed of five segmentation levels, we choose the best result based on the five standard evaluation measures out of the five levels for each image. [2] similarly outputs a hierarchical segmentation of 99 segmentation levels. We use the best layer as evaluated in their paper (threshold = 0.54). Although our method is unsupervised, for fair comparison, we also evaluate it on the 654 test images. As can be seen in Table 1, the performance of our method is very close to theirs. This is very surprising for at least three reasons: 1) Our method is unsupervised, while the method in [1, 2] are supervised. 2) Our method is much simpler than the methods in [1, 2]. 3) Our segmentation result sometimes shows more details than the ground truth, since it is not restricted to known object classes, which incorrectly lowers our accuracy.

In order to visually compare supervised segmentation results [1, 2] with our unsupervised segmentation results, we list 8 different samples in the Fig. 3. in varieties of scene categories such as bookstore, living rooms, offices, classrooms and so forth. As can be seen the segmentation of our result is very competitive. Our approach is robust to the variation of illumination, even when scenes are dark (eg. the scene in the bathroom) or when scenes are extremely bright, e.g., the blinds of the living room in Fig. 1 and the surface of the blackboard in the conference room, or when shades are projected on objects, e.g., the shades on the floor and wall of the bedroom scene. Our approach also works well in very cluttered indoor scenes, like the scenes in the bookstore and the lady's office.

The results in Fig. 2 also demonstrate that depth information is really helpful in our framework for distinguishing objects with similar colors but different locations from each other. As can be seen in the kitchen scene, the surface of the

Method	PRI	GCE	VI	BDE	JI
RGB [8]	0.889	0.178	2.253	9.236	0.527
RGBD [14]	0.875	0.298	2.165	11.381	0.488
RGBD [1]	**0.917**	0.122	1.706	**7.509**	0.605
RGBD [2]	0.916	0.162	**1.501**	7.808	**0.622**
Ours RGBD	0.914	**0.120**	1.891	8.488	0.583

Table 1: Segmentation accuracy evaluated on 654 test RGB-D images in the NYU Depth Dataset V2 [1], since methods in [1] and [2] are supervised. The values are: PRI (larger is better), VI (smaller is better), GCE (smaller is better), BDE (smaller is better) and JI (larger is better).

table, the wall, and the refrigerator have similar white color, and in the living room scene, the sofa and the blanket on the floor also have similar color. So when only RGB information is used, different objects are inclined to be segmented as one superpixel. However, when the depth information is added, all of them become correctly separated.

The average run time per image segmentation is listed in Table 2. It was evaluated on a PC computer with AMD Eight-core CPU @ 3.1HZ and 16GB RAM. Except for [14] which runs in C++, our method is much faster than GPb-OWT-UCM and other two supervised methods.

[1] in Matlab	our method in Matlab	[14] in C++	[8] in Matlab	[2] in Matlab
122.1	68.8	7.39	301.1	> 300

Table 2: The average run time in seconds to segment a single image.

Parameter setting: The input to our segmentation are superpixels obtained from hierarchical segmentations. As is mentioned in Section 3.1, we obtain segmentations at different levels by changing the value of the strength threshold θ_g which falls between 0 and 1. When θ_g increases, the number of regions segmented is reduced. Experimentally, we find that if the segmentation in each layer is too fine, it may produce many areas that consists of only several pixels. They are not only meaningless but also tend to increase the burden of computation. On the other hand, if the segmentation in each layer is too coarse, it also can not provide good candidate regions. Therefore, we set the θ_g to $[0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6]$. For the weights of different gradient signals, we simply set them as $w_b = 1.0$, $w_c = 0.5$, $w_n = 3.0$, $w_d = 2.0$ and $w_s = 3.0$ since depth information and global

spectral signal are much more reliable than brightness and color. In addition, we set constants α, β, and γ to 0.5, 0.3, 0.2 respectively. The constant η is set to 10 in our experiment.

5 Conclusion

In this paper, we propose an unsupervised segmentation method for RGB-D image segmentation. It integrates both color and depth information effectively and partitions one RGB-D image into several most salient regions without the need to know the number or the size of segments in advance. Our experiments on the NYC depth dataset show that the segmentation accuracy of our method is very competitive with respect to both unsupervised and supervised methods. Also the fact that our method is very efficient due to its simplicity, makes it very suitable for many applications from object to action recognition.

Acknowledgements

This work was in part supported by NSF under Grants IIS-1302164 and OIA-1027897.

References

1. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.:Indoor segmentation and support inference from rgbd images. ECCV (2012)
2. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. CVPR (2013)
3. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Transactions on Graphics (2004)
4. Brice, C., Fennema, C.: Scene analysis using regions. Artificial Intelligence (1970)
5. Horowitz, S., Pavlidis, T.: Picture segmentation by a tree traversal algorithm. JACM (1976)
6. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV (2004)
7. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI (2000)
8. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI (2011)
9. Brendel, W.and Todorovic, S.: Segmentation as maximum weight independent set. NIPS (2010)
10. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI (2002)
11. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. ICRA (2011)
12. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. CVPR (2011)

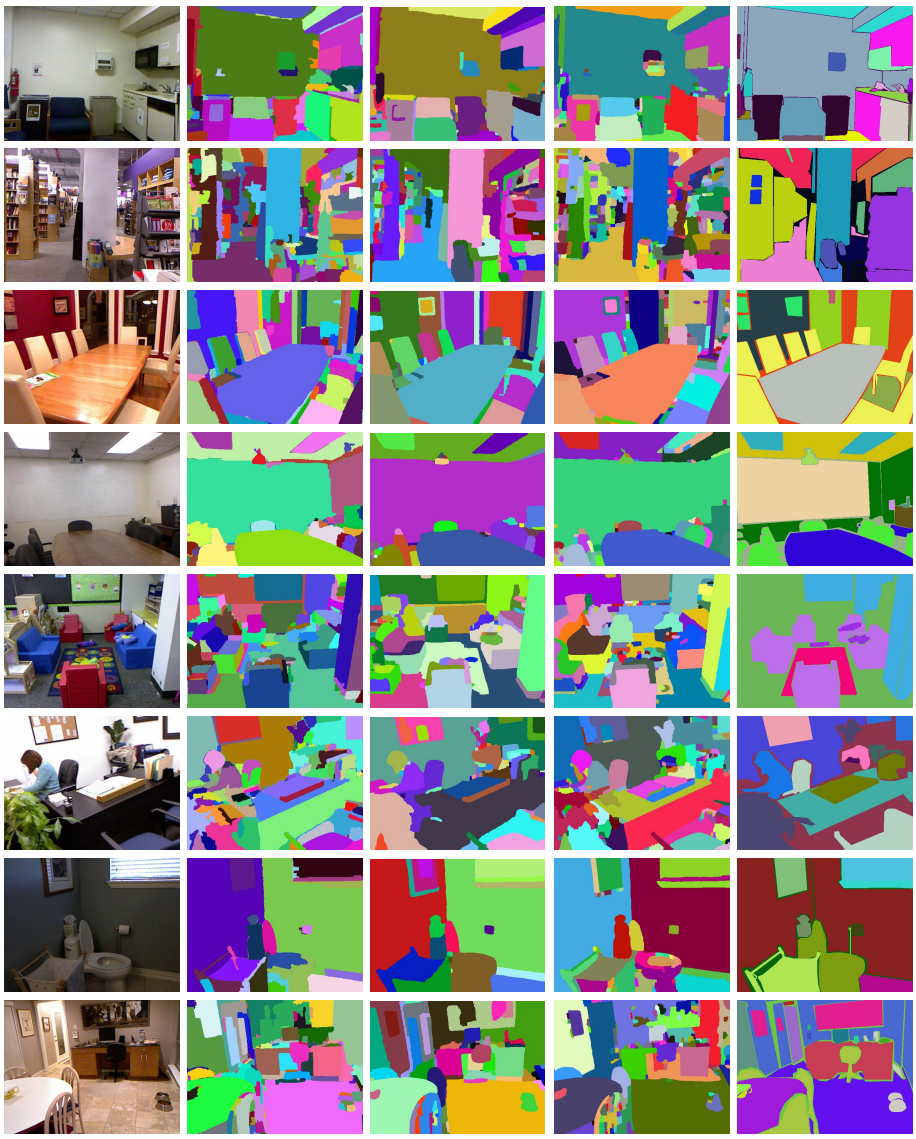


Fig. 3: Examples of unsupervised indoor scene segmentation obtained by our method and supervised methods in [1, 2]. Column 1 shows the original RGB images. Column 2 shows results in [1]. Column 3 shows results in [2]. Column 4 shows our segmentations and last column shows the ground truth.

13. Hoiem, D., Efros, A., Hebert, M.: Recovering occlusion boundaries from an image. IJCV (2011)
14. Strom, J., Richardson, A., Olson, E.: Graph-based segmentation for colored 3d laser point clouds. IROS (2010)
15. Unnikrishnan, R., Pantofaru, C., Hebert, M.: A measure for objective evaluation of image segmentation algorithms. CVPRW (2005)
16. Meila, M.: Comparing clusterings by the variation of information. Learning theory and kernel machines (2003)
17. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. ICCV (2001)
18. Freixenet, J., Munoz, X., Raba, D., Marti, J., Cufi, X.: Yet another survey on image segmentation: Region and boundary information integration. ECCV (2002)
19. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. ICCV (2005)
20. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. ICCV (2009)
21. Hedau, V., Hoiem, D., Forsyth, D.: Thinking inside the box: Using appearance models and context based on room geometry. ECCV (2010)
22. Lee, D., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. CVPR (2009)
23. Lee, D., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. NIPS (2010)
24. Guan, L., Yu, T., Tu, P., Lim, S.: Simultaneous image segmentation and 3D plane fitting for RGB-D sensors - An iterative framework. CVPRW (2012)
25. Erdogan, C., Paluri, M., Dellaert, F.: Planar segmentation of rgbd images using fast linear fitting and markov chain monte carlo. CRV (2012)
26. Taylor, C., Cowley, A.: Segmentation and analysis of rgb-d data. RSS (2011)
27. Taylor, C.J., Cowley, A.: Parsing indoor scenes using rgb-d imagery. Robotics (2013)
28. Ion, A., Carreira, J., Sminchisescu, C.: Image segmentation by figure-ground composition into maximal cliques. ICCV (2011)
29. Cour, T., Benezit, F., Shi, J.: Spectral segmentation with multiscale graph decomposition. CVPR (2005)
30. Meyer, F.: Color image segmentation. Image Processing and its Applications (1992)
31. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. CVPR (2012)
32. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. IJCV (2005)
33. Deng, Y., Manjunath, B., Shin, H.: Color image segmentation. CVPR (1999)
34. Khoshelham, K.: Accuracy analysis of kinect depth data. ISPRS workshop (2011)
35. Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Influence of depth cues on visual saliency. ECCV (2012)
36. Lempitsky, V., Vedaldi, A., Zisserman, A.: Pylon model for semantic segmentation. NIPS (2011)