OBJECT-LEVEL SEGMENTATION OF RGBD DATA

Hai Huang[a]*, Hao Jiang[b], Claus Brenner[c], Helmut Mayer[a]

[a] Institute of Applied Computer Science, Bundeswehr University Munich, Neubiberg, Germany
{hai.huang, helmut.mayer}@unibw.de
[b] Computer Science Department, Boston College, Chestnut Hill, MA, USA
hjiang@cs.bc.edu
[c] Institute of Cartography and Geoinformatics, Leibniz University Hannover, Hannover, Germany
claus.brenner@ikg.uni-hannover.de

KEY WORDS: Segmentation, Point cloud, Scene, Interpretation, Image, Understanding

ABSTRACT:

We propose a novel method to segment Microsoft^{TM}Kinect data of indoor scenes with the emphasis on freeform objects. We use the full 3D information for the scene parsing and the segmentation of potential objects instead of treating the depth values as an additional channel of the 2D image. The raw RGBD image is first converted to a 3D point cloud with color. We then group the points into patches, which are derived from a 2D superpixel segmentation. With the assumption that every patch in the point cloud represents (a part of) the surface of an underlying solid body, a hypothetical quasi-3D model – the "synthetic volume primitive" (SVP) is constructed by extending the patch with a synthetic extrusion in 3D. The SVPs vote for a common object via intersection. By this means, a freeform object can be "assembled" from an unknown number of SVPs from arbitrary angles. Besides the intersection, two other criteria, i.e., coplanarity and color coherence, are integrated in the global optimization to improve the segmentation. Experiments demonstrate the potential of the proposed method.

1. INTRODUCTION

Image segmentation employing depth information has been intensively studied. Silberman and Fergus (2011) use RGB-Depth (RGBD) data from the Kinect sensor to improve the segmentation of indoor scenes. Each segment is classified as one of the seven categories, e.g., bed, wall and floor. Silberman et al. (2012) further extract the support relationship among the segments. In (Koppula et al., 2011), RGBD images are segmented into regions of 17 object classes, e.g., wall, floor, monitor and bed, for office or home scenes. Li et al. (2011) propose a method to segment engineering objects that consist of regular parts from point clouds. They consider the global relations of the object parts. In (Bleyer et al., 2012) the depth is estimated from a stereo image pair and an unsupervised object extraction is conducted maintaining physical plausibility, i.e., 3D scene-consistency.

Additionally, many methods have been proposed to segment geometric primitives , e.g., cubes and cylinders, or regular objects such as buildings, for which reconstruction rules can be derived. An overview of point cloud processing is given by Vosselman (2009). Rabbani et al. (2007) present an approach for the labeling of point clouds of industrial scenes. Geometric constraints are given in the segmentation in the form of the primitives: cylinder, torus, sphere and plane. Current research for 3D building extraction is reported by Lafarge and Mallet (2012) and Huang et al. (2013), in which the buildings are modeled as an assembly of primitive components.

In this paper we propose a novel method to segment RGBD images. First, we calculate "superpixels" in the images using both color and surface normal information and pre-segment the point cloud into relatively small groups: the patches. The "synthetic volume primitive" (SVP) model is then constructed by extending each patch with a synthetic extrusion in 3D. The SVPs vote for a common object via intersection as they actually represent the same underlying solid body. By this means, an object can be assembled from a number of such primitives of arbitrary shape from arbitrary angles.We use Markov Random Field (MRF) to model the SVPs and their relationships. Besides the intersection, two other criteria, i.e., coplanarity and color coherence, are integrated in the global optimization to improve the segmentation.

Compared to other approaches, the proposed method focuses on a full 3D parsing of the scene by working directly on the 3D point cloud derived from the raw data instead of dealing with 2D images with the depth values as an additional channel. No specific physical constraints or top-down modeling is required to ensure plausible results, because the essential 3D spatial constraints have been embedded into the SVP model (cf. Section 3.).

The paper is organized as follows: Section 2. introduces the concept of SVP and its construction. In Section 3. we describe the voting for freeform objects by SVPs. A global optimization of the segmentation is presented in Section 4. Experiments were performed on open source data-set and are presented in Section 5. The paper ends up with conclusions in Section 6.

2. THE SYNTHETIC VOLUME PRIMITIVE – SVP

The point clouds from the Kinect sensor (or other sensors providing RGBD data) only reveal the (partial) surface of a target scene. E.g., if a cube is scanned, the point cloud shows one up to three of its facets. We assume that every planar patch in the point cloud represents (a part of) the surface of an underlying solid body in the 3D world. For full 3D parsing we use SVPs to simulate how different patches of an object interact with each other. A hypothetic volume is generated as an extension along the normal direction. Multiple SVPs may vote for a common object by intersection with each other in 3D space. SVPs provide the following advantages for 3D scene understanding:

1. Freeform objects can be represented by the intersection of SVPs from arbitrary angles.

*Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-3-73-2014

2. There is no limit for the number or size of component SVPs.

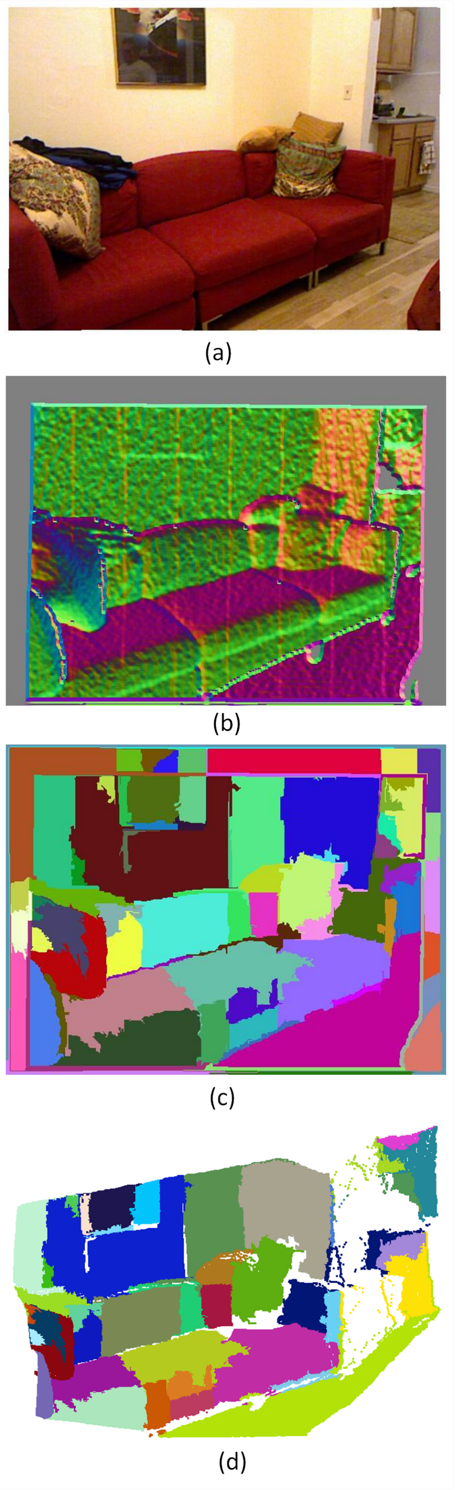

Figure 1: 3D patches: (a) input color image, (b) normal map, (c) superpixel segmentation and (d) a snapshot of the 3D point cloud with labeled patches.

By means of preliminary segmentation, we group a point cloud into a set of relatively small 3D patches. As shown in Figure 1, the segmentation is conducted on the input color and depth images. We use the superpixel segmentation of (Felzenszwalb and Huttenlocher, 2004) with the modifications of adding local normal information to individual points and separating points by either color incoherence or normal direction change. This separates the segments into mostly planar patches. Please note, however, in this work we do not detect any plane. In the SVP, a patch does not have to be planar either. The points in a patch could represent a surface segment of arbitrary shape, which makes an assembly of a freeform object possible. The 3D patches are extracted (Figure 1, d) with the known relationship between the image and the point cloud.

As shown in Figure 2, an SVP is constructed from a 3D patch by an extrusion. We assume that the underlying solid body is always behind the patch. The direction of the extrusion is along the normal vector that points towards the viewing direction. The SVPs vote for a common object by intersection with each other in 3D space. We, thus, use SVP models as “bricks” to build the 3D world.

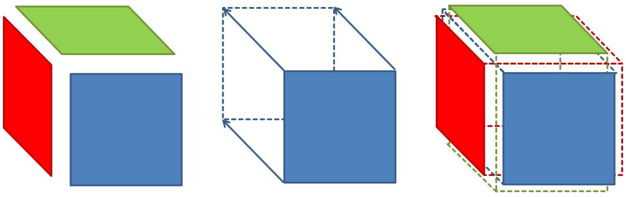

Figure 2: Definition of SVPs: (left) 3D patches, (middle) an SVP with hypothetical extrusion and (right) the intersection of SVPs.

Figure 3 shows the construction of SVPs. Since a surface patch may have an arbitrary shape (top left), in practice we model the extrusion volume with multiple “sticks” (top right) to approximate the shape. The stick model is a discrete representation of the 3D extrusion. For simplification the sticks are given a uniform diameter, which is equal to the average distance of the neighbor points. A tricky problem is how to define a reasonable length for the extrusion. With too short extrusions, the SVPs will fail to intersect with other object components. Too long extrusion, on the other hand, will result in the merging of multiple objects into a large one. Without any prior information, we can only assume a quasi cube shape for a hypothetical 3D body. We empirically found, that it is reasonable to use the average edge length of the patch’s bounding box as the length of the extrusion. An example scene with SVPs is shown in Figure 3 (bottom).

3. FREEFORM OBJECT VOTING

Object-level segmentation is challenging since the objects may have different sizes, colors and shapes. To decide if two 3D patches belong to the same object, a simple way would be to check if they are neighbors and their back sides face to each other. This, however, might fail in many cases, e.g., two patches belong to one object, but are not directly adjacent, two patches are connected with each other but do not belong to the same object, the patch is non-planar, or contains more than one plane, etc. To improve the results, geometric constraints and/or top-down models (with simplified and regular primitives) could be employed to ensure more reasonable segmentation. Even then, such methods are confined to limited, and mostly convex, shapes.

In this paper, we do not consider the various possibilities of patch combination, but enforce a single condition: If two patches belong to the same object, the hypothetical volumes behind them should have a certain overlap as they actually represent (parts of) the same 3D object. This is a simple but reasonable condition which describes the actual 3D relationship of the patches and the underlying object. As shown in Figure 4, with the help of SVPs

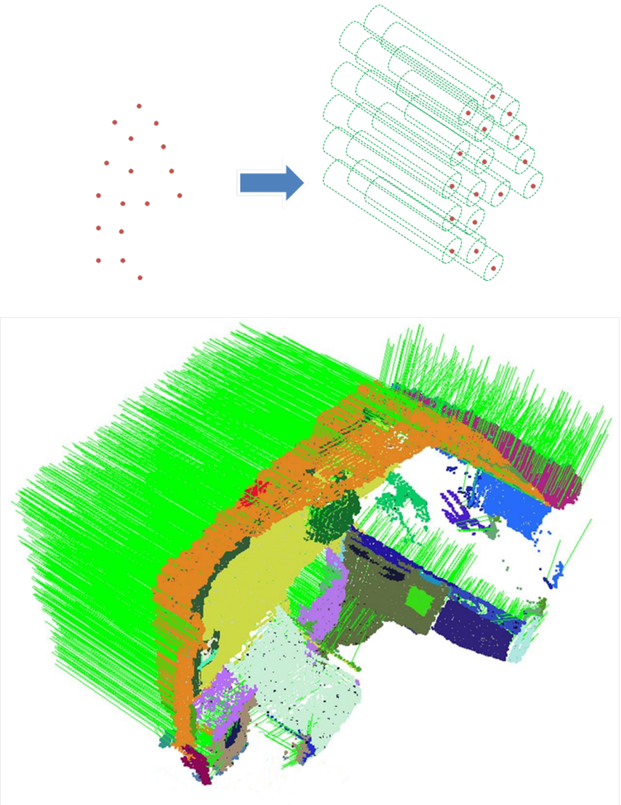

Figure 3: Application of SVPs. Top: using "sticks" (cylinders) to model the extrusion of a freeform patch. Bottom: example scene, in which the sticks are visualized as their axes (green beams).

the hypothetical components are "assembled" from arbitrary angles into an object. Remote object parts or the patches of concave shapes, as shown in Figure 4 (c, d), may not link with some neighbors, but still can be included via other object members from a different direction.

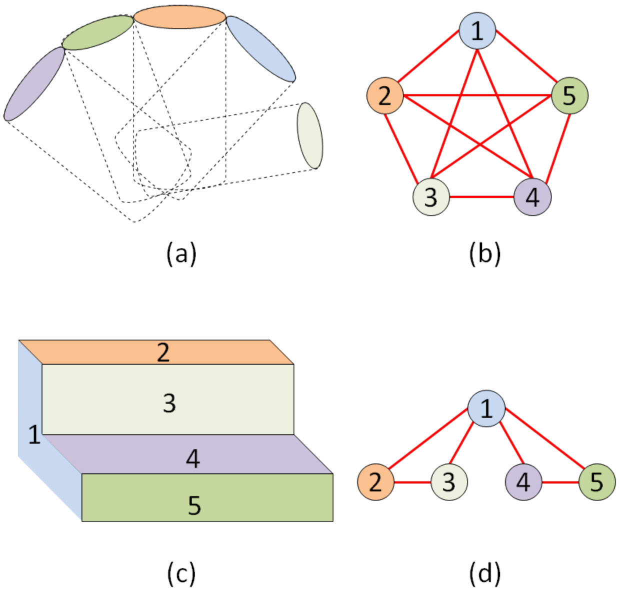

Figure 4: SVPs voting for freeform objects: SVPs can be assembled from arbitrary angles (a) and form concave shapes (c) when they are either fully (b) or partially (d) connected.

Instead of using a binary decision, we use an "intersection degree" which quantifies the interaction of two SVPs by a continuous (floating point) value. Besides a discrete approximation of the 3D volume, the above mentioned stick model also provides an easy way to quantify the intersection of 3D objects by just counting the intersecting sticks of different objects instead of calculating the real 3D volume overlap. We define the degree of intersection for a patch pair i and j as follows:

$$\mathcal{I}(i,j) = max\{\mathcal{J}_{i\to j}, \mathcal{J}_{j\to i}\} \tag{1}$$

with the intersection degree of the individual patch

$$\mathcal{J}_{i\to j} = \frac{m_i}{n_i} \cdot \frac{t}{m_i \cdot n_j^{0.5}} = \frac{t}{n_i \cdot n_j^{0.5}} \quad , \tag{2}$$

where n is the number of sticks, t the total number of intersections and m is the number of sticks involved in the intersection. The intersection degree $\mathcal{J}$ is defined as the product of the percentage of intersected sticks and the degree of intersection depth. We use $n_j^{0.5}$ to approximate the maximum possible depth. The larger value is taken for the patch pair (i, j). This score is then used as the likelihood for a valid grouping of these two patches.

4. GLOBAL OPTIMIZATION

A global optimization is beneficial because various scales of objects in the scenes significantly influence the parameter setting in the segmentation. An MRF is employed to model the SVPs and their relationships. As shown in Figure 5, the SVPs are represented as vertices and their neighborhood relationship as edges. In this undirected graphical model each SVP is only related to its first-order neighbor. The SVPs are defined as neighbors if the corresponding 3D patches are adjacent, which is practically calculated by finding the common boundaries in the 2D image (see Figure 1, c) and checking their 3D distance.

4.1 Coplanarity and Color Coherence

Coplanarity is an important geometric relationship. Patches may be merged if they are coplanar or their SVPs intersect. In this paper we use a simple threshold of 5° for the normal angle difference to determine coplanarity. Color coherence is employed as an additional criterion. We use the YUV distance to evaluate the color similarity. As shown in Figure 5, both of the criteria contribute to the correct segmentation (1-2-3-4-5 and 6-7).

4.2 Binary Energy

We defined the MRF model as:

$$G = (V, E) \tag{3}$$

with $v = v_i, i \in V$ the vertices and $e = e(i,j), \{i,j\} \in E$ the pairwise neighbor relationship. Any pair of non-neighbor vertices are conditionally independent given all other vertices, i.e.:

$$v_i \perp\!\!\!\perp v_j, if\ \{i,j\} \notin E. \tag{4}$$

Please note that there is no unary energy in the MRF. Different from most labeling tasks, e.g., figure/ground separation, the number and types of groups are unknown. There is no likelihood that can be derived from the local features of the individual vertex. In

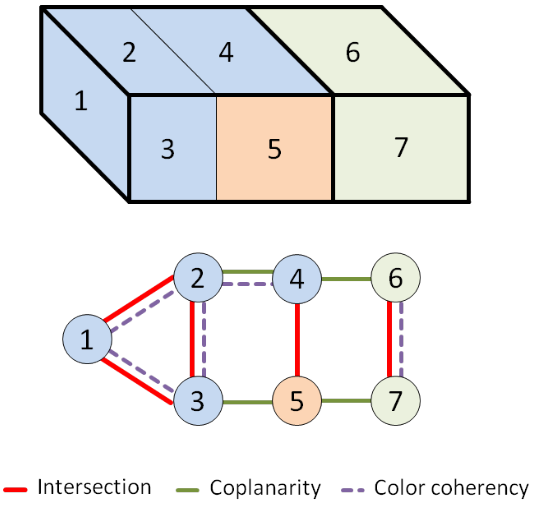

Figure 5: Coplanarity and color coherence are integrated in the neighbor relationship of an MRF and lead to an improved segmentation.

this segmentation task we observe only the binary energy of the pairwise cliques which is defined as:

$$\mathcal{B}(i,j) = \begin{cases} 0.7 \cdot \mathcal{P}(i,j) + 0.3 \cdot \mathcal{C}(i,j) & \forall v_i, v_j \textit{ coplanar} \\ 0.7 \cdot \mathcal{I}(i,j) + 0.3 \cdot \mathcal{C}(i,j) & \forall v_i, v_j \textit{ intersecting} \\ -1 & \textit{otherwise} \end{cases} \tag{5}$$

The binary energy is calculated for the intersection ($\mathcal{I} \in [0,1]$) or the co-plane ($\mathcal{P}$=1, otherwise 0) in combination with the color coherence ($\mathcal{C} \in [0,1]$). For these two cases, the binary energy represents the probability that the pair of patches belongs to the same object. We empirically found it to be advantageous to give 70% of the weight for the geometric relationship and 30% for the color. We assume that in the other cases the patches do not belong to the same object, which will be penalized with “-1” to discourage any group to include this pair.

The goal of the optimization is to find the maximum overall energy $\mathcal{H}$ of the graph model with the configuration $\mathcal{K}$, i.e., the grouping. Let $p(i,j)$ indicate the state of each pair in $\mathcal{K}$ (if connected $p(i,j) = 1$, otherwise 0). The goal function can be expressed as:

$$\widehat{\mathcal{K}} = \underset{\mathcal{K}}{argmax}\{\mathcal{H}\} = \underset{\mathcal{K}}{argmax}\left\{\sum \mathcal{B}(i,j) \cdot p(i,j)\right\} \tag{6}$$

subject to:
i and j are guaranteed to be disconnected if p(i,j)=0.

By this means the main objects in the scenes are segmented without previously giving the number of objects.

4.3 Stochastic Sampling

In the Markov field every edge has its probability to be broken and thus lead to a different grouping result. This makes the segmentation a high-dimensional optimization task. We employ a Markov Chain Monte Carlo (MCMC) sampler with Metropolis-Hastings scheme to solve this high-dimensional optimization task. The positive edge weight has been normalized to $[0,1]$ and quantifies the likelihood of two patches belonging to the same object, i.e., the edge persists conditionally to the sampling result. The negative value will be just treated as 0 in the sampling, i.e., the edge remains broken. The sampling process can be summarized as follows:

1. Initialization: $(\mathcal{M}^{(s=0)}, \mathcal{K}^{(s=0)})$
2. Propose new state $\mathcal{M}'$
   - 2.1 Sample new configuration $\mathcal{K}$ from $\mathcal{B}(i,j)$, for all $\{i,j\} \in E$
   - 2.2 Recover the above broken edges inside the new groups
   - 2.3 Calculate the overall energy $\mathcal{H}'$
3. Accept the new proposal with the probability

$$\mathcal{A}(\mathcal{M}^{(s)}, \mathcal{M}') = \min\left\{1, \frac{p(\mathcal{M}'|\mathcal{D})}{p(\mathcal{M}^{(s)}|\mathcal{D})} = \frac{\mathcal{H}'}{\mathcal{H}^{(s)}}\right\} \tag{7}$$

4. $\mathcal{M}^{(s+1)} = \mathcal{M}'$ if accepted, otherwise $\mathcal{M}^{(s+1)} = \mathcal{M}^{(s)}$

with s the step number and $p(\mathcal{M}|\mathcal{D})$ the likelihood that the state $\mathcal{M}$ fits the data $\mathcal{D}$, which is represented by the overall energy $\mathcal{H}$.

5. EXPERIMENTS AND RESULTS

Experiments are performed on the New York University RGBD dataset (Silberman et al., 2012). Figure 6 shows some examples.

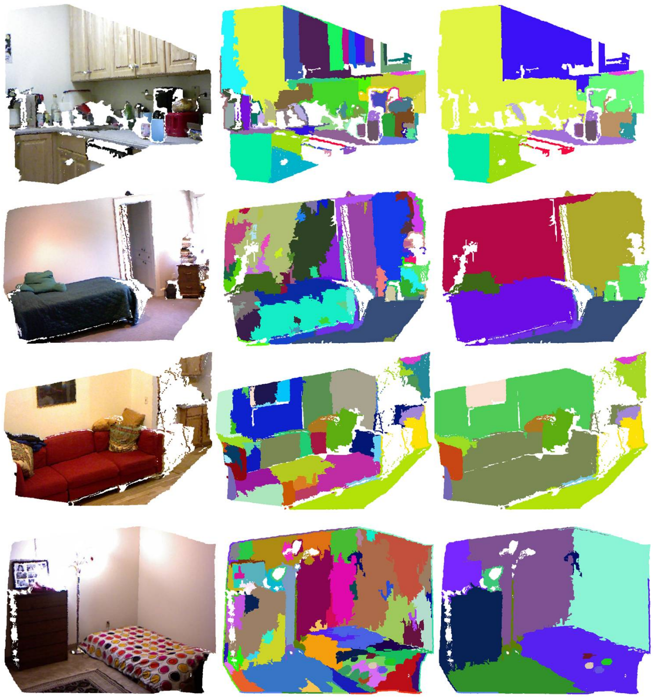

Figure 6: Segmentation examples (snapshots of 3D models): point clouds with color (left), patches/superpixels (middle) and the segmentation results (right).

We use the “accuracy” of region reconstruction:

$$\mathcal{S} = Accuracy_{rec.} = \frac{TP}{TP + FP + FN}, \tag{8}$$

with

TP: True Positive, regions of the object that have been segmented

FP: False Positive, incorrectly segmented regions

FN: False Negative, regions of the object that have not been covered

to quantitatively evaluate the segmentation result of individual objects. For each scene we only consider the (five to ten) dominant objects and calculate the total score as:

$$\mathcal{S}_{scene} = \frac{\sum A_i \cdot \mathcal{S}_i}{\sum A_i} \tag{9}$$

with $\mathcal{S}_i$ the accuracy of individual objects and the ground truth object regions A_i used as weights.

Table 1 shows the segmentation results of the four scenes presented above. The scores are compared with the (best) object segment proposal presented in (Endres and Hoiem, 2010).

Scene	1	2	3	4	...	ave.
objects	6	4	9	5	...	6.3
Prop.	0.8633	0.9182	0.9021	0.8811	...	0.89
SVP	0.9006	0.9234	0.9563	0.9193	...	0.93

Table 1: Test scenes and evaluations (accuracy) in comparison to the work of Endres and Hoiem (2010).

The results of more test scenes are summarized in Figure 7. Again, the results are compared with the work in (Endres and Hoiem, 2010), in which a category-independent method is introduced to generate a number of regions and find good segmentation schemes (proposals) based on the top-ranked ones. The average accuracy values for the 20 test scenes are 0.9289 (SVP) and 0.8910 (the best proposal of Endres and Hoiem (2010)), respectively.

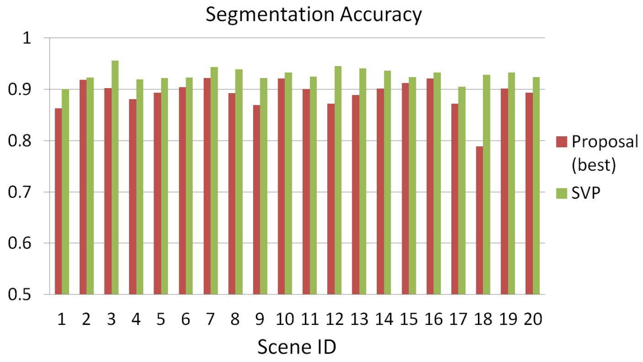

Figure 7: Segmentation accuracy and comparison with the object segment proposals of Endres and Hoiem (2010).

The proposed method shows encouraging object segmentation accuracy and robustness in finding the dominant objects in a 3D scene. Please note that we compare our segments with the best proposals in (Endres and Hoiem, 2010), which are selected from hundreds of unsorted proposals. I.e., our method shows better precision as well. The main source of error are anomalous initial patches from the superpixel segmentation, e.g., patches that already link two objects and curved threadlike patches which are hard to assemble correctly.

Please note that although we have enriched the pre-segmentation of the color image with normal information, cross-object segments can still not be avoided or perfectly planar patches be guaranteed. We conduct RANdom SAmple Consensus (RANSAC) (Fischler and Bolles, 1981) for each patch to estimate more reasonable plane parameters in spite of outliers, which improves the results. Still, the strangely shaped patches, e.g., curved bars, sparse patches (along the viewing direction) and those that cross two objects, lead to errors. One way to reduce this effect would be to segment the image into smaller patches. This, however, will require much more computational effort while it cannot totally avoid false segments.

6. CONCLUSIONS

We have proposed a novel method for object-level segmentation of RGBD data. The synthetic volume primitive – SVP is introduced to parse the 3D geometrical relationships between the pre-segmented data patches. The proposed method demonstrates its potential in finding the dominant objects in indoor scenes including the walls and the floor without using domain knowledge of specific object classes.

The main contributions of this paper can be summarized as follows:

1. The introduction of SVP: A hypothetical quasi-3D model;
2. A novel segmentation scheme for freeform objects based on assembling SVPs;
3. A global optimization integrating both 3D spatial and color consistency constraints.

One possible future work is to improve the pre-segmentation to avoid trivial patches and patches across objects, which are now the main cause of errors in the results. Furthermore, the SVP model could be improved by an adaptable context-sensitive extrusion length instead of a fixed one, which will help the SVP to connect more remote components or avoid wrong ones.

ACKNOWLEDGEMENTS

This project has been partially supported by the United States National Science Foundation (NSF) funding #1018641.

References

Bleyer, M., Rhemann, C. and Rother, C., 2012. Extracting 3d scene-consistent object proposals and depth from stereo images. In: Proceedings of the 12th European conference on Computer Vision - Volume Part V, ECCV'12, Springer-Verlag, Berlin, Heidelberg, pp. 467–481.

Endres, I. and Hoiem, D., 2010. Category independent object proposals. In: Proceedings of the 11th European conference on Computer vision: Part V, ECCV'10, Springer-Verlag, Berlin, Heidelberg, pp. 575–588.

Felzenszwalb, P. F. and Huttenlocher, D. P., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59/2, pp. 167–181.

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6), pp. 381–395.

Huang, H., Brenner, C. and Sester, M., 2013. A generative statistical approach to automatic 3d building roof reconstruction from laser scanning data. ISPRS Journal of Photogrammetry and Remote Sensing 79(0), pp. 29 – 43.

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

Lafarge, F. and Mallet, C., 2012. Creating large-scale city models from 3d-point clouds: A robust approach with hybrid representation. International Journal of Computer Vision 99, pp. 69–85.

Li, Y., Wu, X., Chrysathou, Y., Sharf, A., Cohen-Or, D. and Mitra, N. J., 2011. Globfit: consistently fitting primitives by discovering global relations. ACM Transactions on Graphics 30(4), pp. 52:1–52:12.

Rabbani, T., Dijkman, S., Van den Heuvel, F. and Vosselman, G., 2007. An integrated approach for modelling and global registration of point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 61(6), pp. 355 – 370.

Silberman, N. and Fergus, R., 2011. Indoor scene segmentation using a structured light sensor. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 601 –608.

Silberman, N., Hoiem, D., Kohli, P. and Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. In: Proceedings of the 12th European conference on Computer Vision.

Vosselman, G., 2009. Advanced point cloud processing. In: D. Fritsch (ed.), Photogrammetric Week '09, Heidelberg, Germany, pp. 137–146.