Attention-based Active 3D Point Cloud Segmentation

Matthew Johnson-Roberson and Jeannette Bohg and Mårten Björkman and Danica Kragic

Abstract— **In this paper we present a framework for the segmentation of multiple objects from a 3D point cloud. We extend traditional image segmentation techniques into a full 3D representation. The proposed technique relies on a state-of-the-art min-cut framework to perform a fully 3D global multi-class labeling in a principled manner. Thereby, we extend our previous work in which a single object was actively segmented from the background. We also examine several seeding methods to bootstrap the graphical model-based energy minimization and these methods are compared over challenging scenes. All results are generated on real-world data gathered with an active vision robotic head. We present quantitive results over aggregate sets as well as visual results on specific examples.**

I. Introduction

The state of the art in active vision systems is constantly advancing and giving robots a greater capability for understanding their environment. The active detection of objects is crucial to interaction and manipulation tasks. The work presented here focuses on the problem of segmenting previously unseen objects. In many scenarios, an autonomous system is required to act upon new objects in new environments. In contrast to known objects the segmentation of previously unseen objects cannot rely on preexisting shape and appearance models.

The work presented in this paper poses object understanding as a fully 3D global multi-class segmentation problem. This paradigm allows for the identification of multiple objects from multiple views. The proposed techinque is performed using a state-of-the-art graphical min-cut framework [?].

We begin by gathering a point cloud from stereo vision. The point cloud data for this paper has been generated using an active humanoid head that incrementally builds a scene representation integrating range measurements from stereoscopic cameras. We extend our previous work [?], by labeling multiple objects simultaneously, in a full 3D point cloud, performing a global segmentation. We remove the planar assumptions made in that work where objects are assumed to be on flat surfaces. Also the transition from a disparity representation to fully 3D representation allows for a more principled segmentation of objects occluded in one view.

The paper is organized as follows: The remainder of this section discusses the experimental platform. Section II covers the previous work on the subject. Section III presents the 3D representation and framework for segmentation. Section IV explains the steps in the segmentation procedure. Section V presents the results and experimental evaluation and finally Section VI concludes and discusses future work.

II. Previous Work

The discussion of previous work is divided into three categories: traditional segmentation of point clouds, traditional image segmentation, and random field based methods.

A. Point Cloud Segmentation

Point cloud segmentation is a field of ongoing research. Vosselman [?] distinguishes between two categories of technique. First are methods that attempt to segment based upon properties such as surface normals and color similarity. The second being those techniques that attempt to directly estimate parametric surfaces by clustering the data in parameter space.

Techniques of the first class include scan-line segmentation [?], where regions are split based upon distance criteria. Another example is region growing methods which are direct 3D extensions of the 2D techniques.

Differently, techniques that cluster in parameter space make use of the 3D Hough transform [?]. Similar parameterization techniques for cylinders and spheres exists but generally these techniques do not extend well to arbitrary objects, and as such are not applicable to the work presented here.

B. Image Segmentation

Image segmentation is an extensive field of research and the background is beyond the scope of this paper. However two techniques are directly relevant to this work: Normalized Cuts [?] and Grabcut [?] image segmentation. These techniques are applications of graph cut based global optimization for foreground background segmentation. We build upon these techniques extending them to three dimensions and allowing for multiple object classes.

C. Previous Random Field Point Cloud Work

Several approaches have been proposed for using random fields for point cloud segmentation. Early work by Anguelov et al. [?] presented an Associative Markov Network that segmented point clouds based on simple geometric features. Later Triebel et. al extended this work adding an instance-based classifier [?]. Munoz et. al developed a Conditional Random Field (CRF) based method that relies on strictly geometric feature such as spectral scatter, local tangent, and local normal [?]. Additionally Rusu et. al proposed

This work was supported by EU through the project PACO-PLUS, IST-FP6-IP-027657, and GRASP, IST-FP7-IP-215821 and Swedish Foundation for Strategic Research. The authors are with the Centre for Autonomous Systems and Computational Vision and active Perception Lab, CSC-KTH, Stockholm, Sweden. mattjr,bohg,celle,dani@kth.se

the Fast Point Feature Histogram as the input to a similar framework [?]. All make no use of color and are trained point cloud classifiers as opposed to the work presented here, untrained segmentation. Lim & Suter propose another CRF based classifier including color but operate over 'super-voxels' instead of the complete cloud and again require a training phase [?].

Markov Random Field (MRF) segmentation techniques have previously been applied to the point cloud segmentation problem. Golovinskiy & Funkhouser [?] and Sedlacek & Zara [?] both propose MRF based binary segmentations that are closely related to this work, but do not provide multi-class segmentation and do not utilize color information. Quan et. al proposed a color based 3D algorithm but again restricted results to single class segmentations and relied heavily on user input [?].

III. Segmentation Framework

Building upon previous work we develop an MRF based labeling approach to solve the multi-class segmentation problem.

A. Graphical Model

MRFs are graphical models that provide a framework for labeling problems. This paper utilizes an MRF formulation that allows for a multi-class labeling of a color point cloud.

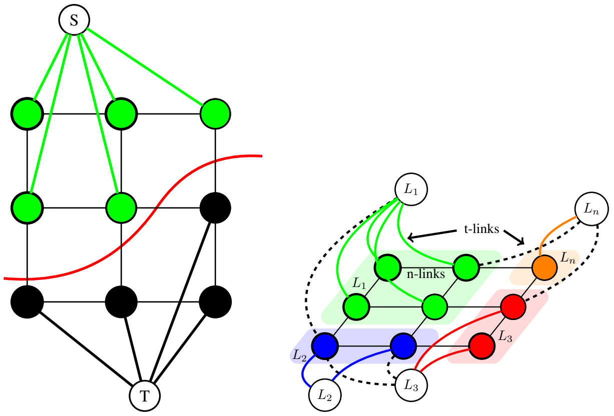
Let us define $G = (V, E)$ to be a graph with nodes $V = (V_1, ..., V_n)$ and edges $E = (E_{i,j} \mid i, j \in V)$ where $E_{i,j}$ is a pairwise relationship between node i and node j. The full formulation can be found in Boykov et. al [?]. In this paper we describe the energy function as the sum of ϕ the unary potential function and ψ the pairwise potential function. The two energies represent the two types of edges in the graph: t-links that denote terminal-links and n-links that denote neighborhood connections between vertices. The n-link energy encourages coherence in regions of property consistency.

B. Multi-class segmentation

In the multi-way cut case as proposed by Boykov et. al [?] additional t-links are created. They span between each node in V and n terminals, one for each label in $L = \{L_1, L_2, ..., L_n\}$. The cut process is now attempting to split the graph into n subsets that contain only one t-link between each member of V and a single terminal. Illustrations of the binary and expanded multiway cut are shown in Figure 1.

C. Optimization

Determining the minimization to the energy function for a multi-way graph cuts formulation is a well-studied problem in computer vision. The α-expansion algorithm with available implementation [?], [?], [?] efficiently computes an approximate solution that approaches the NP-hard global solution. This allows us to optimize the proposed energy function using existing techniques.



(a) Binary Seg (b) Multi-way Cut

Fig. 1. Illustration of traditional binary segmentation cut and multiway cut that allows for multi-class segmentation. In (a) foreground and background are segmented through connection of t-links to a foreground and background node. In (b) a lattice of nodes is connected to the label set $\mathscr{L} = L_1, L_2, L_3, ..., L_n$ only four labels are shown and many initial t-links are omitted for clarity's sake.

IV. Segmentation Procedure

The proposed three-dimensional segmentation technique makes direct use of the input point cloud to perform the segmentation. To do this, points in the point cloud specify the vertices of V in the graphical framework. The process to generate a labeling is as follows:

- The edge relationships in E are computed using a nearest neighbor calculation based upon a kd-tree. The techniques locates the nearest four vertices with respect to the current point. A typical point cloud and associated neighborhood links can be seen in Figure 2.
- An initial seed point for each possible object is generated using one of the three methods described in section IV-A. A small region around this seed point is used to initialize a color model for the object.
- The color models as discussed in Section IV-B are used to determine the unary weights of the points in the scene.
- Using the edge relationships in E the pair-wise weights are calculated as described in Section IV-C.
- An iterative energy minimization is performed as described in Section IV-D.

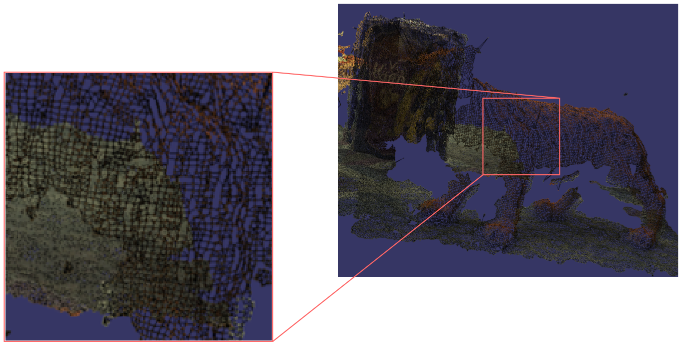


Fig. 2. A sample point cloud and associated links. The inset displays the neighborhood relationship between points. The black lines connect neighboring points. These relationships map directly into the graphical representation discussed in Section III.

A. Segmentation Seeding

As with many segmentation techniques based upon energy minimization, an initial coarse segmentation is required to bootstrap the graph cuts. Here we propose two methods for seeding the segmentation and obtain human ground truth seeds for comparison. We discuss the tradeoffs between the techniques and later in Section V we will present a comparison of the performance of the seeding methods on data gathered with our experimental setup.

1) Geometric Plane Seeding: The first technique is geometric technique that is only applicable to table top environments. It is important to note, only this seeding relies on a planar assumption, the segmentation technique proposed in this paper is independent of any such restriction. The method assumes a table plane supporting a set of objects.The method is as follows:

- The input point cloud is down-sampled and cleaned based upon an occupancy thresholding per voxel in a uniform subdivision of the space.
- The normal vectors for each point are found using a neighborhood of surrounding points computed with a kd-tree.
- The points with normals aligned, within some threshold, to the gravity vector are clustered and a RANSAC plane fit is performed on each cluster [**?**]. The largest cluster, lowest in the scene (assuming the floor height is known), is selected as the table. Additionally any clusters within a threshold of this height are also assumed to be part of the table.
- The points higher than the estimated table plane are clustered based upon color, normal, and distance. These clusters provide the initial segmentation for the proposed energy minimization technique.

An example of an estimated plane fit can be seen in Figure 3.

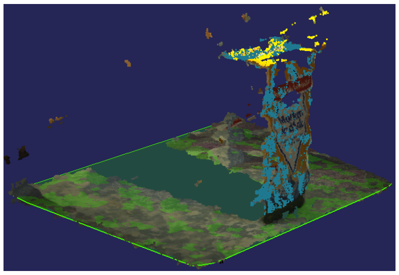
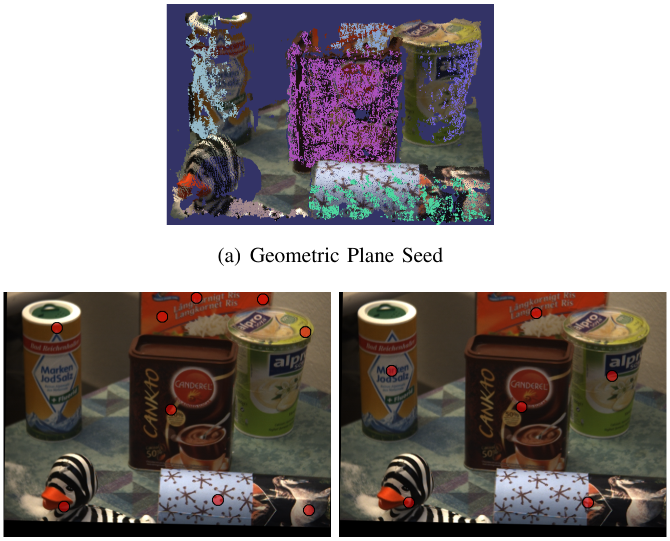


Fig. 3. Example of plane fit on point cloud to provide a seeding for segmentation. The green transparent region is the estimated plane. The blue points are the initial seeding points clustered based on color, normal, and distance.

2) Image Saliency Seeding: The active humanoid head used in these experiments has two sets of cameras. The peripheral view provides a complete view of the tabletop. Using this view, the image saliency techniques proposed by Rasolzadeh et. al [**?**] were applied to generate seed point for the proposed segmentation technique. The salient points provide a set of hypotheses that we project into the point cloud to begin the segmentation process. One limitation of this technique is that it requires a view of the complete scene. This could potentially be stitched together from multiple views, or as in this scenario, gathered from a wide field of view camera. More recent techniques such as Context-Aware Saliency Detection offer the possibility of greater region separation through the exploitation of multiple cues [**?**].



(a) Geometric Plane Seed

(b) Saliency Seed

(c) Human Seed

Fig. 4. Depiction of the seeding of the segmentation based upon the three techniques. (a) the geometric technique. (b) the image saliency based technique. (c) the human selected seed points.

3) Human Selection: Finally to provide a reference, seed points are obtained from a human operator. This very closely mirrors the techniques used in interactive segmentation such as those proposed by Quan et al. [**?**] and Normalized Cuts [**?**], where a user selects the initial seed points. These human generated points allow us to understand the effect of seeding on the final segmentation results. A sample seeding for the three techniques appears in Figure 4. The results of comparing the techniques appears later in Section V-C.

B. Unary Weighting: Color Modeling

Once seed points have been generated it is necessary to create models for the hypothesized objects. For modeling the color properties of an object hypothesis, we adopt Gaussian Mixture Models (GMMs) as utilized in GrabCuts [**?**]. Let us define GMM_p to be the Gaussian Mixture Model for label p. For each of K components in GMM_p we learn μ_i the mean rgb value, $\sum_i^{-1}$ the inverse covariance matrix, and π_i the component weight from the seed point. Then the unary cost $\phi_p(x_n)$ for point n taking label p, is determined by the likelihood of the color c_n of that point belonging to GMM_p. This can be computed for a point n with respect to a seed point p as follows [**?**]:

$$\phi_p(x_n) = -\log \sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{\det \sum_i}} e^{\left(-\frac{1}{2}(c_n-\mu_i)^\top \sum_i^{-1}(c_n-\mu_i)\right)} \qquad (1)$$

This value $\phi_p(x_n)$ is used as the unary weight for the link between n and t_p in the graphical representation.

C. Pairwise Weighting

We use the pairwise function defined in GrabCuts [**?**] modified for point distances where $\psi(x_i, x_j)$ is defined as:

$$\psi(x_i, x_j) = \frac{1}{dist(i,j)}[x_i \neq x_j]e^{(-\beta \|c_i - c_j\|^2)} \tag{2}$$

where $dist$ is the Euclidean distance between vertices i and j, and $[x]$ denotes the function returning 1 if the statement x is true and 0 if false. A full discussion of the calulation of β and its implications can be found in Talbot and Xu [**?**].

One important note is that hard links or vertices that must stay with their initial label are absent from this formulation as we allow for the changing of all labels. This provides greater recovery from poor initial hypotheses at the cost of less topdown control over the resulting segmentation.

D. Iterative energy minimization

For brevity's sake the full procedure for iterative GMM learning will not be described here. Please refer to Rother et al. [**?**] for greater detail on the GMM learning process. The important extension here is the multiple hypothesis GMM initialization.

- The procedure begins by assigning the components of GMM$_p$ to the points currently labeled p (if this is the first iteration then it is simply the seed point). This is done for all labels/seed points.
- Each GMM learns a model from the set of points assigned to it in the previous step as described in [**?**].
- Finally the costs for the graph are assigned and the energy minimization is performed as discussed in Section III-C.
- The process is repeated until convergence which is determined when the energy change between iterations falls below a threshold.

V. Experimental Results

To validate the technique we present a series of experimental results designed to display the performance of the technique on scenes with human ground truth. Complex scenes were chosen to stress the importance of multiple hypotheses in the presence of clutter. Six sets were constructed. Each set consisted of five views of the same object centered in the field of view, with a rearrangement of the background and foreground objects. This design allowed for the direct comparison of a single object segmentation with the proposed technique. The thirty views were labeled by hand specifying all the points belonging to each object in the scene.

A. F_1 Single Object Comparison

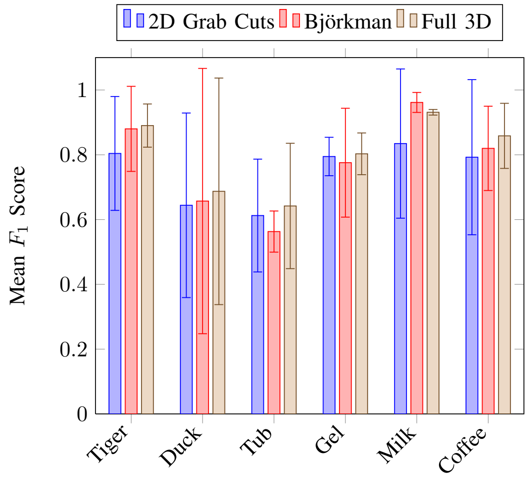
The first experiment performed was to compare the performance of the proposed technique to that of our previous work [**?**]. To run the simplest comparison, both algorithms were seeded by hand with the same single initial point and run until convergence. To benchmark the classification performance of the techniques the F_1 score also know as the F-measure (the harmonic mean of precision and recall) was selected. The F_1 score is a scalar encapsulation of both precision and recall. The results on the set of thirty views appear in Figure 5(a). The classes along the bottom represent the centered object and the mean F_1 score is an average over the five views centered on that object, error bars reflect the standard deviation. The scenes were quite challenging, see Figures 5(d)-5(i), the centered object was partially occluded in many cases and cluttered by other foreground and background objects. The comparison is being performed on a single view and so a traditional 2D image segmentation can also be applied. 2D grab cuts [**?**] was selected to provide a baseline. The proposed technique performs similarly to the other techniques in all cases providing a slight improvement in almost all views.

B. Multi-Object Hypotheses

The strength of the proposed algorithm is the ability to segment multiple objects simultaneously. The result of the complete algorithm is a multi-class classification. Two examples of confusion matrices for the full multi-class segmentation appear in Tables 5(b) & 5(c) as well as visual examples in Figure 6. These tables show examples of typical performance, however to validate the increase in performance multiple object segmentation affords, another quantitive comparison was done. In this comparison the proposed technique and Björkman's algorithm were seeded with the same points, but the proposed technique attempted to segment out all objects. This allowed the labeling energies for all objects, including occluding and neighboring ones, to propagate across the neighborhood links described in Section III. The result of this comparison is shown in Figure 7(a). Five views of the same centered object were used and the proposed technique was compared to [**?**], the accuracy reflects only the segmentation of the centered object as the older technique provides no segmentation for additional objects. The proposed technique is first run with a single object seed then allowed to use all seeds to perform multi-object segmentation. The performance demonstrates the advantage of using multiple object segmentation even when interested in only the centered object. The strength becomes apparent the more difficult the scene. Run three and four proved the greatest challenge for all techniques with the largest amounts of occlusion and objects very near to one another. The complexity of the scenes can be observed in Figures 7(c)-7(g). The most challenging runs showed the greatest improvement when performing the multi-object segmentation.

C. Seed Comparison Results

To understand the sensitivity of the proposed technique to its initial seeding, a benchmark was performed on a large cluttered scene with five views and human ground truth. The scene was seeded with the three proposed techniques and the segmentation was performed. The results of this comparison is displayed in Figure 7(b). It should be noted that the human seeding performs best, as expected, but the



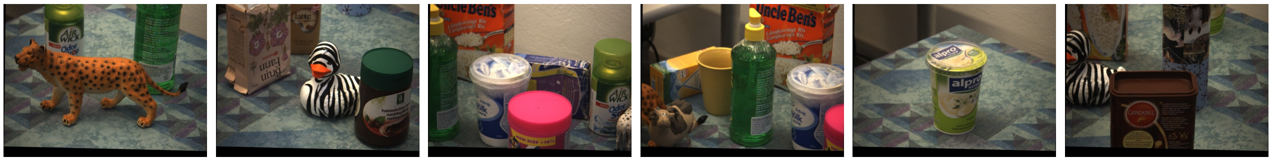
(a) Comparison of Single Label Results Across Five Views of Six Objects

		Actual					
		Bkrd.	Obj 1	Obj 2	Obj 3	Obj 4	Obj 5
Predicted	Bkrd.	**33749**	304	520	170	121	1
	Obj 1	132	**18157**	1	0	0	3
	Obj 2	1288	506	**10614**	10	0	1
	Obj 3	106	80	20	**5880**	0	0
	Obj 4	43	80	86	0	**12452**	0
	Obj 5	92	138	0	0	0	**3562**
	Acc:	95.3%	94.2%	94.4%	97.0%	99.0%	99.9%

(b) Full Multi-Class Confusion Matrix Example 1

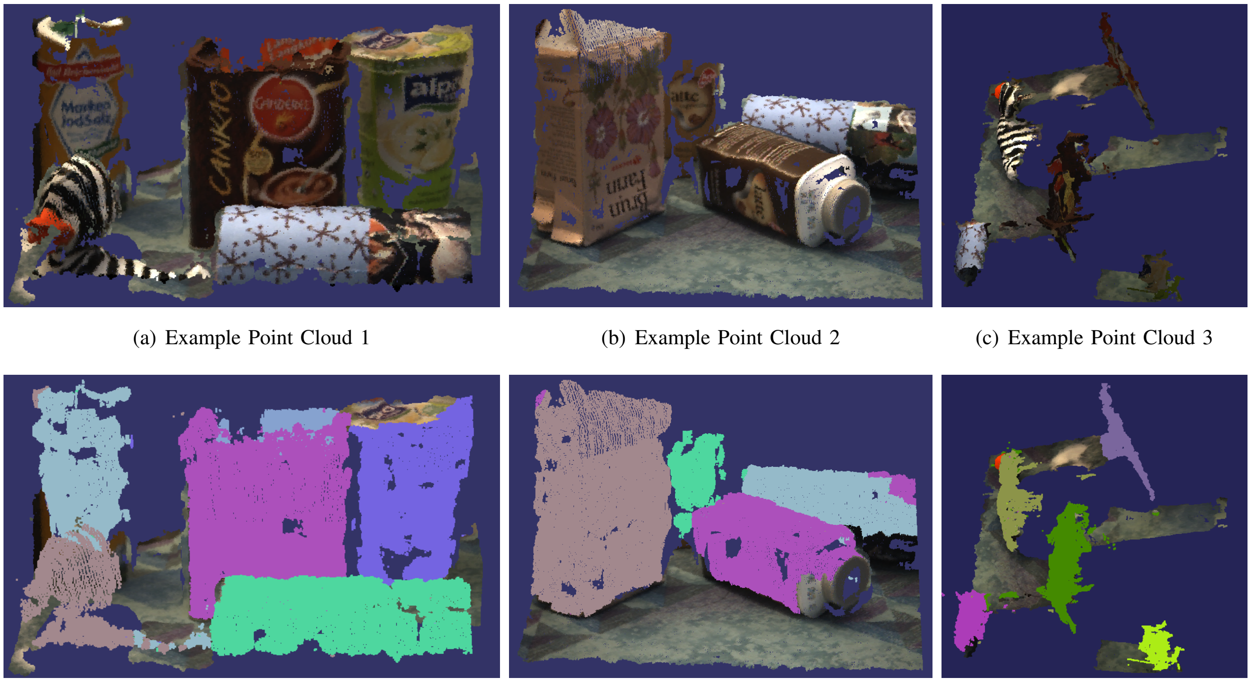
		Actual					
		Bkrd.	Obj 1	Obj 2	Obj 3	Obj 4	Obj 5
Predicted	Bkrd.	**65745**	390	6005	245	0	23
	Obj 1	297	**45064**	1	0	350	6
	Obj 2	1605	1425	**29055**	45	7	37
	Obj 3	394	67	54	**15989**	3	0
	Obj 4	14659	216	121	0	**14547**	0
	Obj 5	115	395	0	0	0	**6700**
	Acc:	79.4%	94.8%	82.5%	98.2%	97.6%	99.0%

(c) Full Multi-Class Confusion Matrix Example 2



(d) 1 of 5 Tiger Scenes (e) 1 of 5 Duck Scenes (f) 1 of 5 Tub Scenes (g) 1 of 5 Gel Scenes (h) 1 of 5 Milk Scenes (i) 1 of 5 Coffee Scenes

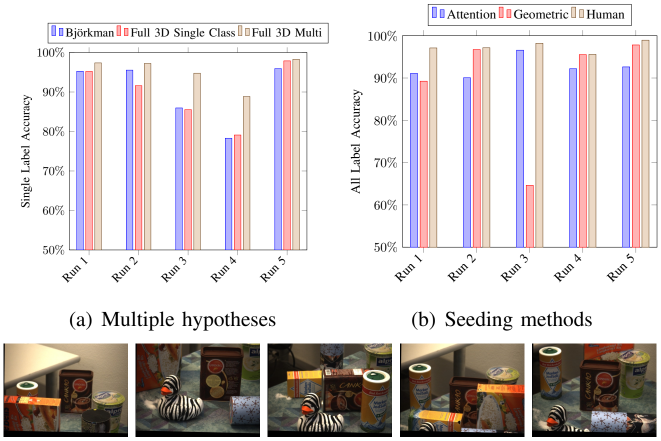
Fig. 5. This figure presents the aggregated result of running the proposed algorithm against our previous work and traditional 2D grabcuts. The tests were run over six object, each appearing in five views. The scenes had significant clutter and varying foreground and background objects. The graph in (a) presents the mean F_1 scores (the harmonic mean of precision and recall), for the binary classification of the centered object from the background across the five views. The standard deviation appears as the errors bars on the graph. This was the most direct comparison between techniques. However it does not take full advantage of the multi-class framework. True multi-class results are shown in (b) & (c) where the confusion matrices show the mislabeling of each point across all labels. To give a sense of the complexity of the scenes, one of the five views for each objects is displayed in (d)-(i).



(a) Example Point Cloud 1 (b) Example Point Cloud 2 (c) Example Point Cloud 3

(d) Example Segmentation 1 (e) Example Segmentation 2 (f) Example Segmentation 3

Fig. 6. Visual examples of typical multi-class segmentations. The input point clouds can be seen in the top row (a)-(c). The bottom row (d)-(f) contains the segmentations with each color reflecting a different label. Large amounts of occlusion and similarly colored objects were used. The top down view (c) highlights the 3D nature of the scenes.



(a) Multiple hypotheses

(b) Seeding methods

(c) Run 1

(d) Run 2

(e) Run 3

(f) Run 4

(g) Run 5

Fig. 7. In depth results for the multi-class segmentation. Our previous work is compared with the proposed technique: run as a binary classifier and in full multi-class mode in (a). The accuracy improves in every case when the full 3D multi-class segmentation is run. The seeding method is compared in (b). The setup for all other scenes in this trials is shown in (c)-(g)

other technique perform closely. One exception is run three (particularly cluttered) where the geometric method seeds two close objects as one and the proposed technique cannot separate them resulting in the complete mislabeling of one object and consequently a poor accuracy score.

VI. Conclusion and Future Work

We have presented a novel point cloud segmentation technique that proposed multiple seeding methods and a global multi-object approach. The technique harnesses the power of a graphical MRF framework to provide a principled approach to multi-segment labeling. We have validated the results on ground truth data and explored the sensitivity of the results to the multiple seeding techniques.

Future direction could explore the improvement of region seeding in the complex environments. New saliency technique offer promise with respect to region separation [**?**]. Additionally more complex connectivity between the graph nodes is an area of open research in the computer vision community as well as improved seed selection and affinity matrix calulation [**?**]. Exploration down these avenues could provide great improvements over current resutls. Additionally exploring other less traditional ways of providing the seeding is an interesting avenue of future research. Human robotic interaction techniques that would allow an operator to flexibly specify seeds would enable greater functionality in real domestic scenarios.