Attention-driven Object Detection and Segmentation of Cluttered Table Scenes using 2.5D Symmetry

Ekaterina Potapova, Karthik M. Varadarajan, Andreas Richtsfeld, Michael Zillich and Markus Vincze
Automation and Control Institute
Vienna University of Technology
1040 Vienna, Austria
{potapova,varadarajan,ari,zillich,vincze}@acin.tuwien.ac.at

***Abstract*— The task of searching and grasping objects in cluttered scenes, typical of robotic applications in domestic environments requires fast object detection and segmentation. Attentional mechanisms provide a means to detect and prioritize processing of objects of interest. In this work, we combine a saliency operator based on symmetry with a segmentation method based on clustering locally planar surface patches, both operating on 2.5D point clouds (RGB-D images) as input data to yield a novel approach to table-top scene segmentation. Evaluation on indoor table-top scenes containing man-made objects clustered in piles and dumped in a box show that our approach to selection of attention points significantly improves performance of state-of-the-art attention-based segmentation methods.**

I. Introduction

Segmentation of objects from a static scene is a crucial step in many robotic tasks. Different approaches have been proposed to tackle the object segmentation problem, which can be broadly classified into two groups: *discriminative* and *agglomerative*. Discriminative segmentation algorithms tend to classify the whole scene at once and assign a label to every pixel [1], [2], [3]. Agglomerative segmentation algorithms grow regions from a seed point to segment the foreground object. Active segmentation or attention-driven segmentation are agglomerative methods that segment images starting from a fixation point or region [4], [5].

Segmentation in cluttered scenes is a critical module in robotics, with the need to find task relevant objects quickly amongst a possibly large number of distractors.

In this paper, we present a novel method for attention-driven segmentation for cluttered table scenes. The contribution of this paper is two-fold: first, we employ a novel object detection and selection algorithm based on attention points from 2.5D symmetry saliency maps first presented in [6]. Secondly, we introduce a segmentation procedure based on clustering of planar surface patches using color similarity and a notion of compactness. We evaluate our approach on two databases consisting of different types of table scenes ranging from simple to complex scenarios. We show that the proposed approach works better than existing approaches for attention-driven segmentation (Fig. 1).

*The research leading to these results has received funding from the Austrian Science Fund (FWF) under grant agreement No. TRP 139-N23 InSitu and from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement No. 600623, STRANDS.

(a) Mishra *et al.* [4] (b) Proposed approach

Fig. 1: *An example of active segmentation for a cluttered table scene.* Fixation points are shown in black with numbering reflecting the order of attention shift. As can be seen, the approach of Mishra *et al.* (a) does not segment all attended objects properly, while the proposed approach (b) successfully deals with the scene complexity.

The paper is organized as follows: In Section II, we review related work. Section III and IV describe the proposed algorithm in detail. The evaluation in Section V shows the benefits of our algorithm. Section VI concludes the paper with a discussion about future directions of research.

II. Related Work

The focus of this paper is segmentation of indoor table scenes typical of robotic task environments. Therefore, we primarily concentrate on the work developed for RGB-D data. A number of discriminative segmentation algorithms use depth information to boost segmentation performance for complex indoor scenarios. Such algorithms include those proposed in [7], [8], [9], [10], [11], [12].

The concept of active segmentation or attention-driven segmentation was first presented by Aloimonos *et al.* [13]. It was argued that the human visual system investigates and observes the scene by a set of fixations that are followed by segmentation. Attention-driven segmentation usually has two stages. During the first stage, a selection mechanism detects candidate object locations. During the second stage, the detected objects are segmented. The attention-driven segmentation approaches in [5], [14], [4] propose different solutions

Preprint submitted to 2014 IEEE International Conference on Robotics and Automation. Received September 5, 2013.

for the two stages. Given that the selection mechanism is directed by a specific search task [15] in the case of robotic applications, attention-driven segmentation finds greater use than discriminative segmentation approaches.

In [14], Mishra *et al.* proposed a framework for active segmentation, where an object is segmented using object boundaries, given a fixation point on the object. Though originally in the paper no strategy for detection of fixation points was proposed, it was discussed that visual attention mechanism [16], [17], [18] can be used for such a selection. Later, [4] extended the active segmentation approach to use the concept of "simple objects and border ownership", which is defined using depth, color and/or motion information about the scene. A new strategy for the calculation of fixation points was also proposed. Kootstra *et al.* [5] proposed an attention-driven graph-cut segmentation. Objects are localized with fixation points extracted from 2D symmetry saliency maps. An energy minimization function is applied to depth and color along with a support plane constraint for segmentation. It is worth mentioning that both the above segmentation algorithms [5], [4] were developed specifically for table top scenes and require information about the support plane. However, both approaches fail to segment objects if the scene consists of multiple occluded and cluttered objects, having several colors and textures. These types of scenes are common in domestic robotic tasks and are needed to be resolved correctly to enable manipulation of objects.

As mentioned earlier, the problem of fixation points selection stays open with numerous solutions. Selecting fixations as attention points of the saliency map is one of the widely used approaches [5], [6]. Potapova *et al.* [6] proposed the use of 2.5D symmetry based saliency maps to extract fixation points for segmentation. It was shown that 2.5D symmetry-based saliency maps capture the properties of the scene better than 2D based saliency maps.

In this paper, we adopt the idea of Potapova et al. [6] for object detection and extend it with a novel attention points selection algorithm. Furthermore, a novel segmentation algorithm is introduced, using the fixation points to enable clustering of planar surface patches, similar to [8], using color similarity and the notion of compactness.

III. Object Detection

In this section, we describe the detection of good object candidate locations in a cluttered scene. Einhauser *et al.* [19] showed that objects attract human attention better than early vision saliencz features. Symmetry is one of the characteristics of many natural as well as human-made objects and at the same time a powerful attentional cue [20]. Therefore, we based our object detection strategy on the calculation of a 2.5D reflective symmetry-based saliency map.

A. Saliency Map from 2.5D Symmetries

We follow the algorithm in [6] to generate the reflective symmetry based saliency map, starting from a 2.5D point cloud, i.e. a rectangular array of depth values. Each point p in the point cloud P is indexed by image coordinates (i, j), has color (r, g, b), and is characterized by a set of values $(x, y, z, \mathbf{n})$, where (x, y, z) are spatial coordinates, and $\mathbf{n}$ is the estimated surface normal at that point. Normals are calculated by locally fitting planes to neighboring points.

Following ideas by Minovic *et al.* [21] and Sun *et al.* [22], Potapova *et al.* [6] proposed to estimate the local amount of symmetry $s(p)$ at point p on the neighborhood $N(p)$. Minovic *et al.* [21] showed that planes of reflective symmetries are perpendicular to the directions of the object's principal axes. The principle axes of a 3D model can be detected from the Extended Gaussian Image (EGI) created from point normals as was proposed by Sun *et al.* [22].

In our scenario the EGI for point p is created from normals of the points in the neighborhood $N(p)$, which is a 10×10 pixel window around p. The principal axes $\{u_1, u_2, u_3\}$ of the local surface in the neighborhood $N(p)$ are estimated using Principal Component Analysis (PCA) on the EGI. The corresponding reflective symmetry planes $\{\pi_1, \pi_2, \pi_3\}$ are planes going through the point p perpendicular to the respective principal axes.

For a given reflective plane π_i $(i = 1, 2, 3)$ the neighborhood $N(p)$ is divided into two neighborhoods $N_{i1}(p)$ and $N_{i2}(p)$, so that $\forall p' \in N(p)$:

$$p' \in \begin{cases} (N_{i1}(p)) & \text{if } d(p', \pi_i) > 0 \\ (N_{i2}(p)) & \text{if } d(p', \pi_i) < 0 \end{cases} \quad (1)$$

where $d(p', \pi_i)$ is the signed Euclidean distance from point p' to the plane π_i.

For each neighborhood N_{ij} $(j = 1, 2)$ the mean point $\overline{p}_{ij}$ and the mean normal $\overline{\mathbf{n}}_{ij}$ are calculated.

The amount of local reflective symmetry for the point p over the reflective plane π_i is then given by:

$$s(p) = \max_{i=1,2,3} \{\Omega_i(N(p), p)\} \quad (2)$$

$$\Omega_i(N(p), p) = e^{-\triangle z_i} e^{-\triangle d_i} \omega_1 \omega_2 \quad (3)$$

$\triangle d_i$ represents the absolute difference between distances from mean points $\overline{p}_{i1}$ and $\overline{p}_{i2}$ to the reflective plane π_i:

$$\triangle d_i = |\, |d(\overline{p}_{i1}, \pi_i)| - |d(\overline{p}_{i2}, \pi_i)|\,| \quad (4)$$

where $d(\overline{p}_{ij}, \pi_i)$ is the distance from mean point $\overline{p}_{ij}$ to the plane π_i. $\triangle d_i$ reflects the fact that we are searching for points where the parts of the neighborhood left and right of the symmetry plane are positioned symmetrically.

$\triangle z_i$ in eq. 3 represents the absolute difference between z-coordinates (depth values) of mean points $\overline{p}_{i1}$ and $\overline{p}_{i2}$ and therefore favors symmetries facing the view point

$$\triangle z_i = \left| z_{\overline{p}_{i1}} - z_{\overline{p}_{i2}} \right| \quad (5)$$

ω_1 in eq. 3 measures the co-planarity between the line $\mathbf{l}_i$ connecting $\overline{p}_{i1}$ and $\overline{p}_{i2}$ and the two mean normals $\overline{\mathbf{n}}_{i1}$ and $\overline{\mathbf{n}}_{i2}$:

$$\omega_1 = \left|\left| \frac{\overline{\mathbf{n}}_{i1} \times \overline{\mathbf{n}}_{i2}}{||\overline{\mathbf{n}}_{i1} \times \overline{\mathbf{n}}_{i2}||} \times \mathbf{l}_i \right|\right| \quad (6)$$

$$\mathbf{l}_i = \frac{\overline{p}_{i1} - \overline{p}_{i2}}{||\overline{p}_{i1} - \overline{p}_{i1}||} \quad (7)$$

Preprint submitted to 2014 IEEE International Conference
on Robotics and Automation. Received September 5, 2013.

(a) Original image (b) Saliency Map (c) Attention points (d) Surface patches (e) Segmentation result

Fig. 2: *Object detection process:* starting from original image (a), saliency from 2.5D symmetries is calculated (b) (shown in green superimposed on the original image); (c) shows attention points from symmetry (red) and skeletal line segments (green) (please note that for visualization purposes both attention points and skeletons were dilated); (d) shows planar surface patches and (e) shows segmentation result with respective attention points.

ω_2 in eq. 3 measures the similarity between mean normal directions based on the symmetry operator from Reisfeld *et al.* [23] and is calculated as follows:

$$\omega_2 = (1 - \cos(\alpha_1 + \alpha_2))(1 - \cos(\alpha_1 - \alpha_2)) \quad (8)$$

where α_j is the angle between mean normal $\overline{\mathbf{n}}_{ij}$ and $\mathbf{l}_i$. This term is largest in regions, where the normals are oriented completely opposite to each other and smallest in regions, where normals have the same orientation (*e. g.* flat surfaces).

We take the product of all four terms in eq. 3, since we are searching for local symmetries with neighborhoods that produce high values for all four quantities.

B. Multi-Scale Symmetry-Based Saliency Map

The above symmetry is defined locally over a neighborhood around a given point. To capture symmetries at different scales we calculate saliency maps on a Gaussian pyramid of depth images and then sum them using across scale addition [17]:

$$S = \bigoplus_{l=L_1}^{L_n} s_l \quad (9)$$

where L_1 and L_n are the finest and the coarsest scales of the pyramid. In our experiments we used four levels of the pyramid. Finally, saliency map S is normalized to the range $[0, 1]$. Figure 2b shows an example of saliency map from 2.5D symmetries.

C. Attention Points from Symmetry

From the multi-scale symmetry-based saliency map S we extract 8-neighbor connected components of pixels with saliency value bigger than zero $\{C_k\}$. The average saliency $\overline{S}_k$ of each connected component C_k is computed as

$$\overline{S}_k = \frac{1}{n_k} \sum_{p \in C_k} S(p) \quad (10)$$

where n_k is the number of pixels in the connected component, $S(p)$ is the saliency value of the point p.

The connected component C_k is considered to be valid only if $\overline{S}_k > \theta_{sal}$. In our experiments we set θ_{sal} to 10% of the maximum saliency value.

The skeleton T_k is extracted from the connected component C_k. Symmetry attention points $\{f_k\}$ are extracted from the skeleton T_k as junction points, if they exist, or as midpoints for simple skeletal line segments. Figure 2c shows examples of attention points $\{f_k\}$ and skeletons T_k.

IV. Object Segmentation

Given attention points we want to segment the scene incrementally. We first cluster points into planar patches based on their normals similar to Richtsfeld *et al.* [8]. We then cluster these patches beginning from the attention points by connecting similar patches as long as a given object-ness measure is valid.

A. Clustering Normals

Neighboring points are clustered to uniform patches without discontinuities using point normals. Normal clustering starts at the point with lowest curvature and greedily assigns neighboring points as long as they fit to the initial plane model. The algorithm iteratively creates planar surface patches until all points belong to some plane or are identified as noise. After normal clustering we obtain a set of planar surface patches $\{\rho_t\}$ (Figure 2d).

B. Clustering Patches

Patches $\{\rho_t\}$ are now greedily clustered into object hypotheses μ_k. Object hypotheses are initialized using the symmetry attention points $\{f_k\}$, which are sorted in decreasing order of $\overline{S}_k$. Given a symmetry point f_k, all patches ρ_t bordering this point (with a 5 pixel radius) form an initial cluster. Patches are then greedily added to the cluster subject to a color and compactness constraint. Once a cluster cannot be extended further, the next cluster is initiated from the next attention point.

1) Color Similarity: A new patch ρ'_t is considered to be a part of the object only if its color model is similar to the already existing model for the object. The color similarity CS_{in} between a new patch ρ'_t and an object μ_k is computed as the Chi-square distance between their HSV color histograms.

$$CS_{in}(\mu_k, \rho'_t) = \chi^2(H(\mu_k), H(\rho'_t)) \quad (11)$$

We also calculate the similarity CS_{out} between the patch ρ'_t and *not*-object, *i. e.* a mask μ'_k surrounding the object, which is defined as the part of the image outside an enlarged

mask μ_k of the object. This enlarged mask is created from the μ_k by doubling its area.

$$CS_{out}(\mu'_k, \rho'_t) = \chi^2(H(\mu'_k), H(\rho'_t)) \tag{12}$$

The color constraint is fulfilled if $CS_{out} < CS_{in}$.

2) Object Compactness: A new patch patch ρ'_t is only added to an object, if its addition does not violate the compactness measure κ of the object. Compactness κ is calculated as the mean of the shortest distances of the object points to the visible surfaces of the object's 3D convex hull. Let a set $\{p_{ki}\}$ be object points, V_k be the corresponding object's convex hull, and v_j be a set of faces facing the viewpoint. Compactness measure κ is then calculated as:

$$\kappa = \frac{1}{n_k} \sum_{p_{ki}} d_{min}(p_{ki}, V_k) \tag{13}$$

where n_k is the number of object points and $d_{min}(p_{ki}, V_k)$ is the shortest distance from the point to any visible face

$$d_{min}(p_{ki}, V_k) = \min_j d(p_{ki}, v_j) \tag{14}$$

The compactness constraint for a patch ρ_t is fulfilled if compactness measure of the object plus patch ρ_t is smaller than the given threshold $\kappa < \theta_{com}$. As shown in the evaluation section the optimal value for the compactness threshold $\theta_{com} = 0.005$. Examples of segmented objects using these constraints are shown in Figure 2e and Figure 4.

V. Results and Evaluation

We evaluated our segmentation algorithm on two publicly available databases: the Table Object Scene Database (TOSD)[1] and the Willow Garage Table Objects Database (Willow)[2].

Other databases as Caltech256, Pascal VOC, LabelMe, Berkeley's B3DO, NYU's Depth Dataset, UW's RGB-D Object Dataset do not cater to our specific task of cluttered table scene segmentation.

TOSD database is targeted towards segmentation evaluation and consists of scenes with varied object configuration complexities. It is composed of images with complex and cluttered scenes, as well as scenes where only several boxes or other simple objects are presented, as shown in Fig. 4. The TOSD database consists of 111 scenes for training and 131 scenes for testing.

The Willow Garage database was originally presented as a benchmark for object recognition for the "Willow Garage: Solutions in Perception Challenge". While the database was created for the task of object recognition, it still serves as a good benchmark for the performance evaluation of segmentation algorithms. The Willow database consists of 175 images taken from the challenge final test set.

Labeling for both databases is in the form of precise segmentation mask contours as opposed to bounding boxes. This makes the evaluation more precise and allows to evaluate how algorithms perform in terms of under-segmentation and over-segmentation.

Evaluation was carried out by varying two specific aspects – namely, object detection and object segmentation. We do not attempt to directly measure the quality of object detection, but instead present the effect of choices in the object detection methodology on our object segmentation approach. In addition, the evaluation was performed to compare the performance of our approach against several state-of-the-art segmentation approaches.

A. Object Detection Strategies

In our work, we applied several strategies for object detection in order to estimate their influence on the object segmentation. As described earlier, the primary object detection strategy used in our pipeline involves the generation of saliency maps from 2.5D symmetries. In this strategy (TJ3D), attention points are selected as points of T-Junctions (or mid-points for simple lines) in symmetry lines extracted from the 2.5D symmetry-based saliency maps (Figure 2c). The second strategy (WTA3D) employed, extracts attention points using Winner-Take-All (WTA) [24] from the 2.5D symmetry-based saliency maps. To see how the use of 2.5D information improves the quality of a detection strategy, we also include attention points using Winner-Take-All from 2D symmetry-based saliency maps [20] (WTA2D).

B. Object Segmentation

To evaluate the attention-based aspect of the algorithm we performed comparison against an attention-driven active segmentation algorithm proposed by Mishra *et al.* [14] (M09), as well as its extension which uses depth information as described in [4] (M11).

Though it is clearly not fair to compare our approach to algorithms that use only color information, it is still interesting to see how the performance differs. Interactive segmentation algorithms [25], [26], [27] require user input (*e. g.* bounding box as in [25]). In scenarios where it is not possible for a user to provide input, the user behavior can be simulated by a computational model of the visual attention system [28], [29]. Therefore, we selected a state-of-the-art interactive segmentation algorithm presented by Gulshan *et al.* [26] (G10) to compare with our algorithm. The algorithm proposed in [26] requires strokes of foreground and background as input. Foreground strokes in our evaluation were simulated as twice dilated skeleton lines from saliency maps. Background strokes were simulated as rectangles near the image border 20% smaller than the size of the original image.

The output segmentation masks are compared to the ground truth labeling in terms of the F-measure defined as $2PR/(P+R)$. We calculated precision P as the fraction of the segmentation mask overlapping with the ground truth and recall R as the fraction of the ground truth overlapping with segmentation mask.

[1] https://repo.acin.tuwien.ac.at/tmp/permanent/TOSD.zip

[2] http://vault.willowgarage.com/wgdata1/vol1/solutions_in_perception/Willow_Final_Test_Set/

Segmentation	TOSD						Willow					
	All		Best		First		All		Best		First	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
M09+TJ3D	0.59	0.07	0.60	0.06	0.59	0.07	0.76	0.05	0.78	0.04	0.76	0.05
M09+WTA2D	0.54	0.07	0.65	0.05	0.52	0.06	0.68	0.07	0.85	0.02	0.72	0.05
M09+WTA3D	0.57	0.07	0.62	0.06	0.58	0.06	0.74	0.05	0.80	0.04	0.75	0.05
M11	0.57	0.08	0.66	0.06	0.62	0.07	0.82	0.06	0.88	0.03	0.86	0.04
G10	0.47	0.08	0.50	0.08	0.47	0.08	0.66	0.08	0.71	0.06	0.68	0.07
Proposed Algorithm	**0.80**	0.05	**0.81**	0.04	**0.80**	0.04	**0.964**	0.010	**0.974**	0.010	**0.970**	0.011

TABLE I: F-score for different segmentation algorithms evaluated on TOSD and Willow datasets.

Segmentation algorithm M09 was evaluated using object detection strategies TJ, WTA2D and WTA3D, mentioned earlier. Segmentation algorithm M11 was evaluated with its own object detection strategy, because this strategy is an intrinsic part of the algorithm. Segmentation algorithm G10 was evaluated using symmetry lines as foreground strokes. Evaluation results are presented in Table I.

Note that the attention mechanism cannot rule out that several attention points come to lie on the same object. In this case, each attention point leads to a possibly different segmentation for an object. Therefore, we calculated three F-scores: the label *first* in Table I refers to the segmentation from the first attention point, *best* refers to the best segmentation w.r.t. ground truth, and *average* refers to the average score over all segmentations for an object. If *first* is lower than *best* this means that the attention points are not optimal. Ideally the first attention point leads to the best segmentation. If *average* is significantly lower than *best* this means that segmentation algorithm depends a lot on the position of the attention point. All F-scores in Table I are averaged over all objects and all scenes.

The proposed segmentation algorithm depends on the value of the threshold θ_{com}. To find the optimal value of the threshold, we evaluated F-scores against threshold values for both databases. As can be seen from Figure 3, the optimal value is 0.005, balancing between over-segmentation (smaller values) and under-segmentation (larger values). Note that the scenes in the Willow Garage database are simpler (isolated standing objects), so that a further increase in θ_{com} does not lead to under-segmentation and performance stays constant. The highest value of F-score obtained was 0.81 for TOSD at this optimal value of θ_{com}. As can be seen from Table I, the primary object detection strategy (TJ3D) proposed in this paper results in improved performance for all types of segmentation algorithms compared to other detection strategies. Evaluation results also show that our combined approach of detection and segmentation performs better on both databases than state-of-art segmentation algorithms. Results for G10 show that color-only segmentation cannot handle complicated table scenarios without good user input. Figure 4 shows visual segmentation outputs for some segmentation strategies. It can be seen that the proposed approach visually gives better results than other attention-driven segmentations.

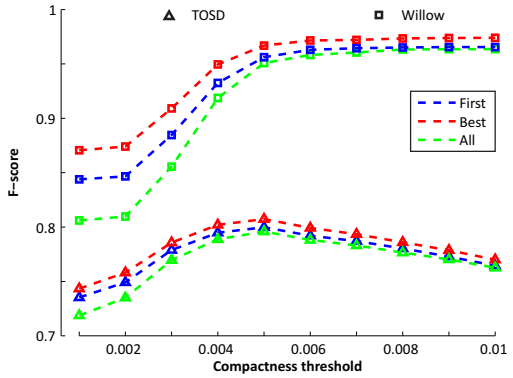

Fig. 3: F-score for proposed segmentation algorithm using different thresholds for compactness θ_{com} for TOSD and Willow datasets. As can be seen from the plot, the optimal segmentation is achieved when threshold $\theta_{com} = 0.005$.

VI. Conclusion and Future Work

In this paper we proposed a novel attention-driven algorithm for cluttered table scene segmentation. We combined a novel object detection strategy using a saliency operator based on 2.5D symmetry with attention points estimation based on symmetry lines and T-Junction points. This was further combined with a segmentation approach based on greedy clustering of planar surface patches using the notion of compactness and color similarity. Our approach shows good results on typical cluttered table scenes containing human made objects with an F-score of 81%. We have shown that our selection of attention points improves performance of attention based segmentation methods and that our combined attention and segmentation approach improves over state-of-the-art attention-driven segmentation approaches. Future work will lie in the area of attention-driven segmentation of more complex scenes directed by task specifications.

References

[1] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[2] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 22, no. 8, pp. 888–905, 2000.

[3] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik, "Semantic Segmentation using Regions and Parts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3378–3385.

Fig. 4: *Visual comparison of different segmentation algorithms*. Row 1 shows segmentation results for Willow database for different segmentation algorithms and rows 2-7 show segmentation results for TOSD database. Segmentation masks and attention points are shown in different colors with respective numbering reflecting the order of attention shift. Our results are shown in the last column. As can be seen, the majority of the existing algorithms have difficulties handling cluttered table scenes, while the proposed algorithm shows promising results.

[4] A. K. Mishra and Y. Aloimonos, "Visual Segmentation of Simple Objects for Robots," in *Robotics: Science and Systems*, 2011.

[5] G. Kootstra, N. Bergström, and D. Kragic, "Fast and Automatic Detection and Segmentation of Unknown Objects," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*, dec. 2010, pp. 442 –447.

[6] E. Potapova, M. Zillich, and M. Vincze, "Local 3D Symmetry for Visual Saliency in 2.5D Point Clouds," in *Computer Vision ACCV 2012*, ser. Lecture Notes in Computer Science, K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, Eds. Springer Berlin Heidelberg, 2013, vol. 7724, pp. 434–445.

[7] A. Ückermann, R. Haschke, and H. Ritter, "Real-Time 3D Segmentation of Cluttered Scenes for Robot Grasping," in *12th IEEE-RAS International Conference on Humanoid Robots*, 2012.

[8] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of Unknown Objects in Indoor Environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 4791 – 4796.

[9] N. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Automatic 3D Object Segmentation in Multiple Views using Volumetric Graph-cuts," in *British Machine Vision Conference*, vol. 28, 2007, pp. 530–539.

[10] J. Wan, T. Xia, S. Tang, and J. Li, "Robust Range Image Segmentation Based on Coplanarity of Superpixels," in *21st International Conference on Pattern Recognition (ICPR 2012)*, 2012, pp. 3618–3621.

[11] D. Sedlacek and J. Zara, "Graph-Cut Based Point Cloud Segmentation for Polygonal Reconstruction," in *7th International Conference on Computer Vision Systems*, 2009, pp. 218–227.

[12] J. Strom, A. Richardson, and E. Olson, "Graph-Based Segmentation for Colored 3D Laser Point Clouds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2010, pp. 2131–2136.

[13] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active Vision," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 333–356, 1988.

[14] A. Mishra, Y. Aloimonos, and C. L. Fah, "Active Segmentation with Fixation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 468–475.

[15] S. Frintrop, P. Jensfelt, and H. I. Christensen, "Attentional Landmark Selection for Visual SLAM," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, 2006.

[16] D. Walther and C. Koch, "Modeling Attention to Salient Proto-Objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[17] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] N. D. B. Bruce and J. K. Tsotsos, "Saliency, Attention, and Visual Search: An Information Theoretic Approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[19] W. Einhauser, M. Spain, and P. Perona, "Objects Predict Fixations Better than Early Saliency," *Journal of Vision*, vol. 8, no. 14, pp. 1–26, 2008.

[20] G. Kootstra, A. Nederveen, and B. d. Boer, "Paying Attention to Symmetry," in *Proc. of the British Machine Vision Conference*. BMVA Press, 2008, pp. 1115–1125.

[21] P. Minovic, S. Ishikawa, and K. Kato, "Symmetry Identification of a 3D Object Represented by Octree," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 507–514, 1993.

[22] C. Sun and J. Sherrah, "3D Symmetry Detection Using The Extended Gaussian Image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 164–168, 1997.

[23] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context Free Attentional Operators: the Generalized Symmetry Transform," *International Journal of Computer Vision*, vol. 14, pp. 119–130, 1995.

[24] D. K. Lee, L. Itti, C. Koch, and J. Braun, "Attention Activates Winner-Take-All Competition Among Visual Filters," *Nature neuroscience*, vol. 2, no. 4, pp. 375–381, 1999.

[25] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts," in *ACM SIGGRAPH 2004 Papers*, ser. SIGGRAPH '04. New York, NY, USA: ACM, 2004, pp. 309–314.

[26] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman, "Geodesic Star Convexity for Interactive Image Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[27] Y. Boykov and M.-P. Jolly, "Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, 2001, pp. 105–112 vol.1.

[28] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging Stereopsis for Saliency Analysis," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 454–461.

[29] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.