Learning What Matters: Combining Probabilistic Models of 2D and 3D Saliency Cues *

Ekaterina Potapova, Michael Zillich and Markus Vincze

Automation and Control Institute
Vienna University of Technology
{*potapova,zillich,vincze*}*@acin.tuwien.ac.at*

Abstract. In this paper we address the problem of obtaining meaningful saliency measures that tie in coherently with other methods and modalities within larger robotic systems. We learn probabilistic models of various saliency cues from labeled training data and fuse these into probability maps, which while appearing to be qualitatively similar to traditional saliency maps, represent actual probabilities of detecting salient features. We show that these maps are better suited to pick up task-relevant structures in robotic applications. Moreover, having true probabilities rather than arbitrarily scaled saliency measures allows for deeper, semantically meaningful integration with other parts of the overall system.

Key words: 3D saliency cues, cue integration, probabilistic learning

1 Introduction

Vision in complex real world scenarios, especially unconstrained segmentation of objects, is a notoriously difficult problem and robotics has realised the importance of attention for robotic systems [24]. Vision in a robot is part of a larger system, which has specific tasks to solve. These tasks allow to derive constraints for the vision system to keep vision problems tractable. These constraints come in the form of attention operators that highlight those parts of the scene most promising for the task at hand.

The range of robotic tasks we consider for this paper includes manipulation, grasping and tracking. We therefore assume objects to appear in various locations and configurations, partly occluded, surrounded by clutter, but typically located on a supporting surface, such as a table or shelf.

What we essentially want is the system to segment objects that can be picked up, or if that is not possible due to clutter or occlusion, we want to at least detect

* The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement IST-FP7-IP-215821 GRASP 2008-2012.

good initial grasp points. These tend to be located somewhere on parts sticking out from the scene. Pre-grasp manipulation of such parts might free the object from the pile.

Scene segmentation is one of the most researched topics in computer vision, and many different approaches have been proposed [3, 5, 18], but no generic solution suitable for every task exists. Recent state-of-the-art research in this field suggests the use of seed points to guide the segmentation process [19, 15, 23]. This leads to the problem of identifying good seed points. Inspired by pre-attentive vision theory recent research has suggested the use of attention points, which can be extracted from saliency maps, using for example a winner-takes-all (WTA) algorithm [16].

Many well-known and widely acknowledged models for computation of saliency maps, such as [11, 14, 13, 12, 1] use only 2D information about the scene. Itti-Koch-Niebur (IKN) [14] is a generic cue inspired by physiological models, and has proven its efficiency in 2D images. Fig. 1,e) shows saliency map based on IKN cue for the image in Fig. 1,a). Several recent extensions to 3D take advantage of the increased availability of 3D sensing equipment, such as inexpensive laser or time-of-flight sensors and RGB-D cameras [10, 17, 22, 2].

However, classical cues indicate only outliers in the scene, while we want regions with specific properties be poped out, no matter to what particular object they belong. One can see this problem as top-down attention task described in [21, 9], while our current goal in to build bottom-up attention system tuned to identifying particular properties of the visual search space. Finally, given that there is a number of intuitively plausible saliency cues (2D and 3D) there is no model for combining these cues in a principled manner with respect to desired task, without using top-down specific features of required objects or parts of visual space.

We address the above issues with a learning based approach, which can be in future extended to top-down search tasks. With the help of Kinect sensor we have created an RGB-D image database, consisting of different types of table scenes which are challenging for segmentation, owing to the presence of fully and partially occluded objects, multi-colored objects etc. The database consists of four types of scenes: a) isolated free-standing objects (IFSO), b) occluded objects (OO), c) objects placed in a box (BO) and d) a box containing objects and surrounded by other objects (BOSO). For each type of scenes multiple configurations of objects are presented. Totally there are 86 RGB-D images in the database. Task regions were hand-labeled by outlining them with a polygon. In our problem task relevant regions are whole objects, that is why labeling was done by only one person, whose task was to segment objects in the scenes as precisely as it was possible. For BOSO objects we are interested only in objects situated directly in the box, that is why objects around the box were not labeled at all. Fig. 1, a)-d) show examples of labeled images.

The main novelty of this paper lies in the area of understanding how and what preattentive cues should be combined in a specific robotics task of calculating attention points for segmentation of graspable objects.

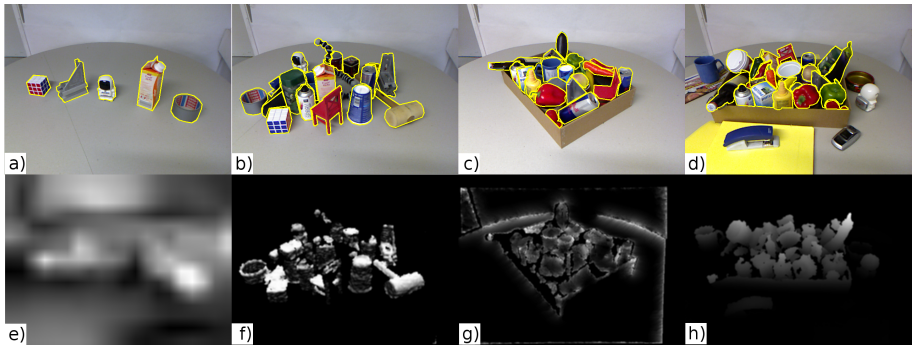

Fig. 1. Four pairs of images and saliency maps (a) and e), b) and f), c) and g), d) and h)). Images a)-d) show examples of images along with labeling for isolated free-standing objects, occluded objects, objects placed in a box and a box containing objects and surrounded by other objects respectively. Images e)-h) show examples of saliency maps based on IKN cue, RSO cue, OE cue and SH cue respectively.

2 Investigated Cues

Inspired by findings from preattentive human vision [7, 6, 20] we investigated several 3D cues, e.g. based on surface height (SH), relative surface orientation (RSO) and occluded edges (OE) and combined them with cues obtained from 2D information (color, orientation and intensity).

2.1 Surface Height Cue

For the task of picking up objects in a cluttered scene, the simplest way to start grasping is first to pick up all objects that stick out from the clutter. These objects are good candidates for initial grasping attempts, and they should therefore be considered more interesting than the rest. These objects can be pointed out by attention points derived from the surface height preattentive cue, which is based on a height map of the scene. Fig. 1,h) shows the saliency map based on the SH cue for the image in Fig. 1,d).

As input we have a point cloud of the table scene. To calculate height of the objects we need to determine the ground plane or absolute zero – the supporting plane, which is the plane on which the box rests (i.e. table). We use RANSAC [8] to determine the plane coefficients $Ax + By + Cz + D = 0$. Considering that our robotic system has the specific task of grasping objects from a table scene we do not pay much attention here to cases where supporting plane cannot be determined at all, or many supporting planes exist in a scene. For every point in the point cloud the distance to the supporting plane is calculated. We set d_{max} to be the distance between the ground plane and the most remote point in the point cloud. Values of the SH cue are calculated according to:

$$SH(i,j) = f(D(i,j)) \tag{1}$$

where $D(i,j)$ is a distance from correspondence point in the point cloud to the supporting plane. We furthermore scale height values non-linearly according to

$$f(x) = ax^2 \tag{2}$$

to obtain more pronounced salient regions, where a is determined such that $f(d_{max}) = 1$

2.2 Relative Surface Orientation Cue

The surfaces of objects parallel to the supporting plane often present good candidates for first grasping positions, because they usually indicate top-surfaces of simple objects that can be easily grasped. One of our 3D preattentive cues aims to identify top-surfaces based on surface orientation. We calculate relative orientation between local surface normals and supporting plane normal. Fig. 1,f) shows saliency map based on RSO cue for the image in Fig. 1,b).

We use the same supporting plane coefficients $Ax + By + Cz + D = 0$ as during calculation of the SH saliency map.

The vector with coordinates $\overrightarrow{N} = \{A, B, C\}$ represents the normalized plane normal. For every point $p(i,j)$ in the point cloud P local surface normal vectors are determined by fitting a local plane and estimating plane normal:

$$\forall p(i,j) \in P : \overrightarrow{n(i,j)} = \{n_x(i,j), n_y(i,j), n_z(i,j)\} \tag{3}$$

Values of the RSO cue are calculated according to:

$$RSO(i,j) = |g(\overrightarrow{n(i,j)})| \tag{4}$$

where $g(\cdot)$ calculates the cosine between two vectors:

$$g(\overrightarrow{n}) = (\overrightarrow{n}, \overrightarrow{N}) \tag{5}$$

2.3 Occluded Edges Cue

The success of the segmentation based on seed points depends a lot on the position of the seed point. The more central the location of the seed point with respect to the object, the higher is the probability that the object will be properly segmented. To this end we designed a cue based on occluded edges. The cue is derived from the depth map of the scene. Fig. 1,g) shows example of saliency map based on OE cue for the image in Fig. 1,c). Using the Canny operator [4] an edge map EM is calculated from the depth map. From every point $p(i_0, j_0)$ that belongs to one of the edges we create a potential field $P(\cdot)$ according to the following formula:

$$P(d) = a\frac{1}{d} - b \tag{6}$$

where d is the distance from the current point $p(i,j)$ to the initial edge point $p(i_0, j_0)$ whose influence we are calculating, a is set to 0.5 and b is set to 0.01

in our experiments. The influence is expanded only in directions of decreasing values of the depth map, i.e. the object side of the occluding edge.

The value of the point $p(i,j)$ in the OE map is equal to:

$$OE(i,j) = \sum_{\forall (i_0,j_0): EM(i_0,j_0) \geq 0} P(\sqrt{(i-i_0)^2 + (j-j_0)^2}) \quad (7)$$

2.4 Cue Combination

We investigated two approaches for cue combination to obtain a final saliency map SM. The first approach is similar to cue combination used in IKN method – the final saliency map SM_S is equal to the sum of normalized separate cues:

$$SM_S(i,j) = N(IKN(i,j)) + N(SH(i,j)) + N(RSO(i,j)) + N(OE(i,j)), \quad (8)$$

where $N(\cdot)$ - linear normalization operator to the range [0,1].

The second combination method uses multiplication instead of summation, so that we obtain SM_M as normalized multiplication of separate cues:

$$SM_M(i,j) = N(IKN(i,j))N(SH(i,j))N(RSO(i,j))N(OE(i,j)). \quad (9)$$

Fig. 6 e)-h) and Fig. 6 m)-p) show examples of SM_S and SM_M combination types for different types of the scenes.

3 Probabilistic Learning

Combining cues according to Eq. 8 or 9 does not take into account the relative importance of cues. One way to address this is to weight the individual cues and learn these weights. Another possibility is to directly learn probabilistic models of cues and then combine these. We used a labeled database to train a probabilistic model of relevance for each saliency cue. For each cue c_i we learned the probability of observing that for given cue a pixel was marked as task relevant salient ($s = true$) - situated inside labeled polygons, or non-salient ($s = false$) - situated outside labeled polygons.

$$\begin{aligned} p(c_i \,|\, s = true) \\ p(c_i \,|\, s = false) \end{aligned} \quad (10)$$

We estimated parameters for normal distributions for every type of cue separately and for two types of cue combination: addition and multiplication. Note that our labels essentially mark whole objects, with parts of them being salient (different parts for different cues) and other parts not salient, i.e. we use generic labels, rather than labeling for each cue individually. But this means that estimating the above probabilities directly from the labeled images would essentially learn that inside a region labeled as salient, all sorts of saliency values can appear. But we know that inside labeled regions we are only interested in what

makes part of that region salient, not the fact that not all of it is salient. To this end, during estimation of the normal distribution, we weight pixels with intensity I according to $w(I) = I^2$. Note that this measure was not necessary with marked regions, precisely outlining salient regions for each cue individually. We chose this method however, because we want one set of generic labels, applicable to various different cues, picking up saliency somewhere inside those regions.

Fig. 2 shows estimated normal distributions of intensity values (in the range $[0, 1]$) for the IKN cue constructed for occluded objects scenes and for the RSO cue constructed for a box with objects surrounded by other objects scenes (a) and b) respectively). We can clearly see that distributions are well separated, allowing distinction of salient from non-salient regions. Following Bayes rule we can then infer the posterior probability of saliency as

$$\begin{aligned} p(s \mid c_i) &= \frac{p(c_i \mid s)\ p(s)}{p(c_i)} \\ &= \frac{p(c_i \mid s)\ p(s)}{\sum_{k \in \{t,f\}} p(c_i \mid s = k)} \end{aligned} \tag{11}$$

with $p(s)$ being the prior probability of saliency. This could be obtained from top level context information, but is simply assumed 1 here, as we are more interested in the relative differences between cues and $p(s)$ is the same for all.

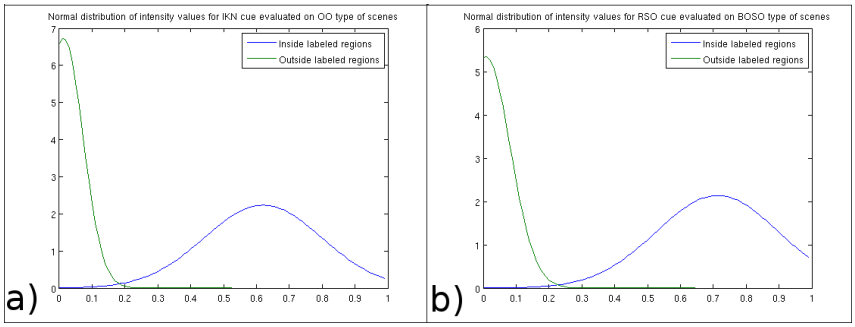

Fig. 2. Normal distribution of intensity values inside and outside labeled regions: a) for IKN cue for occluded objects scenes, b) for RSO cue for a box with objects surrounded by other objects scenes

Fig. 3 a)-d) shows the posterior probabilities of salient values for different cues and combinations of cues for different types of scenes. The smaller slope of the IKN as well as OE cues over all types of images indicates that for our type of scenes they are less distinctive than the others. This means that these cues cannot precisely distinguish regions belonging to different objects.

Based on evaluated parameters of the normal distributions, posterior probability images were built for a validation set. The relative sizes of training set and test set were 0.8:0.2.

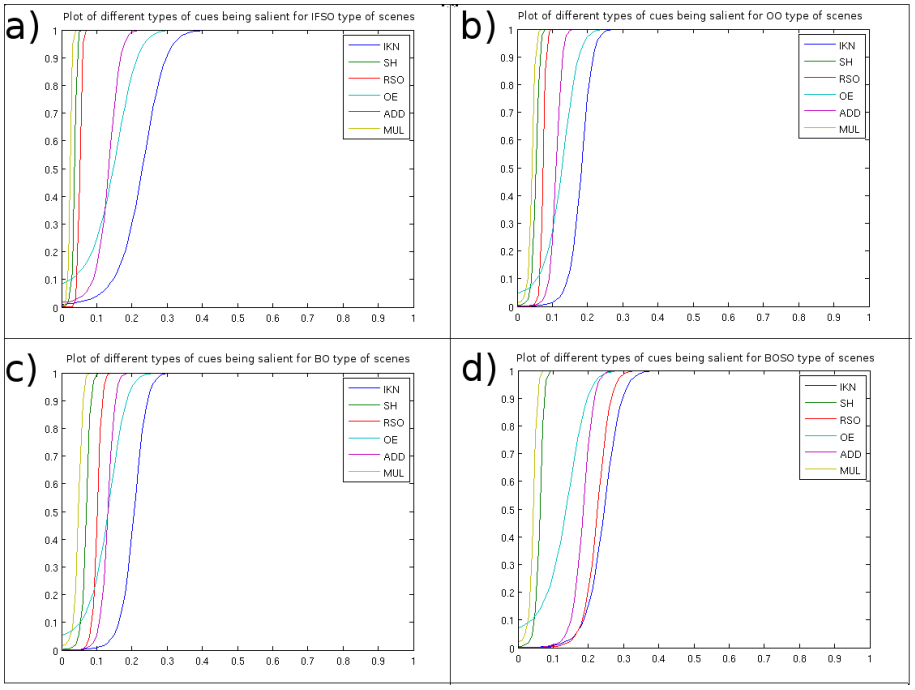

Fig. 3. Probability of salient regions being situated inside labeled regions for different types of scenes: a) isolated free–standing objects b) occluded objects c) objects placed in a box, and d) a box containing objects and surrounded by other objects for different individual cues and cue combinations (for all plots probability via salient value).

Fig. 4 shows examples of posterior probability images for different types of cues and cue combinations for the image shown in Fig. 1 d). For an ideal probabilistic image regions of different objects should have the highest intensity values (in our case 1) and be separated from each other. As we can see from Fig. 4 among individual cues RH and RSO cues show the best performance, while combination by multiplication performs better that combination by summation.

As can be seen from the Fig. 4 the IKN cue for such complex scenes assigns high probability values to areas, which do not belong to any object. This is because IKN cue does not take into consideration 3D spatial positions of the objects, and thus cannot distinguish objects with e.g. similar color. Probability images give us insight into how cues can be combined in terms of top-down attention for a specific task of segmentation for grasping.

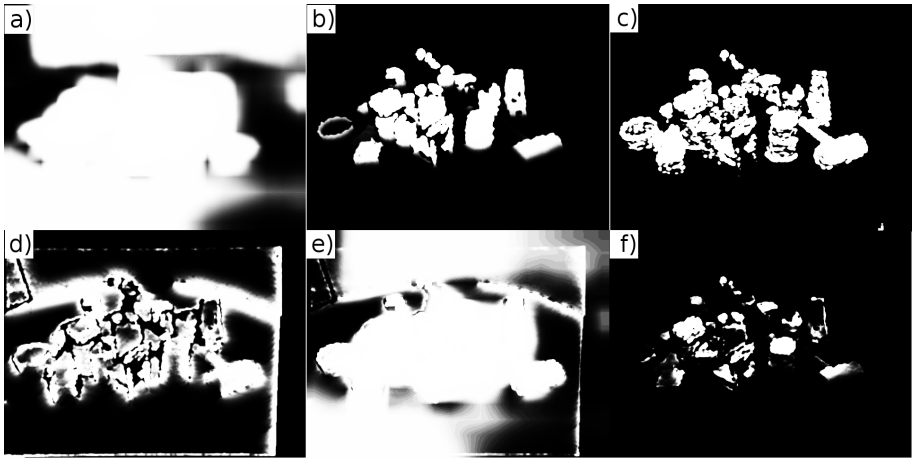

Fig. 4. Posterior probability images for image shown in Fig. 1 d) for a) IKN cue, b) SH cue, c) RSO cue, d) OE cue, e) SM_S cue and f) SM_M cue.

4 Evaluation and Results

To evaluate individual cues as well as the cue combinations, we calculated the ratio of first five WTA [16] attention points from saliency map being situated inside labeled regions of a hold-out set of training images. Averaged results are presented in Fig. 5. Results indicate that especially for complicated scenes with occluded objects 3D saliency cues based on surface height and relative surface orientation perform better than simple 2D cues. Furthermore the cue based on occluded edges did not prove to be a useful cue for our tasks.

Evaluation results go along with distributions obtained from probabilistic learning, while there is still an open question what cue combination is the best for the given task and more experiments on that should be provided.

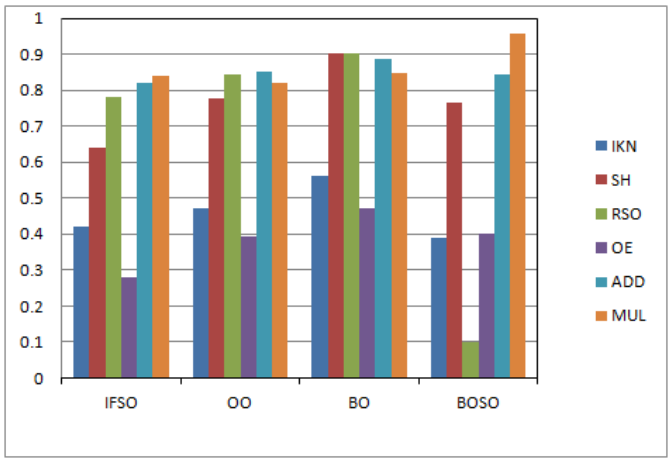

Fig. 5. The ratio of the first five attention points being situated inside different labeled ROIs (IFSO - single standing objects, OO - occluded objects, BO - objects in a box, BOSO - a box with objects which is situated among other objects.

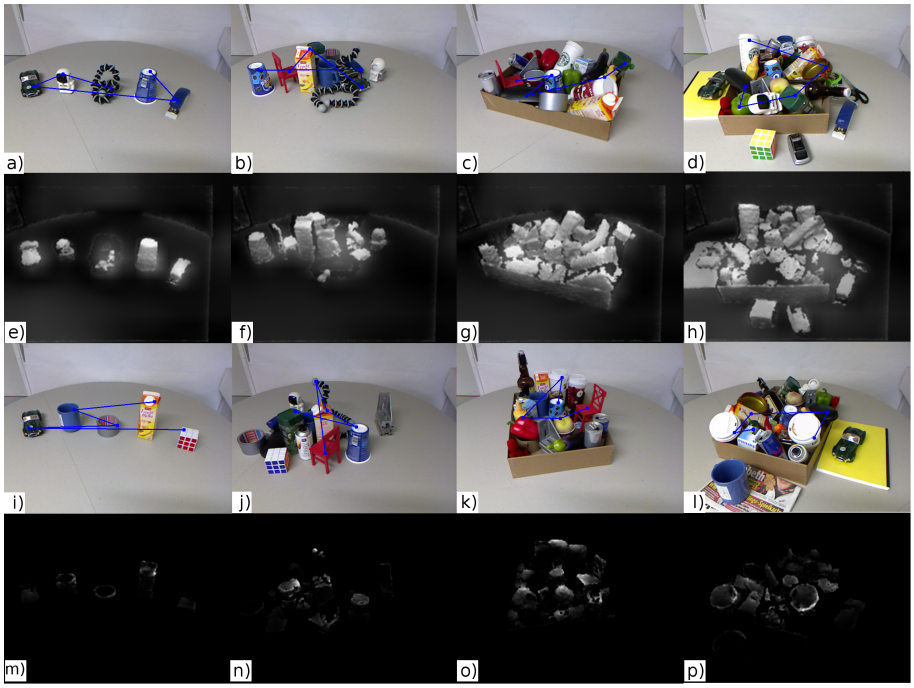

Fig. 6. Results of WTA algorithm for different types of images (left to right: IFSO, OO, BO, BOSO) and corresponding saliency maps: a)-d) present WTA results on SM_S and e)-h) are corresponding saliency maps; i)-l) present WTA results on SM_M and m)-p) are corresponding saliency maps.

Fig. 6 shows examples of images with first five attention points indicated in blue color and corresponding saliency maps.

5 Conclusion and Future Work

In this paper we investigated the use of 3D cues to obtain attention points that can be used as seed points for segmentation of objects for robotic grasping tasks. We implemented three 3D cues to compete against the standard IKN model [14]. Scenes with growing complexity (isolated free-standing objects, occluded objects, objects in a box, and a box containing objects and surrounded by other objects) were evaluated against each cue and two types of cue combination – summation and multiplication. We furthermore estimated probabilistic models over the whole set of images for every type of cue. We could show that height and relative surface orientation cues considerably improve performance in calculating attention points on potential objects for grasping over the standard IKN model [14]. In the most complex cases the combination of both 3D cues gives clearly the best results. This indicates that 3D cues deserve more attention when moving out into the real world with robots.

Our future work will lie in the area of implementing and evaluating more types of 3D preattentive cues and using the results in actual grasping scenarios.

References

1. Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Ssstrunk. Salient region detection and segmentation. In Antonios Gasteratos, Markus Vincze, and John Tsotsos, editors, *Computer Vision Systems*, volume 5008 of *Lecture Notes in Computer Science*, pages 66–75. Springer Berlin / Heidelberg, 2008.
2. Jonker P. Akman O. Computing saliency map from spatial information in point cloud data. *Advanced Concepts for Intelligent Vision Systems, Lecture Notes in Computer Science*, 6474/2010:290–299, 2010.
3. Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 1:105–112, 2001.
4. J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679698, 1986.
5. D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, may 2002.
6. J. T. Enns and R. A. Rensink. Influence of scene-based properties on visual search. *Science (New York, N.Y.)*, 247(4943):721–723, feb 1990.
7. James T. Enns and Ronald A. Rensink. Sensitivity to three-dimensional orientation in visual search. *Psychological Science*, 1/5:323–326, 1990.
8. Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24:381–395, 1981.

9. Simone Frintrop, Gerriet Backer, and Erich Rome. Goal-directed search with a top-down modulated computational attention system. In *DAGM-Symposium*, pages 117–124, 2005.
10. Simone Frintrop, Erich Rome, Andreas Nchter, and Hartmut Surmann. A bimodal laser-based attention system. *Computer Vision and Image Understanding*, 100:124–151, 2005.
11. Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 19:545–552, 2007.
12. Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.
13. L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, mar 2001.
14. Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
15. Byoung Chul Ko and Jae-Yeal Nam. Object-of-interest image segmentation based on human attention and semantic region clustering. *J. Opt. Soc. Am. A*, 23(10):2462–2470, Oct 2006.
16. D. K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nature neuroscience*, 2(4):375–381, apr 1999.
17. A. Maki, P. Nordlund, and J. O. Eklundh. A computational model of depth-based attention. *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, 4:734–739 vol.4, 1996.
18. Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, jun 2001.
19. Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. Active Segmentation with Fixation. In *Twelfth IEEE Int. Conf. on Computer Vision, Kyoto*, 2009.
20. Ken Nakayama and Gerald H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, 1986.
21. V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2049 – 2056, 2006.
22. N. Ouerhani and H. Huegli. Computing visual attention from scene depth. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 375 –378 vol.1, 2000.
23. Nabil Ouerhani, Neculai Archip, Heinz Hgli, and Pierre-Jean Erard. Visual attention guided seed selection for color image segmentation. In Wladyslaw Skarbek, editor, *Computer Analysis of Images and Patterns*, volume 2124 of *Lecture Notes in Computer Science*, pages 630–637. Springer Berlin / Heidelberg, 2001.
24. John K. Tsotsos and Ksenia Shubina. Attention and Visual Search : Active Robotic Vision Systems that Search. In *The 5th International Conference on Computer Vision Systems*, 2007.