Saliency based method for object localization

Maja Rudinac and Pieter P. Jonker

Delft Biorobotics Lab, Faculty of Mechanical Engineering, Delft University of Technology

m.rudinac@tudelft.nl, p.p.jonker@tudelft.nl

Keywords: Saliency detection, object localization, interest points clustering, robotic application

Abstract

In this paper we present a scene exploration method for the identification of interest regions in unknown indoor environments and the position estimation of the objects located in those regions. Our method consists of two stages: First, we generate a saliency map of the scene based on the spectral residual of three color channels and interest points are detected in this map. Second, we propose and evaluate a method for the clustering of neighboring interest regions, the rejection of outliers and the estimation of the positions of potential objects. Once the location of objects in the scene is known, recognition of objects / object classes can be performed or the locations can be used for grasping the object. The main contribution of this paper lies in a computationally inexpensive method for the localization of multiple salient objects in a scene. The performance obtained on a dataset of indoor scenes shows that our method performs good, is very fast and hence highly suitable for real-world applications, such as mobile robots and surveillance

1. Introduction

Dealing with unknown environments is one of the main demands and challenges of modern robotics. Robots should be able to scan and process their surrounding in real time, to act upon that information and to learn from obtained results. An active vision approach is very suitable for robotic applications since it tends to mimic the ability of the human visual system which uses foveal vision to select just the most relevant (salient) regions and process them. Recent experiments on human subjects [1] validated this hypothesis that the purpose of selective attention in the human visual system is to maximize task efficiency by fixating relevant ("salient") objects in the scene.

We aim at developing a robust vision system for our low-cost service-robot platform which is able to perform tasks such as navigation and object grasping. In order to make it possible for our robot to deal with unknown environments and to learn from them, we first need to design a method that can detect interesting regions in a scene, i.e. where potential items or obstacles might be located, and then to estimate the approximate position of those objects. This paper presents a description of our method. Knowing locations of objects, we will then be able to focus the robot's attention to them and perform recognition of either the specific object or the object class. Additional requirements are that all processing should be fast enough to be used on our mobile robot. Also, only a single low resolution web camera should be used to lower the cost of the system.

Several methods for detecting salient information have been proposed recently. The saliency detection model of Itti et al. [2] builds saliency maps for three features: Color channels, intensity and orientation, and they compute each feature using a set of linear center-surround operations from fine to coarse scales, analogous to visual receptive fields. A model presented by Park et al. [3] expands the previous method by introducing additional feature maps of edges and symmetry. We extended these approaches and used saliency information to localize objects in unknown environments. From all existing methods for image segmentation based on saliency, we found that the best results are obtained by Achanta et al [4]. Their method for saliency detection is based on the spatial frequency content of saliency maps. The authors also proposed a segmentation approach based on adaptive tresholding which had good segmentation results, but due to its high computational complexity, it cannot be used in real-time robotic applications.

Several state of the art robotic platforms have already developed visual attention systems for localizing objects in the real world. One of the best performing platforms is Curious George [5], a mobile robot with a very complex vision framework that shows good performance in building up a detailed semantic representation of its environment. It combines a peripheral fovea vision system that joins bottom-up visual saliency with structure from stereo vision, and a localization and mapping system. Another representative mobile robot, developed by Ekval et al.[6] uses a visual attention approach to detect predefined objects and to estimate their position in the environment, while it integrates this with a localization module to automatically put the objects in a generated map. For this, in addition to a CCD camera a laser scanner is used.

However, the aforementioned approaches are not suitable for our application as they use 3D information obtained from stereo vision or laser scanning to select the objects. We aim at a low cost system that uses one simple webcam only. Consequently, in this paper we propose a computationally inexpensive object localization method that can be applied on low resolution images from a single web camera.

2. Saliency detection in the scene

The first step in our approach is to find salient regions in the scene. Hence, first a saliency map is generated and then regions of various sizes are detected using the Maximally Stable Extremal Region (MSER) method [7].

2.1. Saliency map generation

Since our system is expected to work fast enough for robotic applications, we use the spectral residual approach as proposed by Hou and Zhang [8], which is the fastest method for saliency map generation we found. Similar to [9] we combine this approach with color information, and compute the log spectrum representation *L(f)* of three channels: intensity, red-green and yellow-blue. The log spectrum is obtained as *L(f)=log(A(f))* , where A(f) represents the amplitude of the averaged Fourier spectrum of the channel. To reduce computation time, the channels are computed on a down-sampled image of 128 x 128 pixels. If we presume that *r, g, b* represent channels in RGB space, the three channels are defined similar to [10]:

1. Intensity channel:
$$M_I = \frac{r+g+b}{3} \quad (1)$$

2. Red-Green channel:
$$M_{R-G} = \frac{r-g}{\max(r,g,b)} \quad (2)$$

3. Yellow-Blue channel:
$$M_{B-Y} = \frac{b-\min(r,g)}{\max(r,g,b)} \quad (3)$$

The spectral residual *R(f)* of every channel is defined as: *R(f)=L(f)-A(f),* where *A(f)* is the amplitude of the averaged Fourier spectrum of the channel. Using the Inverse Fourier Transform, the saliency map of every channel is constructed. The total saliency map is calculated by summing the saliency maps of every channel. An example of a generated saliency map of the scene in Figure 1a is shown in Figure 1b.

2.2. Detecting interest points in the saliency map

The regions with a high saliency are now detected in the saliency map using a MSER detector [7]. We calculate the detector to find - high saliency - bright on dark regions with the restrictions that the region must be more than 10% smaller than any nested region and that they have a maximal variation of 0.3. We present every region with a contour lying around all pixels belonging to that region. An example of detected Maximally Stable Extremal Regions (yellow points) in the scene of Figure 1a is shown in Figure 1c. The contours represent interesting regions in the scene. In order to reject outliers and to estimate the positions of potential objects in the scene, in a next step the clustering of contour centers close to each other must be performed.

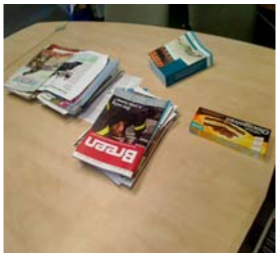


Figure 1a: Original scene

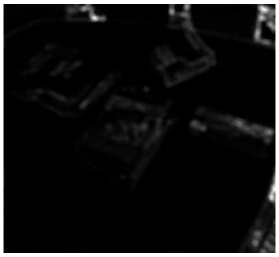


Figure 1b: Saliency map

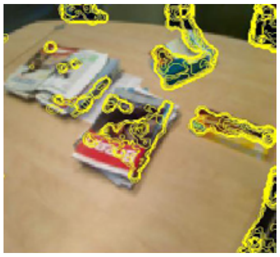


Figure 1c: Detected interest points

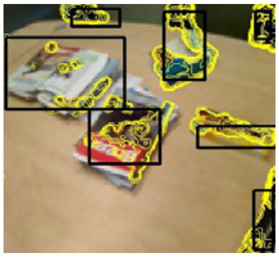


Figure 1e: Objects localized

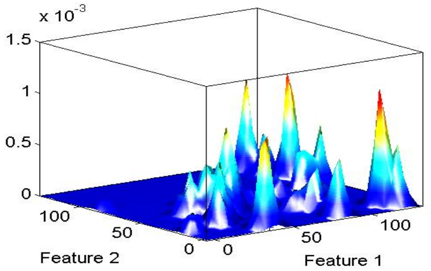


Figure 1d: 3D map of estimated probabilities

3. Clustering of the interest points

In every contour centre of the MSER regions, we first calculate the Gaussian probability in order to make our clustering more robust to outliers. The problem that we encountered was that we had to manually determine the width of every Gaussian kernel since its optimal value depends on the number of objects in the scene. To overcome this problem, we use a Parzen window estimator since it does not assume any function form of the unknown PDF and allows its shape to be entirely determined by the data without having to choose a location of the center. The PDF is now estimated by placing a well-defined kernel function on each contour centre and then determining a smoothing parameter, similar to Tax [11]. The estimated probability (4) is defined as the sum of all Gaussian kernels, multiplied by a scaling factor. In the equation (4), x_i and σ represent the kernel centre and the kernel width respectively, while N is the number of contour centers and d=2 since every contour centre has two coordinates, X and Y.

$$p(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{(2\pi)^d\sigma^d}}\exp(-\frac{\left\|x-x_i\right\|^2}{2\sigma^2}) \qquad (4)$$

In order to reject outliers, we discard the points with low probability values that satisfy the condition:

$$\log(p(x))<\log(\frac{1}{N}\sum_{i=1}^{N}p(x_i))-3*\mathrm{var}(\log(\frac{1}{N}\sum_{i=1}^{N}p(x_i)))$$

(5)

If we now look at the 3D plot of the estimated probabilities (Figure 2d) of the scene from Figure 2a, we notice that the dominant peaks represent the objects in the scene. We then form clusters around these peak probability values using mean shift clustering. For each of N 3D points (two coordinates of the contour center and its estimated Parzen probability) a multivariate kernel density is calculated, using Gaussian kernel as *K* (6) and mean-shift vector is computed (7).

$$\hat{f}_K=\frac{1}{Nh^3}\sum_{i=1}^{N}K(\frac{y-y_i}{h}) \qquad (6)$$

$$m(y_i)=\frac{\sum_{i=1}^{n}y_i g(\left\|\frac{y-y_i}{h}\right\|^2)}{\sum_{i=1}^{n}g(\left\|\frac{y-y_i}{h}\right\|^2)}-y \qquad (7)$$

The term g denotes the derivative of the selected kernel profile while h is the bandwidth parameter that defines the radius of the kernel, which must be manually tuned for the optimal value. A compromise must be achieved because a small value of h leads to a large number of small clusters, while large value of h gives a small number of too large clusters and hence close objects are grouped together. After experimental analysis the value of h was set to 16 as the optimal value. As the improvement step, mean shift algorithm can be performed with several different values for h which will provide better clustering result but with high increase in computation time.

Finally, the location of the objects is estimated with a region of interest formed by the maximal and minimal values of the pixel coordinates X and Y for contour centers in every cluster. An example of object localization in the scene of Figure 1a is shown in Figure 1e. In the next section we provide a detailed analysis of our method and its experimental results.

4. Results, evaluation and analysis

We tested our method in real world indoor settings, which usually contain multiple salient objects per scene. Unfortunately, there are no standardized and widely adopted datasets of such kind and hence we made a custom database of 100 indoor scenes. The main purpose of this database is that the localization of objects in real world scenes can be detected by a robot that has to spot objects while navigating through the scene. The robot can take a closer look at potential regions of graspable objects if it needs more information.

We divided the scenes into 7 categories, such as hallway, kitchen, office … in order to obtain a better understanding of the environmental conditions that dominate their image processing. We acquired the scenes with a webcam. The average number of objects per scene is 5.18 and it varies from 8 for kitchen scenes to 4 for coffee corner scenes. In a pre-processing step all images are down-sampled to 128 x128 pixels in order to speed up the computation. In order to evaluate the method, in all images position of the salient objects is manually labeled and further used in testing as the ground truth data.

The overall precision-recall results are presented in Table I. and show very promising results for our application. High precision results as well as very low processing time are found for “table” and “kitchen” categories, which contain a large number of occluding salient objects. The lowest rates, as well as high processing times, are obtained in the case of images from the “hall-way” category, probably caused by very intensive light from outdoor, which inflicts the saliency detection step. In future we will filter out these images.

In Table II and Table III we compare the performance of our method to the one from Achanta et al [4] (FTRD). In overall, our system performs better demonstrating higher precision and recall rates on 4 out of 6 categories, while on two categories it performs slightly worse. In Table II precision-recall rates are shown for the entire database and our method demonstrates higher rates. In Figure 2, several examples of localized objects from different categories are presented comparing our results with the ones given by FTRD method.

A speed analysis depicted in Figure 3 shows the comparison of the computation time of our approach (in blue) with the time of the approach by Achanta et al [4] (FTRD: in red). A large benefit of our method is the low computation time, and although it is still not real-time (2 frames per second using Matlab on a standard laptop) it is much faster than the representative system of Achanta et al [4] with the equivalent or better precision and recall rate. We expect that a C++ implementation of our system will decrease the computation time an order of magnitude. Hence we can conclude that with a C++ implementation, our method is effectively useful for vision based robot navigation.

Category	Precision	Recall	F-measure
Blurred images	0.900	0.750	0.818
Coffee corner	0.943	0,846	0.892
Entrance	0.809	0.809	0.809
Hallway	0.850	0.872	0.861
Kitchen	0.908	0.831	0.831
Office	0.883	0.855	0.869
Table	0.934	0.925	0.923

Table I: Precision-recall results for different indoor scene categories – our method

For entire Database	Precision	Recall	F-measure
Our method	0.897	0.847	0.871
FTRD	0.850	0.815	0.832

Table II: Precision-recall results for the entire database

Category	Precision	Recall	F-measure
Blurred images	0.766	0.681	0.720
Coffee corner	0.850	0.872	0.860
Entrance	0.705	0.738	0.721
Hallway	0.864	0.950	0.905
Kitchen	0.944	0.810	0.872
Office	0.906	0.762	0.828
Table	0.862	0.903	0.882

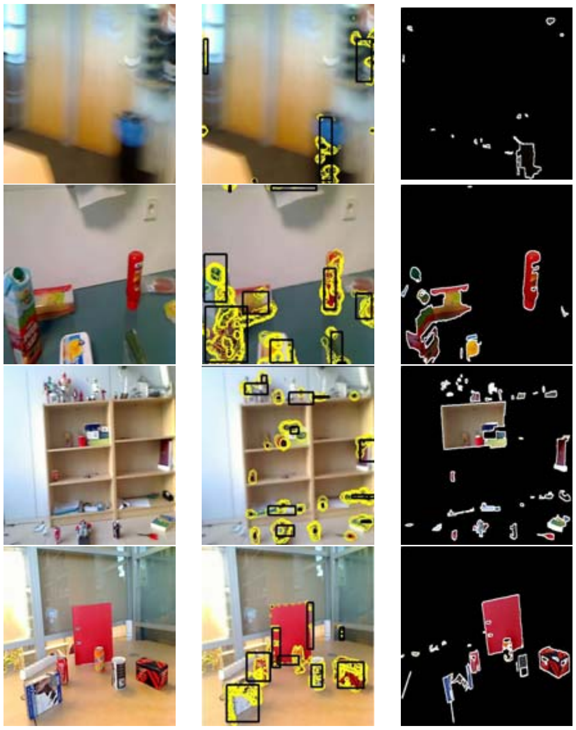
Table III: Precision-recall results for different indoor scene categories – FTRD method

5. Conclusion

In this paper we present a two stage method for object localization. The main idea is to apply visual attention and detect only the most interesting regions in the scene after which object recognition can be performed on the salient regions only. In the first stage of our method a saliency map of the scene is generated on three color channels using a spectral residual approach. Then, interest points are extracted on the saliency map and contours are generated around every region. In the next stage we use a method for clustering the contour centers and we estimate the positions of salient objects. Finally, the location of objects in the scene is estimated, which can be further used as input for robot navigation and grasping.

Our method shows promising results on our database of 100 indoor scenes and is inexpensive compared to representative approaches we found in literature while having a similar or better performance. In future we plan to use this algorithm on a robot for object localization and novel object learning

Acknowledgments: This work has been carried out as part of the FALCON project under the responsibility of the Embedded Systems Institute. This project is partially supported by the Netherlands Ministry of Economic Affairs under the Embedded Systems Institute (BSIK03021) program.



Original image Our method FTRD method

Figure 2: Examples of localized objects

.

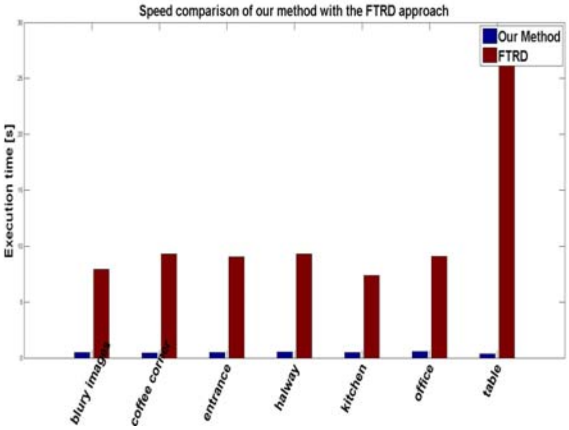


Figure 3: Speed evaluation

References

[1] R. L. Canosa, Real-world vision: Selective perception and task. ACM Transections on Applied Perception. 6, 1-34, 2009.

[2] A L. Itti, U. Koch, and E. Niebur, “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis”, In: IEEE Trans. on PAMI, 20, 11, 1254 – 1259, 1998Fergus, R., Perona, P. and Zisserman, A. Object Class Recognition by Unsupervised Scale-Invariant Learning Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2003.

[3] J.S.Park, J. K. Shin, and M. Lee, “Biologically Inspired Saliency Map Model for Bottom-up Visual Attention”, In: LNCS 2525, Springer, pp. 418 – 426, 2002Tuytelaars, T. and Mikolajczyk, K. 2008. Local invariant feature detectors: a survey. Found. Trends. Comput. Graph. Vis. 3, 3 (Jan. 2008), 177-280

[4] R.Achanta,S.Hemami, F.Estrada, S.Susstrunk. Frequency-tuned salient region detection,,In Proc of Conference on Computer Vision and Pattern Recogntion , 1597-1604, 2009.

[5] D. Meger, P. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. Little and D.G. Lowe, Curious George: An attentive semantic robot. *Robot. Auton. Syst.* 56, 6, 503-511, 2008.

[6] S. Ekvall, D. Kragic, and P. Jensfelt. Object detection and mapping for service robot tasks. Robotica: International Journal of Information,Education and Research, 25, 2, 175-187, 2007M. Stark and B. Schiele, "How good are local features for classes of geometric objects," in Proceedings of the International Conference on Computer Vision, 2007.

[7] J.Matas, O.Chum, M Urba, and T.Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in Proc.of British Machine Vision Conference, 2002.

[8] Hou and L. Zhang. Saliency detection: A spectral residual approach. CVPR 2007, pp. 1–8, 2007.

[9] P. E. Forssén, D. Meger, K. Lai, S. Helmer, J. J. Little, D.G.Lowe: Informed visual search:Combining attention and object recognition. ICRA 2008, 935-942, 2008.

[10] D.Walther, C.Koch: Modeling attention to salient proto-objects. Neural Networks, 19 (9), 1395-1407, 2006.

[11] D.M.J. Tax, One class classification, Phd Theses, Delft University of Technology, 2001.