

A SEGMENT-BASED IMAGE SALIENCY DETECTION

O. Muratov, P. Zontone, G. Boato, F. G. B. De Natale

DISI, University of Trento
Via Sommarive, 14 I-38123 Povo, Italy

ABSTRACT

This paper presents a novel method of visual saliency detection. The use of saliency promises benefits to multimedia applications. However, up to now just few reasonable applications of saliency exist. It is clear that limited accuracy is one of the possible reasons for this. Another reason could be that in general saliency allows us to detect salient regions of the image rather than objects. To fill this gap we study to what extent the integration of segmentation into saliency detection allows the estimation of saliency of objects. In this paper we propose a method that operates with segments rather than with separate pixels. The comparison with state-of-the-art methods shows that our method is successful in highlighting the mass of the object of interest. Finally, we discuss possible directions for the further work.

Index Terms— Image analysis, visual attention, visual saliency

1. INTRODUCTION

In the recent years the amount of user-generated content (UGC) has started to rise rapidly. Photo sharing services collect millions of images each day. Search within this data is a sophisticated task due to the complexity of the description of this data. While for humans it is not a problem to give a description of an image and find an object of interest, for computer-based systems this task is tough. For this reason image search results often contain irrelevant images.

To address the problem of region of interest detection, saliency has been proposed. The use of saliency allows us to estimate the importance of a particular object included in the image. Thereby we can say whether this object was captured in the frame by chance and represents an unimportant part of the background, or in opposite this object is the main subject of the image and gives a relevant description of the image.

The majority of previous works on visual saliency detection were aimed at acquiring salient pixels of images. As a result output maps included only a small part of the object of interest; thereby possible object-wise applications were limited. However, it would thus be of interest to learn whether

the integration of image segmentation in saliency detection could lead to a method that outputs the map of salient objects.

This paper is structured as follows. Section 2 describes the state of the art in visual saliency detection. In Section 3 the proposed method is described. After that, the preliminary results are presented in Section 4. Finally, our conclusions are drawn in Section 5.

2. THE STATE OF THE ART

Since the first noticeable paper on visual saliency detection by Itti et al.[1] was published, various concurrent methods applying different techniques of visual saliency detection were proposed by other researches. The basic idea underlying saliency detection is that ganglion cells are insensitive to uniform signals. Due to this reason color contrast, luminance contrast, as well as orientation dissimilarity are natural features for saliency detection, thereby they are employed by the majority of saliency detection models. These features are responsible for bottom-up attention model. In [2] the model based on multi-scale contrast was proposed. The peculiarity of this method is that the final saliency map is created using a segmentation map, by assigning to each segment a saliency value using thresholding.

Another group of methods use statistics of the image to compute saliency. In [3] the log spectrum curve of the image was proposed as a feature for saliency detection. Considering the analysis of over 1000 images the authors have shown that statistical singularities in the log spectrum are responsible for regions of interest. According to this method, the saliency map is computed by applying inverse Fourier transform (IFT) to a residual that is a difference between an average log spectrum curve of natural images and the log of the given image. In [4] a method based on Shannon's self-information measure was proposed. This method computes saliency as a local likelihood of each image patch considering the basis function learned from natural images.

The most recent methods take advantages of modern machine learning techniques and employ sophisticated feature spaces. In [5] the authors have defined four levels of features for saliency detection: low-level, mid-level, high-level and prior information. The low level employs features proposed in [1]; the mid-level includes a horizon line detector;

the high-level includes face and person detectors; prior information includes the dependence of saliency on the distance from the center of the image. The output map is generated using a support vector machine (SVM) classifier trained on a database acquired using an eye-tracker. In [6] the model based on multi-scale contrast, center-surround histograms and color-spatial distribution was presented. The output map is computed using a conditional random field (CRF) classifier trained on hand-labeled images.

3. MODEL DESCRIPTION

Examining the state of the art methods it is clear that the requirements in terms of saliency in multimedia applications, like image retrieval, scene detection and others, are not fully satisfied. The main problem is that as a rule the output map produced by a saliency detector highlights only small parts of objects of interest like edges and high-contrast points. This kind of maps sufficiently matches with maps obtained from experiments with eye trackers. The human vision system (HVS) has unattainable performance thus is able to recognize objects having incomplete data. Unlikely in computer-based systems it is essential to have the whole object of interest to achieve high accuracy. The most feasible solution for the extraction of objects from images is to perform image segmentation. In case of saliency detection there are two possible ways of applying segmentation: i) by computing a saliency map and deriving an average saliency value over a segment, and ii) by computing directly the saliency value of each segment. The advantage of the first method is that we can use any available method of saliency detection and simply apply segmentation to the output map. The advantage of the second method is that a more accurate estimation of saliency could be achieved due to the consideration of relationships of segments/objects rather than pixels. The choice of the segmentation method is an extremely crucial part of the research since the accuracy of output maps greatly depends on the accuracy of segmentation. At the same time the development and implementation of a segmentation method is an effortful task.

For this reason we have employed the public available segmentation tool [7]. Although we decided to work on the segment-wise level, the use of global features is inevitable as they describe parts of the image in the connection with each others. The global features we included are color information, luminance contrast and the center-surround histogram map. In the following these features will be described in detail.

Colors have a great impact on the perception of the image. In our work we exploit color-saliency dependency investigated in [8]. This dependency is represented by a rank table containing 12 colors versus caused saliency. We utilize a simplified table with respect to the one represented in the original work: a saliency value is defined as the position of a color, thus having the range [1, 12], whereas in [8] non-linear

weights were used. Specifically the input image is converted to CIE Lab colorspace whereupon saliency of each pixel is computed by matching its color to the closest table color and consequently to the corresponding saliency value. In addition we compute color distribution of the image. The reasoning is that the dominant color is very unlikely to be salient, as well as a very rare color is very likely to be noise. This statistics is computed in the same way as one mentioned above with the difference, that the corresponding value is defined as the ratio of the number of pixels of that color to the total number of pixels.

As it was mentioned above human attention is sensitive to contrast. For this reason luminance contrast is included into our model. Before contrast is measured, the input image is downscaled by factor 8. The motivation is that maximum contrast is usually observed on edges and glare spots, while downscaling allows the decrease in this effect. We compute luminance contrast LC as follows:

$$LC(x, y) = \sum_m \sum_n \frac{|L(x, y) - L(x + m, y + n)|}{\sqrt{m^2 + n^2}}, \quad (1)$$

where $L(x, y)$ is the luminance value of the pixel with coordinates (x, y) , and $m, n = \{-2, -1, 1, 2\}$ denote relative coordinates of neighbor pixels.

The idea to measure the distance between foreground and background for saliency detection was used in several previous works. The underlying idea is that usually the histogram of the foreground object has a larger extent than its surroundings. In our work we employ center-surround histogram filter from [6] with slight modifications. The input image is scanned by two rectangular windows R_f and R_s , both having a similar area and R_s encloses R_f (thus R_f is a notch inside the window R_s). We use the following size ratios of windows: [0.1, 0.3, 0.5], which were defined experimentally with respect to the minimum image dimension, as well as the following three aspect ratios: [1.0, 0.75, 1.5]. Specifically the distance of foreground and surrounding histograms is computed as follows:

$$dist(R_s, R_f) = \frac{1}{2} \sum \frac{(R_f^i - R_s^i)^2}{R_f^i + R_s^i}, \quad (2)$$

where R_s^i, R_f^i are surrounding and foreground histograms, respectively. Histogram distances are computed at each scale and aspect ratio. Then, they are normalized and summed into a single map. Finally, after the computation of these features, we assign an average value of each global feature to each segment of the input image.

In addition to the global features, we compute two segment-wise features: location and size. The location feature has been included into the scheme due to the fact that photographers generally place the object of interest to the center of images. The location M_S of the segment S is

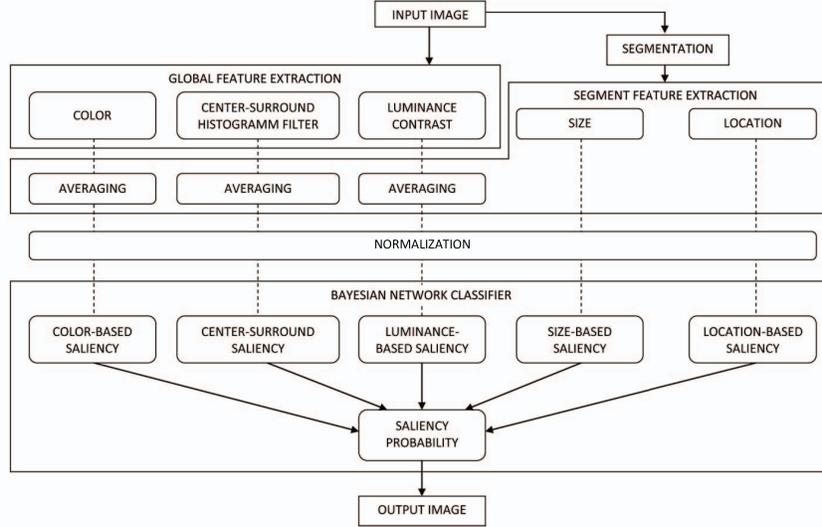


Fig. 1. The structure of the model.

computed as follows:

$$M_S = ((\sum_x \sum_y f(x,y)p(x))^2 + (\sum_x \sum_y f(x,y)q(y))^2)^{\frac{1}{2}}, \quad (3)$$

with

$$f(x,y) = \begin{cases} 1 & \text{if } x \in S \text{ and } y \in S \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

$$p(x) = \frac{mx}{2} - x, \quad (5)$$

$$q(y) = \frac{my}{2} - y, \quad (6)$$

and mx, my are the corresponding dimensions of the image. Considering size, the object of interest usually occupies a significant portion of the image. Therefore, the size of the segment could give relevant information about its saliency. It is important to note that on landscape pictures natural environment objects like sky, grass and forest occupy large area; thereby it is very unlikely that the dominant object is salient.



Fig. 2. Training images and corresponding ground truth saliency maps.

For each segment, all these features are given as input to a Naive Bayesian classifier, see Fig.1. For the learning phase we have collected a database consisting of over 500

images. The acquisition of ground truth data was performed using an application where users have been asked to select the image segments that in their opinion were salient. Thereby the ground truth data are represented by binary masks, see Fig.2. Since the database consists of a relatively small number of images, not all combinations of input features are presented in it. In order to estimate saliency probability of an unobserved combination an expectation-maximization (EM) learning method is applied. Color features node is discrete, whereas others are continuous. All continuous features are normalized to the range $[0, 1]$. The output of the classifier represents the saliency probability of a given segment. After that, the output saliency map is created by associating to the pixels of each segment the corresponding saliency probability.

4. EXPERIMENTAL RESULTS

We have compared our method with state-of-the-art methods proposed by Judd et al. [5], Achanta et al. [2] and Itti et al. [1], whose implementations are public available. The comparison has been performed on the Berkeley¹ and [6] databases, and the results are presented on Fig.3. In the results white color on maps represents the most salient region and black color the least one. The output of the method by Achanta et al. is an image filtered with saliency map, thus in order to obtain saliency maps we have filled non-masked regions of the images. As it can be seen from the results saliency maps generated by our method successfully highlight mass of objects of interest. This trend satisfies the main goal of our research. The comparison with the method proposed by Judd et al. which employs in total 33 features,

¹<http://eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/>

shows that the set of five features employed in our model sufficiently well describes the object of interest. In comparison with the method proposed by Achanta et al. which also employs segmentation, our method produces maps that contain several levels of saliency rather than binary maps, and allows several images to be scored according to saliency of the object of interest. At the same time, as in the method by Achanta et al., the segmentation method used in our model suffers from oversegmentation. For this reason some parts of objects of interest were misclassified (e.g., see the head of the bear on the top picture in Fig.3) and the estimation of saliency over an object of interest differs.

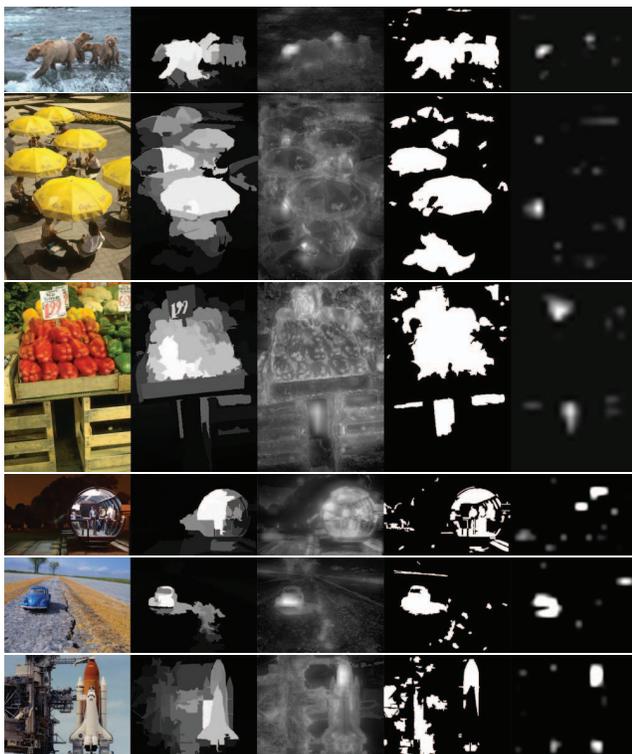


Fig. 3. Output maps. From left to right: original image, saliency maps: by the proposed method, by Judd et al. [5], by Achanta et al. [2] and by Itti et al. [1].

5. CONCLUSIONS

In this paper we presented a segment-based method of saliency detection with in-depth description of employed features. In addition we presented the comparison of our scheme with some of the state-of-the-art methods. The results show that the integration of segmentation into saliency detection allows the assessment of saliency of objects. This facility makes it possible to use this method in multimedia applications like image retrieval, scene detection and others. However, a lot of improvements can be carried out in the

proposed scheme. Concerning the problem of oversegmentation, a possible solution is to use saliency map for estimation of initial foreground and background seeds, and then use it for merging operation. In addition, on the current state the model operates only with low-level features and considers only bottom-up attention model. The addition of higher level features as well as insertion of top-down information could significantly increase prediction accuracy.

6. ACKNOWLEDGMENTS

This work has been partially supported by the E.U., Seventh Framework project LivingKnowledge (IST-FP7-231126).

7. REFERENCES

- [1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [2] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *ICVS'08: Proceedings of the 6th international conference on Computer vision systems*, Berlin, Heidelberg, 2008, pp. 66–75, Springer-Verlag.
- [3] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07)*. *IEEE Computer Society*, 2007, pp. 1–8.
- [4] J. K. Tsotsos and N. D. B. Bruce, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., MIT Press, 2006, pp. 155–162, MIT Press.
- [5] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [6] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, 2011.
- [7] C. M. Christoudias, B. Georgescu, P. Meer, and C. M. Georgescu, "Synergism in low level vision," in *International Conference on Pattern Recognition*, 2002, pp. 150–155.
- [8] E. D. Gelasca, D. Tomasic, and T. Ebrahimi, "Which Colors Best Catch Your Eyes: a Subjective Study of Color Saliency," in *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, Arizona, USA*. 2005, ISCAS, SPIE.