Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Hypergraph-based saliency map generation with potential region-of-interest approximation and validation

Zhen Liang
Hong Fu
Zheru Chi
Dagan Feng

Journal of Electronic Imaging 21(1), 013012 (Jan–Mar 2012)

Hypergraph-based saliency map generation with potential region-of-interest approximation and validation

Zhen Liang
The Hong Kong Polytechnic University
Department of Electronic and Information Engineering
Center for Multimedia Signal Processing
Hung Hom, Kowloon, Hong Kong
E-mail: liangzhen163@gmail.com

Hong Fu
The Hong Kong Polytechnic University
Department of Electronic and Information Engineering
Center for Multimedia Signal Processing
Hung Hom, Kowloon, Hong Kong
and
Chu Hai College of Higher Education
Department of Computer Science Hong Kong

Zheru Chi
The Hong Kong Polytechnic University
Department of Electronic and Information Engineering
Center for Multimedia Signal Processing
Hung Hom, Kowloon, Hong Kong

Dagan Feng
The Hong Kong Polytechnic University
Department of Electronic and Information Engineering
Center for Multimedia Signal Processing
Hung Hom, Kowloon, Hong Kong
and
The University of Sydney
School of Information Technologies
Sydney, Australia

Abstract. *A novel saliency model is proposed in this paper to automatically process images in the similar way as the human visual system which focuses on conspicuous regions that catch human beings' attention. The model combines a hypergraph representation and a partitioning process with potential region-of-interest (p-ROI) approximation and validation. Experimental results demonstrate that the proposed method shows considerable improvement in the performance of saliency map generation. © 2012 SPIE and IS&T.*
[DOI: 10.1117/1.JEI.21.1.013012]

1 Introduction

To efficiently interpret complex scenes, visual attention selects conspicuous regions in a saliency-driven manner. Inspired by the visual attention, computational models have been widely developed to stimulate the visual search progress. However, it remains a challenging task to automatically generate saliency maps for generalized tasks without a prior knowledge of image contents. Generally, bottom-up based saliency map generation methods can be categorized into three types. In a method of the first type, images are processed in terms of pixels, where each pixel is assigned a saliency value under a consideration of the contrast with nearby pixels. For example, Itti et al. proposed a biological-based model to topographically identify the visual saliency of images by respectively considering the center-surround contrasts in terms of intensity, color, and orientation.[1] Moreover, Achanta et al. introduced a frequency-tuned approach to estimate the saliency of pixels through computing the Euclidean distance between the Lab pixel vector in a Gaussian filtered image with the average Lab vector for the input image.[2] However, pixel-based methods may overlook the global associations of image

Paper 11186 received Jul. 18, 2011; revised manuscript received Jan. 9, 2012; accepted for publication Feb. 2, 2012; published online Mar. 22, 2012.

content. In a method of the second type, images are processed in terms of regions, where a pre-segmentation is carried out. For example, Fu et al. proposed a region-based attention-driven method to pop out conspicuous regions from images.[3] To better present the relationships among the regions, Park et al. introduced a pair-wise graph with a region-based center-surround distance measurement and generated the saliency maps by merging nearby regions through the region-based incremental center surround distance process.[4] However, a simple pair-wise graph representation may not be rich enough to describe a region's complex relationships and would lead to a loss of information. In a method of the third type, images are processed in frequency domain, such as the spectral residual approach proposed by Hou and Zhang.[5] One of the main drawbacks of frequency analysis is that the results would be quite sensitive to noise in the background. In this paper, a novel saliency model is proposed to estimate image saliency, where a concept of potential region-of-interest (p-ROI) is proposed to approximate hypothetic ground truths of images and a hypergraph structure is introduced to describe the spatial relationships among an arbitrary number of regions. The saliency map is then generated based on the hypergraph partitioning with p-ROI validation. Fig. 1 shows the block diagram of our proposed model.

2 Approximation of Potential Region-of-Interest

Generally, the distribution of boundaries of salient regions would tend to be continuous and concentrated. In other words, image edge detection could approximate the location of p-ROI. Let I_m be an input image. Canny edge detector is applied to detect edges and to produce an edge image denoted as I_m. Then, two filters are utilized to process the detected edges in terms of edge length and distribution. The first filter is defined as

$$A^1 = \left\{ \forall_{a_i \in A} |a_i| > \frac{M+N}{t_\alpha} \right\}, \qquad i = 1, \ldots, n, \tag{1}$$

where M and N are the width and height of I_m, t_α is a length control parameter, and $|a_i|$ is the length of edge a_i. Let the edge pixels on A^1 be $\{\vec{p}_1, \ldots, \vec{p}_m\}$. The second filter is defined as

$$A^2 = \{\forall_{\vec{p}_i, i=1,\ldots,m} \vec{p}_i \in G_1 \text{ of } \mathbf{D}\}, \tag{2}$$

where $\mathbf{D}$ is the Euclidean distances between $\vec{p}_i (i = 1, \ldots, m)$ and the centroid of A^1. Histogram analysis is used here to cluster $\mathbf{D}$ into two clusters and only the edge pixels belonging to the class G_1 with a small distance are retained. Finally, based on the retained long and concentrated edges, a Delaunay triangulation-based convex hull algorithm is used to form a smallest convex polygon which is termed as p-ROI and is denoted by R.

3 Hypergraph Representation

Psychologists have shown that people view a scene is in an object-based manner.[6] To process images in the similar way as the human visual system without any prior knowledge on the objects of interest, a pre-segmentation step is incorporated to segment images into a set of disjoint regions before feature extraction and saliency evaluation. Here, a widely accepted segmentation method, JSEG,[7] is employed, which has been demonstrated to be realizable and efficient for image interpretation.[3] Assume that I_m is represented as regions $\{r_1^{l_1}, \ldots, r_{s_1}^{l_1}; \ldots; r_1^{l_\gamma}, \ldots, r_{s_\gamma}^{l_\gamma}\}$, where $r_i^{l_j}$ is the ith segmented region at the jth scale. A hypergraph structure is used to describe the spatial relationships among an arbitrary number of regions. The hypergraph is constructed as follows: vertices represent regions; a hyperedge is formed by a region with its ρ nearest neighbor regions in I_m; the weight of a hyperedge is determined by the similarity measures among these regions.

As color is one of the most important features to characterize images, the most discriminating color representation is selected. The discriminating power of a color channel c_i is evaluated in terms of the between-class variance (S_b^i) and the within-class variance (S_w^i) defined as

$$\begin{cases} S_b^i = \sqrt{(C_{\text{dom}} - \bar{C}_{\text{dom}})^2} \\ S_w^i = \frac{\sum_{j=1}^{Q} \sqrt{(\alpha_j - C_{\text{dom}})^2}}{Q} \end{cases}, \tag{3}$$

where C_{dom} and $\bar{C}_{\text{dom}}$ are the dominant color values in R (a p-ROI) and its compliment $\bar{R}$ ($\bar{R} = I_m - R$), Q is the size of R, and α_j is the color value of the jth pixel in R. The most discriminating color representation c^* is determined by

$$c^* = \text{argmax}_{c_i \in C} \frac{S_b^i}{S_w^i}, \tag{4}$$

where C is a set of color channel candidates containing RGB, hue-saturation-value (HSV), opponent colors, and Lab. Note that the dominant color is selected as the color value which

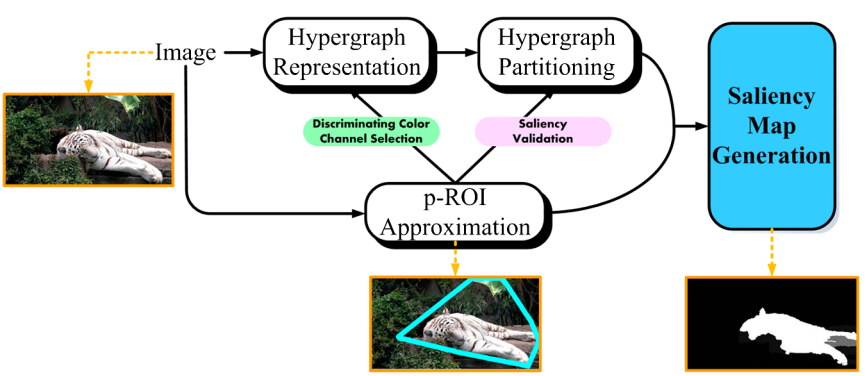

Fig. 1 The block diagram of our proposed saliency map generation method.

maximizes the occurrence probability in a 10-bin color histogram. Then, I_m is represented in c^* and a feature vector, including color histogram, dominant color, color texture, and spatial feature, is extracted from each vertex (region). Finally, a hyperedge weight is defined as

$$w(e_i) = \frac{\sum_{v_j \in e_i, v_j \neq v_i} S(v_i, v_j)}{\rho}, \tag{5}$$

where $S(v_i, v_j) = e^{-\mathrm{dist}(v_i, v_j)}$ is the similarity between vertices v_i and v_j, where $\mathrm{dist}(v_i, v_j)$ is a normalized Euclidean distance of the feature vectors extracted from v_i and v_j.

4 Generation of Saliency Map

Saliency map generation can be considered as a hypergraph partitioning problem, which seeks for the hyperedges with strong connectivity. A hypergraph Laplacian Δ was derived in Ref. 8. Conventional hypergraph partitioning is to find the first several eigenvectors with the smallest non-zeros eigenvalues in Δ and classify the hypergraph into a fixed number of clusters. In our method, an improved version of hypergraph partitioning, Iterative Spectral Hypergraph Partitioning (ISHP), is proposed. Instead of using a fixed number of clusters, Δ is iteratively cut into 2 to K clusters and the corresponding clustering results are given by

$$\{M_i^1, \ldots, M_i^j, \ldots, M_i^i\}, \quad i = 2, \ldots, K, \quad 1 \le j \le i, \tag{6}$$

where M_i^j is the jth clustering result when Δ is cut into i clusters. To validate the saliency of the clustering results by using R (a p-ROI), a validation function is defined as

$$F_{M_i^j} = \frac{\beta S_M}{\langle R, M_i^j \rangle + \langle R, O \rangle + \zeta}, \tag{7}$$

where $O = M_i^j \cap R$ is an overlapped area of M_i^j and R, and S_R, S_M and S_O are respectively sizes of R, M_i^j and O. The ratio $\beta = \frac{S_O}{S_R}$ is used to measure the portion of the overlapped region O over R, and $\zeta = S_M - S_O + 1$ is to measure a mismatched degree between M_i^j and R, where 1 is used here to avoid a division by zero in Eq. (7). $\langle R, M_i^j \rangle$ and $\langle R, O \rangle$ are the compactness properties of M_i^j and O with respect to R, which is measured by summing the Euclidean distances from every pixel to the centroid of R. As R is considered as a hypothetic ground truth of the image, a good clustering result should be compact and overlap with R as much as possible. Thus, from the clustering results $\{M_i^1, \ldots, M_i^i\}$, the one with the highest validation value, M_i^*, is selected:

$$M_i^* = \mathrm{argmax}_{j=1,\ldots,i} F_{M_i^j}. \tag{8}$$

As Δ is iteratively cut into 2 to K clusters, there are in total $(K-1)$ selected clustering results $\{M_2^*, \ldots, M_K^*\}$. The final Saliency Map SM is generated by

$$SM = \frac{\sum_{i=2}^{K} M_i^* \times F_{M_i^*}}{K-1}. \tag{9}$$

The steps for saliency map generation is illustrated in Fig. 2.

Step 1:	Derive the hypergraph Laplacian Δ;
	For $i = 2 : K$
Step 2:	Partition the hypergraph Laplacian Δ into i clusters;
Step 3:	Evaluate each clustering result by Eq. 7;
Step 4:	Select the clustering result with the highest saliency validation by Eq. 8, when Δ is cut into i clusters;
	End
Step 5:	Generate the final saliency map by the weighted average of all the selected clustering results by Eq. 9.

Fig. 2 Detailed steps for saliency map generation.

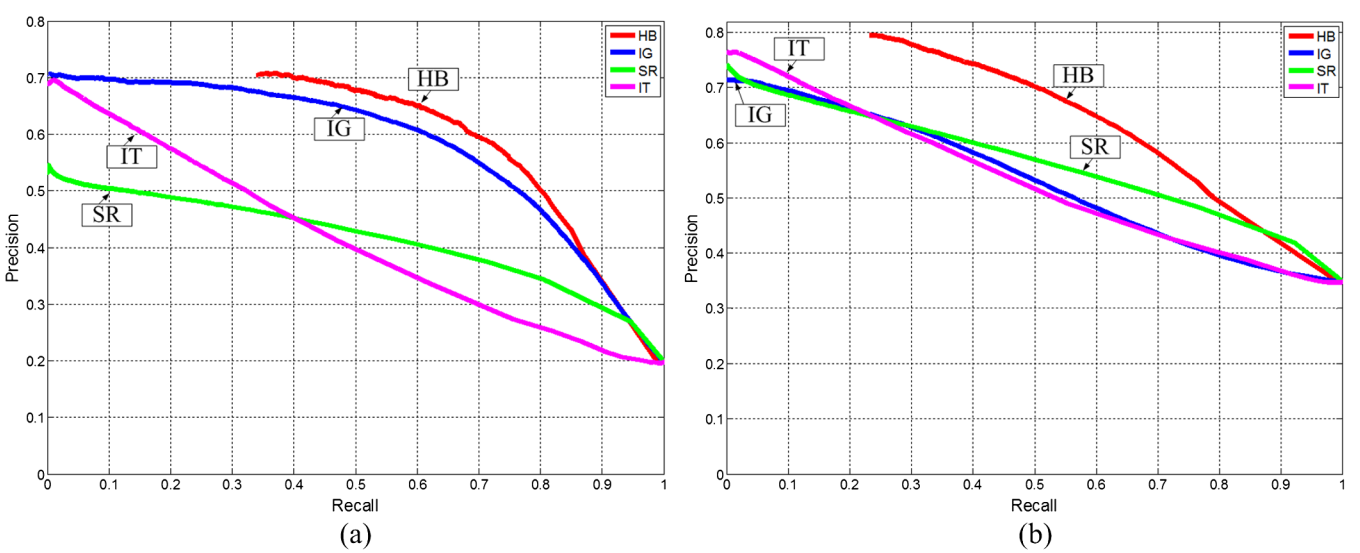

Fig. 3 Objective evaluations by using (a) an accurate object contour-based ground truth of 1000 images, and (b) the median of nine users' labelled rectangle-based ground truth of 5000 images.

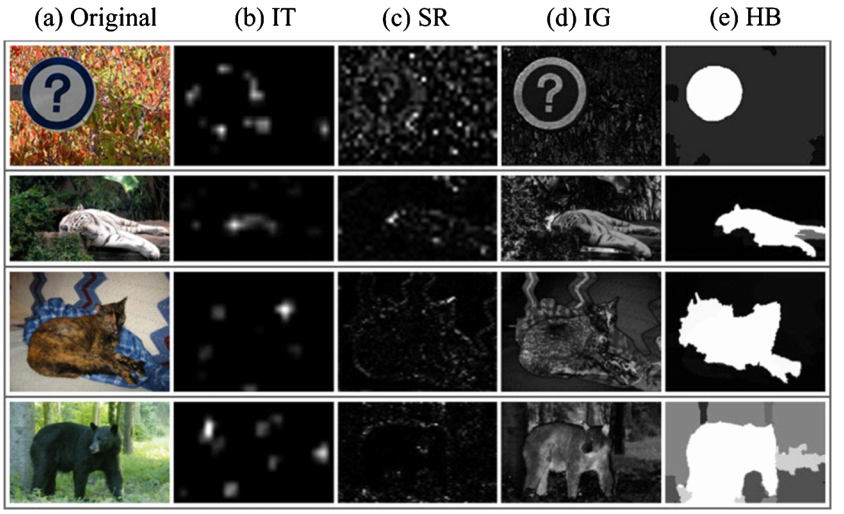

Fig. 4 A visual comparison of saliency maps generated by (b) Itti's model,[1] (c) the spectral residual method,[5] (d) the frequency-tuned method,[2] and (e) our method.

5 Experimental Results and Discussion

Two databases with 1000 and 5000 images, respectively, were used in our experiments to objectively evaluate the performance of our proposed method. In the first database with 1000 images, the ground truth is accurate object-contours created by Achanta et al.[2] For the second database with 5000 images (MSRA database) constructed by Liu et al.,[9] the ground truth is the median of nine users' labelled rectangles. To reliably and objectively evaluate the performance, a fixed threshold ranging from 0 to 255 is used and a corresponding precision-recall curve is obtained, which provides a reliable comparison of how well various saliency maps highlight salient regions in images.[2] Objective evaluation, in terms of precision-recall curve, is conducted against three state-of-the-art methods. The corresponding objective evaluation results of two databases are shown in Fig. 3, where the methods compared are IT,[1] SR[5] and IG,[2] and our proposed method is labelled as HB. It is observed that our proposed method performs better, especially in the larger database. On the other hand, a visual comparison of the generated saliency maps is also shown in Fig. 4. More comparisons between our proposed method and three existing methods are available on our website,* Experimental results demonstrate that our proposed model can generate more precise and complete saliency regions and is less sensitive to ambiguous backgrounds compared with the three existing methods.

The main reason for a superior performance of our method over the three existing methods is that in our model images are interpreted as regions and a hypergraph structure is introduced to successfullty represent the complex relationships among an arbitrary number of regions. For an ideal result, a high and constant saliency value should be assigned over the entire object but without any saliency on the background, such as the tiger image shown in the second row of Fig. 4. However, for some images, we may not be able to obtain ideal saliency maps. For example, the bear image shown in the last row of Fig. 4, we can see that the saliency is distributed over the whole image, but still the saliency of the object is slightly higher than that of the background.

6 Conclusion

A novel bottom-up based saliency map generation method is proposed in this paper. Our method which combines a hypergraph representation and a partitioning process with p-ROI approximation and validation, can reliably detect conspicuous region(s) in natural images. Experimental results show the effectiveness and robustness of our proposed method.

Acknowledgments

The work reported in this paper is substantially supported by the Hong Kong Polytechnic University (Project code: G-U750).

References

1. L. Itti, C. Koch, and E. Neibur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998).
2. R. Achanta et al., "Frequency-tuned salient region detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, 20–25 June, pp. 1597–1604 (2009).
3. H. Fu, Z. Chi, and D. Feng, "Attention-driven image interpretation with application to image retrieval," *Pattern Recogn.* **39**(9), 1604–1621 (2006).
4. M. Park, A. C. Loui, and M. Kumar, "Saliency detection using region based center-surround distance increase," in *IEEE International Symposium on Multimedia*, California, 5–7 December, pp. 249–256 (2011).
5. X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, 17–22 June, pp. 1–8 (2007).
6. A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *J. Vis.* **10**(8), 1–19 (2010).
7. Y. Deng and B. Manjunath, "Unsupervised segmentation of color-texture regions in images and videos," *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(8), 800–810 (2001).
8. D. Zhou, J. Hung, and B. Scholkopf, "Learning with hypergraphs: clustering, classification, and embedding," in *Annual Conf. on Neural Inform. Process. Syst. (NIPS)*, Vol. **19**, , Vancouver, B.C., Canada, 4–9 December, pp. 1601–1608 (2006).
9. T. Liu et al., "Learning to detect a salient object," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, 17–22 June, pp. 1–8 (2007).

*https://sites.google.com/site/janezhenliang/saliencymap