Object Popping-out and Characterization Based on the Human Visual Mechanism

Hong Fu[1], Zheru Chi[1], and Dagan Feng[1,2]

[1]Center for Multimedia Signal Processing
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[2]School of Information Technologies
The University of Sydney, NSW 2006, Australia

[1]{enhongfu, enzheru, enfeng}@eie.polyu.edu.hk
[2]feng@it.usyd.edu.au

Abstract

Semantic image understanding is the basis of a well-performed image management system on a large database whereas current image representations are neither powerful nor semantic-based. In this paper, the human vision mechanism from psychological studies is employed in order to understand an image at a higher level by constructing a layered representation structure based on the image segmentation result. An iterative object popping-out algorithm is proposed to locate the visually attentive objects from a downsized image by maximizing a global attention function designed according to the human's attention model. The extracted objects are then characterized by using the original image while leaving the background as a rough description so as to construct a layered representation of the image. Promising results that agree with human's intuition are obtained, showing the potential of our approach in image retrieval and other related aspects of image management.

Keyword: Human's attention model, object detection, image representation.

I. Introduction

The rapid development of multimedia technologies has resulted in a vast growing of digital images. In the past decades, many researches have been focused on the management of large scale image databases [1], whose performance relies on a meaningful understanding of the image. Image segmentation, as a preliminary procedure of image understanding, partitions an image into regions with local similarities. As a result, a real object is usually over-segmented due to some trivial varieties on color or texture caused by the slight change of illumination or the un-uniformity of the object itself. Moreover, the segmentation usually does not have semantic meaning by itself, i.e., it is normally impossible to identify which segment is object and which segment is background even if the segmentation happens to be right.

Visual attention is an intuitive process that enables human beings to focus on certain objects of a scene with a mass of information. Disclosed by psychologists, it is composed of two procedures: pre-attentive and attentive stage [2-4]. At the un-capacity limited pre-attentive stage, the original stimulus is processed in parallel and a coarse representation is obtained. Then the attentive stage is processed to extract a certain area from the coarse representation for the more detailed analysis. The extracted area is called attended area or attended object. One question frequently asked is what attracts our attention. It is believed that distinct properties of an object that makes it different from the surroundings captures our eyes, e.g., color contrast, texture contrast, movement, etc. Wolfe and Horowitz summarized the attributes that might influence the deployment of attention [5]. Many attempts have been made to model and simulate the attention process [6-10, 13-14]. The investigations reported in [6] demonstrated that the bottom-up attention was useful for object recognition to some extent. By fusing centre-surround differences of multi-scale features, Itti *et al* [7, 8] produced a saliency map that indicated the saliencies of pixels. A dynamical neural network was then used to select attended locations. Itti et al's model was shown to be able to locate some objects such as a traffic sign from a natural image. Han *et al* [9] proposed an object extraction scheme by using a seed growing technique monitored by a saliency-based measure. A novel "hierarchical selectivity" mechanism for object-based visual attention was presented by Sun *et al* [13, 14], wherein both object-based and space-based selections are integrated to give a visual attention mechanism that has multiple and hierarchical selectivity.

In this paper, the human vision mechanism resulted from the relevant psychological studies is employed in order to understand an image at a higher level by recomposing the image segmentation result. An iterative object popping-out algorithm is proposed to find the visual attractive objects from an image by maximizing a global attention function designed according to the human's attention model. The extracted objects are then characterized in detail by using the original image again, driven by the "spotlight" function of human vision [2]. As a result, a meaningful layered representation that emphasizes the remarkable objects while degrades the unimportant background is constructed.

II. Iterative Object Popping-Out Process

Attention, as an intuitive aid of the human vision, provides the primary but important interpretation on the stimulus. An iterative object popping-out algorithm is proposed to extract the remarkable areas sequentially which are probably the real objects in the photographer/user's mind, as shown in Fig. 1. The input image is firstly segmented into several regions by a state-of-art image segmentation algorithm, JSEG [11]. Then the adjacent matrices are calculated to describe the relationships between adjacent regions. An iterative attention process is proposed to search the combination of regions with maximal attention value in each cycle.

A. Image Segmentation

Image segmentation is to group the pixels according to their similarities, which is a pre-clustering for the sub-sequential attention process. JSEG [11], a segmentation algorithm considering both color and texture information, is adopted currently. It is supposed that the

segmentation result is at least over-segmented. That is to say, one of the combinations of the regions should be an object, or it is possible to construct an object using the existing regions.
Suppose a color image I is segmented into N regions, i.e.

$$I = \{p_i\}, i = 1,..., N \tag{1}$$

B. *Adjacent Matrices*

Adjacent matrices include boundary length matrix and feature matrix, which is introduced for describing the relationships among regions. The former records the spatial relationship and the latter records the difference of features between regions.

Boundary length matrix. The boundary length matrix reflects the length of all boundaries among regions, whose diagonals are the length of outer boundary of a region and the non-diagonals are the length of boundary among adjacent regions. The definition of the boundary length matrix L is

$$L = \{l_{i,j}\}, \tag{2}$$

where
$l_{i,j}$ is the length of boundary between region i and region j, if $i \neq j$;
$l_{i,j}$ is the length of the outer boundary of region i, if $i = j$.
If two regions i and j are not adjacent, $l_{i,j} = 0$.

Feature matrix. The adjacent feature salient matrix is to tell how different a region to its surroundings in terms of features including color and texture. The feature adjacent matrix is

$$A^k = \{a^k_{i,j}\} \tag{3}$$

where A^k is the k^{th} feature matrix;
$a^k_{i,j}$ is the difference between region i and j in terms of the k^{th} feature; $k = 1,..., K$.

Here $K = 3$, i.e. pixel color, region color and region texture are used as three features. Color could be the property of either a pixel or a region. The element of the first adjacent feature matrix A^1 is the mean color difference between adjacent boundary pixels belong to two regions. The element of the second adjacent feature salient matrix A^2 is the difference between the mean colors of two adjacent regions. The color difference is defined as the Euclidean Distance in the HSV space. Texture is the feature of a region, so the element of the third adjacent feature matrix A^3 is the difference between the regions of both sides in terms of texture.

C. *Iterative Object Popping-Out*

The reason an object is attractive is that it is different from its surroundings. Here “different” means that the differences both inside the object and its surroundings are small and that the difference between the object and its surroundings is large.
Let G be the set of all the combinations of N regions, find one valid combination C whose attention value $F(C)$ is the maximum.

$$F(C) = D(C, \overline{C}) - \frac{D(C) + D(\overline{C})}{2} \tag{4}$$

where
$G = \{C_1, C_2,..., C_M\}$, M is the number of all the combinations of N regions. [12]
$C \in G$ is a valid combination which satisfies the following conditions.

1) $C \neq \Phi$;
2) $C \neq \{1,2,...,N\}$;
3) C is a connected combination since an object is assumed to be a continuous area.

$\overline{C}$ is the supplementary set of C;

$D(C)$ and $D(\overline{C})$ is the intro-difference of C and $\overline{C}$, respectively;

where

$$D(X) = \frac{1}{K}\sum_{k}\frac{\sum_{i}\sum_{j} a_{i,j}^{k} l_{i,j}}{\sum_{i}\sum_{j} l_{i,j}},\ i, j \in X, i \neq j; \tag{5}$$

$D(C,\overline{C})$ is the inter-difference between C and $\overline{C}$;

where

$$D(C,\overline{C}) = \frac{1}{K}\sum_{k}\frac{\sum_{i}\sum_{j} a_{i,j}^{k} l_{i,j}}{\sum_{i}\sum_{j} l_{i,j}},\ i \in C, j \in \overline{C}. \tag{6}$$

The iterative scheme is

Step I: Find the combination of regions C^t whose attention value $F(C^t)$ maximum. Pop out C^t as the t^{th} object;

Step II: Delete the lines and columns of L and A corresponding to C^t;

Step III: If t>=T, stop; else go to Step I, t=t+1.

In this paper, T is set to 2, i.e., two objects will be detected at most, because we suppose that the number of leading actors in a common picture doesn't exceed two. The iterative attention process on the image "Big Ben Clock Tower and Street Lamp" is illustrated as Fig. 2. At the first iteration, the clock tower is popped out from 38 possible combinations with attention value 0.31. Then the number of left possible combinations is reduced to 11 after getting rid of the first object. The second object is found with attention value 0.14.

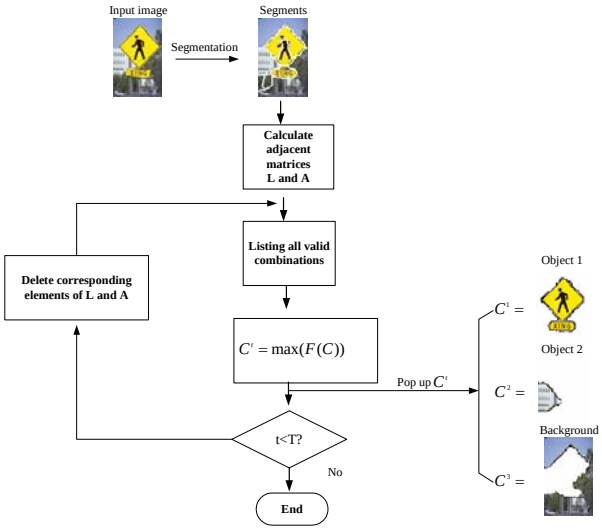

Fig. 1. The iterative object popping-out process.

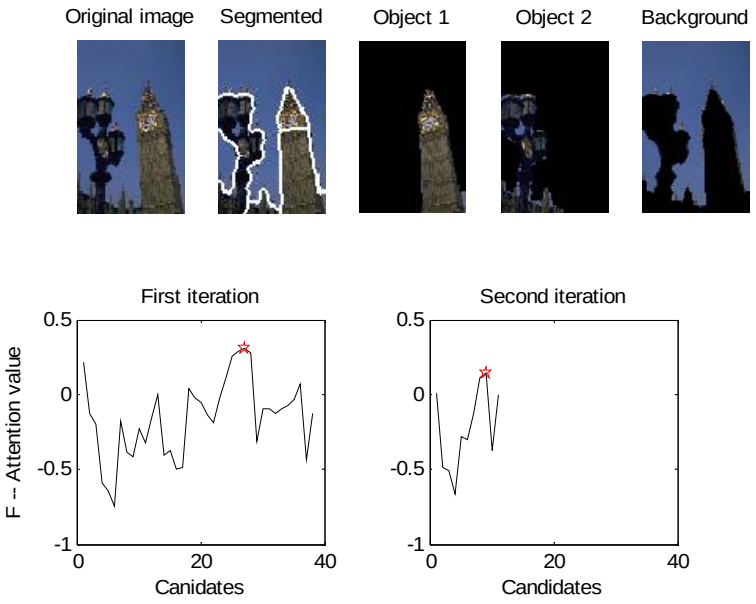

Fig. 2. The iterative attention process on the image "Big Ben Clock Tower and Street Lamp".

III. Layered Representation

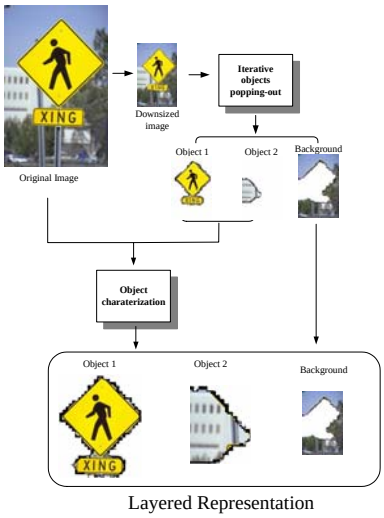

Fig. 3. Layered representation of image contents.

Layered Representation strategy that imitates the "spotlight" procedure of human vision [2] is proposed, as shown in Fig. 3. At first, the image is downsized and inputted into the object popping-out process. The precise representations of the objects are obtained by going back to the original image to extract their features. On the other side, the background is roughly represented with the information from the downsized image. The reasons that the downsized image instead of original one is used for object popping-out are,

- Before focusing on a certain object, people will not pay attention to the details of the image;
- A downsized image helps to save computational time.

The advantage of this layered representation strategy is obvious: it highlights attractive aspects while degrades insignificant aspects of an image as a human does, which may be potentially meaningful for the sub-sequential operations, such as image retrieval.

IV. Experimental Results and Discussion

We tested our object popping-out algorithm on the Hemera images. The images are resized to smaller ones with a width of 100 pixels. The average computing time is about 10 second with MATLAB 6.0 on a Pentium IV 3GHz PC. The average running time of each step is listed in Table 1. Some examples are illustrated in Fig. 4. We can see that the results accord with human's intuition pretty well.

Table 1. Average running time (seconds) of 700 randomly selected images

Iterative object popping-out process			Object characterization	Total
Segmentation	Computing adjacent matrices	Iterative object popping-out		
0.24	0.91	0.69	8.3	10.14

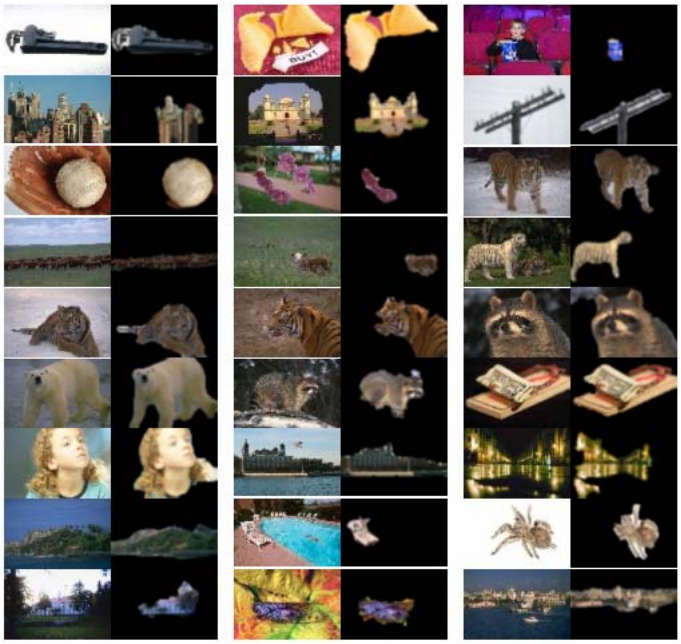

Fig. 4. Some results obtained by using our object popping-out algorithm.
The input images (odd columns) and the first found objects (even columns) are shown.

Subjective Evaluation: We perform evaluations on 100 images randomly selected from 7,376 Hemera images, covering 19 categories such as "agriculture & industry", "education", "nature", "people", etc. Two subjects were invited to evaluate the quality of the object popping-out algorithm. They were shown the original downsized image and the first popped-out object, and asked to give a grade from 0 to 1 (0: least satisfactory, 1: most satisfactory). Since the performance of our method depends on the segmentation result, we also grade the quality of the segmentation result from 0 to 1

(0: impossible to construct an object by the obtained segments, 1: possible to construct an object by the segments). The result is shown in Table 2. Our method takes effect on 59% of the testing images. For the unsatisfactory cases, possible reasons are:

- Poor segmentation. An example is shown in Fig. 5.
- Variations caused by the preferences of different subjects. Different subjects may be interested in different objects contained in the same image. For example, for the image shown in Fig. 6, one subject may be interested in the red apple while another subject considers the girl as the main object in the image.
- There is no obvious object in the image, such as some landscape images shown in Fig. 7.

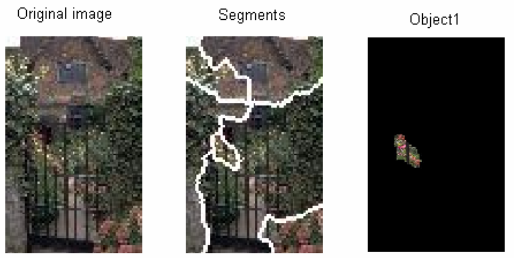

Fig. 5. Failed case caused by poor segmentation.

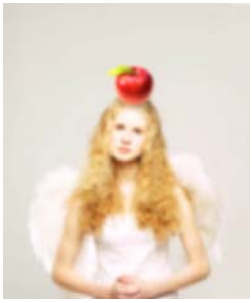

Fig. 6. Which one is the main object, the apple or the girl?

Fig. 7. Images without obvious objects.

Table 2. Performance evaluation

		Score of image segmentation		Total
		0~0.5	0.5~1	
Score of object popping-out	0~0.5	11	30	41
	0.5~1	4	55	59
Total		15	85	100

V. Conclusion

Driven by the human vision mechanism, an iterative object popping-out algorithm and an object-based layered image representation are presented in this paper. Promising results that agree with human's intuition are obtained by testing our proposed approach on the Hemera images. Our future works will focus on attention-oriented image retrieval and object refinement responding to user's relevance feedback.

Acknowledgements

The work reported in this paper is partially supported by a research grant from the Hong Kong Polytechnic University (Project code: A-PE45)

References

[1] Long, F., Zhang, H. and Feng, D.: Fundamentals of Content-Based Image Retrieval. In: Feng, D. Siu, W.C., and Zhang, H. (eds.): Multimedia Information Retrieval and Management, Springer-Verlag. (2003) 1-26

[2] Theeuwes, J.: Visual Selective Attention: A Theoretical Analysis. Acta Psychologica. Vol. 83. (1993) 93-154

[3] Steinman, S., and Steinman, B.: Chapter 14: Computational Models of Visual Attention. In Hung, G. and Ciuffreda, K. (eds.): Models of the Visual System, Kluwer Academic/Plenum Publishers. (2002) 521-563

[4] Wolfe, J.: 9: The Level of Attention: Mediating Between the Stimulus and Perception. In Harris L. and Jenkin M. (eds.): Levels of Perception. Springer. (2002) 169-191

[5] Wolfe, J. and Horowitz, T.: What Attributes Guide the Deployment of Visual Attention and How Do They Do It?. Nature Reviews Neuroscience. Vol. 5. June (2004) 1-7

[6] Rutishauser, U., Walther, D., Koch, C. and Perona, P.: Is Bottom-up Attention Useful for Object Recognition?. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'04). Vol. 2. July (2004) 37-44

[7] Itti, L., Koch, C., and Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20. No. 11. November (1998) 1254-1259

[8] Itti, L. and Koch C.: Computational Modeling of Visual Attention. Nature Reviews Neuroscience. Vol. 2. March (2001) 194-203

[9] Han, J., Li, M., Zhang, H. and Guo, L.: Automatic attention object extraction from images. International Conference on Image Processing. Vol. 2. September (2003) II 403-406.

[10] Dickinson, S., Christensen, H., Tsotsos, J., and Olofsson, G.:Active Object Recognition Integrating Attention and Viewpoint Control, Computer Vision and Image Understanding. Vol. 67., No. 3. September (1997) 239–260

[11] Deng, Y. and Manjunath, B.S.: Unsupervised Segmentation of Color-Texture Regions in Images and Video. IEEE transactions on Patten Analysis and Machine Intelligence. Vol. 23. No. 8. August (2001) 800-810

[12] Joseph Straight, H.: Combinatorics, An Invitaion. Pacific Grove. (1993)

[13] Sun, Y. and Fisher, R.: Hierarchical selectivity for object-based visual attention. Proc. 2nd Biologically Motivated Computer Vision Workshop (BMCV 2002), Tuebingen, Germany, November 2002, pp 427-438, Aka Springer LNCS 2525.

[14] Sun, Y. and Fisher, R.: Object-based visual attention for computer vision. Artificial Intelligence, Vol. 146, pp. 77-123, 2003.

Ms. Hong Fu received her Bachelor and Master degrees from Xi'an Jiaotong University in 2000 and 2003, respectively. From Dec. 2002 to Nov. 2003, she worked as a Research Assistant in the Department of Electronic and Information Engineering at the Hong Kong Polytechnic University. Since Dec. 2003, she has been a full-time PhD research student in the same department at the Hong Kong Polytechnic University. Her research interests include image processing, pattern recognition, and artificial intelligence.

Dr. Zheru Chi received his BEng and MEng degrees from Zhejiang University in 1982 and 1985 respectively, and his PhD degree from the University of Sydney in March 1994, all in electrical engineering. Between 1985 and 1989, he was on the Faculty of the Department of Scientific Instruments at Zhejiang University. He worked as a Senior Research Assistant/Research Fellow in the Laboratory for Imaging Science and Engineering at the University of Sydney from April 1993 to January 1995. Since February 1995, he has been with the Hong Kong Polytechnic University, where he is now an Associate Professor in the Department of Electronic and Information Engineering. Since 1997, he has served on the organization or program committees for a number of international conferences. His research interests include image processing, pattern recognition, and computational intelligence. Dr. Chi has authored/co-authored one book and nine book chapters, and published more than 140 technical papers.

Prof. (David) Dagan Feng received his ME in Electrical Engineering & computing Science (EECS) from Shanghai JiaoTong University in 1982, MSc in Biocybernetics and Ph.D in Computer Science from the University of California, Los Angeles (UCLA) in 1985 and 1988 respectively. After briefly working as Assistant Professor at the University of California, Riverside, he joined the University of Sydney at the end of 1988, as Lecturer, Senior Lecturer, Reader, Professor and Head of the Department of Computer Science/ School of Information Technologies. He is currently Associate Dean of the Faculty of Science at the University of Sydney; Honorary Research Consultant, Royal Prince Alfred Hospital, the largest hospital in Australia; Chair-Professor of Information Technology, Hong Kong Polytechnic University; Advisory Professor, Shanghai JiaoTong University; Guest Professor, Northwestern Polytechnic University, Northeastern University and Tsinghua University. His research area is Biomedical & Multimedia Information Technology (BMIT). He is the Founder and Director the BMIT Research Group. He has published over 400 scholarly research papers, pioneered several new research directions, made a number of landmark contributions in his field with significant scientific impact and social benefit, and received the Crump Prize for Excellence in Medical Engineering from USA. He is a Fellow of ACS, ATSE, HKIE, IEE, and IEEE, Special Area Editor of IEEE Transactions on Information Technology in Biomedicine, and is the current Chairman of IFAC-TC-BIOMED.