Region-Based Visual Attention Analysis with Its Application in Image Browsing on Small Displays

Huiying Liu[1, 2], Shuqiang Jiang[1], Qingming Huang[1, 2], Changsheng Xu[3], Wen Gao[1, 4]

[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, Beijing 100080, China

[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

[3]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore

[4]Institute of Digital Media, Peking University, Beijing 100871, China

Email: {hyliu, sqjiang, qmhuang, wgao}@jdl.ac.cn, {xucs}@i2r.a-star.edu.sg

ABSTRACT

Visual attention has been a hot research point for many years and many new applications are emerging especially for wireless multimedia services. In this paper a novel region-based visual attention is proposed to detect the Regions of Interest (ROI) of images. In the proposed method, density based image segmentation is first performed by regarding region as the perceptive unit, which makes the model robust to the scale of ROIs and contains more perceptive information. To generate region saliency map to detect ROI, global effect and contextual difference are covered in the form of distance factor and adjacency factor respectively. Since different ROIs may have different importance for different purposes, a ROI ranking algorithm is designed for browsing large images on small displays. Experimental results and evaluation reveal that our method works effectively to detect ROIs from images and the users are satisfied with the browsing sequence on small displays.

Categories and Subject Descriptors

I.4.9 [**IMAGE PROCESSING AND COMPUTER VISION**]: Applications; I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding –*Perceptual reasoning*

General Terms

Algorithms, Experimentation

Keywords

Visual attention, ROI ranking, Image adaptation

1. INTRODUCTION

Visual attention is one of the most important features of Human Visual System (HVS). It can filter the signals received by the HVS and select the important ones to be processed. Visual attention has two mechanisms, bottom-up and top-down, correspond to stimulus-driven and object-driven respectively. The bottom-up mechanism is closely related to human sensory system, which is sensible to the contrast including both global effect and contextual difference. The top-down mechanism is closely related to human brain, which is a much complex mechanism. In this paper, we focus on the bottom-up mechanism.

Computational attention model enables computers to understand images and videos in the manner of HVS. It has been a hot research point for years and has been successfully applied in many areas including image coding, progressive image transformation, and scene analysis. Recently, with the development of mobile devices and multi-mode internet services, some new applications are emerging, such as adaptive video browsing [1, 2] and image browsing [3, 4] on small displays, which come forth to overcome the gap between the high resolution videos (or images) and the small mobile displays. To improve the displaying effect on mobile devices, attention model is employed to detect the most interesting regions to be displayed [4]. Recently, a novel method was proposed [3] to aid or automate common image browsing tasks on mobile devices, in which attention model is used to obtain the minimal perceptible time and the optical browsing path to mimic human watching process. This method can improve the display effect of large images with dispersive ROIs.

Most current computational attention models place emphasis on bottom-up mechanism [5-9] for its general applications. Bottom-up models are mainly saliency based, which calculate the saliency of the perceptive units to construct saliency maps. Here perceptive unit is defined as a set of pixels perceived as a union in attention analysis. Pixel can be chosen as the perceptive unit using the statistical information of the image [5] and a coherent computational approach [6] to analyze bottom-up attention. The perceptive unit can also be a block [7, 8] or a region [9, 10]. In the case of block Gaussian pyramid is used to yield multi-scale feature extraction [7]. A region can be of any size, and can be extracted using subspace analysis [9] or K-means clustering [10]. In these models, either global effect [5, 9] or contextual difference [6-8] is considered, but they are not integrated together in the literatures. To improve the performance of ROI detection, the top-down attention is integrated with bottom-up models [8, 11, 12]. For example, face detection is used to improve the model's performance in general applications [8], and machine learning based object detection is used to optimize detection speed [11] or to predict gaze transition [12].

In this paper we try to analyze visual attention more perceptively by choosing region as the perceptive unit. Different from the methods of [9, 10], image segmentation is first performed in our

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

method to obtain the region information. Then the saliency of each region will be calculated to obtain the region saliency map. Finally, there may be multiple ROIs with different attractiveness in an image, which will be stared at one by one in the order of descending attractiveness. So based on the ROI detection result, an ROI ranking method is proposed to forecast the human watching process and is used in adaptive image browsing on small displays. The framework of the proposed model and its application is illustrated in Figure 1.

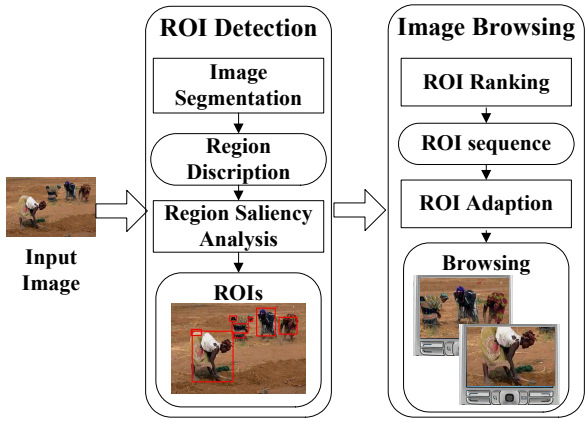


Figure 1.The framework of the proposed method.

The contribution of this paper can be summarized in the following 3 aspects. Firstly, region is chosen as the perceptive unit, which makes the method more effective in terms of perception. Since a region can be of any size, the method can detect the ROIs of any scale. Secondly, the coverage of both global effect and contextual difference is closer to the property of HVS and makes the method robust to complex background. Finally, a ROI ranking method is proposed and used in adaptive image browsing, which generates an image sequence from the input image and displays it on small screens.

The rest of this paper is organized as follows: Section 2 presents the proposed attention analysis method. Section 3 provides the ROI ranking method and image adaptation for image browsing. Section 4 shows the experimental results and in section 5 we will conclude the work.

2. Region-Based Attention Analysis

It is easy to find out that while watching an image we are looking at it neither pixel by pixel nor block by block, but concentrating on the objects in it. An object may be something we are familiar with, or something we just know it is an "object" but do not know what it actually is. In color images, an object is composed of one or more regions, or in other words, a region is a unit between a pixel (or a block) and an object. It contains more perceptive information than a pixel or a block and can be obtained by image segmentation, which is much easier than object detection. Therefore we choose region as the perceptive unit in our model.

2.1 Density-Based Image Segmentation

The segmentation method in [13] is adopted for its effectiveness in terms of human perception. This segmentation method integrates the spatial connectivity and color feature of the pixels to cluster them into different groups. It uses density-based clustering to discover the spatial connectivity and measures the color similarity in Munsell color space to ensure the perceptive smoothness of color change in regions.

2.2 Region Saliency Map

After image segmentation, each region is represented as its average feature, which in this paper is the RGB color. A region's saliency is determined by its position factor and the sum of its contrast compared with the other regions. The contrast between two regions may be simply feature contrast [7, 8]. In this paper a general feature contrast is needed, hence the one used in [8] is adopted, which is the Gaussian distance between two colors and defined as:

$$FD_{i,j} = \left(1 - \exp\left(-d_{i,j}/2\sigma^2\right)\right) \times 255 \quad (1)$$

where $d_{i,j}$ is the Euclidian distance between color i and color j.

However, in addition to the feature contrast, HVS has other sensitive factors including area factor, global effect and contextual difference, and central effect. These factors are presented below.

Area factor: It is obvious that larger regions will have lager effect to others. This is represented as area factor which is simply the ratio of the area of the region to the area of the image and is represented as:

$$\theta_1(A_i) = \frac{A_i}{AreaOfImage} \quad (2)$$

Global effect: As discussed in Section 1, a region's saliency includes both global effect and contextual difference. Global effect is achieved through statistical information in [5] and through cumulative projection in [9]. In these two methods, each unit, the pixel in [5] and the region in [9], has the same weight to other units, which is inconsistent with HVS. In HVS, the attractiveness of a unit is affected by the nearer neighbors much more than by the farther ones. This property is represented as a distance factor, which is a Gaussian function of the spatial distance between units, mean regions here. The factor is referred as $\theta_2(SD_{i,j})$, which is calculated as:

$$\theta_2(SD_{i,j}) = 1 - \exp\left(-SD_{i,j}^{\ 2}/2\sigma^2\right) \quad (3)$$

where $SD_{i,j}$ is the relative spatial distance between Region i and Region j, normalized to $[0,1]$.

Contextual difference: Different from the pixel based or block based method, in which a unit's contrast is determined equally by the neighboring units, the contextual difference between regions has a coefficient of weight proportional to the adjacency degree between them. An adjacency factor is used to describe this mechanism, calculated as:

$$\theta_3(E_{i,j}) = 1 + \frac{E_{i,j}}{Edge_i} \quad (4)$$

where $E_{i,j}$ is the number of neighbor pixels between the two regions and $Edge_i$ is the total number of edge pixels of Region i.

Central effect: It is referred as central effect that while watching an image observers have a general tendency to stare at the central locations. In [6] this phenomenon is modeled to evaluate the

position of each pixel and in [9] cumulative projection provides higher attention value to the central regions. Here a position factor is used to evaluate the central effect, which is represented as a Gaussian function:

$$\theta_0(P_i) = 1 - \exp\left(-\frac{P_i^2}{2\sigma^2}\right) \quad (5)$$

where P_i is the relative distance of the region away from the center of the image and is normalized to $[0,1]$, and σ determines the saliency of marginal regions.

Considering all the above factors, a region's saliency can be calculated as:

$$S_i = \theta_0(P_i) \times \sum_{j=0}^{N-1} \left(FD_{i,j} \times \theta_1(A_j) \times \theta_2(SD_{i,j}) \times \theta_3(E_{i,j}) \right) \quad (6)$$

Figure 2 shows an example of the saliency map and the ROI detection result obtained from this method.

(a) (b) (c) (d)

Figure 2. ROI detection result. (a) original image; (b) segmentation result; (c) saliency map; (d) final result.

3. Adaptive Image Browsing

Visual attention plays an important role in adaptive image browsing, which is used to detect the ROIs [4] and to guide the watching process [3]. The method of [3] can display a large image progressively and automatically and allows the users to stop the automatic process at any time and choose where to look at and resume the browsing process afterwards. However, it is time consuming for its smooth transition between the optical focus. We prefer to displaying an image sequence cropped from the large image and adapted to the size of the screen. In the proposed method, ROIs are first detected and then ranked in descending order of attention value. Then the ROIs are adapted to the size of the screen and displayed instead of the input image.

3.1 ROI Ranking

ROI ranking is to predict the gaze transition. If there are multiple ROIs of different attractiveness in an image, human being will pay attention to them one by one in the order of descending attractiveness. It may have many applications such as image browsing, camera control in video surveillance and object tracking.

A ROI is defined as the minimum rectangular box which contains the connective attended regions, described as

$$\{(Left, Top, Right, Bottom), AV\}$$

where AV is used to represent the attention value of the ROI, which is simply the weighted sum of its regions' saliency, with the area factor as the coefficient of weight. It is calculated as:

$$AV_i = \sum_{A_k \in ROI_i} S_k A_k \quad (7)$$

Then the ROIs can be ranked according to their attention value.

3.2 ROI Adaptation

The ROIs obtained from section 3.1 are rectangular of any ratio and of any size, thus they need to be adapted to fit the screen. A ROI is adapted as follows:

(1) If it is smaller than the size of the screen, it will be extended towards the orientation of higher saliency.
(2) If it is larger than the size of the screen, it will be cropped or resized according to its minimal perceptive size [3].
(3) If its aspect ratio is different from the ratio of the screen, it needs to be reshaped by cropping and extending according to its minimal perceptive size [3].

In extending, two ROIs might be merged. If that is the case, the saliency of the final ROI should be the sum of the two merged ones. After adaptation the ROIs should be ranked again and then an image sequence is generated, which starts with the overall image including all the ROIs and followed by the ROIs in the order of descending saliency. Figure 3 is an example of image browsing.

(a) (b) (c) (d)

Figure 3. Image displaying sequence, (a) original image; (b) overall image; (c)the first ROI; (d) the second ROI.

4. EXPERIMENTS

The proposed methods are tested on the standard Coral Draw Library. 1000 images are randomly chosen to test both the ROI detection and the image browsing methods. We design two experiments to test the ROI detection method and evaluate the image browsing result respectively.

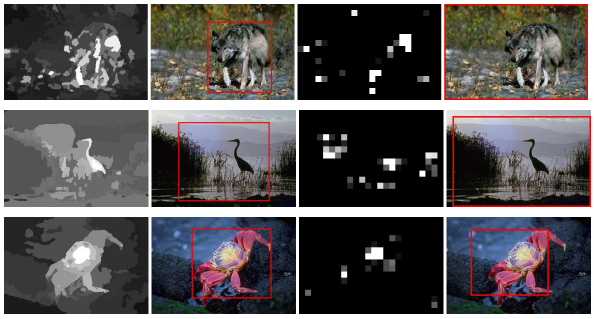
4.1 ROI Detection

We compare our method with the acknowledged visual attention model of [7], which uses Gaussian pyramid to extract multi-scale features and a "winner-take-all" (WTA) neural network to detect the ROIs from the saliency map. The results of the two methods on simple images are similar, however, on complex images our method performs much better. 3 results are illustrated in Figure 4. It can be seen that the saliency maps of our method contain much region information, which may help to detect the ROIs more precisely. The results confirm that the region as perceptive unit and the combination of global effect and contextual difference make the method effective and robust.

4.2 Adaptive Image Browsing

The proposed adaptive image browsing method is tested on 12 images with multiple objects. For image browsing, there is not an objective criterion yet to evaluate the effectiveness of the method so we adopt a subjective evaluation. 10 users, including 7 males and 3 females, aging from 23 to 30, are invited to evaluate the experimental results. The users are requested to score the results based on the following criteria: (1) are the ROIs really the attentive regions? and (2) does the browsing order meet the watching process? The scores are scaled from 1 to 5 to represent the satisfactory degree with 1 represents not satisfying at all and 5 represents very satisfying. Each result's mean and variance are illustrated in Figure 5. The mean says that each of the results is above acceptable and 9 of them are satisfying. 11 of the variances

are below 1, which mean the evaluation is reliable. Figure 6 shows the 12th and 9th results, which have the highest and lowest average scores respectively.



(a) (b) (c) (d)

Figure 4. ROI detection result, (a) our saliency map; (b) ROI from (a); (c) saliency map of [7]; (d) ROI from (c).

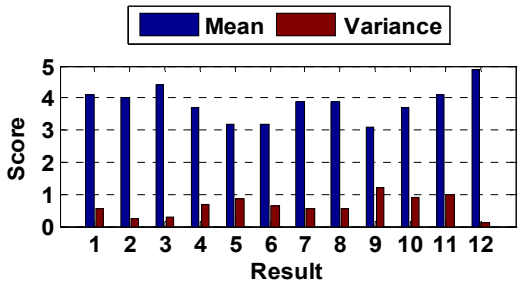


Figure 5. The mean and variance of the scores of each image browsing result.

Figure 6. Browsing sequences of the 12th (first row) and the 9th (second row) results.

5. CONCLUSIONS

This paper proposes a region based visual attention analysis method which is able to detect ROIs of any scale and is more effective from the human perception point of view. In this method both the global effect and the contextual difference are covered. Then, based on the ROI detection result, a ROI ranking method is proposed and used to generate the browsing sequence on mobile displays. Experimental results confirm the effectiveness of the attention analysis and image browsing methods.

In this paper the visual attention is analyzed spatially. We are extending it to temporal field, and are also planning to apply spatial-temporal visual attention in video content analysis.

6. ACKNOWLEDGMENTS

This work was supported by National High-Tech Research and Development Program (863 Program): 2006AA01Z117, in part by Science100 Plan of Chinese Academy of Sciences: 99T3002T03, Beijing Natural Science Foundation: 4063041, and National 242 Project: 2006A09.

7. REFERENCES

[1] W-H. Chen, C-W. Wang, J-L. Wu. Video Adaptation for Small Display Based on Content Recomposition, *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 17, No. 1, pp: 43-58, JAN 2007.

[2] F. Liu, M. Gleicher. Video Retargeting: Automating Pan and Scan, *Proceedings of the 14th annual ACM international conference on Multimedia*, pp: 241-250, 2006.

[3] X. Xie, H. Liu, W-Y Ma, H-J Zhang. Browsing Large Pictures Under Limited Display Sizes, *IEEE Trans on Multimedia*, Vol. 8, No. 4, pp: 707-715, 2006.

[4] Y. Hu, L-T. Chia, D. Rajan. Region-of-Interest based Image Resolution Adaptive for MPEG-21 Digital Item, *Proceedings of the 12th annual ACM international conference on Multimedia,* pp: 340-343, 2004.

[5] Y. Zhai, M. Shah. Visual Attention Detection in Video Sequences Using Spatiatemporal Cues. *Proceedings of the 14th annual ACM international conference on Multimedia*, pp: 815-824, October 2006.

[6] O L. Meur, P L. Callet, D. Barba, D. Thoreau, A Coherent Computational Approach to Model Bottom-up Visual Attention, *IEEE Trans on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 5, pp: 802-816, MAY 2006.

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

[8] Y-F. Ma, H-J. Zhang. Contrast-based image attention analysis by using fuzzy growing, *Proceedings of the 11th annual ACM international conference on Multimedia.* pp: 374-381, 2003.

[9] Y. Hu, D. Rajan and L-T Chia. Robust subspace analysis for detecting visual attention regions in images, *Proceedings of the 13th annual ACM international conference on Multimedia*, pp: 716-724, 2005.

[10] Y. Li, Y-F. Ma, H-J. Zhang. Salient Region Detection and Tracking in Video, *International Conference on Multimedia and Expo,* Vol. 2, pp: 269-272, 2003.

[11] V. Navalpakkam, L. Itti, An Integrated Model of Top-down and Bottom-up Attention for Optimizing Detecting Speed, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Vol. 2, pp: 2049-2056, 2006.

[12] R J. Peters, L. Itti, Beyond bottom-up Incorporating task-dependent influences into a computational model of spatial attention, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2007.

[13] Q. Ye, W. Gao, W. Zeng, Color Image Segmentation Using Density-Based Clustering, *International Conference on Multimedia and Expo*, Vol. 2, pp: 401-403, 2003.