TOWARDS UNSUPERVISED ATTENTION OBJECT EXTRACTION BY INTEGRATING VISUAL ATTENTION AND OBJECT GROWING

Junwei Han[1,2], King N. Ngan[1], Mingjing Li[3], Hongjiang Zhang[3]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
[2]School of Computer Engineering, Nangyang Technological University, Singapore 639798
[3]Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, China

ABSTRACT

Content-related functionalities of image/video applications call for efficient tools that can automatically extract meaningful objects from images. However, traditional methods generally fail to capture objects of user interest because they totally neglect human visual attention perception. Aiming to address this problem, this study proposes a generic model for unsupervised extraction of viewer's attention objects from color images. We formulate the attention objects as a Markov random field (MRF). Then, the MRF is expressed in the form of a Gibbs random field with an energy function. The energy minimization that integrates visual attention and object growing provides a practical way to obtain attention objects. The proposed model works in a manner analogous to humans and has great promise to be a basic tool for content-based image/video applications. Experimental results show the effectiveness of the proposed model.

1. INTRODUCTION

Object-based image representation opens the door for content-based image/video applications. Therefore, a great amount of efforts [1-2] have been devoted to the problem of object extraction. However, most existing methods mainly focus on how to extract objects. They totally ignore whether the extracted objects can attract the users' attention. This often results in many uninteresting objects being produced by expensive labors but supply little information to the users. In practice, when humans view a picture, they may not pay equal attention to all the objects in the picture and generally only attend to a few objects. Hence, there is an urgent need for a system that is able to extract objects that more likely attract the user's attention.

Visual attention is a neurobiological conception. It implies the concentration of mental powers upon an object by close or careful observing. After James [3] first founded the visual attention theory, many computational attention models [4-5] have been proposed by the psychophysical experiments. So far, the model built in [5] is the most successful model of visual attention. Its principal idea is based on the human perceptual characteristic that locations with distinctive features to their surroundings are often perceived. By combing three multi-scale contrast-based image features into a saliency map, the attended points are detected.

Although the mechanism of visual attention is not fully understood yet, some confirmed conclusions have been applied to a few fields. In [6], Salah applied a selective attention model to handwritten digit recognition and face recognition. In [7], Ma investigated a video summarization system by linear combination of multiple attention cues. In [8], Stiefehagen took advantage of gaze directions and sound sources to model people's focus of attention in a meeting situation. Recently, Chen et al. [9] discussed an image adaptation model using user attention.

In this paper, we attempt to develop a generic model to automatically extract viewer's attention objects by means of visual attention mechanisms. Without the need to have full semantic image understanding, it works in a manner analogous to humans. We hope the proposed model will become a basic tool for a wide range of content-based multimedia applications.

The unsupervised attention object extraction model is fulfilled in a two-stage process. (1) Itti's visual attention model [5] is employed to generate the saliency map, which encodes the attention value at every location in the image. (2) Attention objects are extracted by an MRF model that combines the saliency map and object growing techniques.

The rest of this paper is organized as follows. Section 2 introduces the saliency map. The attention object extraction by an MRF model is described in Section 3. The experimental results are shown in Section 4. Finally, conclusions are given in Section 5.

2. THE IMAGE SALIENCY MAP

The *saliency map* (SM) indicates the saliency at every location in the visual field by a scalar quantity. Itti et al. [5] implemented the SM under a general observation that

0-7803-8554-3/04/$20.00 ©2004 IEEE.

locations which stand out from their surrounds are likely to attract viewers' attention. This paper utilizes the similar idea of [5] to produce the SM. Fig. 1(b) displays a set of SM samples. In the SM, the larger the pixel brightness is, the more likely the pixel attracts the observer's attention.

3. ATTENTION OBJECT EXTRACTION

In our model, an attention object (AO) is described as: (1) It is a meaningful, physical entity in the image, such as a car, a tiger, etc. (2) It is more likely to attract viewers' attention than other objects in the image.

Seeded region growing (SRG) [1, 10] is a powerful technique for image segmentation. One of its nice characteristic is that high-level semantic knowledge can be easily putted into the segmentation process by the choice of meaningful seeds, which is very attractive for our work. Motivated by the idea of SRG, this paper presents an *attention object growing* (AOG) algorithm, which is performed in five steps. Firstly, the colors in the image are quantized to several representative classes and a class-map of the image is formed. Secondly, a local homogeneity map is obtained directly from the class-map, in which the high and low values correspond to the possible boundaries and interiors of the objects. Then, a number of attention seed areas are determined by means of both the SM and the local homogeneity map. Next, starting from those attention seeds, the AOs are grown using an MRF model. Finally, two post-procedures are carried out to fill the holes in the AOs and merge the similar objects.

There are three novel characteristics that distinguish the proposed AOG from the traditional SRG. (1) The semantic knowledge of human attention is incorporated into the process of object extraction. (2) In SRG, a number of initial seeds grow in a parallel manner. In contrast, AOG grows the objects one by one. By the order of decreasing attention value, the attention seeds sequentially grow their objects. The novel growing manner may facilitate the extraction of the full object. (3) A prior knowledge learned from the human segmented objects is adopted to guide the AO growing.

3.1. Color quantization and class map

As in [10], the task of the color quantization is to extract only a few representative colors that enable us to differentiate the neighboring objects in the image. The perceptual color quantization algorithm [10] is used in our implementation. Then, the image pixels quantized to the same color are assigned a same color class label. The image pixel colors are replaced by their corresponding color class labels, which result in a class-map (CM) [10].

3.2. Homogeneity measure and local homogeneity map

Natural objects are rich in color and texture. Many SRG techniques only work well on homogenous color regions, but can not handle texture information. Deng et al. [10] recommended a good color-texture homogeneity measure (HM) based on CM. It measures the spatial distribution of colors in the image. When we apply HM to an object, for the case of this object consisting of several homogeneous color patterns, the color classes are separated further from each other and the value of HM is large. On the contrary, if all color classes are uniformly distributed over the entire object, the value of HM tends to be small. When we apply HM to a local area of the image, it is a good indicator of whether that area is in the object interiors or near object boundaries. A grey-scale map whose pixel values are the HM values calculated over small windows centered at the pixels may be created. In this map, the brighter a pixel is, the more likely that it is close to an object boundary. To differentiate against the HM of an object, we refer to this map as *local homogeneity map* (LHM). Please see [10] for the details of HM. Fig. 1(c) shows a set of LHM samples.

3.3. Seed of attention object

Generally speaking, the seeds for object growing should be the representative parts of the objects. As for an AO, a reasonable seed area should not only be the focus of the viewer's attention, but also be located near its center. Concerning SM and LHM, it is easy to see that, the attention seed areas should correspond to minima of local HM values as well as maxima of local saliency values.

Let P be a pixel in the image, and R be a small window centered at P. We first calculate the average of the local HM values in R, denoted by μ_{LHM}. Then, the average of the saliency values in R, denoted by μ_S is computed. The area attention of R, A_R, is defined by

$$A_R = \mu_s - \mu_{LHM} \tag{1}$$

Essentially, the area attention determines the priority of an area that is considered as the seed to start an AO growing. The higher the area attention value, the more possible is the area associated with an AO. The optimal size for R is empirically set according to the image size. Please see Table 1 for details.

3.4. Attention object growing

Assume that an attention object, AO, grows from an initial attention seed, R. Subsequently, those unassigned adjacent pixels P of AO that satisfy a similarity test are repeatedly added to it, which is formulated as:

$$ST(P \to AO) = \begin{cases} \text{true} & \text{if } |C(P) - MEAN(AO)| \le \delta \\ \text{false} & \text{otherwise} \end{cases} \tag{2}$$

where $C(P)$ is the color label of P, $MEAN(AO)$ is the average color label of AO, and δ is a threshold.

3.4.1. Problem formulation

We model the AOG as a conditional probability problem. Given the SM, LHM, and HM of the object, we compute the maximum a posteriori probability (MAP) estimation of the AO. That is, find the optimal AO^* which satisfies:

$$AO^* = \arg\max_{AO} p(AO \mid SM, LHM, HM; \theta) \quad (3)$$

Here, θ denotes the parameters. The objects in an image can be modeled as an MRF, and the MRF is expressed by the Gibbs formulation with an energy function. Thus:

$$p(AO \mid SM, LHM, HM; \theta) = \frac{1}{Q}\exp\{-E(AO \mid SM, LHM, HM; \theta)/T\} \quad (4)$$

where $E(AO \mid SM, LHM, HM; \theta)$ is the energy function defined for an AO, Q is a normalization factor, and T is the temperature.

3.4.2. Energy function definition

In our model, the energy function for AO is defined as:

$$E(AO \mid SM, LHM, HM; \theta) = \theta_1 E_{Attention}(AO \mid SM) + \theta_2 E_{Edge}(AO \mid LHM) + \theta_3 E_{Homogeneity}(AO \mid HM) \quad (5)$$

where $E_{Attention}(AO \mid SM)$ is the attention energy, $E_{Edge}(AO \mid LHM)$ the edge energy, $E_{Homogeneity}(AO \mid HM)$ the homogeneity energy, and $\theta = (\theta_1, \theta_2, \theta_3)$ are the parameters.

The *attention energy* is first calculated using the SM:

$$E_{Attention}(AO \mid SM) = AV(I)/AV(AO) \quad (6)$$

where $AV(I)$ is the sum of saliency values in the image I and $AV(AO)$ is the sum of saliency values in an AO.

The *edge energy* is defined using the LHM. First, the LHM is binarized and the binarized LHM may be regarded as the edge map of the image. Then,

$$E_{Edge}(AO \mid LHM) = BL(AO)/EN(AO) \quad (7)$$

where $BL(AO)$ is the boundary length of the AO, and $EN(AO)$ is the number of pixels that are on the boundary of AO as well as edge pixels in the binarized LHM.

Finally, the *homogeneity energy* is defined by introducing a new concept of "learning from good objects". Without doubt, the guide by some prior knowledge on the object is beneficial to the object extraction. In this work, HM is a good feature to characterize the objects. Thus, "learning from good objects" refers to estimating the homogeneity probability distribution (HPD) of good objects. The learned HPD may be used as the prior knowledge. Of late, a database of human marked segmentations has been established [11] by a group in UC Berkeley. We select 4287 objects from this database as the "good objects". For each "good object", its HM is calculated. The HPD of good objects can be estimated from these 4287 training data. The homogeneity energy is defined as:

$$E_{Homogeneity}(AO \mid HM) = -\log P_{Homogeneity}(AO) \quad (8)$$

where $P_{Homogeneity}(AO)$ predicts the probability that AO is a good object, which is evaluated by the learned HPD.

3.4.3. Optimization and implementation

The MAP estimates are obtained by minimizing the energy function. Actually, after the attention seed is fixed, its AOG may be transformed to a problem of finding an optimal threshold δ to end the growing process. That is to say, the optimal AO can be obtained through finding the optimal threshold δ^* to lead the minimum energy. Given a value of $\delta, \delta \in [0, M]$, an AO is grown controlled by (2), and then the energy of this AO is calculated by (5). From all possible δ, the optimal δ^* is chosen, which satisfies:

$$\delta^* = \arg\min_{\delta=0,1,...,M} \{E(AO^{\delta} \mid SM, LHM, HM; \theta)\} \quad (9)$$

where AO^{δ} means the AO is grown under the control of δ. In our current implementation, $M = 6$.

3.4.4. Post-processing

After an AO has been extracted, two post-processes are required to refine its results. A "close" operation is first employed to fill the holes in the AO. Then, neighboring AOs are merged provided that their color distributions are sufficiently close to each other.

3.4.5. Number of attention objects determination

An algorithm is proposed to adaptively decide the number of AOs in an image based on the attention intensity. Let $AO_1, ..., AO_n, ...$ be attention objects in an image I, and their corresponding attention seeds are $R_1, ..., R_n, ...$. The number of AOs in I is determined as the minimum n, $n \geq 1$, which meets either of the following conditions:

$$\text{(i)} \quad \sum_{i=1}^{n} [E_{Attention}(AO_i \mid SM)]^{-1} \geq 0.5$$

$$\text{(ii)} \quad \mu_S^{R_{n+1}} \leq \frac{2}{3}\mu_S^{R_1} \quad (10)$$

where $\mu_S^{R_{n+1}}$ and $\mu_S^{R_1}$ denote the average of the saliency value in the R_{n+1} and R_1, respectively.

3.4.6. The flow of attention object growing

In summary, the unsupervised AOG is implemented by following steps: (1) Find an appropriate seed whose corresponding area attention value is maximal from the remaining unlabeled pixels; (2) Starting with the selected

seed area, grow the AO by means of the MAP estimation and MRF model; (3) Perform the post-processing to improve the extracted results; (4) If either item of (10) holds, end the AOG; otherwise, return to Step (1).

4. EXPERIMENTAL RESULTS

The proposed model has been tested on 880 general-purpose images from the Corel Gallery. Our test images cover a variety of categories. All of them are processed using one set of fixed parameters determined by empirical analysis ($\theta_1 = 0.25, \theta_2 = 0.5, \theta_3 = 0.25$). Fig. 1 shows some experimental results.

Due to the very strong subjectivity of visual attention, we constructed a user subjective test to assess the quality of our model. 16 human subjects were invited to take part in this test. 100 images and their corresponding results randomly selected from 880 experimental images were used for the test. In each test, we let the subject look through the original image first, and its extracted AO. Every subject was required to give his/her subjective scores to each experimental result. In order to obtain a quantitative assessment, we provided the subjects with following 5 scores: (1) 5: the extracted objects are what I want to focus on, and the object extraction results are perfect; (2) 4: the extracted objects are what I want to focus on, and the object extraction results are not very good but acceptable; (3) 3: the extracted objects are what I want to focus on, but the object extraction results are inaccurate; (4) 2: the less important AOs are missed but not the important ones; (5) 1: failed.

The statistical results of the subjective test are listed in Table 2 by percentages averaging over all subjects and images in each type of scores. The promising evaluation results illustrate the effectiveness of the proposed model.

5. CONCLUSIONS

As computer techniques evolve towards the new era of user-centric mode, the concept of learning human abilities gains increasing interest. By integrating visual attention and object growing, this work has built up an attention object model that mimics the human visual system. This model does not rely on completely understanding the semantic content of the image, and is promising to be a basic tool for many content-based multimedia applications.

Table 1 Window sizes at different image scales (pixels)

Image size	Window size of attention seed
256×384	17×17
512×768	37×37

6. REFERENCES

[1] J.P. Fan, X.Q. Zhu, and L.D. Wu, "Automatic Model-Based Semantic Object Extraction Algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 1073-1084, 2001.

[2] Y.W. Xu, E. Saber, and A. M. Tekalp, "Object Segmentation and Labeling by Learning from Examples," *IEEE Trans. Image Processing*, vol. 12, pp. 627-638, 2003.

[3] W. James, "The Principles of Psychology," *Harvard University Press*, 1890.

[4] Y. R. Sun and R. Fisher, "Object-Based Visual Attention for Computer Vision," *Artificial Intelligence*, vol. 146, pp. 77-123, 2003.

[5] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1254-1259, 1998.

[6] A.A. Salah, et al., "A Selective Attention-Based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 420-425, 2002.

[7] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li, "A User Attention Model for Video Summarization," in *Proc. ACM Int. Cof. Multimedia*, France, Dec. 2002, pp. 533-542.

[8] R. Stiefehagen, J. Yang, and A. Waibel, "Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues," *IEEE Trans. Neural Networks*, vol. 13, pp. 928-938, 2002.

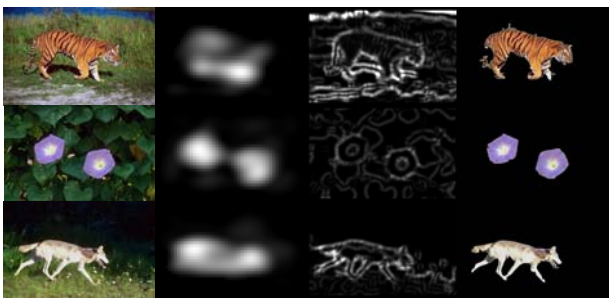
[9] L.Q. Chen, et al., "A Visual Attention Model for Adapting Images on Small Displays," *Multimedia systems*, vol. 9, pp. 353-364, 2003.

[10] Y.N. Deng et al., "Unsupervised Segmentation of Color-Texture Regions in Images and Video," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 800-810, 2001.

[11] D. Martin et al., "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics, " in *Proc. IEEE Conf. Computer Vision*, Canada, 2001, pp. 416-425.

Table 2: The statistical results of the subjective evaluation

	1	2	3	4	5
Percentage	22.9%	37.4%	28%	4.9%	6.8%
Average score	3.65				



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

Fig. 1 Some experimental results of attention object extraction: col.(a) original images, col.(b) saliency maps, col.(c) local homogeneity maps, col.(d) extracted attention objects.