Edge Detection with Automatic Scale Selection Approach to Improve Coherent Visual Attention Model

Jiayu Liang and Shiu Yin Yuen
Department of Electronic Engineering
City University of Hong Kong
Hong Kong, China
{jiayu.liang@student.cityu.edu.hk, kelviny.ee@cityu.edu.hk}

Abstract

An automatic scale selection approach is developed to improve the coherent visual attention model (Le Meur, O., Le Callet, P., Barba, D., Thoreau, D., 2006. A coherent computational approach to model bottom-up visual attention. IEEE Trans. Pattern Anal. Machine Intell. 28 (5), 802-817). The new approach uses linear summation to combine the automatic scale selection attention model with the coherent visual attention model. It is biologically more plausible because two important properties of human vision (i.e. edge detection and scale selection) are used. Its performance is evaluated by a large human fixation dataset. The t-test indicates that the improved model outperforms the original one highly significantly ($p<0.01$), and thus the new approach furnishes a more accurate model for visual attention prediction.

1. Introduction

Computational visual attention models (also known as saliency prediction models) predict where humans look when they attend to images. A successful prediction method can efficiently reduce the complexity of a computer vision system and aid in modeling higher level human vision behaviors. As a result, visual attention models are applied to many areas such as surveillance system [1], robotic vision [1], advertising [2], video compression [2] and game production [3].

Because visual attention is widely used, it has been a hot research topic in the past decade. Visual attention models which consider only low-level features are called bottom-up models. A representative work based on the Feature-Integration Theory (FIT) [4] is [5]. Their work proposes a multi-scale approach which uses color, orientation and intensity as basic descriptors to predict human fixations. Their method produces good results even when intense noise is added to the original images. A more recent work [6] makes use of psychological observations on human visual system. It considers space (pixels) as the basic unit for attracting attention. It uses Krauskopf's color space, contrast sensitivity function and visual masking as novel features for modeling visual attention. Rosin [7] proposes using only edge feature as the descriptor to predict salient regions. Although his method is simple, it produces good results on some popular datasets.

Among these visual attention models, the coherent visual attention model [6] has the best performance when tested on a popular eye-tracking dataset [2]. However, this method can be further improved because it ignores two important properties of early vision in humans. The first is that edge feature is important for attracting human attention. The observation is supported by the psychological evidence in [8]. They found a strong relationship between human fixations and edge density on images. The second observation is that cross-channel combination occurs in human vision system [9], and hence a multi-scale approach should be considered.

Because of the reasons mentioned above, we adopt the automatic scale selection edge detection in [10] to the coherent model in our work. A unique property that distinguishes [10] from other edge detection methods is that it is based on the Gaussian scale-space theory, in which only different orders of Gaussian derivative operators are used in the algorithm. Recent studies on neurophysiology show that receptive field profiles which can be modeled by Gaussian derivative operators exist in the mammalian retina and visual cortex [11], and human vision system may apply the same mechanism. For this reason, the biological plausibility of Lindeberg's method is strongly supported in [12].

We test the new model on a large dataset, and the result indicates that the new model outperforms the coherent model highly significantly ($p<0.01$).

The rest of the paper is organized as follows. Section 2 discusses our proposed method. Section 3 presents results and analysis. Section 4 draws the conclusions

2. The proposed method

2.1. Pre-processing: color space conversion

Based on the opponent color theory which claims that two opponent channels (red/green channel and blue/yellow channel) exist in human vision [9], images represented in RGB color space are converted to IUV color space by Equation (1), as described in [11]:

$$\begin{cases} I = (R + G + B)/3 \\ U = R - G \\ V = B - (R + G)/2 \end{cases} \quad (1)$$

2.2. Edge detection

After the pre-processing step, Lindeberg's edge detection with automatic scale selection [10] is applied to each channel of the IUV image. The three generated edge maps are summed up and normalized to form the final edge map of the given image.

The idea of scale selection in [10] is to find the intersection between the edge surface and the surface defined by the locally maximal normalized edge strength in

scale-space [10]. The edge surface is formed by edge points at all scales in the scale-space representation of an image I [10]. Given that

$$L = I * G_t \tag{2}$$

in which G_t denotes a Gaussian kernel with standard deviation $\sigma = \sqrt{t}$ and t is the scale parameter, an edge point (intersection) at a certain scale of an image satisfies [10]:

$$\begin{cases} L_{vv} = 0 \\ L_{vvv} < 0 \end{cases} \tag{3}$$

in which v denotes the local coordinate axis parallel to the gradient direction, and L_{vv} and L_{vvv} denote the second and third order derivatives of L in the v direction respectively. The normalized edge strength is defined to be

$$\mathcal{G}_{\gamma-\text{norm}} L = t^{\gamma}\left(L_x^2 + L_y^2\right) \tag{4}$$

in which $\gamma = 1/2$ is to maximize the characteristic scale of the edge [13], and L_x and L_y denote the first order partial derivative of L in the x and y direction of the global coordinate respectively.

Our implementation of Lindeberg's method follows that in [13].

2.3. Gaussian average

Based on [8], edge density for the edge map $EM(x,y)$ used as an estimator for visual attention. Edge density on a given pixel (x,y) is calculated as

$$ED(x,y) = \frac{\sum_{(u,v)\in A(x,y)} EM(u,v)}{\text{pixel number in } A(x,y)}, \tag{5}$$

in which $A(x,y)$ denotes a 1 degree visual angle area centered on pixel (x,y). Visual angle V is defined as

$$V = 2\tan^{-1}\frac{S}{2D}, \tag{6}$$

in which S denotes the viewing length and D denotes the viewing distance as in Figure 1.

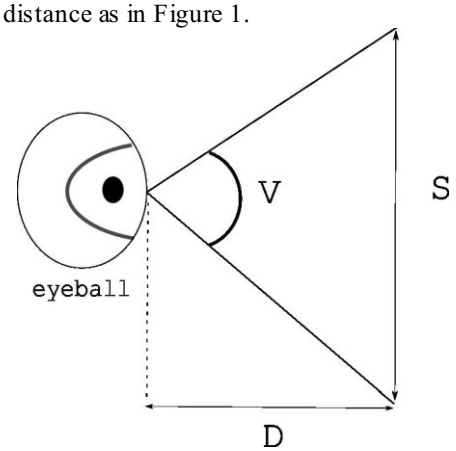

Figure 1. Illustration of visual angle

However, based on the idea presented in [14], where the authors propose that regions close to features have more effect of drawing attention and the effect can be modeled by Gaussian distribution, a Gaussian average is used instead of the equal weighting average mentioned above, and hence Equation (5) is rewritten as

$$ED(x,y) = \frac{\sum_{(u,v)\in A(x,y)} G(u,v)\,EM(u,v)}{\sum_{(u,v)\in A(x,y)} G(u,v)}, \tag{7}$$

in which $G(u,v)$ denotes a 2D Gaussian filter whose mean is zero and standard deviation equals to 0.5 degree of visual angle, which is the radii of $A(x,y)$. In actual implementation, the filter length is taken to be one standard deviation in order to balance prediction accuracy and computation speed.

2.4. Combination

The coherent model [6] exhibits strong biology plausibility. First, it converts the RGB image into the Krauskopf's Color Space, which is developed based on psychological and physiological experiments. The converted image is then measured by the contrast sensitivity function to evaluate the visibility. Finally visual masking is implemented to determine the in-context visibility. The obtained data is coherently normalized and mapped to a psychovisual space for the saliency map.

The coherent model can be further improved by integrating two important descriptors (i.e. the edge detection and the scale selection). We use a simple weighting function to combine our model with the coherent model. A linear function is chosen for simplicity. The combined saliency map reads

$$CS(x,y) = \alpha LM(x,y) + (1-\alpha)\,ED(x,y) \tag{8}$$

in which α is determined by trial and error on a small human fixation dataset provided in [6], and $LM(x,y)$ denotes the saliency map predicted by the coherent model. We found that when $\alpha = 0.66$, the performance of the combined saliency map is maximized.

2.5. Weighting

Both theoretical and practical works show that humans tend to focus more on the center of an image than on other parts [15, 6]. Recent studies indicate that this center-bias may be derived from instinct or postnatal learning [15]. In order to model this effect, a weighting function (WF) is used after the combined saliency map is generated. As described in [6], the WF reads

$$W(x,y) = e^{-\left(\frac{(x-x_0)^2}{2\sigma_x^{e2}} + \frac{(y-y_0)^2}{2\sigma_y^{e2}}\right)}, \tag{9}$$

in which

$$\sigma_y^e = \sigma_x^e\left(\frac{R_x}{R_y} Ind(R_x < R_y) + \frac{R_y}{R_x} Ind(R_x > R_y)\right),$$
$$\sigma_x^e = 2.5 \text{ degrees visual angle} \tag{10}$$

and σ_x^e is a tuned parameter which generates the highest prediction rate for the dataset in [6], R_x and R_y denote the width and height of the image respectively, and $Ind()$ is the indication function. As a result, the final saliency map is

$$S(x,y) = W(x,y)CS(x,y), \tag{11}$$

3. Experimental results

3.1. Dataset

The test dataset is from [2]. There are three reasons for

using this dataset. First, it is a large dataset which contains 1003 images[1], and hence is suitable for test. Second, its salient regions are produced by an eye-tracking apparatus. All images are free-viewed by 15 participants in random order for 3 seconds [2], and fixation patterns are recorded by the eye-tracking apparatus. The final salient regions are calculated by these fixation patterns. As a result, the dataset of [2] better represent human visual behaviors. Third, the dataset in [2] contains images of different kinds, and hence it ensures that the evaluation is not biased to a particular type of scene.

3.2. Results

Prediction results of the coherent model are provided by Le Meur [6]. We hereby refer to the coherent model as LM (named after the author Le Meur), our automatic scale selection model as AE (Automatic scale-selection Edge detection) and the combined model as LM+AE respectively.

3.2.1 Qualitative results

To investigate the properties of our method, qualitative results are examined, and they are given by image comparisons (Figure 2 and Figure 3). Figure 2 shows that LM+AE is better than LM, while Figure 3 identifies some limitation of LM+AE.

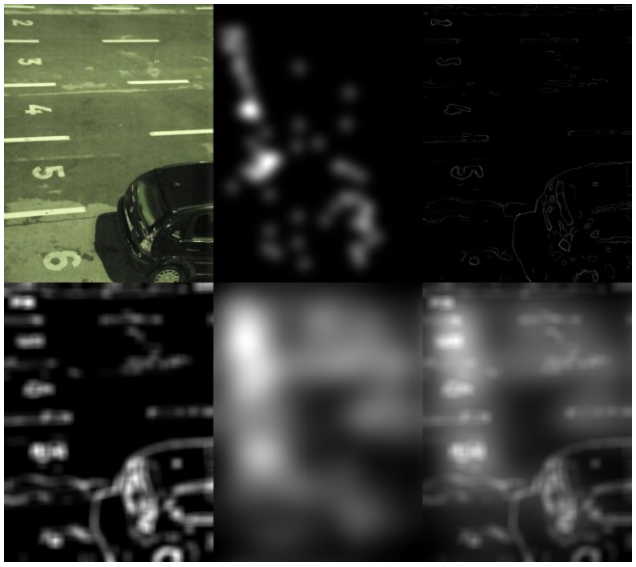

Figure 2. Image comparison 1 in which LM+AE outperforms LM (top row from left to right: original image, human fixation density map and automatic scale selection edge map[2]; bottom row: predicted saliency maps of AE, LM and LM+AE respectively)

By inspecting the results, we find that LM+AE predicts better than LM when edges constitute the whole object (the numbers in Figure 2) or the inside of an object contains complex edge features that draw human attention. In either case, AE can highlight the important edge features and attenuate the noise to improve the prediction, while LM cannot accurately locate the fixation positions. On the other hand, LM+AE fails to predict well when the inside of an object contains no significant edges, such as the lemons in Figure 3. In this case, human attention is directed to the center of the object, which AE is unable to highlight. We will discuss this observation in the Conclusions section and point out future research directions.

It should be noted that LM is easily affected by texture and background noise which humans usually ignore, such as the texture of the ground in Figure 2. However, because of the scale-selection algorithm, AE can remove this kind of noise and remain focused on what humans pay attention to.

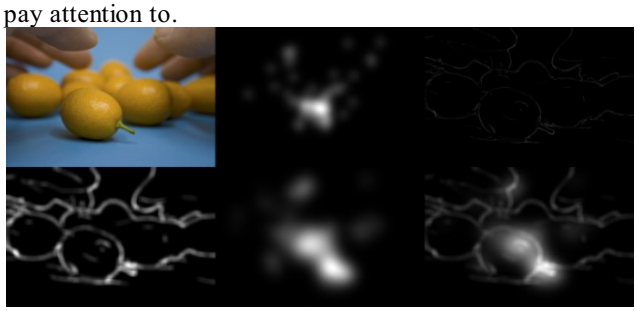

Figure 3 Image comparison 3 in which LM outperforms LM+AE (images arrangement is the same as Figure 2)

3.2.2 Quantitative results

The quantitative results are evaluated by Linear Correlation Coefficient (LCC) between the human fixation density map and the predicted saliency map. LCC measures the linear dependence between two variables. It has been applied to performance evaluations on image registration and object recognition [6], and is also suitable for comparing saliency prediction results [6, 16].

LCC is defined as

$$cc(p,h) = \frac{cov(p,h)}{\sigma_p \sigma_h} \tag{12}$$

in which p and h denote the predicted and human fixation density maps respectively, $cov(\cdot,\cdot)$ denotes the covariance operator and $\sigma_{\cdot}$ denotes the standard deviation operator [6]. LCC ranges from -1 to +1. It indicates a perfectly negative or positive linear relationship between the two compared maps when it is -1 or +1 respectively [6]. +1 is preferred in our paper since it implies an accurate prediction of human fixations.

Table 1 shows the average LCC for all three methods when WF is not used. It can be seen that AE alone does not give a high prediction accuracy, but the combined model LM+AE outperforms the reference model LM highly significantly in one-tailed paired t-test ($p<0.01$), which demonstrates that AE is a good complement to the coherent model.

Table 2 shows the average LCC for all three methods when WF is used. Although the average LCC of LM+AE

[1] We use only 961 images in our work. The reason for this is that AE and/or LM fail to give prediction for some images. AE fails to give prediction for 25 images because of memory limitation of the code using MATLAB. In our experiments, we use the saliency maps of LM provided by the correspondence author Le Meur (2012 personal communication). LM gives no prediction (generate "black" images) for 20 images. Referring to Equation (12), for "black" images, the denominator of LCC becomes zero which invalidates LCC. We abandon these 42 images (there are 3 images that cannot be predicted by either AE or LM).

[2] Please refer to the online version for a clearer view of the edge map.

is only 0.003 larger than that of LM, the new model still outperforms LM highly significantly ($p<0.01$) because we use a large dataset for test. The proposed model also outperforms WF highly significantly.

Table 1 LCC for prediction results without weighting

	Average	t-test p value when compared to LM+AE
LM	0.315	4.15E-17
AE	0.214	6.13E-64
LM+AE	0.324	-

Table 2 LCC for prediction results with weighting

	Average	t-test p value when compared to LM+AE
WF	0.326	1.77E-24
LM with WF	0.351	1.00E-03
AE with WF	0.285	1.78E-88
LM+AE with WF	0.354	-

4. Conclusions

This paper introduces a novel method to improve the coherent visual attention model. The new method is based on Lindeberg's edge detection with automatic scale selection. We use a simple linear summation to combine the two models AE and LM together to generate a better predictor, but the actual model is believed to be more complicated. Quantitative results show that the combined model outperforms the reference model highly significantly in both the non-weighting case and weighting case.

The result of this work indicates that edge feature and scale selection are important for drawing human attention. These two properties should be considered when building biologically plausible visual attention models.

Though edge detection with automatic scale selection is important, human fixation density maps shown in Section 3.2.1 suggest that humans tend to focus more on the center of a Gestalt object than on dense edge regions. Actually, this observation is consistent with the neurobiology finding [17] that objects are the basic units of attention selection. Combining the above facts with the general belief that edges have the ability of implying the existence of objects [9], we consider that it is not edge density but the objects perceived via edge density that draw human attention. Moreover, it can be seen from Figure 2 and Figure 3 that humans usually focus on the center of an object, which is also the "gravity center" of edge density. Based on this finding, we consider that post-processing AE with a symmetry detector may significantly improve the performance of AE on saliency prediction

For future work, based on the above analysis, we will focus on exploring the relationship between visual attention and symmetry detected by edge features.

Acknowledgements

The work described in this paper was supported by a grant from City University of Hong Kong [Project No. 7002746]. The first author is supported by the Research Studentship of the university.

References

[1] M. Begum, F. Karray: "Visual attention for robotic cognition: a survey," IEEE Transaction on Autonomous Mental Development, vol. 3, no. 1, pp. 92-105, 2011.

[2] T. Judd, K. Ehinger, F. Durand, and A. Torralba: "Learning to predict where humans look," in Proceedings of International Conference on Computer Vision, 2009.

[3] M. El-Nasr and S. Yan: "Visual attention in 3D video games," in Proceedings of ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, 2006.

[4] A. M. Treisman and G. Gelade: "A feature-integration theory of attention," Cognitive Psychology, vol. 12, no. 1, pp. 97–136, 1980.

[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, no. 11, pp. 1254–1259, 1998.

[6] O. Le Meur, P. Le Callet, D. Barba and D. Thoreau: "A coherent computational approach to model bottom-up visual attention," IEEE Trans. Pattern Anal. Machine Intell., vol. 28, no. 5, pp. 802-817, 2006.

[7] P. Rosin: "A simple method for detecting salient regions," Pattern Recognition, vol. 42, pp. 2363–2371, 2009.

[8] S. K. Mannan, K. H. Ruddock and D. S. Wooding: "The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images," Spatial Vision, vol. 10, no. 3, pp. 165-188, 1996.

[9] J. P. Frisby and J. V. Stone: Seeing: The Computational Approach to Biological Vision. MIT Press, Cambridge, MA, 2010.

[10] T. Lindeberg: "Edge detection and ridge detection with automatic scale selection," International Journal of Computer Vision, vol. 30, no. 2, pp. 117–154, 1998.

[11] T. Lindeberg: "Scale-space," in Benjamin, W.W. Encyclopedia of Computer Science and Engineering 4th Edition, John Wiley and Sons, Hoboken, New Jersey, pp. 2495-2504, 2009.

[12] M. A. Georgeson, K. A. May, T. C. A. Freeman and G. S. Hesse: "From filters to features: scale–space analysis of edge and blur coding in human vision," Journal of Vision, vol. 7, no. 13(7), pp. 1-21, 2007.

[13] I. Kokkinos, P. Maragos and A. Yuille: "Bottom-up and top-down object detection using primal sketch features and graphical models," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[14] Y. Sun and R. Fisher: "Object-based visual attention for computer vision," Artificial Intelligence, vol. 146, no. 1, pp. 77-123, 2003.

[15] D. Parkhurst, K. Law and E. Niebur: "Modeling the role of salience in the allocation of overt visual attention," Vision Research vol. 42, no. 1, pp. 107-123, 2002.

[16] G. Kootstra, B. de Boer and L. R. B. Schomaker: "Predicting eye fixations on complex visual stimuli using local symmetry," Cognitive Computation, vol. 3, pp. 223–240, 2011.

[17] K. M. O' Craven, P. E. Downing and N. Kanwisher: "fMRI evidence for objects as the units of attentional selection," Nature, vol. 401, no. 6753, pp. 584-587, 1999.