Comparison Research on Bottom-up Visual Attention Models

Guiliang Chen
Chengdu Aircraft Design and Research Institute, Chengdu, China

Yongheng Li
Northwestern Polytechnical University, Xi'an, China

ABSTRACT: In recent years, tremendous progresses have been made in research on basic principle of visual cortex information processing, which bring wide attention to bottom-up visual attention system, some models have already been established successfully and been applied to image processing and machine vision. This essay aims to expatiate the seven kinds of currently most representative bottom-up visual significance algorithm models, summarize and compare the advantages and disadvantages of these algorithm models, explain the results of these algorithm models on different categories of images, and provide some thinking and suggestion on possible future development and research direction of bottom-up visual attention models.
KEYWORD: Bottom-up; Significance; Visual attention model

1 INTRODUCTION

Visual attention system is an important research topic in computer vision field, which is widely used in image compressing and subdivision, target detecting and identification, image retrieval, and active machine vision, etc. Complicated as visual attention system is, it covers many subjects, such as cognitive science, neuroscience, biology and computer science, etc. So far people have not yet completely understood the process of visual attention. The current visual models are mainly focusing on bottom-up model which is data driven. Although it succeeded to some extent, still there are some disadvantages. Therefore, comparison research on these visual attention models is significant.

This essay firstly summarizes two categories of significance testing, and introduces seven kinds of the most representative computational models, including residual spectrum algorithm, local entropy algorithm, natural statistical algorithm based on Bayesian structure, sliding window integral histogram algorithm, quaternion image model and multi-scale quaternion Fourier transform model. Afterwards it explains the critical technique related to such algorithm, and then discusses actual application result and advantage/disadvantage of such algorithm on different types of images. Finally, it provides some thinking and suggestion on future development and research direction of bottom-up visual attention system.

2 MAIN ALGORITHM OF VISUAL SIGNIFICANCE

As per process of simulation of human eye attention processing, existing visual attention models can be divided into two categories as follows:

(1) One is top-down method, which depends on current task.

(2) The other is bottom-up method, which depends on input image.

Top-down visual attention is related to high-level perceptual process, of which people have limited knowledge and understanding. Therefore, few researches have been conducted on this attention system, and most of these researches are combined with bottom-up visual attention system based on specific task.

According to tested object, visual attention models can be divided into:

(1) Monitoring object tester, aiming to find specific categories such as people, cars, etc. This method brings about high performance, but the disadvantage is that the interested object must stay in training sample and must be an available predefined category. Besides, testing performance depends on training data.

(2) Common significance detector, e.g. Itti model. The inspiration of these methods comes from the fact that human visual system can rapidly focus on common significant target without training. This technique is suitable to the situation that target and imaging condition is not known in advance.

So far, there is no unified consistent definition of significance in literatures. Significance models are various. Next we will introduce the seven most representative visual attention algorithm idea respectively as per route of technique development.

2.1 *Visual Structure Significance Model (Itti)*

(Itti, et al. 1998) The significance model brought forward is the earliest, which is also the algorithm model most commonly used for comparison in following work. This model is expanded and realized based on the basic idea by Koch and Ulman. It was elicited from literatures on visual attention, e.g. feature integration theory, ensuring this model structure is feasible on neurobiology. This model divides input image into 3 channels: intensity, color and direction. Through center-surrounding algorithm operators, viz. filters response difference value in different scales a group of characteristics pattern are generated. Afterwards, the characteristics pattern of each channel is normalized, and combined cross-scale and cross-directionally, to create prominent image of each channel. These channels linearly combine to form significance image of the whole. This model has already proved useful in predicting human attention and target detecting.

2.2 *Local Entropy Model (AE)*

(Kadir, et al. 2001) Local entropy of brightness images was brought forth. (Renninger, et al. 2005) Entropy was put forward of local linear histogram of brightness image, and for most significance points at any given time, provide maximum information gain under acquired knowledge condition during previous fixation point period. The basic thinking of taking entropy model as significance model is that, one entropy of local area feature distribution can be used to measure area richness and variety, and intuitively, if an area contains many features of different direction and intensity, this area should be significant.

2.3 *Spectral Residual (SR)*

(Hou, et al. 2007) The famous spectral residual method was raised which was regarded as a fast method for spatial domain structure universal significance detector. It is based on this idea: when facing unpredictable visual illustrations and uncountable categories, significance detector should achieve better detection result with statistics knowledge of minimum reference object. This method is to resolve the issue of tracing shared background of target object in different classification. This article resolves this issue with utilization of information theory, that image information is divided into 2 parts: (Innovation) innovative image information and (Priori Knowledge) systematic redundancy information. After researching the logarithmic spectrum of large amount of images, it is discovered that although each logarithmic spectrum has statistic singularity, different image logarithmic spectrum has similar trend, which hinted redundancy information. Visual attention focuses on none other than innovative information that is statistic singular point in spectrum. It proves again that natural images are not random but conform to altitude predictable distribution which also shows that image redundancy information is a partial linearity in averaged logarithmic spectrum, thus residual spectrum should contain image innovative information.

2.4 *Natural Statistical Model Based on Bayesian Structure (SUN)*

(Zhang, et al. 2008) The self-information significance detection model based on Bayesian frame is raised after referring to AIM algorithm. The model starts with a simple assumption, stating that an important purpose of significance detection is to predict the fixation point when human beings freely watch image and establishes an optimal Bayesian probability calculation frame to realize this target and the bottom-up significance is shown in this frame. Instead of entirely defining bottom-up significance algorithm in image being watched as previous, such model definition uses statistic image (acquired from a group of natural images) to represent the visual experience obtained by living organisms when they are watching. In Bayesian frame derivation, assuming that the position is constant (no beforehand potential target position information is available) and position prior knowledge is elliptical. In this condition, only two items left in this model, self-information and logarithmic likelihood item. They can combine as point-to-point mutual information to present the relationship between visual features and target. This article assumes that when human is watching freely (not intended to positively search specific objective), human's fixation point should be on potential target in visual domain directly. Actually, logarithmic likelihood is unknown and this item is also elliptical in this article when calculating significance model (this means for the targets in different classification, their features distribution may be uniform). In this condition, the final significance model becomes the item of self-information. Based on simulation of primary cell, this article refers to AIM method and obtains image visual feature vectors from natural image set by ICA and then complete estimating the features distribution of complex cell with zero-mean generalized Gaussian Distribution, which is also called as exponential distribution.

2.5 *Sliding Window Integral Histogram (SSO)*

(Esa, et al. 2010) Regularized integral histogram and image shaping method were ingeniously utilized and combined to solve significance. The recommended significant action in this article is, based on an image with a sliding window applied. (Rahtu & Heikkila 2009, Mahadevan & Vasconcelos 2010) This article uses rectangle window as reference. One rectangle window W is divided into two separated parts, one rectangle inner window K (kernel) and boundary B. It is assumed that points in K are significant and points in B are part of background. Significance is comparing the contrast between each feature distribution of inner window and boundary window, and then developing the significance expression with Bayesian formulae. Since the window size is fixed, significance can be expressed by histogram. To avoid the major change of histogram due to minor change of feature distribution, this article uses regularized histogram and states that a larger regularized Gaussian kernel against target feature will reduce the feature weight in generated significant image. It is beneficial for significance detection.

2.6 *Quaternion Image Significance Model (PQFT)*

(Guo, et al. 2010) The method was raised of establishing a quaternion image significance model with quaternion Fourier transform. Different from previous method of separately processing color channel, this method processes color image as an entirety. It does not need to individually process image channel, but tears up mutually-linked color information. When regarding scalar part of quaternion as the 4^{th} channel to integrate a component motion, this method can also be used to process video sequence. Because quaternion provides a strong algebra to realize hyper-complex Fourier transform, the image multi-channel frequency-domain expression can be achieved.

2.7 *Multi-scale Quaternion Fourier Transform (Ms-EigenSR)*

(Boris, et al. 2012) Multi-scale model was adopted. (Peters, et al. 2008) The effect of quaternion algorithm was evaluated. Significance algorithm of quaternion Fourier transform based on quaternion feature axis and phase was raised and evaluated. Herein, the quaternion Fourier transform is different with FFT for which the PQFT is used to directly solve each component of quaternion image. This method solves FFT through a plural expression of quaternion transform axis which is obtained by finding out a quaternion orthogonal basis coefficient which is perpendicular to feature axis. For obtained frequency-domain, phase expressions of frequency spectrum residual and PQFT are used to establish image significance. For most of significance model in previous method, only fixed and single scale is considered; however, when image vision is significant, scale is an important parameter as well as an integral part of significance model. This algorithm calculates a multi-scale saliency map by combining spectrum significance of image under different image scale.

3 KEY TECHNOLOGY DISCUSSION OF VISUAL ATTENTION MODEL

3.1 *Research Method Based on Fourier Frequency Spectrum*

Spectrum Residual (SR) is that logarithm spectrum subtracts average spectrum, and logarithm is solved with FFT.

$$R(f) = S(f) - h_n(f) * S(f) \tag{1}$$

$$S(x) = g(x) * F^{-1}[\exp(R(f) + \mathrm{i}*P(f))]^2 \tag{2}$$

In which, S(f) is image logarithm extent spectrum. P(f) is image phase spectrum. $h_n(f)$ is 2 dimensional averaging filter, $h_n(f)*S(f)$ is average spectrum, which can be approached by integral convolution input image. R(f) is residual error spectrum. g(x) is Gaussian filter used to smooth salient figure.

3.2 *Partial Entropy and Partial Self-information and ICA Independence Assumption*

Partial entropy,

$$t_1 = -\sum_{i=1}^{N} D_i \log(D_i) \tag{3}$$

is a measure of partial mobility.
Partial self-information,

$$t_2 = -\log(p(x)) \tag{4}$$

is a measure of partial contrast.
Assuming ICA analytical independence means

$$p(w_1 = v_1, w_2 = v_2, \square, w_n = v_n) = \prod_{i=1}^{n} p(w_i = v_i) \tag{5}$$

It allows estimation of getting n-dimensional space from n one-dimension probability density functions.

3.3 *Assuming ICA analytical independence Quaternion Image Saliency Model (PQFT)*

Seek quaternion character axis r (i.e. unit quaternion)

$$r = \frac{q}{|q|} = \frac{q}{\sqrt{q \cdot \bar{q}}} \tag{6}$$

In which, is

$$q = -1*i - 1*j - 1*k \tag{7}$$

Create a quaternion image QIR,

$$QIR = 0 - R*i - G*j - B*k \tag{8}$$

Seek 2 dimensional FFT for QIR,

$$FQIR = FFT2(QIR) \tag{9}$$

Seek angle (phase) of four figures FQIR on quaternion character axis r,

$$\theta = angle(FQIR, r) \tag{10}$$

Saliency based on phase is

$$S = \left|\text{IFFT2}\ (e^{(\gamma*\theta)})\right|^2 \tag{11}$$

4 EXPERIMENTAL RESULT COMPARISON AND ANALYSIS

The above mentioned 7 algorithmic models are experimented with different types of images, experimental result as shown in (Fig.1). The first two lines of Fig. 1 is biological visual image, while the next two lines of Fig. 1 is natural image. For the first original drawing, we need algorithm to detect the line which has outstanding color and shape, in the second original drawing, outstanding red circle is required to be detected; in the third and forth drawing people on the beach, sheep and vehicle is to be detected.

As shown in the figure, Itti model effect is acceptable, but it can only be an evaluation methodology. Part of the reason is that the overall goal (i.e. what it is designed to optimize) of the system is not specified, and it has many parameters that require hand setting. The effect of AE algorithm is the worst. It can not judge salient target exactly whether for biological visual image or natural image. For these models, a common criticism is that at high tactile sensation area, however their background, they are always salient. For example, human observer finds an egg in a nest highly salient, but partial entropic algorithm finds that nest is more salient than egg. SR model has the simplest calculation and the fastest operating speed among all models, but the disadvantage is that it lacks of biological reference, and it can only calculate visual salience figure of image brightness channel. This can be easily seen from the result of biological visual image experiment. The reason why the result of SSO model is precise, is because at after-treatment of algorithm, condition random field is used to accurately divide up the image. It should be unfair to compare such result with those from other algorithms. The effect of SUN model is ordinary when detecting biological visual image, but experimental result on natural image is very good because this algorithm uses natural image assembly for statistics conditioning, thus it has natural resolving ability for natural image; PQFT algorithm can quickly calculate time space salience figure of color video frame. The disadvantage is also the lack of biological reference, and the motive of adoption quaternion for calculation in PQFT method is not clear. Calculation method based on Fourier transformation only shows their salience detect ability by experimental result, whereas their principle of visual salience calculation is hard to receive acceptance; Ms-EigenSR algorithm has the best effect among all algorithm, especially in biological visual image where other algorithm has poor performance. This is inseparable with the fact that this algorithm treats all color channels as a whole and simulate human eye to treat with different visual measure.

In summary, by means of experiment with different types of images, it is found that MS-EigenSR algorithm has the best effect and SR model has the simplest calculation and the fastest operating speed.

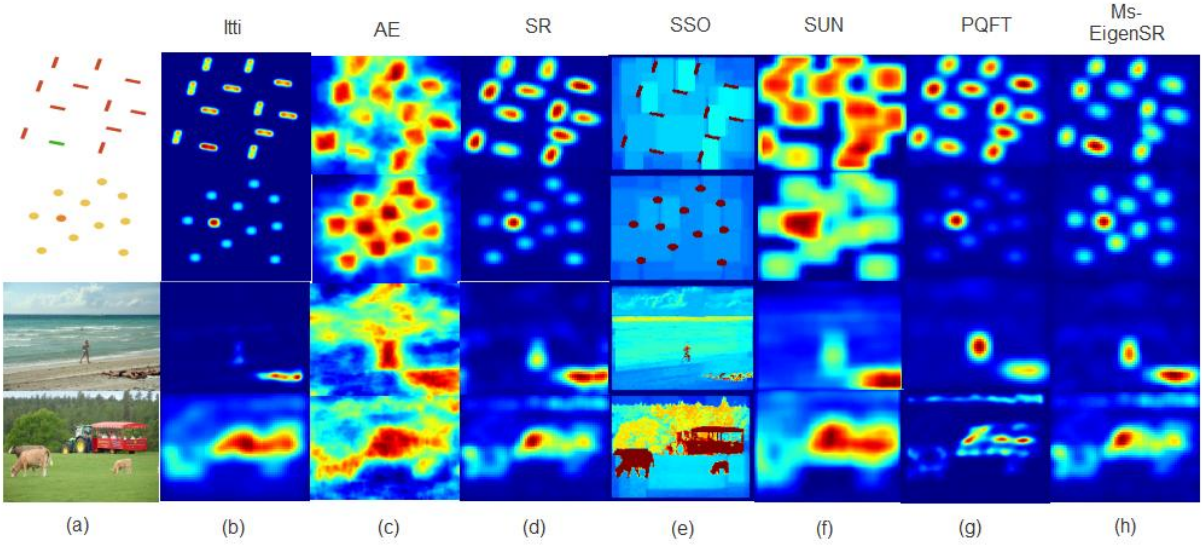

Figure 1. Experimental result of 7 algorithmic models

5 PERSPECTIVE AND RESEARCH

Lastly, observation and suggestion on development of visual attention in the future is given. Current visual attention research hotspot is that quaternion method is applied to combine multiple color channels, salient area of image is analyzed with the method of multiple measures. This is embodied in many high level articles published recently. Future trend of visual attention technology will be to continue combining the basic principle of biology, statistics, informationism, and at the same time using the most leading-eadge method in classification technique and filter technique, to get the best detection result.

REFERENCES

[1] Guo, C. & Zhang, L. 2010. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Transactions on image processing* 19(1): 185-198.

[2] Hou, X. & Zhang, L.Q. 2007. Saliency Detection: A Spectral Residual Approach. 2007 *IEEE Conference on Computer Vision and Pattern Recognition*. ISBN: 1-4244-1179-3. pp: 1-8.

[3] Itti, L. Koch, C. & Niebur, E. 1998. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Maching Intelligence* 20(11): 1254-1259.

[4] Kadir, T. & Brady, M. 2001. Saliency, Scale and Image Description. *International journal of computer vision* 45(2): 83-105.

[5] Mahadevan, V. & Vasconcelos, N. 2010. Spatiotemporal Saliency in Dynamic Scenes. *IEEE Transactions on Pattern Analysis and Maching Intelligence* 32(1): 171-177.

[6] Peters, R. & Itti, L. 2008. The Role of Fourier Phase Information in Predicting Saliency. *Journal of Vision* 8(6): 879-879.

[7] Rahtu, E. & Heikkila, J. 2009. A Simple and Efficient Saliency Detector for Background Subtraction. *the 9th IEEE International Workshop on Visual Surveillance*, 1137-1144.

[8] Rahtu, E., et al. 2010. Segmenting Salient Objects from Images and Videos. *Computer Vision–ECCV*, part V, LNCS 6315:366-379.

[9] Renninger, L.W. et al. 2005. An Information Maximization Model of Eye Movements. *Advances in neural information processing systems* 17: 1121-1128

[10] Schauerte, B. & Stiefelhagen, R. 2012. *Quaternion-based spectral saliency detection for eye fixation prediction*. in *Computer Vision–ECCV*. Springer

[11] Zhang, L. et al. 2008. SUN: A Bayesian Framework for Saliency Using Natural Statistics. *Journal of Vision* (Impact Factor: 2.48). 02/2008; 8(7): 32.1-20. DOI: 10.1167/8.7.32.