Saliency Detection: A Spectral Residual Approach

Conference Paper *in* Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition · June 2007

DOI: 10.1109/CVPR.2007.383267 · Source: DBLP

2 authors:

Xiaodi Hou
California Institute of Technology
9 PUBLICATIONS **6,087** CITATIONS

SEE PROFILE

Liqing Zhang
Shanghai Jiao Tong University
348 PUBLICATIONS **11,478** CITATIONS

SEE PROFILE

Saliency Detection: A Spectral Residual Approach

Xiaodi Hou and Liqing Zhang
Department of Computer Science, Shanghai Jiao Tong University
No.800, Dongchuan Road, Shanghai
http://bcmi.sjtu.edu.cn/~houxiaodi, zhang-lq@cs.sjtu.edu.cn

Abstract

The ability of human visual system to detect visual saliency is extraordinarily fast and reliable. However, computational modeling of this basic intelligent behavior still remains a challenge. This paper presents a simple method for the visual saliency detection.

Our model is independent of features, categories, or other forms of prior knowledge of the objects. By analyzing the log-spectrum of an input image, we extract the spectral residual *of an image in spectral domain, and propose a fast method to construct the corresponding* saliency map *in spatial domain.*

We test this model on both natural pictures and artificial images such as psychological patterns. The result indicate fast and robust saliency detection of our method.

1. Introduction

The first step towards object recognition is object detection. Object detection aims at extracting an object from its background before recognition. But before performing recognitive feature analysis, how can a machine vision system extract the salient regions from an unknown background?

Traditional models, by relating particular features with targets, actually convert this problem to the detection of specific categories of objects[3]. Since these models are based on training, the expansibility become the bottleneck in generalized tasks. Facing unpredictable and innumerable categories of visual patterns, a general purpose saliency detection system is required. In other words, the saliency detector should be implemented with the least reference on statistical knowledge of the objects.

How is the saliency detection process achieved in human visual system? It is believed that two stages of visual processing are involved: first, the parallel, fast, but simple *pre-attentive* process; and then, the serial, slow, but complex *attention* process. Properties of pre-attentive processing have been discussed in literature [27, 24]. In this stage, certain low level features such as orientation, edges, or intensities can "pop up" automatically. From a viewpoint of object detection, what pops up in the pre-attentive stage is the candidate of an object. In order to address a candidate that has been detected but not yet identified as an object, Rensink introduced the notion of *proto objects* in his coherence theory [15, 13, 14].

To find the "proto objects" in a given image, models had been invented in the field of machine vision. Based on Treisman's integration theory [24], Itti and Koch proposed a saliency model that simulates the visual search process of human [8, 6, 7]. More recently, Walther extended the saliency model, and successfully applied it to object recognition tasks[26]. However, as a pre-processing system, these models are computationally demanding.

Most of the detection models focus on summarizing the properties of target objects. However, general properties shared by various categories of objects are not likely to exist. In this paper, we pose this problem in an alternative way: to explore the properties of the backgrounds.

In Section 2, the *spectral residual* is introduced. Starting from the principle of natural image statistics, we propose a front-end method to simulate the behavior of pre-attentive visual search. Different from traditional image statistical models, we analyze the log spectrum of each image and obtain the spectral residual. Then we transform the spectral residual to spatial domain to obtain the *saliency map*, which suggests the positions of proto-objects. In Section 3, we also demonstrate multiple object detection based on the spectral residual approach.

To evaluate the performance of our method, in Section 4.1, we compare our method with [8] and human-labeled results. The result indicates that our method is a fast and reliable computational model form early stage visual processing.

2. Spectral Residual Model

Efficient coding is a general framework under which many mechanisms of our visual processing can be interpreted. Barlow [1] first proposed the efficient coding hy-

pothesis that removes redundancies in the sensory input. A basic principle in visual system is to suppress the response to frequently occurring features, while at the same time keeps sensitive to features that deviate from the norm [9]. Therefore, only the unexpected signals can be delivered to later stages of processing.

From the perspective of information theory, effective coding decompose the image information $H(\text{Image})$ into two parts:

$$H(\text{Image}) = H(\text{Innovation}) + H(\text{Prior Knowledge}),$$

$H(\text{Innovation})$ denotes the novelty part, and $H(\text{Prior Knowledge})$ is the redundant information that should be suppressed by a coding system. In the field of image statistics, such redundancies correspond to statistical invariant properties of our environment. These properties have been comprehensively discussed in literature pertaining to natural image statistics [4, 25, 17, 18]. Now it is widely accepted that natural images are not random, they obey highly predictable distributions.

In the following sections, we will demonstrate a method to approximate the "innovation" part of an image by removing the statistical redundant components. This part, we believe is inherently responsible to the popping up of proto-objects in the pre-attentive stage.

2.1. Log spectrum representation

Of the invariant factors of natural image statistics, scale invariance is the most famous and most widely studied property [20, 17]. This property is also known as $1/f$ law. It states that the amplitude $\mathcal{A}(f)$ of the averaged Fourier spectrum of the ensemble of natural images obeys a distribution:

$$E\{\mathcal{A}(f)\} \propto 1/f. \tag{1}$$

On a log-log scale, the amplitude spectrum of the ensemble of natural images, after averaging over orientations, lies approximately on a straight line.

Although the log-log spectrum is theoretically matured and has been widely used, it is not favored in the analysis of individual images because: (1) the scale-invariance property is not likely to be found in individual images; (2) the sampling points are not well-proportioned, the low frequency parts span sparsely on the log-log plane, whereas the high frequency parts draw together, suffering from noise [25].

Instead of the log-log representation, in this paper, we adopt the log spectrum representation $\mathcal{L}(f)$ of an image. Log spectrum can be obtained by $\mathcal{L}(f) = \log\big(\mathcal{A}(f)\big)$. The comparison between log-log and log spectrum representation is shown in Fig.1.

The log spectrum representation has been used in a series of literature pertaining to statistical scene analysis

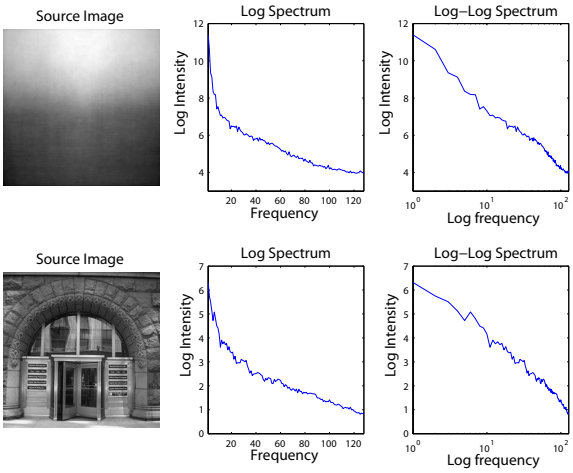

Figure 1. Examples of log spectrum and log-log spectrum. The first image is the average of 2277 natural images.

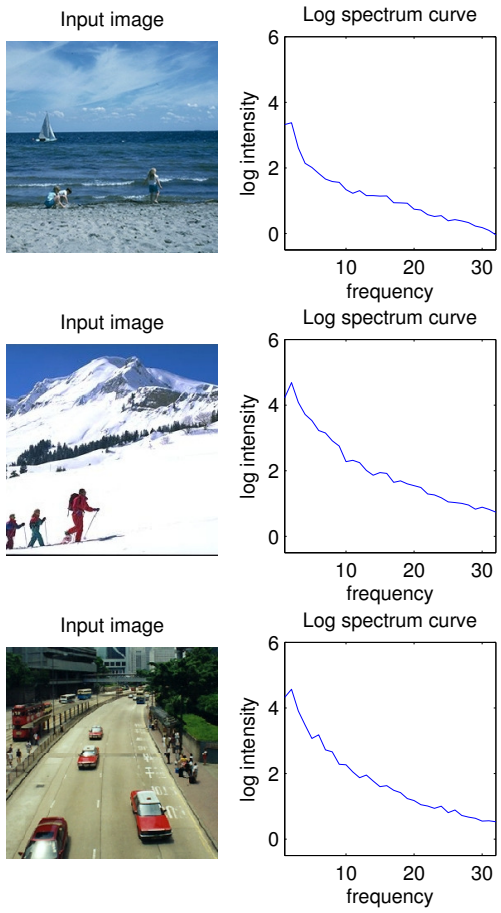

Figure 2. Examples of orientation averaged curves of log spectra. These curves share similar shape. The log spectrum is computed from down-sampled image. The size of each log spectrum is 64×64

[22, 23, 21, 11]. In the following section, we will exploit the power of log spectrum in saliency detection tasks. Ex-

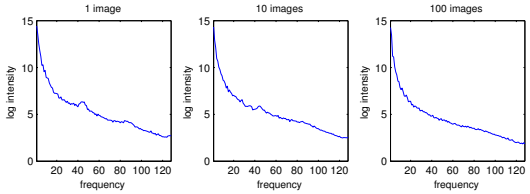

Figure 3. Curves of averaged spectra over 1, 10 and 100 images.

amples of the log spectra are presented in Fig.2. We find that the log spectra of different images share similar trends, though each containing statistical singularities. Fig.3 shows the curves of averaged spectra over 1, 10 and 100 images, respectively. This result suggests a local linearity in the averaged log spectrum.

2.2. From spectral residual to saliency map

Similarities imply redundancies. For a system aiming at minimizing the redundant visual information, it must be aware of the statistical similarities of the input stimuli. Therefore, in different log spectra where considerable shape similarities can be observed, what deserves our attention is the information that jumps out of the smooth curves. We believe that the statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects are popped up.

Given an input image, the log spectrum $\mathcal{L}(f)$ is computed from the down-sampled image with height (or width) equals 64 px. The selection of the input size is related to visual scale. The relationship between visual scale and visual saliency is discussed in Section 3.1.

If the information contained in the $\mathcal{L}(f)$ is obtained previously, the information required to be processed is:

$$H(\mathcal{R}(f)) = H\big(\mathcal{L}(f)|\mathcal{A}(f)\big), \tag{2}$$

where $\mathcal{A}(f)$ denotes the general shape of log spectra, which is given as prior information. $\mathcal{R}(f)$ denotes the statistical singularities that is particular to the input image. In this paper, we define $\mathcal{R}(f)$ as the *spectral residual* of an image.

Shown in Fig.3, the averaged curve indicates a local linearity. Therefore, it is reasonable to adopt a local average filter $h_n(f)$ to approximate the shape of $\mathcal{A}(f)$. In our experiments, n equals 3. Changing the size of $h_n(f)$ alters the result only slightly (see Fig.5). The averaged spectrum $\mathcal{A}(f)$ can be approximated by convoluting the input image:

$$\mathcal{A}(f) = h_n(f) * \mathcal{L}(f), \tag{3}$$

where $h_n(f)$ is an $n \times n$ matrix defined by:

$$h_n(f) = \frac{1}{n^2}\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

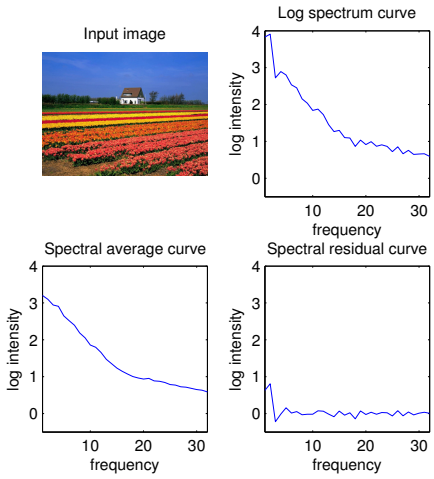

Figure 4. The shape information $\mathcal{A}(f)$ is removed from the original log spectrum $\mathcal{L}(f)$. The uniform distribution of spectral residual $\mathcal{R}(f)$ is desirable since similar response is expected in the neural representation of images [19].

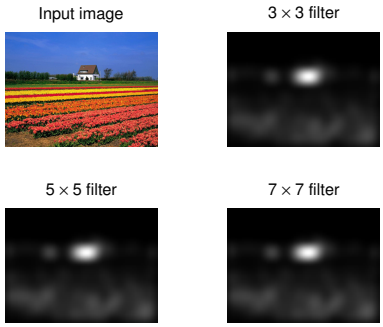

Figure 5. An example of using different average filter $h_n(f)$ in Eq.3. The size of $h_n(f)$ affects the result only slightly.

Therefore the *spectral residual* $\mathcal{R}(f)$ can be obtained by:

$$\mathcal{R}(f) = \mathcal{L}(f) - \mathcal{A}(f). \tag{4}$$

In our model, the spectral residual contains the innovation of an image. It serves like the compressed representation of a scene. Using Inverse Fourier Transform, we can in spatial domain construct the output image called the *saliency map*. The saliency map contains primarily the nontrivial part of the scene. The content of the residual spectrum can also be interpreted as the unexpected portion of the image. Thus, the value at each point in a saliency map is then squared to indicate the estimation error. For better visual effects, we smoothed the saliency map with a gaussian filter $g(x)$ ($\sigma = 8$).

In sum, given an image $\mathcal{I}(x)$, we have:

$$\mathcal{A}(f) = \Re\Big(\mathfrak{F}\big[\mathcal{I}(x)\big]\Big), \quad (5)$$

$$\mathcal{P}(f) = \Im\Big(\mathfrak{F}\big[\mathcal{I}(x)\big]\Big), \quad (6)$$

$$\mathcal{L}(f) = \log\big(\mathcal{A}(f)\big), \quad (7)$$

$$\mathcal{R}(f) = \mathcal{L}(f) - h_n(f) * \mathcal{L}(f), \quad (8)$$

$$\mathcal{S}(x) = g(x) * \mathfrak{F}^{-1}\Big[\exp\big(\mathcal{R}(f) + \mathcal{P}(f)\big)\Big]^2. \quad (9)$$

where $\mathfrak{F}$ and $\mathfrak{F}^{-1}$ denote the Fourier Transform and Inverse Fourier Transform, respectively. $\mathcal{P}(f)$ denotes the phase spectrum of the image, which is preserved during the process.

3. Detecting proto-objects in a saliency map

The saliency map is an explicit representation of proto-objects, in this section, we use simple threshold segmentation to detect proto-objects in a saliency . Given $\mathcal{S}(x)$ of an image, the object map $\mathcal{O}(x)$ is obtained:

$$\mathcal{O}(x) = \begin{cases} 1 & \text{if } \mathcal{S}(x) > \text{threshold}, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Empirically, we set $threshold = E\big(\mathcal{S}(x)\big) \times 3$, where $E\big(\mathcal{S}(x)\big)$ is the average intensity of the saliency map. The selection of threshold is a trade-off problem between false alarm and neglect of objects. A brief discussion of this problem is provided in Section 4.1.

While the object map $\mathcal{O}(x)$ is generated, proto-objects can be easily extracted from their corresponding positions in the input image. Multiple targets are extracted sequentially.

3.1. Selection of visual scales

A visual system works under certain scales. For example, in a large scale, one may perceive a house as an object, but in a small scale, it is very likely that the front door of the house pops up as an object. The selection of scale in our experiment is equal to the selection of the the input image size. When the image is small, detailed features are omitted, and the visual search is performed in a large scale. However, in a finer scale, large features becomes less competitive to the small but abrupt changes in the image. Changing the scale leads to a different result in the saliency map. This property can be illustrated in Fig.7.

The visual scale is tightly related to the optical ability of the visual sensors. For a pre-attentive task, it is reasonable to adopt a constant factor as an estimation of the visual scale. Since the spatial resolution of pre-attentive vision is very limited [5]. Without a slow process of scrutinizing, human are not likely to perceive the details of an image

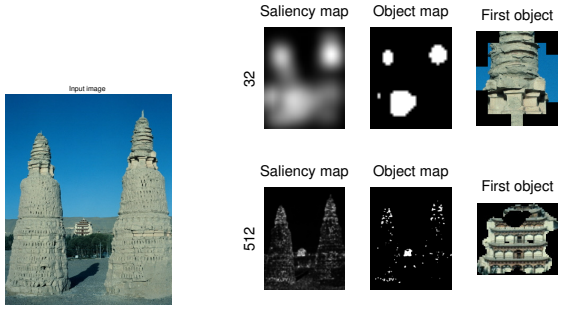

Figure 7. An example of attention in different scales.

which corresponds to the high frequency parts in the Fourier spectrum[12]. According to the simulation experiments, we find that 64 px of the input image width (or height) is a good estimation of the scale of normal visual conditions.

4. Experiments and analysis

It is not easy to evaluate the performance of an object detection system. One of the widely used measurements is the recording of eye movements [7]. However, this method is not applicable in our experiments, because an eye tracker records only positional information – sizes and shape of attended regions cannot be recorded. Furthermore, covert attention plays a role in object detection, proto-objects can be perceived without apparent eye motion.

4.1. Evaluating the result

In our experiment, we provide 4 naïve subjects with natural scene images. These images are taken from [11], [10], and [26]. Each subject is instructed to "select regions where objects are presented". If each of the subject reported impossible to define an object in a certain image, that image would be rejected from the data set. At last, 62 images are collected to test the performance of our method.

The purpose of the experiment is different from segmentation [10]. The main concern in segmentation tasks is the abrupt changes in space. But in our task, hand labelers concentrate only on the edges between the foreground and the background.

For each input $\mathcal{I}(x)$, the binary image obtained from k^{th} hand-labeler is denoted as $\mathcal{O}_k(x)$, in which 1 denotes for target objects, 0 for background. Given the generated saliency map $\mathcal{S}(x)$, the *Hit Rate* (HR) and the *False Alarm Rate* (FAR) can be obtained:

$$HR = E\big(\prod_k \mathcal{O}_k(x) \cdot \mathcal{S}(x)\big), \quad (11)$$

$$FAR = E\big(\prod_k (1 - \mathcal{O}_k(x)) \cdot \mathcal{S}(x)\big). \quad (12)$$

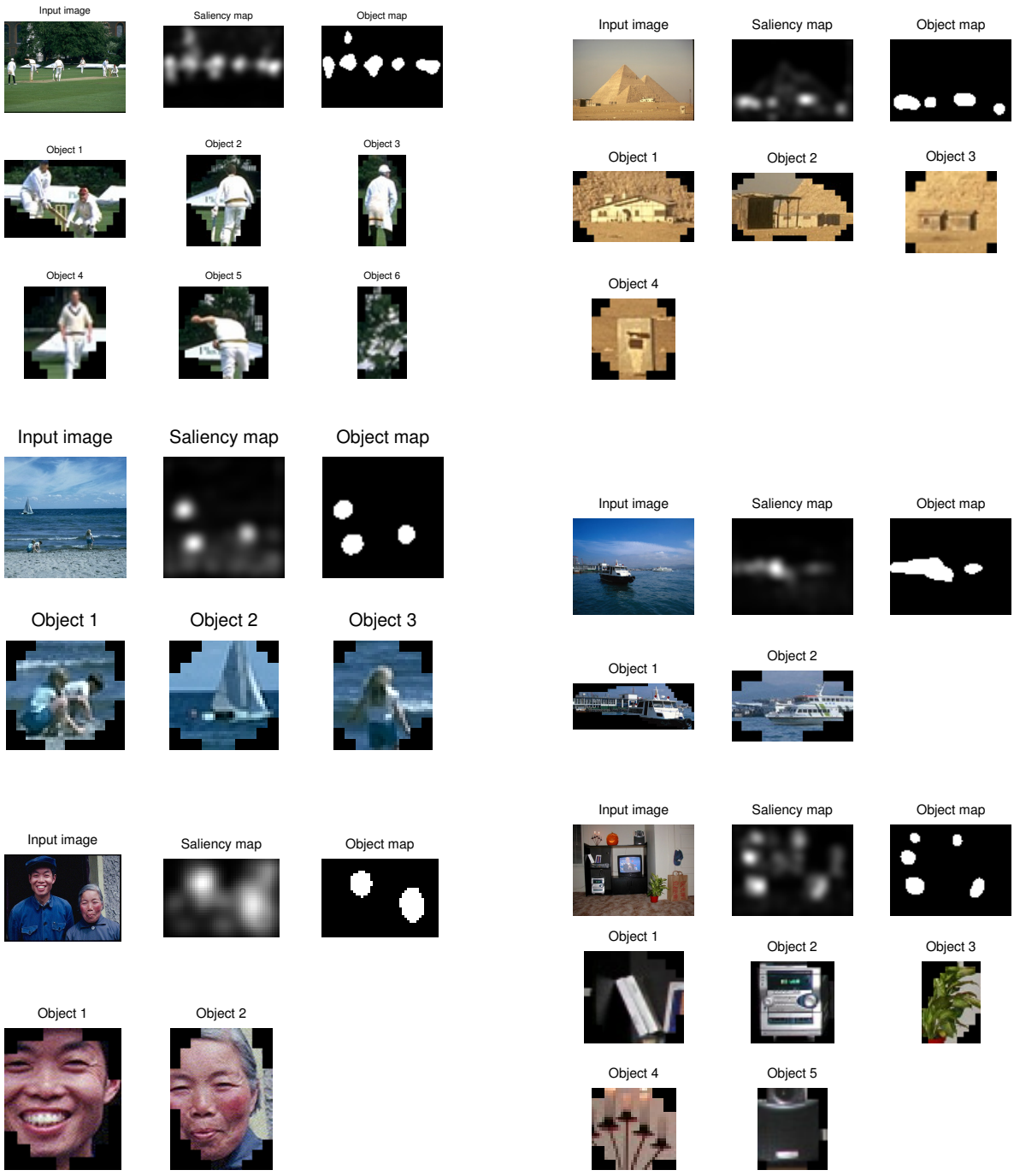

Figure 6. Detecting objects from input images. Objects are popped up sequentially according to their saliency map intensity.

This criterion states that an optimal saliency detection system should response low in regions where no hand-labeler suggests proto-object, and response high in region where most labelers meet at an consensus of proto-objects.

We compare our result with previous methods in the field, we also generate the saliency maps based on Itti's well known theory [8] as a control set. The MATLAB implementation of this method can be downloaded from `http://www.saliencytoolbox.net`. The image is down-sampled to 320×240 for Itti's method. For spectral

residual method, each color channel is processed independently. In order to make a comparison, we must set either FAR or HR of the two methods equal. For instance, given the FAR of the spectral residual saliency maps, we can adjust the saliency map of Itti's method $\mathcal{S}(x)$ by a parameter c:

$$\hat{\mathcal{S}}(x) = c \cdot \mathcal{S}(x), \tag{13}$$

and use $\hat{\mathcal{S}}(x)$ instead of $\mathcal{S}(x)$ to compute FAR and HR in Eq.11 and Eq.12. Similarly, given the HR of Itti's method, we linearly modulate the saliency maps generated by spectral residual.

Table 1. Performance of the two methods

	Spectral Residual	Itti's Method
HR	0.4309	0.2482
FAR	0.1433	0.1433
HR	0.5076	0.5076
FAR	0.1688	0.2931
Total time	4.014s	61.621s

From the result, we observe that our method provides overall better performance than Itti's method. Computationally, the cost of performing FFT is relatively low – this brings considerable advantage for a saliency detector, making it easier to implement on an existed system.

4.2. Responses to psychological patterns

We also test our method with artificial patterns. These patterns are adopted in a series of attention experiments [24, 27] in order to explore the mechanisms of pre-attentive visual search.

It is widely accepted that certain complex features are beyond the capability of pre-attentive perception, the more delicate and time-consuming search process must be employed to distinguish singularities in patterns such as "closure" in Fig.9. Correspondingly, our method fails to find out the unique circle among "c"'s.

5. Discussion

We proposed a method for general purpose object detection. This method is based on the log spectra representation of images. Our major contribution is the discovery of spectral residual and its general ability to detect proto-objects.

5.1. The prospect of spectral residual approach

One of the advantages of the spectral residual approach is its generality. The prior knowledge required for saliency detection is not necessary in our system. In addition, this all-in-one definition of saliency covers unknown features

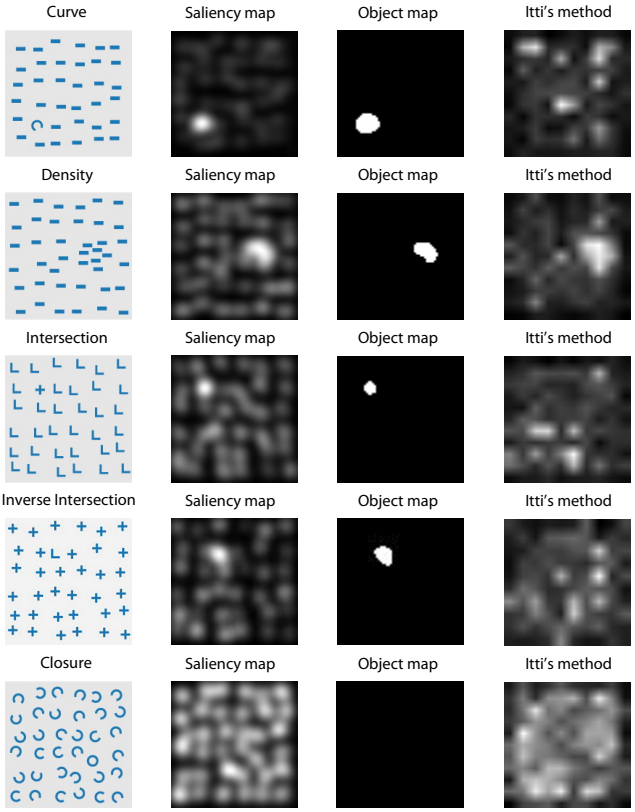

Figure 9. Responses to psychological patterns. In the figure of "Closure", no proto object is detected since no pixel has a output higher than $E\big(\mathcal{S}(x)\big) \times 3$.

such as "curve" in Fig.9. Also, the spectral residual resolves the problem of weighting features from different channels (for example, shape, texture, and orientations). The result of our system, in contrast with its simple implementation, is demonstrated effective. Finally, compared with other detection algorithms, the computational consumption of our method is extremely parsimonious, providing a promising solution to real time systems.

5.2. Further work

Is the striking similarities of our results and performance of human visual system, especially, the response to psychological patterns, all comes in a coincidence, or if there is biological implications of the human visual system and the spectral residual? It has been reported that different objects with similar frequency spectra interfere with each other [2]. More recent studies also indicate that a visual target takes more time to be identified when the spectrum of background is carefully tuned to mask the spectrum of the foreground [28]. More work is required to discover the spectral properties of early vision.

In this paper, our discussion is limited to static images. Although it is possible to compute the saliency map for each

Figure 8. The result of our method in comparison with Itti's method and the result of human labelers. In each group, we present 1) the input image, 2) saliency map generated by spectral residual, 3) saliency map generated by Itti's method, and 4) labeled map of the four labelers. In the labeled map, the white region represents the hit map, where $\prod \mathcal{O}_k(x) = 1$; the black region represents the false alarm map, where $\prod(1 - \mathcal{O}_k(x)) = 0$; and the gray region is selected by some labelers but rejected by others.

frames of a video sequence without considering their continuity, incorporating motion features will greatly extend the application of our method. Due to the particularity of motion features, a unified model of features has not yet been proposed. Yet, we are glad to see that efforts have been made in incorporating motion into a general framework of features [16].

Another potential work is to cooperate our method with segmentation techniques. Segmentation is an independent area of research whose primary goal is to separate borders. In comparison, our method overlooked the spatial homogeneity of an object. For instance, in the last example of Fig.8, the poloists and their horses are separated. In order to achieve the general purpose object detection, further efforts should be done to delimit a clear border of an object.

6. Acknowledgement

The work was the National High-Tech Research Program of China (Grant No.252006AA01Z125) and supported by the National Basic Research Program of China (Grant No. 2005CB724301). The first author would like to thank Deli Zhao, Dirk Walther, and Yuandong Tian for their valuable discussions.

References

[1] H. Barlow. Possible Principles Underlying the Transformation of Sensory Messages. *Sensory Communication*, pages 217–234, 1961. 1

[2] H. Egeth, R. Virzi, and H. Garbart. Searching for Conjunctively Defined Targets. *Journal of Experimental psychology: Human Perception and Performance*, 10(1):32–39, 1984. 6

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Proc. CVPR*, 2, 2003. 1

[4] J. Gluckman. Order Whitening of Natural Images. *Proc. CVPR*, 2, 2005. 2

[5] J. Intriligator and P. Cavanagh. The Spatial Resolution of Visual Attention. *Cognitive Psychology*, 43(3):171–216, 2001. 4

[6] L. Itti and C. Koch. A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention. *Vision Research*, 40(10-12):1489–1506, 2000. 1

[7] L. Itti and C. Koch. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. 1, 4

[8] L. Itti, C. Koch, E. Niebur, et al. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 1, 5

[9] C. Koch and T. Poggio. Predicting the Visual World: Silence is Golden. *Nature Neuroscience*, 2(1):9–10, 1999. 2

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. *Proc. ICCV*, 2, 2001. 4

[11] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 2, 4

[12] A. Oliva, A. Torralba, and P. Schyns. Hybrid Images. *ACM Transactions on Graphics (TOG)*, 25(3):527–532, 2006. 4

[13] R. Rensink. Seeing, sensing, and scrutinizing. *Vision Research*, 40(10-12):1469–87, 2000. 1

[14] R. Rensink and J. Enns. Preemption Effects in Visual Search: Evidence for Low-Level Grouping. *Psychological Review*, 102(1):101–130, 1995. 1

[15] R. Rensink, J. ORegan, and J. Clark. To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5):368–373, 1997. 1

[16] S. Roth and M. Black. On the Spatial Statistics of Optical Flow. *Proc. ICCV*, 1, 2005. 7

[17] D. Ruderman. The Statistics of Natural Images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994. 2

[18] D. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3395, 1997. 2

[19] E. Simoncelli and B. Olshausen. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. 3

[20] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu. On Advances in Statistical Modeling of Natural Images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003. 2

[21] A. Torralba. Modeling Global Scene Factors in Attention. *Journal of the Optical Society of America*, 20(7):1407–1418, 2003. 2

[22] A. Torralba and A. Oliva. Depth Estimation from Image Structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1226–1238, 2002. 2

[23] A. Torralba and A. Oliva. Statistics of Natural Image Categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003. 2

[24] A. Treisman and G. Gelade. A Feature-Integration Theory of Attention. *Cognitive Psychology*, 12(1):97–136, 1980. 1, 6

[25] A. van der Schaaf and J. van Hateren. Modelling the Power Spectra of Natural Images: Statistics and Information. *Vision Research*, 36(17):2759–2770, 1996. 2

[26] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional Selection for Object Recognition – a Gentle Way. *Lecture Notes in Computer Science*, 2525(1):472–479, 2002. 1, 4

[27] J. Wolfe. Guided Search 2.0: A Revised Model of Guided Search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994. 1, 6

[28] J. Wolfe, A. Oliva, T. Horowitz, S. Butcher, and A. Bompas. Segmentation of Objects from Backgrounds in Visual Search Tasks. *Vision Research*, 42(28):2985–3004, 2002. 6