Visual Saliency Based on Conditional Entropy

Yin Li, Yue Zhou, Junchi Yan and Jie Yang

Institute of Image Processing and Pattern Recognition,
Department of Automation, Shanghai Jiaotong University
{happyharry,zhouyue,yanesta,jieyang}@sjtu.edu.cn

Abstract. By the guidance of attention, human visual system is able to locate objects of interest in complex scene. In this paper, we propose a novel visual saliency detection method - the conditional saliency for both image and video. Inspired by biological vision, the definition of visual saliency follows a strictly local approach. Given the surrounding area, the saliency is defined as the minimum uncertainty of the local region, namely the minimum conditional entropy, when the perceptional distortion is considered. To simplify the problem, we approximate the conditional entropy by the lossy coding length of multivariate Gaussian data. The final saliency map is accumulated by pixels and further segmented to detect the proto-objects. Experiments are conducted on both image and video. And the results indicate a robust and reliable feature invariance saliency.

Key words: Saliency Detection, Conditional Entropy, Lossy Coding Length

1 Introduction

In the real world, human visual system demonstrates remarkable ability in locating the objects of interest in cluttered background, whereby attention provides a mechanism to quickly identify subsets within the scene that contains important information. Besides the scientific goal of understanding this human behavior, a computational approach to visual attention also contributes to many applications in computer vision, such as scene understanding, image/video compression, or object recognition.

Our ultimate goal is to develop a computational model to locate objects of interest in different scenes. The human visual system follows a center-surround approach in early visual cortex [1, 2]. When the surrounding items resemble close to the central item, the neuron response to the central location is suppressive modulated; but when they differ grossly, the response is excitatory modulated [3]. Therefore, the regions in the visual field, which most differ from their surroundings in certain low level features would automatically pop up. The candidates popped up here are called the proto-object [4, 5], which could be further grouped into coherent units to be perceived as real objects. How can a computer vision system benefit from this procedure?

Many efforts have been focused on this topic. Most of them follow a center surround and bottom up approach [6–10]. Typically, these methods are based on a set of biological motivated features, followed by center-surround operations that enhance the local stimuli, and finally a combination step leading into a saliency map. In [11], a model of overt attention with selection based on self information is proposed, where the patches of a image are decomposed into a set of pre-learned basis and kernel density estimation is used to approximate the self-information. In [12, 13], saliency models based on spectral information are proposed and provide a novel solution.

Inspired by the aforementioned research, we propose to measure the saliency as the the minimum conditional entropy [14], which represents the uncertainty of the center-surround local region, when the surrounding area is given and the perceptional distortion is considered. The minimum conditional entropy is further approximated by the lossy coding length of Gaussian data, which balances between complexity and robustness. The final saliency map accumulated by pixels is an explicit representation of the proto-objects. Thus, we simply segment the proto-objects by thresholding. Fig.1 demonstrates the procedure of our method.

Fig. 1. The procedure of our method: original image, saliency map and possible proto objects (from top to bottom).

Overall, we summarize our main contribution into two folds. First, the perceptional distortion is introduced. And the conditional entropy under the distortion is proposed to measure the visual saliency. Second, due to the strictly local approach, our method is insensitive to ego-motion or affine transform, applicable to both image and video, and free from prior knowledge or pre-training.

The paper is organized as follows. A local saliency model based on conditional entropy is presented in Section 2. In Section 3, the saliency map is extended to the proto-object detection. Experimental evaluations and result analysis are performed in Section 4. Finally, Section 5 concludes the paper.

2 Conditional Saliency

2.1 Lossy Coding Length of Multivariate Gaussian Data

We first present a brief introduction to the lossy coding length of multivariate Gaussian data [15], which is used to calculate the visual saliency in the following section. Assume a set of vectors $W = (w_1, w_2, ..., w_m) \in R^{n \times m}$ is presented, a lossy coding scheme $L(\cdot)$ maps W to a sequence of binary bits. Thus, the original vectors can be recovered up to an allowable distortion $E[\|w_i - \tilde{w_i}\|^2] \leq \varepsilon^2$. If the data are i.i.d. samples from a multivariate Gaussian distribution, the length of the encoded sequence is denoted by $L(W)$

$$L_\varepsilon(W) \doteq \frac{m+n}{2} \log_2 \det(I + \frac{n}{m\varepsilon^2} \bar{\boldsymbol{W}} \bar{\boldsymbol{W}}^T) + \frac{n}{2} \log_2(1 + \frac{\mu^t \mu}{\varepsilon^2}) \tag{1}$$

where $\mu = \frac{1}{m} \sum_1^m w_i$ and $\bar{\boldsymbol{W}} = [w_1 - \mu, w_2 - \mu, ..., w_m - \mu]$.

Note that the first term of the equation stands for the coding length required to code $\bar{\boldsymbol{W}}$, and the second term of the equation stands for the additional coding length of the mean vector. The equation also gives a good upper bound for degenerated Gaussian data or subspace-like data [15]. Moreover, the coding length has proven to be effective for clustering [16] and classification [15].

2.2 Saliency as Conditional Entropy

One basic principle in visual system is to suppress the response to frequently occurring input patterns, while at the same time keeping sensitive to novel input patterns. Thus, we could define the saliency of the center area as the uncertainty of the center-surround local region when the surrounding are is provided. Such uncertainty is naturally measured by the conditional entropy in the information theory.

Assume image I is divided into small spatial patches. Denote $c \in I$ as a patch of image I. Let $S(c) = [s_1, s_2, ...s_m]$ be the spatial or spatio-temporal surrounding patches of c and $SC(c) = S(c) \bigcup c$ be the center-surround area, overlapping is allowed for surroundings in order to capture the structure information of the image (Fig.2). We further assume a feature transform F is performed across the image. Let vector $\boldsymbol{x} = Fc \in R^n$ identify the center patch by sticking its columns. Let $\boldsymbol{S} = FS = [Fs_1, Fs_2, ...Fs_m] \in R^{n \times m}$ and $\boldsymbol{SC} = FSC(c)$ identify the surroundings and center-surround area, respectively.

Suppose distortion ε exists across the local region, i.e. what we perceive $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{S}}$ is not exact what what we see $\boldsymbol{x}$,$\boldsymbol{S}$. Thus, the input of our method is some latent variable $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{S}}\boldsymbol{c}$ rather than $\boldsymbol{x}$, $\boldsymbol{S}$ and $\boldsymbol{SC}$. Even if we do not know the latent variable, we can expect a certain constrain i.e. the tolerance of perception:

$$\begin{aligned} &E[\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2] \leq \varepsilon^2 \\ &E[\|Fs_i - \tilde{Fs_i}\|^2] \leq \varepsilon^2 \end{aligned} \tag{2}$$

The distortion ε stands for the tolerance of perception. Two reasons account for ε: 1)there exists distortion during the acquisition of the image, thus the

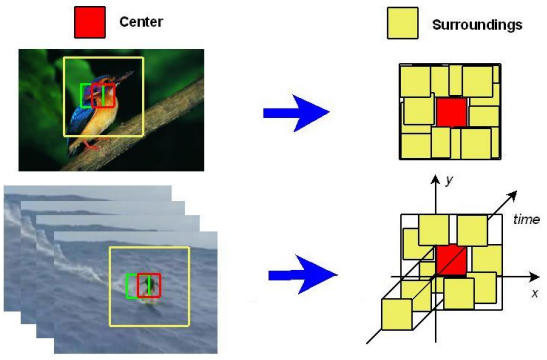

Fig. 2. Our center-surround architecture: overlapping surroundings of the center patch.

observed data contains noise; 2)there exists distortion during the perception of the human vision system: we are still able to recognize the objects of interest even if the distortion is severe [2]. We believe it is the first time that the perceptional tolerance of distortion is considered in visual saliency.

Inspired by the center-surround modulation in human vision system [3], we define the local saliency as the minimum conditional entropy $H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}})$ under distortion ε:

$$\begin{aligned} \inf_{Q_{\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}})} H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}) \quad s.t. \quad & E[\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2] \le \varepsilon^2 \\ & E[\|Fs_i - \tilde{Fs_i}\|^2] \le \varepsilon^2 \end{aligned} \tag{3}$$

where $H(\tilde{x}|\tilde{S})$ is the conditional entropy by

$$\begin{aligned} H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}) &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Q_{\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}) P_{\tilde{\boldsymbol{S}}}(\tilde{\boldsymbol{S}}) \log_2(Q_{\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}))d\tilde{\boldsymbol{x}}d\tilde{\boldsymbol{S}} \\ &= H(\tilde{\boldsymbol{SC}}) - H(\tilde{\boldsymbol{S}}) \end{aligned} \tag{4}$$

$Q_{\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}})$ is the conditional probability density function, $H(\cdot)$ is the information entropy [14].The equation (3) is an extension to the concept of rate-distortion function [14] in information theory.

Intuitively, the whole local region contains $H(\tilde{\boldsymbol{SC}})$ bits of information: we need $H(\tilde{\boldsymbol{SC}})$ bits of information to reconstruct its state. If we know the surroundings $\boldsymbol{S}$, we have gained $H(\tilde{\boldsymbol{S}})$ bits of information, and the local region still has $H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}})$ bits remaining of uncertainty. The larger the uncertainty is, the more salient the center would be, vice verse. $H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}) = 0$ if and only if the center c can be completely determined by its surroundings $S(c)$, which indicates high similarity. Conversely, $H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}) = H(\tilde{\boldsymbol{x}})$ if and only if $\boldsymbol{x}$ and $\boldsymbol{S}$ are drawn from two independent random variables, which indicates high saliency.

2.3 Gaussian Conditional Saliency

The conditional saliency model is simple in concept, intuitive, and highly flexible. However, the equation (3) is computational intractable. Instead of calculating (3) directly, we approximate the entropy by the lossy coding length. We could simplify the problem by making the following assumption:

Assumption: $\boldsymbol{S}$ *and* $\boldsymbol{SC}$ *are both multivariate Gaussian data.*

Extensive research on the statistics of natural images has shown that such a density can be approximated by Generalized Gaussian Distribution [17]. For simplicity, we use a Gaussian distribution instead. Although such a simplified assumption is hardly satisfied in real situations, experimental results indicate a robust solution to the visual saliency. In fact, the assumption can be extended to degenerated Gaussian or sub space like data [15].

By this assumption, the lossy coding length $L_\varepsilon(\boldsymbol{SC})$ and $L_\varepsilon(\boldsymbol{S})$ are reasonable estimations for $H((\tilde{\boldsymbol{SC}}))$ and $H((\tilde{\boldsymbol{S}}))$ under the distortion ε, since the optimal coding length is exactly the information entropy $H(\cdot)$ [14]. Thus, the minimum of the conditional entropy can be approximated by

$$\begin{aligned} H(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{S}}) &= H(\tilde{\boldsymbol{SC}}) - H(\tilde{\boldsymbol{S}}) \\ &\doteq L_\varepsilon(\boldsymbol{SC}) - L_\varepsilon(\boldsymbol{S}) \end{aligned} \tag{5}$$

where $L_\varepsilon(\boldsymbol{S})$ and $L_\varepsilon(\boldsymbol{SC})$ are given by

$$\begin{aligned} L_\varepsilon(\boldsymbol{S}) &= \frac{m+n}{2}\log_2\det(I+\frac{n}{m\varepsilon^2}\bar{\boldsymbol{S}}\bar{\boldsymbol{S}}^T) + \frac{n}{2}\log_2(1+\frac{\mu_s^t\mu_s}{\varepsilon^2}) \\ L_\varepsilon(\boldsymbol{S}C) &= \frac{m+n+1}{2}\log_2\det(I+\frac{n}{(m+1)\varepsilon^2}\bar{\boldsymbol{SC}}\bar{\boldsymbol{SC}}^T) + \frac{n}{2}\log_2(1+\frac{\mu_{sc}^t\mu_{sc}}{\varepsilon^2}) \end{aligned} \tag{6}$$

μ_s and μ_{sc} are the mean vector of $\boldsymbol{S}$ and $\boldsymbol{S}C$ respectively. And $\bar{\boldsymbol{S}} = [Fs_1 - \mu_s, ..., Fs_m - \mu_s]$, $\bar{\boldsymbol{SC}} = [x - \mu_{sc}, Fs_1 - \mu_{sc}, ..., Fs_m - \mu_{sc}]$.

The final saliency of the center can be calculated by

$$sal(c, S(c)) \;=\; L_\varepsilon(\boldsymbol{SC}) - L_\varepsilon(\boldsymbol{S}) \tag{7}$$

The only user defined parameter of our model is the distortion ε, which is discussed in Section 4. Note that the equation (7) corresponds to the incremental coding length in [16], where the coding length is used for classification. Since overlapping is allowed in the surroundings, we always have $n < m$. Thus, the $sal(c, S(c))$ can be computed in $O(n^2)$ time for each patch c. The total time of our algorithm is $O(Kn^2)$, where K is the number of patches in the image. Moreover, the parallel nature of the our algorithm is similar to the early human visual system.

3 Proto-Object Detection from Saliency Map

The saliency map is an explicit representation of the proto-object, we adopt a simple threshold to segment the proto-object. Given the saliency map $sal(I)$ of image I, we get the proto object map $P(I)$ by

$$P(I) = \begin{cases} 1 & \text{if } sal(I) \geq threshold \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

We set $threshold = 3 * E(sal(I))$ according to our pre-experiments, where $E(sal(I))$ is the average value of the saliency map.

3.1 The Role of Features

Traditional center surround saliency deals with a set of biology motivated features, such as intensity, orientation, color and motion [6–10]. In general, dozens of feature channels are extracted, and finally combined into a master saliency map. Nevertheless, during the experiment of our method, we find the method insensitive to the features. Though dozens of features may slightly improve the performance in some images, our method provides a fairly good solution with only the pixel value of the image/video. Fig.3 demonstrate a comparison of the features and raw data.

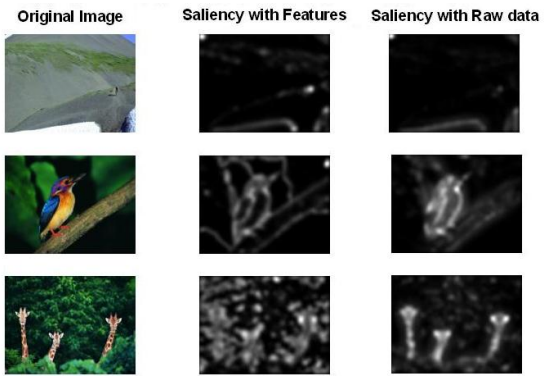

Fig. 3. Though feature integration may improve the performance(see the first row), it may introduce some new problems(see second and third rows).

Certain feature invariance may be explained by the invariance of the coding length under some transforms, e.g. the orthogonal transform will surely preserve the coding length [15]. It may indicate that the traditional feature integration introduce extra redundancy for visual saliency. Further experiment could be found in Section 4. And the reason behind this phenomenon is left as our future work.

4 Experiment and Result Analysis

To evaluate the performance of our method, three different kinds of experiments including both image and video are conducted. And the results are compared to the main stream approaches in the field [6, 11, 12].

We set the saliency map of our method at 160×120 in all experiments. The center patch is set 4×4 pixel, where its surroundings are set 7 times the size of the center. For simplicity, we also normalize the pixel value into [0 1]. All the tests are executed in Matlab on the platform of linux.

4.1 Evaluation of the Results

One problem with visual saliency in general, is the difficulty in making impartial quantitative comparisons between methods. Since our goal is to locate the objects of interest, the key factor is the accuracy, which could be shown as the Hit Rate(HR) and False Alarm Rate(FAR). The candidate of correct object are labeled by affiliated volunteers from a voting strategy. For each input image $I(x)$, we have a corresponding hand-labeled binary image $O(x)$, in which 1 denotes target objects while 0 denotes background. Given the saliency map $S(x)$, the Hit Rate(HR) and False Alarm Rate(FAR) could be obtained by

$$\begin{aligned} HR &= E(O(x) \cdot S(x)) \\ FAR &= E((1 - O(x)) \cdot S(x)) \end{aligned} \tag{9}$$

A good saliency detection should have a high HR together with a low FAR. However, if we adjust the saliency map $sal(I)$ by $C \cdot sal(I)$, we would get $C \cdot HR$ and $C \cdot FAR$ consequently. Therefore, we evaluate the accuracy by the following accuracy rate(AR):

$$AR = \frac{HR}{FAR} \tag{10}$$

Higher AR indicates a high HR v.s a low FAR, vice verse.

4.2 Natural Images

In order to give a comprehensive comparison among the methods, 62 natural images are used as a test set, which are also used in [12, 18]. First, we tested our method without features, namely using only the pixel values of the images. Next, we tested our method within the framework of feature integration in [6].

The only parameter of our method is the distortion ε. Fig.4 shows the effect of distortion in both situations. As we mentioned in Section 3, the features may even slightly decrease the performance of our method in general. The results also indicate that when ε is sufficiently large, it does little positive effect on the final results. Therefore, we set $\varepsilon^2 = 0.6n$ in the following experiments. Fig.5 presents some of our detection results.

We also compare our conditional saliency(CS) with Itti's theory [6], self information [11] and spectral residues [12](see Fig.6). For Itti's method, all the

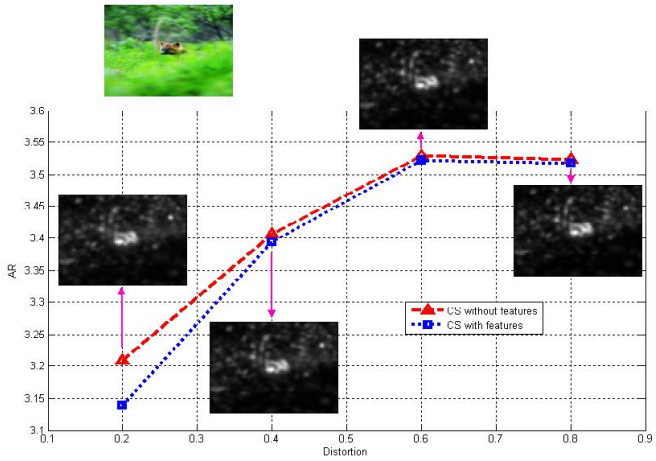

Fig. 4. Quantitative analysis of ε^2: When the distortion is around 0.6, the experiment shows a high accuracy rate. And it also indicates that feature integration would generally deteriorate the performance in our method.

parameter are set as default [6]. For self information, patches with size 8×8 are used. Also 200 natural images are used for training. Furthermore, we set the resolution of its input at 320×240 and resize the output into 160×120, in order to avoid the boundary effect [11]. For spectral residues, we resize the input into 160×120, since its original input resolution is 64×64.

Itti's method	*Spectral residue*	*Self information*	*CS with features*	*CS without features*
1.7605	3.2092	3.5107	3.5224	3.5295

Table 1. Comparison between [6, 11, 12] and our method. Note that to achieve a high accuracy rate, the method in [11] is trained by 200 natural images in a higher resolution(320×240).

From Fig. 5 and Fig. 5, we can see that our method generally produce a more acute and discriminant result than previous methods [6, 11, 12]. Moreover, the method tends to preserve the boundary information, which is beneficial in further perceptional grouping in the post attention stage. That is to say, the proto objects by our method is easier to be grouped into real objects.

Table.1 compares the performance in a quantitative way as mentioned in Section 4.1, our method outperforms state-of-the-art methods in the field. The introduction of the perceptional distortion should be responsible for the performance.

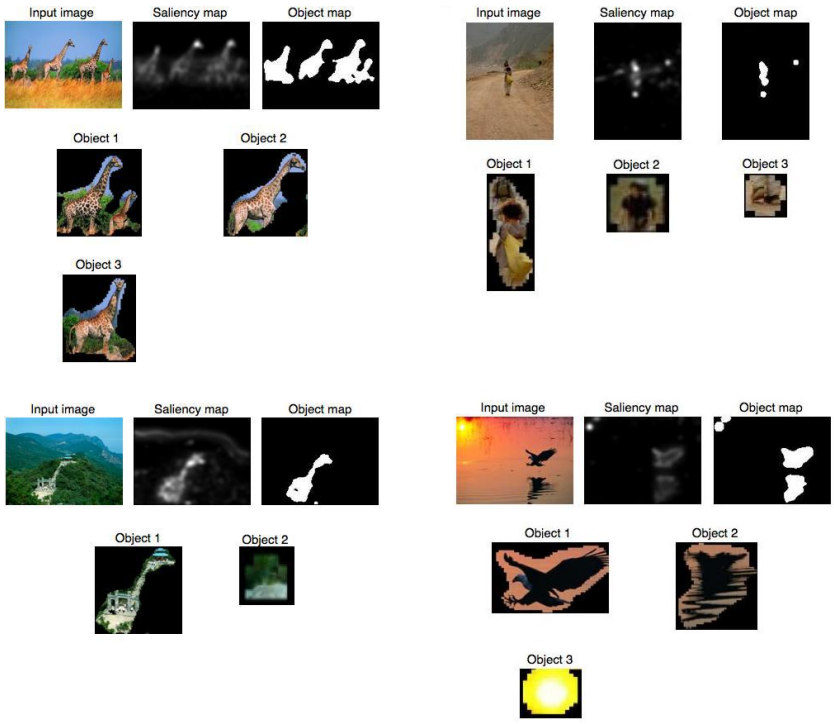

Fig. 5. Detection results of our method: our method is able to capture the possible proto objects in high accuracy.

4.3 Psychological Patterns

We also test our method on psychological patterns, which are adopted in a series of attention experiments to explore the mechanisms of pre-attentive visual search [19]. The preliminary results can be found in Fig.7. Our method is able to deal with simple psychological patterns.

4.4 Video Sequences

Two videos collected from the internet with the resolution 160×120 are used to test the performance of our method. One has a stationary background with moving pedestrians and vehicle; the other one has a highly dynamic background with surfers riding a wave. The center-surround architecture provides our method insensitivity to ego-motion. The parameter is similar to the natural image, except that extra 10 frames are used as the spatio-temporal surroundings of the center patch. Fig.8 presents the results of our methods. Our method shows pretty good performance.

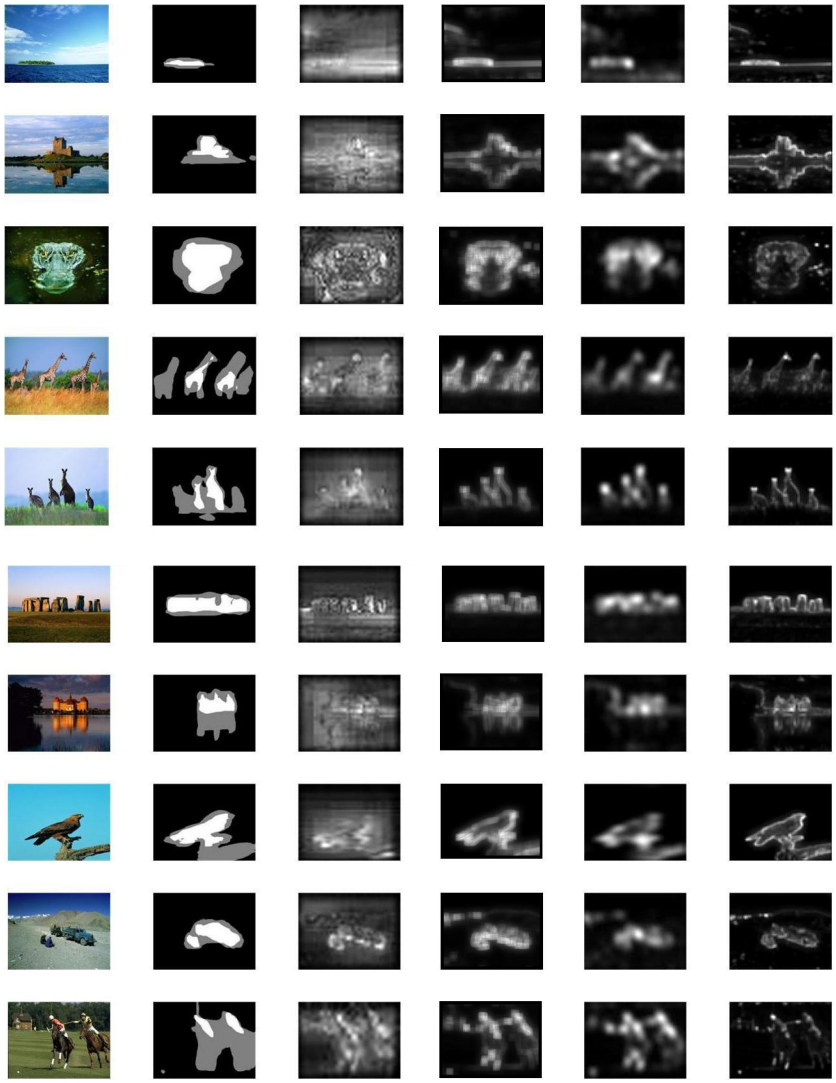

Fig. 6. Comparison of our method with [6, 11, 12],the first column is the original image, the second column is the hand-labeled image, the third to sixth column are results produced by [6, 11, 12] and our method, respectively.

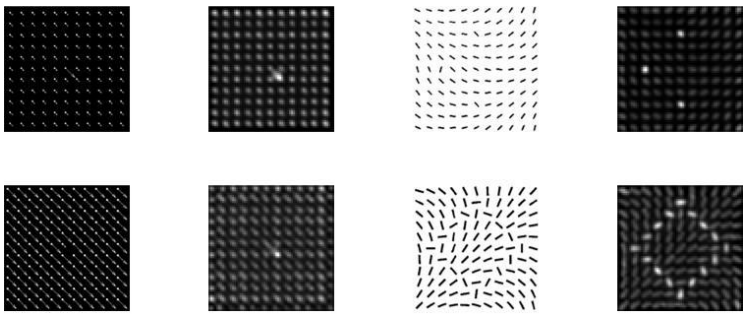

Fig. 7. Saliency map produced by our method on psychological patterns.

Fig. 8. Saliency map produced by our method in videos with different scenes.

5 Conclusion and Future Work

In this paper, we propose a novel saliency detection method based on conditional entropy under distortion. Though the visual saliency is still defined locally, we extend the concept of rate-distortion function to measure the saliency of the local region, where perceptional tolerance is considered. For the future work, we would investigate the roles of different features and try to find the reason behind the feature invariance in our method. Moreover, we would like to further explore the application of our method in computer vision.

6 Acknowledgement

The authors would like to thank the reviewers for their valuable suggestions. This research is partly supported by national science foundation, China, No.60675023, No.60772097; China 863 High-Tech Plan, No.2007AA01Z164.

References

1. Cavanaugh, J.R., B.W.M.J.: Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. Journal of Neurophysiol **88** (Nov. 2002) 2530–2546
2. J. Allman, F. Miezin, E.M.: Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. Annual Review of Neuroscience **8** (1985) 407 – 430
3. Vinje, W.E., Gallant, J.L.: Sparse coding and decorrelation in primary visual cortex during natural vision. Science **287**(5456) (2000) 1273–1276
4. Rensink, R.A.: Seeing, sensing, and scrutinizing. Vision Research **40**(10-12) (2000) 1469–1487
5. Rensik, R.A., Enns, J.T.: Preemption effects in visual search: Evidence for low-level grouping. Psychological Review **102** (1995) 101–130
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on **20**(11) (Nov 1998) 1254–1259
7. L Itti, C.K.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40**(10-12) (2000) 1489 – 1506
8. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on **1** (June 2005) 631–637 vol. 1
9. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 545–552
10. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: ICCV. (2007) 1–6
11. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA (2006) 155–162
12. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on (June 2007) 1–8
13. Mahadevan, V., Vasconcelos, N.: Background subtraction in highly dynamic scenes. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (June 2008) 1–6
14. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379C423, 623C656
15. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy data coding and compression. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(9) (Sept. 2007) 1546–1562
16. Wright, J., Tao, Y., Lin, Z., Ma, Y., Shum, H.Y.: Classification via minimum incremental coding length (micl). In: Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA (2008) 1633–1640
17. Mallat, S.: A theory for multiresolution signal decomposition: The wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on (1989)
18. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19**(9) (2006) 1395 – 1407 Brain and Attention, Brain and Attention.
19. Wolfe, J.: Guided search 2.0, a revised model of guidedsearch. Psychonomic Bulletin and Review **1**(2) (1994) 202–238