2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Visual Saliency via Loss coding

Hao Zhu and Biao Han

***Abstract*— A novel and effective bottom-up saliency model inspired by the recent findings of the early vision system is proposed. The lossy coding length, which resembles the neural cost in the hierarchical structure of human vision system, is exploit to measure saliency. We show that the proposed efficient coding network can be considered as the coding process in the early vision system. The sparse coding process in simple cells of the primary visual cortex and a dimensionality reduction process via the principal component analysis are integrated in the proposed network. The saliency value at each image pixel is computed based on the residual of the coding process. The proposed biological-inspired saliency model is evaluated on two different eye-tracking datasets against several state-of-the-art algorithms. Experimental results demonstrate the effectiveness, efficiency as well as robustness of the proposed model, and bear out the hypothesis of lossy coding for visual saliency.**

I. Introduction

Rich visual information of any scene entails the human visual system (HVS) to constantly process enormous amount of data. According to the classical findings by Kelly [14], retinal receptors receive information at an estimated rate of 10^9 bits per second. As a key process in the HVS, visual attention helps humans understand the scenes and recognize objects by rapidly selecting the highly relevant information. In the meanwhile, the principle of efficient coding is the widely accepted for explaining visual perception in the HVS. Extensive research work in neuroscience, psychology and computer vision has been done, with focus on not only to explain visual attention mechanisms but also understand the ways to perceive the world. In addition, the study of how humans process such astronomical amount of visual data is of great interest and importance to computer vision with numerous applications.

In the context of modeling, efficient coding and visual attention can be respectively considered as the process of feature extraction and feature selection. The efficient coding step transforms the input data from retinal receptors into a reduced representation set of features which facilitates further visual processing [1]. This process is by its nature can be modeled by lossy coding in HVS for great dimensionality reduction of visual data. Visual attention, which can be differentiated according to its states as "overt" and "covert" [25], selects a subset of relevant information for the further processing. Overt attention directs the high resolution fovea sensors towards the selected stimulus source, and covert attention enhances a particular part of the sensory panorama in the neural process. Unlike most computer vision algorithms, there is no particular order in the "feature extraction" and "feature selection" process. Efficient coding and visual attention work closely to deal with enormous amount of visual information and help humans perceive the world.

Hao Zhu is with the 3M Cogent Beijing R&D Center, (email: ahzhu@mmm.com), and Biao Han is with Université de Toulouse, Centre de Recherche Cerveau et Cognition, (email: biao.han@cerco.ups-tlse.fr)

Koch and Ullman propose a saliency map model based on the difference of a stimulus from its surroundings [15]. This saliency map is developed based on the feature integration theory [21] that models visual attention mechanism by combining information from feature maps (e.g., color, orientation, movement, etc.). In computer vision, Itti and Koch [12] propose the classic saliency model which is derived from this center-surrounding model. Numerous saliency models have since been proposed based on information theory [3], [10], [16], [26], spectral theory [9], [5], graph theory [7], [24], and feature learning [13]. While these algorithms perform well in detecting salient objects in scenes, they can hardly explain the relationship between these computational models and real neural systems.

In this work, we propose a saliency model which aims to explain the relationship between efficient coding theory and bottom-up saliency map in the early vision system. The proposed model exploits the spatial slowest component, which can be considered as the neural cost in the hierarchical structure of the HVS, to measure the potential entropy loss of the efficient coding process. We first model the sparse coding process in simple cells of the primary visual cortex (V1). The results of this coding process suggest strong correlations among signals which can be compressed for efficient computation. A dimensionality reduction step based on principal component analysis (PCA) is thus incorporated. Finally, the lossy coding cost in this coding network is used to measure visual saliency.

Figure 1 shows the main steps of the proposed model. First, an input image is divided into non-overlapping patches of fixed size and each is filtered with a number of sparse coding bases to model the responses of simple cells. Second, the response of simple cells is further compressed by PCA. Third, lossy coding length which measures the reconstruction residual in this coding network, is used to compute the saliency value. Patches with high reconstruction residuals indicate these are distinct areas which cannot be easily coded, and thus exhibit strong saliency.

Our contributions of this work are three-fold. First, we propose a novel computational model to model the bottom-up attention process in neural mechanisms. Second, the proposed model exploits the relationship between efficient coding and bottom-up saliency. Third, we show that the proposed model outperforms several state-of-the-arts methods, which can be applied to real-world vision tasks.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

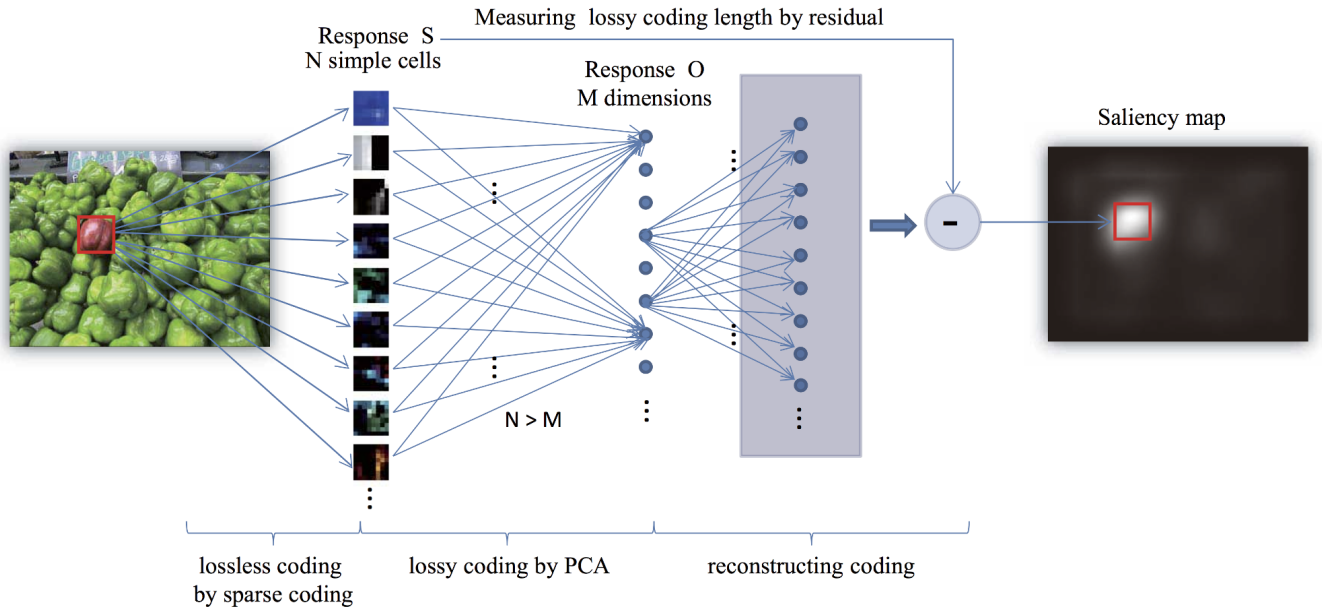


Fig. 1. Proposed framework. An patch in location (x, y) is filtered by sparse coding basis functions to obtain the corresponding response in the receptive field of simple cells. This response is then encoded using PCA to reduce the redundancy. By measuring the lossy coding length, we compute the saliency value of the patch at location (x, y).

II. Saliency Model

V1 cells plays a vital role in processing huge amount of visual information. To process visual data efficiently and effectively, it is crucial to reduce the data dimensionality. There are two main strategies for dimensionality reduction, data compression with minimum information loss and data representation with explicitly selected lossy information, and it can be related to the coding process and attention in the HVS. In [17], a bottom-up saliency model is proposed based on V1 cells. Based on this model, we compute the information loss in V1 cells as indication of saliency. In this section, we exploit efficient coding in simple cells of V1 and beyond, and from the information loss the proposed saliency model is constructed.

A. Efficient Coding in Simple Cells

The efficient coding theory [1] shows that the HVS exploits the statistical regularities or redundancies in natural scenes for data compression with minimum information loss using limited neural resources (limited number of neurons and power consumption by neural activities). To reduce data redundancy, it is modeled that the HVS transforms the original input $S = \{s_1, s_2, \cdots s_N\}$ in the neurons (e.g., photo-receptors and simple cells of receptive field) to signals $O = \{o_1, o_2, \cdots o_M\}$ in other neurons (e.g., simple cells of receptive field or cortical neurons). The problem is simplified by approximating the neural transform as a linear function K,

$$O = K(S) \text{ without regard to noise.} \tag{1}$$

Hence the optimal encoding, which balances the neural resources and information extraction, is to find the transform function with minimal loss

$$E(K) = \text{neural resources} - \lambda I(O; S), \tag{2}$$

where the parameter λ balances the information extraction $I(O; S)$ and the neural resource (explained in the following section). The mutual information $I(O; S)$ is defined as:

$$\begin{aligned} I(O; S) &\equiv H(S) - H(S|O) \\ &= \sum_{O,S} P(O, S) \log \frac{P(S|O)}{P(S)}, \end{aligned} \tag{3}$$

where $P(O, S)$ is the joint probability of the output signal O and the input signal S, and $P(S|O)$ is the conditional probability. Olshausen [19] proposes the receptive field of simple-cells in the primary visual cortex encode neural responses by sparse coding. The sparse coding theory is then developed to extract the intrinsic structure of natural images for efficient coding [19], [22]. These studies show that an image can be sparsely represented by a linear combination of small number of bases. The transform function can be represented by

$$S(x, y) = \sum_i O_i \phi_i(x, y) + \varepsilon. \tag{4}$$

where the coefficient O_i obeys a Laplace distribution. These coefficients and residual error can be computed by its corresponding filter basis $\phi_i(x, y)$.

Similar to Equation 2, the neural cost and information extraction of Equation 4 needs to be balanced by

$$E(\phi) = [\varepsilon]^2 - \lambda[\text{sparseness of } O_i], \tag{5}$$

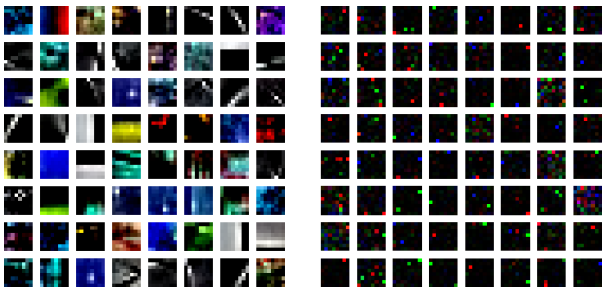
where the first term measures information loss and the second term indicates the power consumption by neural activities.

The coefficient O_i can be computed by the corresponding filter function $\phi_i^{-1}(x, y)$ as,

$$O_i = \phi_i^{-1}(x, y) \cdot S(x, y), \tag{6}$$

where $\phi_i^{-1}(x, y)$ is the inverse or pseudo-inverse of $\phi_i(x, y)$.

We learn a set of basis functions which yields a sparse representation of natural image patches by independent component analysis (ICA) [2] which has been shown to produce sparse codes [8]. For computational efficiency and coding effectiveness, we learn sparse coding basis by using 120,000 $8 \times 8 \times 3$ RGB image patches randomly extracted from natural images. A set of 192 basis functions are obtained using ICA [2]. Figure 2(a) shows some of the leaned basis functions. We use the inverse of the trained bases as the filter functions. Figure 2(b) shows some learned filter functions.



(a) (b)

Fig. 2. (a) 64 basis functions from the set of 192 learned basis functions. (b) the corresponding filter functions from the 64 basis functions.

B. Efficient Coding beyond Simple Cells

In the sparse coding stage, visual information is compressed. Nevertheless, there exist strong correlations between two different sparse coding components as shown in Figure 3. These results show that sparse coding process does not decorrelate different components perfectly. In other words, redundant information still exists in multiple channels. Hence, more efficient coding can be obtained (e.g., decorrelation or dimensionality reduction) for neurons in later stages.

For this reason, the neural resource term is defined as $\sum_i \langle O_i^2 \rangle$ [27], which is the trace of the output correlation matrix R^O with elements $R_{ij}^O = \langle O_i O_j \rangle = \langle K(S)_i K(S)_j \rangle = 0$ when $i \neq j$. Thus, Equation 2 can be expressed as

$$\begin{aligned} E(K) &= \text{neural resource} - \lambda I(O; S) \\ &= \sum_i \langle O_i^2 \rangle - \lambda I(O; S), \end{aligned} \tag{7}$$

where $\sum_i \langle O_i^2 \rangle = Tr(R^O) = Tr[K(R^S)K^T]$ and $Tr(\cdot)$ denotes the trace of a matrix.

To minimize the above functions, further dimensionality reduction is required. Therefore, the reduced-dimensionality representation of signal S (from the N-dimensional space) is projected down to the reduced M-dimensional space (i.e., $N > M$). In this work, we use PCA for this process, and the value of M is obtained by retaining 90% of total variance.

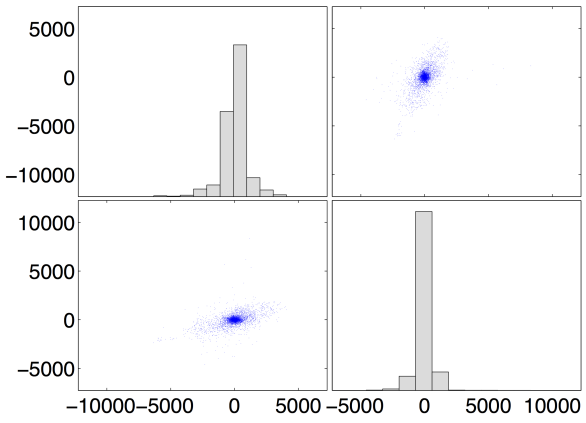


Fig. 3. The plots the distribution in two different channels of simple cells responses. It is easy to observe that each channel is subject to Laplace distribution and they are strongly correlated.

C. Lossy Coding Length and Visual Saliency

Lossy coding occurs beyond simple cells in V1. In the following paragraphs, we will argue the relationship between lossy coding and visual saliency. Approximating the residual error between response of simple cells and reconstructed one as Gaussian, the extracted information at the output is $2^{-I(O;S)} \propto \langle (PS)^\top (PS) \rangle$ and coding uncertainty is described by $H(S|K^{-1}O) \approx \log\langle (PS)^\top (PS) \rangle$ [27]. The projection P is defined as following:

$$P = I - KK^\top \tag{8}$$

where K is a rectangular matrix of low rank whose columns are the k eigenvectors having the largest eigenvalues of $Tr(R^O)$. From the information theory perspective, efficient coding results can be decomposed into two parts [6]:

$$\begin{aligned} &H(\text{Simple Cells Response}) \\ =&H(\text{Redundancy}) + H(\text{Saliency}) \\ =&I(O; S) + H(S|K^{-1}O), \end{aligned} \tag{9}$$

where $H(\text{Redundancy})$ denotes redundant information that can be interpreted by a coding system, and $H(\text{Saliency})$ is related to lossy coding length. We define the residual to measure saliency in order to simplify the computation by

$$\text{Lossy Coding Length} = (PS)^T (PS) \tag{10}$$

Based on this formulation, we compute the residual for each patch at image location (x, y), and then construct a residual map. Finally the saliency map is generated from the normalized residual map.

III. Experiments

In this section, we evaluate the performance of the proposed model with extensive experiments on two public eye-tracking datasets: the Bruce dataset [3] and the Judd dataset [13]. On each eye-tracking dataset, we compare

our model with ten state-of-the-art models including Itti et al. [12], attention by information maximization (AIM) [3], discriminant saliency (DS) [4], incremental coding length (ICL) [10], saliency using natural statistics (SUN) [26], Judd et al. [13], saliency detection by self-resemblance (SDSR) [20], Li et al. [16], saliency by site entropy rate (SER) [24], and scale integration (SI) [18]. The computational efficiency of these methods is also demonstrated using the Bruce dataset. The relationship between scale, execution time and performance are also analyzed in the experiments. These results show that the proposed saliency model is robust, effective, and efficient.

A. *Qualitative Evaluation*

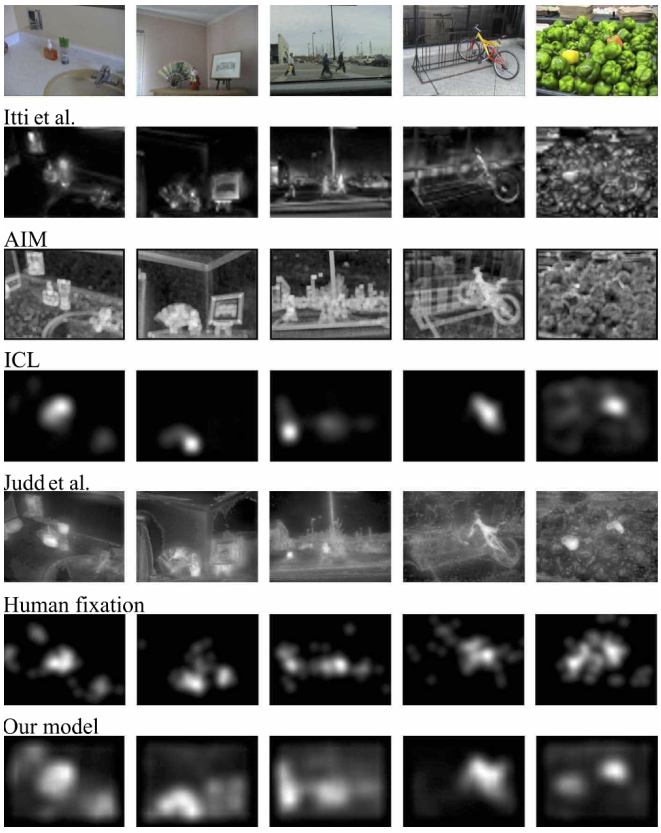


Fig. 4. Representative experimental results of our method with comparisons to four state-of-the-art methods and human fixation density maps using the Bruce dataset. The rows from top to down are: the original stimulus image, saliency maps generated by Itti et al. [12], AIM [3], ICL [10], Judd et al. [13], human fixations density maps and our saliency maps

Experiments with the Bruce Dataset: The Bruce dataset [3] consists of 120 images in a variety of indoor and outdoor scenes. The eye-tracking data points in this set are collected by showing images on a 21-inch CRT monitor with a 4 seconds interval at a distance of 0.75 meters from the subject. This dataset has been widely used to benchmark results of different saliency models.

For fair evaluations, we use the default parameters of existing algorithms from the original source codes. Instead of using the original iNVT toolkit or the Saliency Toolbox [23], we use the implementation by Harel [7] to evaluate the model by Itti et al. [12] which performs faster and more accurately in fixation prediction.

We evaluate the performance of saliency models in terms of their qualitative results. Representative saliency maps from four state-of-the-art algorithms [12], [3], [10], [13] and human fixation density maps are shown for comparisons with our model in Figure 4. The fixation density maps are obtained by overlapping Gaussian kernels in every fixation map. Overall, our saliency model performs favorably against other methods and matches human fixation maps well.

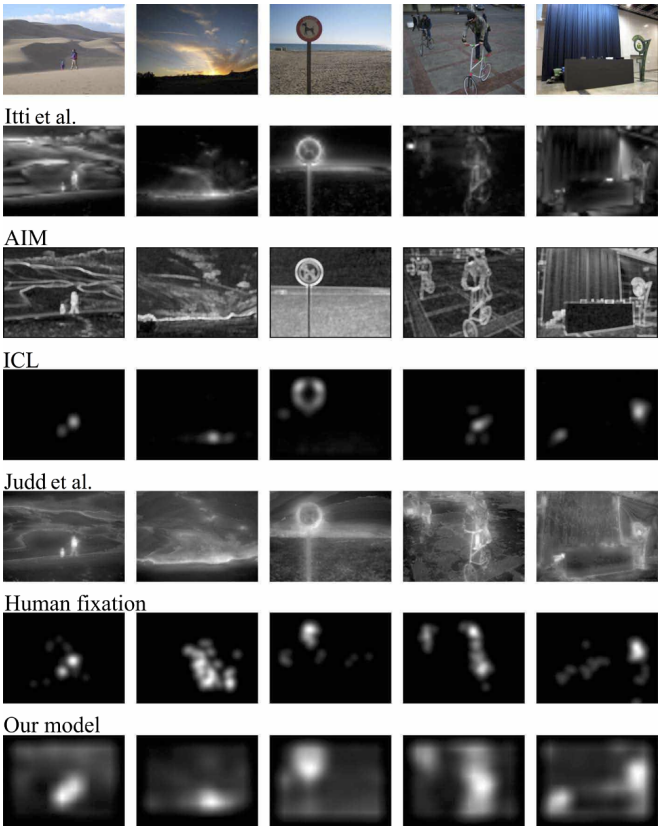


Fig. 5. Representative experimental results of our method with comparisons to four state-of-the-art methods and human fixation density maps using the Judd dataset. The rows from top to down are: the original stimulus image, saliency maps generated by Itti et al. [12], AIM [3], ICL [10], Judd et al. [13], human fixations density maps and our saliency maps

Experiments with Judd Dataset: The Judd dataset [13] contains 1003 images with 779 landscape and 228 portrait images. The eye-tracking data in this dataset is collected by showing images on a 19-inch monitor at the resolution of 1280×1024 pixels with a 3 seconds interval separated by 1 second gray screen at a distance of about 0.6 meters from the subject.

Most of the saliency models in existing literature simply use the Bruce dataset for evaluations. However, this set just has a small number of images and there are no portrait images while the image orientation could affect the eye-tracking

results significantly. Thus, we also evaluate the proposed model against other algorithms using the Judd dataset [13] which contains more and complex images. Representative saliency maps from four state-of-the-art algorithms [12], [3], [10], [13] and human fixation density maps are presented in Figure 5. Our saliency model performs well against other methods and matches human fixation maps closely.

B. Quantitative Evaluation

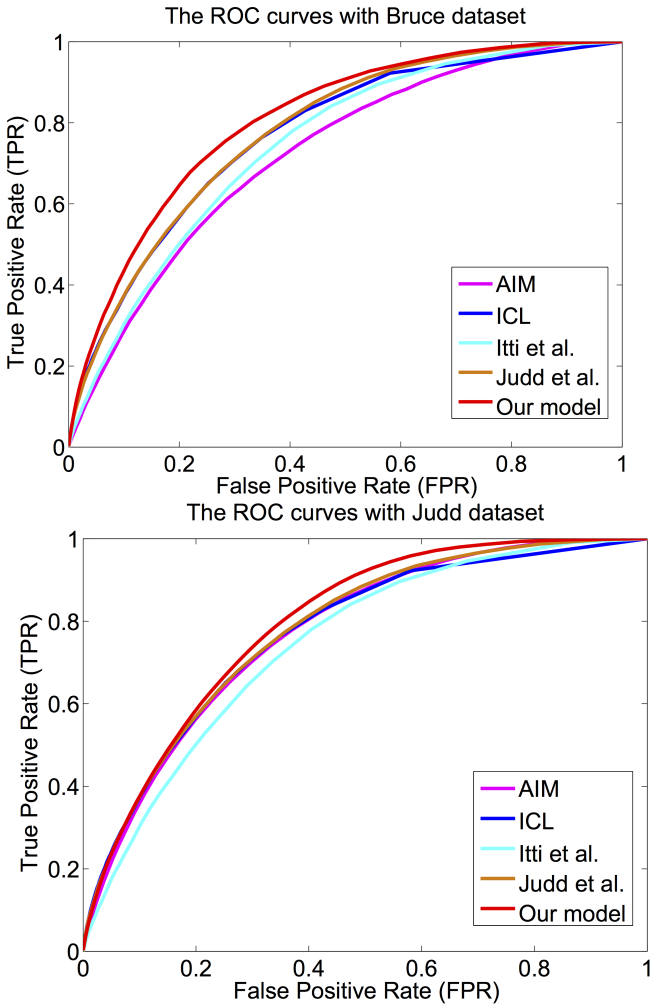


Fig. 6. The ROC curves of our model and the other four state-of-the-art approaches. (a) using the Bruce dataset. (b) using the Judd dataset.

For quantitative evaluation, we compute the Receiver Operator Characteristic (ROC) curves and the Area Under Curves (AUC) of each approach using both datasets. The ROC curves and AUC are generated by computing the mean value of the output from toolbox provided by Harel [7] with human fixation and saliency maps for each image in the dataset. There are minor differences between the showed results and their reported results in the literature as the settings of ROC algorithm are different. The results by Itti et al. [12] are better than those from AIM [3] as we use the implementation by Harel [7]. The ROC curves of state-of-the-art approaches and the proposed algorithm using both

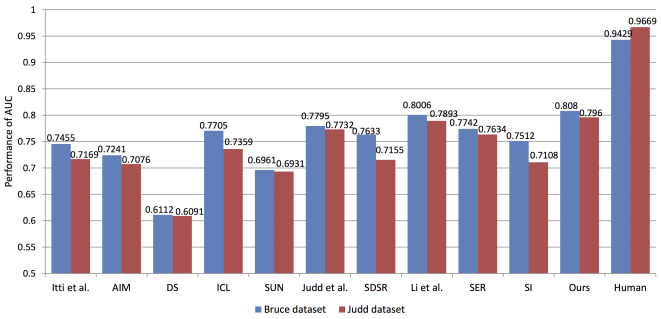


Fig. 7. The AUCs with both datasets. Larger AUC values indicate better performance.

datasets are shown in Figure 6. In addition, the AUCs of the evaluated algorithms are are presented in Figure 7. Overall, the proposed saliency model performs well against state-of-the-art methods on both datasets.

C. Computational Cost and Image Scale

Method	Run time per image (s)	Implementation
Itti et al. [12]	0.28	Matlab&C
AIM [3]	41.64	Matlab
DS [4]	243.7	C
ICL [10]	0.2	Matlab
SUN [26]	0.57	Matlab
Judd et al. [13]	12.35	Matlab&C
SDSR [20]	2.4	Matlab
Li et al. [16]	9.62	Matlab
SER [24]	5.2	C
SI [18]	6.71	Matlab
Proposed	**0.15**	Matlab

TABLE I
RUN-TIME PER IMAGE COMPARISON IN THE EXPERIMENT WITH THE BRUCE DATASET.

For real-world applications, the computational cost is important. Table I shows the average execution time of evaluated methods using the Bruce dataset. These algorithms are implemented in different languages including C, Matlab, and Matlab with C (i.e., MEX). The experiments are carried out on a PC with CORE2 dual processors and 4G memory. Our algorithm performs more efficiently than other methods with the Matlab implementation and no code optimization.

The scale of an input image is the key factor of computational cost in our saliency model and other methods. The computational cost grows rapidly when the scale is increased. In addition, the image scale also affects the performance of our saliency model. With smaller scale images, the AUC values tend to be larger. Figure 8 shows the effects of image scale to the proposed saliency model.

To select proper image scale for best performance, we develop a biologically-inspired method. As the perceived image scale depends on several factors including size of input image and eyes as well as the relative distance between a monitor and the human subject. In our model, each patch is processed by an independent receptive field. Thus, the input

image should be resized based on the receptive field diameter, image size, monitor size and the distance between subject and monitor. By solving trigonometric functions, we estimate the view angle of the image diagonal is 40.75 degrees in the Bruce dataset and the diagonal length of images is 851 pixels. Therefore, each pixel we use in the Bruce dataset occupied 0.034×0.034 degrees while the receptive field is 0.25×0.25 degrees to 0.5×0.75 degrees [11]. For computational efficiency, we rescale the input image to 0.06125 of the original size for the experiments with the Bruce dataset. In the Judd dataset, the monitor diagonal is about 48 degrees. The view angle of the image diagonal is about 37 degrees and the diagonal length of images from the Judd dataset is 1280 pixels using a 1024×768 monitor. Each pixel we use in the Judd dataset thus occupies 0.021×0.021 degrees. For efficiency, we rescale the input image to 0.05 of the original size for the experiments with the Judd dataset. These scale settings are used in all experiments of this work.

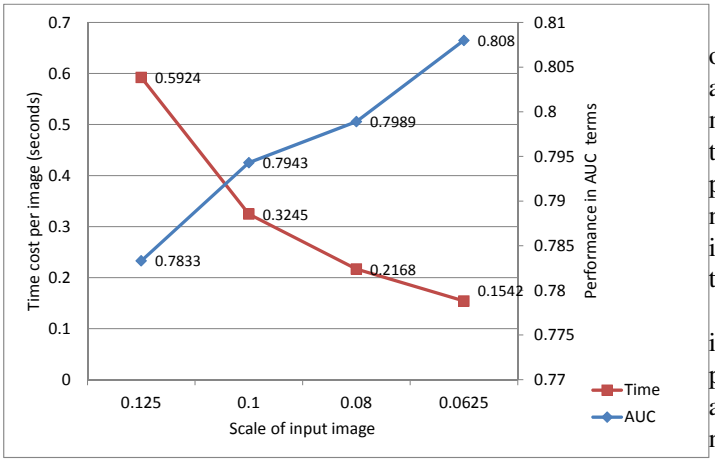


Fig. 8. The effect of scale to computational time and performance on our method using the Bruce dataset.

D. Discussion

As Figure 4 and Figure 5 show, our saliency maps look like the human fixation density maps. It could be seen as that our saliency maps could show the characteristics of human fixation very well.

Our model outperforms all other models both in the Bruce and Judd datasets by comparing the AUCs as described in Figure 7.

The ROC curves show a good performance as we can see in Figure 6. Our model could perform a 0.8 true positive rate at the false positive rate of about 0.3.

It is also very fast. As we can see in Table I, we only use 0.15 seconds to compute the saliency map for one input image which is slightly shorter than the time used in the implementation of the model by Itti et al. [12] while our model is not optimized for speed. In the eye-tracking datasets, the performance of all models is decreased when the input image number is increased. Nevertheless, our model and the method by Judd et al. [13] are least affected. This indicates that the proposed model is more robust and applicable to tasks with large and complex scenes. In addition, we develop a simple and effective method to determine optimal image scale for better performance.

In two different eye-tracking datasets, performances have decreased with the increase of amount of the images in different models. Our model and the method by Judd et al. [13] are least affected. This indicates that the proposed model is robust, and large scale and complex dataset is possible to comprehensively test models.

Both the model by Judd et al. [13] and the algorithm of Itti et al. [12] implemented by Harel [7] apply a prior to focus on central areas for performance improvements. However, there is no obvious evidence that the center prior exists in visual attention. Thus, we do not use this prior in the experiments described above although it may significantly improve the performance of our model.

IV. Concluding Remarks

In this paper, we propose a novel bottom-up model based on lossy coding length. It also helps us understand the architecture of the early vision system since each step in our model corresponds to the functions of the primal visual cortex. The proposed biologically-inspired model is also simple, practical and computationally efficient. As demonstrated in numerous experiments, the proposed method shows promising performance in two different datasets which include eye tracking data.

Based on a simple two-layer coding network, as shown in Figure 1, we develop a saliency model with demonstrated performance. It is plausible that bottom-up saliency is only an additional function for the coding network while there are no particular neurons for it.

In our future work, we will propose methods which introduce more reasonable bottom-up neural mechanism and useful top-down cues. In this way, we could accomplish more works which could be a more reasonable way to simulate human visual attention and perhaps it is a better way to narrow the distance between human beings and computers. Our future work will focus on more saliency models with top-down cues to explain purposeful visual search. The top-down, task-based modulation of saliency has been shown in several studies which can be applied to computer vision tasks such as contextual priming and object detection. We aim to extend the proposed saliency model within the supervised learning framework for these tasks. Our future work will focus on more saliency models with biologically-plausible bottom-up neural mechanism to explain visual attention and perception. The bottom-up, data-driven modulation of saliency is one of the most important topics in psychology and neuroscience, and is of great potential for computer vision applications. We aim to extend the proposed saliency model within the supervised learning framework for these tasks.

References

[1] H. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234, 1961.

[2] A. Bell and T. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. *NIPS*, 18:155–162, 2006.

[4] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. *NIPS*, 20:497–504, 2007.

[5] C. Guo, Q. Ma, and L. Zhang;. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. *CVPR*, 2008.

[6] B. Han and H. Zhu. Bottom-up saliency based on weighted sparse coding residual. In *ACM MM*, pages 1117–1120, 2011.

[7] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *NIPS*, 19:545–552, 2007.

[8] S. Hornillo-Mellado, R. Martín-Clemente, and J. Górriz-Sáez. Connections between ica and sparse coding revisited. *Computational Intelligence and Bioinspired Systems*, pages 1035–1042, 2005.

[9] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007.

[10] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 21:681–688, Jan 2008.

[11] D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

[12] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *ICCV*, pages 2106–2113, 2009.

[14] D. Kelly. Information capacity of a single retinal channel. *IRE Transactions on Information Theory*, 8(3):221–226, 1962.

[15] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–27, 1985.

[16] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang. Incremental sparse saliency detection. *ICIP*, pages 3093–3096, 2009.

[17] Z. Li. A saliency map in primary visual cortex. *Trends in cognitive sciences*, 6(1):9–16, 2002.

[18] N. Murray, M. Vanrell, X. Otazu, and C. Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR*, pages 433–440, 2011.

[19] B. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[20] H. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 2009.

[21] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[22] W. Vinje and J. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

[23] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–407, Nov 2006.

[24] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In *CVPR*, pages 2368–2375, 2010.

[25] R. D. Wright and L. M. Ward. *Orienting of attention*. Oxford University Press, 2008.

[26] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[27] L. Zhaoping. Theoretical understanding of the early visual processes by data compression and data selection. *Network: Computation in neural systems*, 17(4):301–334, 2006.