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

Bottom-up Model of Visual Saliency: A Viewpoint based on Efficient Coding Hypothesis

Hao Zhu and Biao Han

Abstract— **This paper proposes a novel bottom-up saliency model based on the mechanism of the early vision system. A relationship between the efficient coding theory and bottom-up saliency map in primate visual cortex is established. In this paper, we make a distinction of neural response between activated and inactivated by sparse coding, and define the saliency as uncertainity of internal representation. Beyond the definition of saliency, our model also accounts for the issue of why we need such a saliency map. Finally, we test this model on artificial images such as psychological patterns and two different scale datasets. Although it is only a simple model of bottom-up saliency, the experiment results show it outperforms other state-of-the-art methods.**

I. Introduction

Human beings have the remarkable ability to effectively recognize objects in complex scenary. However, since the amount of visual information surpasses the processing capability of the visual system, for these "overloaded information", there must be an optimization system to select the major parts from the original input. Selective visual attention would remove massive redundancy from visual input data and save a few important information. It is generally recognized that the attention could function as selecting and filtering visual input information through two methods: bottom-up and top-down.

In the research [11], Koch et.al propose the existence of saliency map in the visual system and the relationship between saliency map and attention. Based on feature integration theory [20], the saliency model combining information from several abstract feature maps (e.g. local contrast, orientations, color), is designed as an input to control mechanism for converting visual selective attention. Here we only focus on biologically plausible computational model of bottom-up saliency since little is known about the neural instantiation of the top-down, volitional component of attention [7].

Many existing literatures have proposed bottom-up saliency models based on information theory [3], spectral theory [8], and graph theory [6]. But these models don't concern the real neural mechanism. So far as we know, there is no literature proposing that the sparse coding leads to bottom-up saliency. In this paper, we will elaborate the relationship between efficient coding and bottom-up saliency based on the framework of free energy theory.

The free energy principle could account for action, perception and learning. The principle suggests that agents are able to suppress free energy by changing sensory input by acting on the world [5]. The efficient coding[1], as a theoretical model of the brain sensory coding and special case of free energy principle, considers that the responses of the sensory system form neural codes for efficient sensory information expression. It means neurons in the visual system suppose to optimally code input as an efficient representation [12]. There are many studies showing that filters optimized for coding natural images lead to filters that resemble the receptive fields of simple-cells in V1[16], [17].

In this paper, we propose a biological inspired bottom-up saliency model based on free energy theory. The attention can be understood as inferring the level of uncertainty. More specifically, in the work we try to substantiate this point in primary visual cortex using neuronal simulation of sparse coding. For a given image, the free energy is defined as the cost function of learning sparse coding for it. This definition simulates that the visual system optimizes the uncertainity of probabilistic representation which is concerned with attention. And then the salient value is defined as the uncerntainity between internal representation of neurons and sensory input.

Our major contribution is building a new computational model to simulate the bottom up saliency in primary visual cortex. This model is compatible with the psychophysics of human pre-attentive vision. More specifically, it is shown that the proposed model replicates various fundamental properties of human pre-attentive vision. The experimental comparison addresses the ability to predict human eye fixations on natural scenes. Our model surpasses other 4 state-of-the-art models, though some of them even considering the top-down factor.

II. The Proposed Saliency Model

In this section, we will present our saliency model based on the efficient coding hypothesis in the framework of free energy theory. For the sake of understanding this hypothesis, first we will introduce free energy theory and efficient coding principle. And then our model based on sparse coding is proposed further.

A. *Efficient Coding Hypothesis*

The efficient coding principle suggests that the brain is to optimize that mutual information between sensorium and its internal representation. This theory is formalized later in term of infomax principle [13]. In the view of information theory, efficient coding decomposes an input image into two parts:

$$H(Image) = H(Representation) + H(Uncertainity)$$

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)

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

where $H(Representation)$ denotes the information that can be interpreted by the internal representation. $H(Uncertainity)$ is minor, salient areas that are hard to be represented by internal representation. In other words, frequent patterns only lead to tiny uncertainity between the internal representation and sensory input while irregular patterns lead to much more uncertainity. As shown in the Figure 1, (c) and (d) represent the uncertainity of (a) in different internal representation. Either (c) or (d) show that the representation of the tower in (a) has larger uncertainity than other surrounding textures such as the sky and the grass land. We will explain the phenomenon more specifically in the next section by introducing sparse coding. In the following section, we will demonstrate the sparse representation which is used to estimate salient area by removing the statically redundant components.

In primary visual cortex, the uncertainity between sensorium and its internal representation is regarded as a type of free energy. Thus, the gaze moving is induced by the free energy to suppress the energy by changing sensory inputs. The action changes the location of foveal. The receptive fields of foveal is able to decompose the input more precisely than receptive fields out of foveal. Thus the free energy is suppressed.

There are evidences showing that only a few of early visual neurons out of a large set will be activated as stimulating by a scene [2]. To simulate the property of simple cells in the primary visual cortex, the sparse coding theory is proposed to extract the intrinsic structure of natural images for efficient coding [16], [21]. These studies show that an image can be sparse represented by linear combination with few bases. The input image can be divided into sliding overlap patches $I = [y_1, \ldots, y_n]$, n is the amount of patches. And then the corresponding sparse representation for patches are proposed as following:

$$y_i = \sum_{k=1}^{m} \alpha_{ik} d_k + r_i \tag{1}$$

where r_i is the residual of sparse coding for y_i and m is the number of bases. Then y_i can be decomposed to the sparse linear combination $D = [d_1, ..., d_m]$ by coefficient $\alpha_i = [\alpha_{i1}, ..., \alpha_{im}]$. Supposing we have learned the over-complete basis set, and d_k is a basis of dictionary D which is learning from the raw patches.

For simulating the optimal coding, we choose the setting of learning different dictionaries for different images. Because this setting all input images have similar capability of internal representation. In order to train the over-complete basis set D for present images, the online dictionary learning method [14] is employed to learn different basis sets for each input image. This process in one image is under constraint conditions as following:

$$\min_{\mathcal{D},\alpha_\rangle} \sum_{i=1}^{n} \|y_i - D\alpha_i\|_2^2 + \lambda \|a_i\|_1 \tag{2}$$

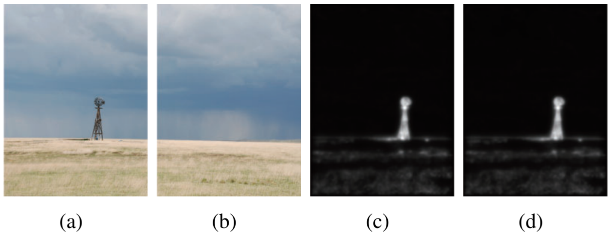

Fig. 1. Residual maps using different basis sets. (a) and (b) are divided from one photo with same size. (c) and (d) are the residual map of input stimuli (a) and reconstruction result of (a) using different basis sets. (c) used the basis sets learned from (a). (d) used the basis sets learned from (b). The images are modified from the photo taken by *Ronnie Pitman* from Flickr creative commons[1].

where $\{y_1, \ldots, y_n\}$ are sliding overlap patches divided from a given image. The above optimization problem solves an over-complete basis set by minimizing the sum of reconstruction errors in every patch and regularizing the representation.

To demonstrate the tendency in learning basis set with sparse constraint, as shown in the Figure 1, we divided one image into two same size sub-images and trained the sparse basis set from them respectively. The two different images have similar textures and background, but the salient region is just in Figure 1(a). We learned two different basis sets from different sub-images and respectively reconstructed the sub-image contained salient region by the different basis sets. Figure 1(c) and (d) show the residual maps using different basis sets. The two maps are almost identical. This fully explained why unsupervised learning with sparse constraint only pays close attention to the redundant information. It is an optimal coding strategy for data compression with minimum information loss in constraint of limit neuron activated. So for the complex part of the image, the internal representation needs more resource (i.e. more bases in the dictionary). In the global optimization, giving up the rare and complex part seems to work better.

B. The definition of saliency from Sparse Coding

This principle of information maximization suggests that the human eyes tend to focus on the most informative points on an image. But it is still a problem that how to define the informative points in the image. In the view of information theory, the smaller the probability for one sample in the dataset is, the more information it contains. Nevertheless, it is hard to compute the probability in high dimensions by statistical methods, and to find the real informative sample because the probability perhaps uniform distributed or all patches perfectly represent by few bases but no similar patches each others. In the paper, we propose the definition of saliency is the reconstruction error in the sparse coding (i.e. r_i in Eq.1). In the following part of this section, we will show the purpose of the definition in the viewpoint of information theory. In summary, this definition is a way to measure the uncertainity in the coding process.

By sparse coding, the huge amount of patches is represented by $A = [\alpha_1, \ldots, \alpha_n]$ for intermediate latent variables

$D = [d_1, ..., d_m]$ (i.e. by over-complete basis set). The patches $[y_1, \ldots, y_n]$ employ sparse representation to reconstruct self-information as Eq.1, we denote y_i as $O(\text{observer})$ $\sum_{k=1}^{m} \alpha_{ik} B_k$ as $\hat{O}(\text{reconstruction})$ and $\varepsilon(\text{residual})$ as reconstruction error. Hence Eq.1 was reformulated as:

$$O(\text{observer}) = \hat{O}(\text{reconstruction}) + \varepsilon(\text{residual}) \quad (3)$$

The goal of reconstruction is minimizing the cost function

$$E = -I(O; \hat{O}) + \lambda \sum_{k=1}^{m} |\alpha_{ik}|. \quad (4)$$

Hence it needs maximizing $I(O; \hat{O})$ where

$$I(O; \hat{O}) = H(O) - H(O|\hat{O}) \quad (5)$$

Since $\hat{O}$ is completely determined by over-complete basis set $D = \{d_1, \ldots, d_n\}$, then

$$I(O; \hat{O}) = I(O; \alpha_i) \quad (6)$$

Assuming that the residual ε is independent of O hence

$$H(\varepsilon) = H(O|\hat{O}) = H(O) - I(O; \alpha_i) \quad (7)$$

Therefore, the $H(\varepsilon)$ implies the extent of uncertainty when knowing coding α_i to predict $O(\text{observer})$ The higher the $H(\varepsilon)$ is the more salient in the patch is. Here we assume the residual ε is gaussian distribution. By the assumption, the $H(O|\hat{O})$ approximating the residual as

$$H(O|\hat{O}) \approx \log \|y_i - \sum_{k=1}^{m} \alpha_{ik} B_k\|^2 \quad (8)$$

$$\|y_i - \sum_{k=1}^{m} \alpha_{ik} B_k\|^2 \propto 2^{-I(O;\alpha_i)} \quad (9)$$

Therefore by above formula we know it is closely related between residual and $H(O|\hat{O})$. The minimizing residual is closely related to maximizing $I(O; \alpha_i)$, hence the high residual means that the corresponding patch in the image is hard to be represented sparsely. Thus we define the saliency as reconstruction error (i.e. Eq.9).

C. Implementation

In our model, the sparse representation residual contains the salient information. Sparse representation is used to remove representation for scene. By computing the sparse representation residual of each patch, a residual map is built. After normalization, we get our saliency map. The model is consist of following three steps:

(1) An image input is divided into patches of the same size and then learn the corresonding over-complete basis set to ensure such an image can be sparsely coded in a optimal way.

(2) Patches are sparsely coded with the over-complete basis set, just as the mechanism in primary visual cortex.

(3) The reconstruction error is used to measure the bottom-up saliency. The higher the energy is, the more the salient degree is.

III. Experiments

We conduct experiments on psychological stimuli and two public image datasets to evaluate the performance of the proposed model. On different public image datasets we compare our model with four state-of-the-art approaches: AIM[3], Itti[9], Judd[10], and SMVJ[4]. These approaches are carefully selected. Commonly-used evaluation criterion of receiver operating characteristic curve (ROC) is used to evaluate the performance of each approach.

A. Experimental Setup

All input images are down-sampled to $\frac{1}{4}$ and $\frac{1}{8}$ of the original size. We learn three sets of sparse coding over-complete bases by SPAMS[15] with default setup from patches of the *L*a*b** channel respectively in each scale.

The patches are extracted from the upper-left corner with 1 pixel overlap in each direction and normalized to range $[-1, 1]$. The bases are learned from these 7×7 pixels patches like the setting of [3]. Each over-complete basis set includes 128 sparse bases learned from the patches which are almost as much as pixel number. By computing the residual map of each scale and each channel, we get 6 sub-maps for each input image. We resize, normalize and average out all 6 sub-maps to generate the saliency map.

In each dataset, we conduct the experiments in two different settings: "with *Center Prior*" and "without *Center Prior*", to avoid the effect of *Center Prior*, which explained in [10] as humans naturally tend to gaze interesting object near the center of the image when receive a visual stimulus. *Center Prior* is added into AIM, Itti's and our algorithm in the setting "with *Center Prior*". The *Center Prior* is simulated by using a two dimensional anisotropic Gaussian function with standard deviations (σ_x, σ_y):

$$f(x, y) = exp\{-\frac{1}{2}(\frac{x - x_0}{\sigma_x^2} + \frac{y - y_0}{\sigma_y^2})\} \quad (10)$$

where (x_0, y_0) denotes the center of the image and $\sigma_y = H$ and $\sigma_x = W$. H and W denote the height and the width of the image. This function is used to convolve the saliency maps as a weighted sum to generate saliency map with *Center Prior*. *Center Prior* has already been added into SMVJ and Judd's algorithms in their own code. As [8], we also smoothed the saliency map with a Gaussian filter $g(x)(\sigma_{gauss} = 8)$.

We plot the ROC curve by computing the mean value of the output from toolbox provided by [6]. The True Positive Rate (TPR) and False Positive Rate (FPR) are computed by following equations:

$$\begin{aligned} TPR &= \frac{TP}{P} = \frac{TP}{TP + FN} \\ FPR &= \frac{FP}{N} = \frac{FP}{FP + TN} \end{aligned} \quad (11)$$

where true positive (TP) denotes hit, true negative (TN) denotes correct rejection, false positive (FP) denotes Type I error and false negative (FN) denotes Type II error. We also compute the Areas Under Curve (AUCs) for the comparison of different methods.

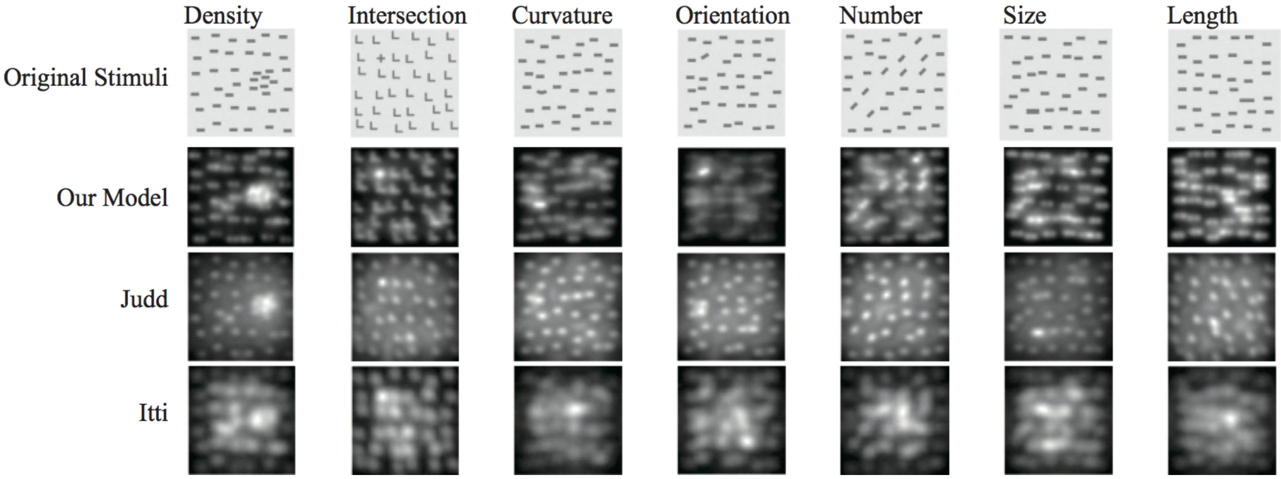

Fig. 2. Comparison of seven psychological stimuli between our model and other approaches.

B. Experimental Results

In this section, we present the results of our method on extensive experiments on psychological stimuli and two public eye-tracking datasets. We did experiments on psychological stimuli adopted in a series of attention experiments [22], [20] and made a comparison of the saliency map accuracy, using two datasets, Bruce dataset [3] and Judd dataset [10], one with 120 images and other one with more than 1000 images. For each dataset we compare our model with the state-of-the-art approaches. In psychological stimulus experiments, the results fully illustrate our method's generality in different patterns. By the quantitative analysis with ROC curve and area under ROC, proposed method shows a good performance in different datasets.

We have the experiments on several psychological stimuli which had been identified as pre-attentive. In the area of psychology, these psychological stimuli are used to perform the some pre-attentive visual tasks such as target detection, boundary detection, counting and estimation. These artificial patterns include “orientation”, “length”, “size”, “curvature”, “density”, “intersection” and etc. . We select seven of these to verify performance of the proposed method.

As shown in Figure 2, we compare our results with others from Judd and Itti by showing the saliency map.The figure clearly shows that our method predicts the good saliency spots in seven different psychological stimuli though the “size” and “length” pattern detect other salient regions for local density change. However, Itti's method completely fails in “length”, “curvature”, “orientation”, and “length” patterns. Even Judd's method also fails in “curvature”, “orientation” and “length” patterns.

1) Experiments on small-scale dataset: We use the eye-tracking dataset collected by [3] as the small-scale dataset in the experiments to test our model. This dataset usually serves as the benchmark dataset for comparing the results of saliency detection. All 120 images are shown on a 21 inch CRT monitor with a 4 second interval at a distance of 0.75 m from the subject. The eye-tracking data are collected from 20 different subjects for the full set of 120 images.

In this experiment, we evaluate our model by both the saliency map visual effect and the ROC computed by saliency map compared with human fixation density. In the comparison of ROC, as Figure 4 shown, both in “without *Central Prior*” and the “with *Central Prior*” group, our model shows competitive performance. Itti's algorithm obtains better results than original edition because of adopting Harel's implementation. Our model achieves the best performance in both group with and without center prior. It can be seen clearly that our model outperforms other methods in Bruce dataset.

2) Experiments on large-scale dataset: To test the robustness of our algorithm, we use a bigger eye-tracking dataset collected by [10] as the large-scale dataset in the experiment section. This dataset is one of the largest eye tracking datasets of natural images available for the vision and graphics community on open website.

The dataset contains 1003 stimulus images from Flickr creative commons and LabelMe [18]. The Eye tracking data are recorded from 15 subjects who free viewed these images. The longest dimension of each image is 1024 pixels. Images are shown on a 19 inch monitor with a resolution 1280×1024 with a 3 second interval separated by 1 second gray screen at a distance of approximately two feet from the subject. The subjects are in a dark room and use a chin rest to stabilize their head.

In the comparison of ROC, as shown in Figure 5, both in the “without *Central Prior*” and the “with *Central Prior*” settings, our model is also much better than the other algorithm. The results of this dataset are lower than the small-scale dataset because the scale of the dataset. It also can be seen clearly that our model performs well in the Judd dataset.

We show some of saliency maps in Figure 6. It shows clearly that our model achieves very good results. The saliency map is able to represent salient area accurately. The AUCs comparison on different datasets and different

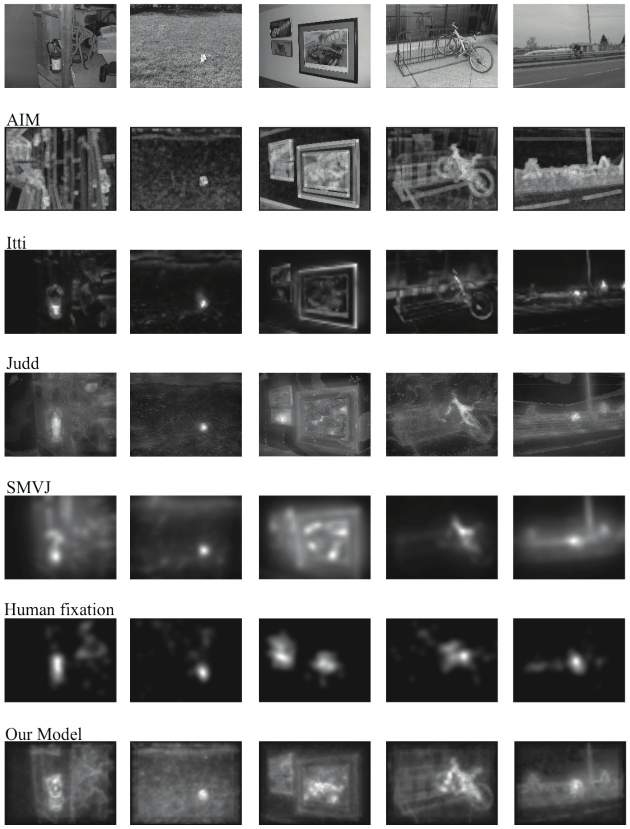

Fig. 3. The comparison with five state-of-the-art methods, human fixation density maps and the proposed model in small-scale dataset . The rows from top to down are: the original stimulus image, saliency maps generated by AIM, Itti, Judd, SMVJ, Human fixations density maps of stimulus images and saliency maps generated by Our model.

center prior condition is also shown in Table I, and SMWO denotes Small dataset without center prior, SMW denotes Small dataset with center prior, SMWO denotes Large dataset without center prior, LGW denotes Large dataset with center prior. It obviously shows that our model outperform other ones. These experiments prove that our algorithm is consistent in different datasets with different scales.

IV. Conclusion and Discussion

In this paper, we proposed a method for saliency detection based on the sparse representation residual of images. The sparse representation residual (SR^2) model has several advantages in different perspectives:

(1) It is a generic algorithm to different raw visual input such as psychological patterns and natural scene images. SR^2 uses self information to construct feature, it doesn't need to compute prior features such as orientations, color, etc.

(2) SR^2 is easy to understand under the architecture of the early vision system. Each step in our model corresponds to a function model in the primary visual system. We build it all based on acknowledged computational models.

(3) It is a simple, efficient computational model. As the illustration in experiments section, the proposed method shows the exciting performance in two different dataset which include eye tracking data. Compared with these methods based on learning or complex model, our method need not dataset for supervised learning and high level feature, however it still outperforms others methods in the benchmark. In future work we are interested in how to put the model into a neural network architecture which can be easily understood and computed. Actually, these is a related work, giving a reasonable interpretation with a similar model [19].

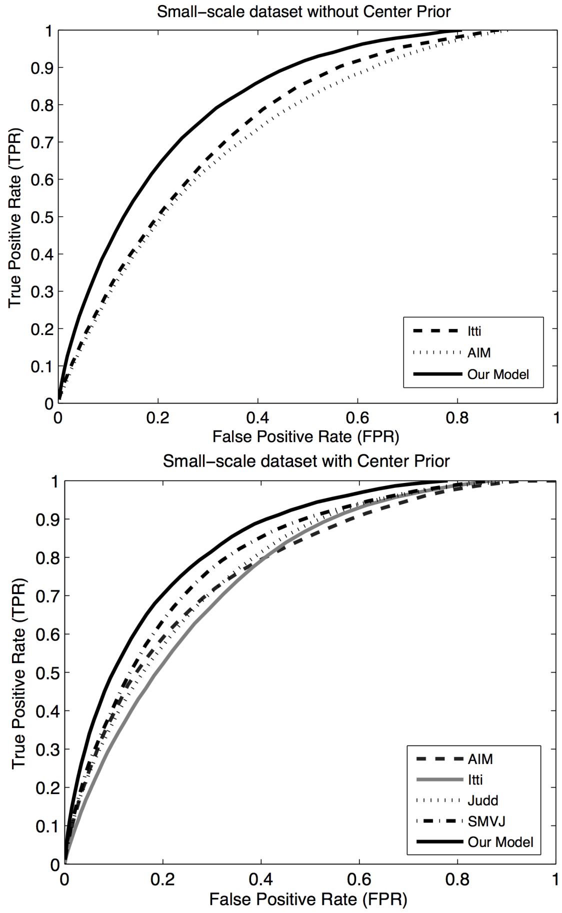

Fig. 4. The ROC curves of our model and the other state-of-the-art approaches on the Bruce image dataset.

For the problem of setting of learning a different set of basis for every image, we think that the dictionary learning is difficult to simulate the performance of primary visual cortex (V1) because there are too many neurons in the V1. Thus, learning a different set of basis will lead to less reconstruction error which is more likely to simulate the representation in V1. Actually, if we use the unified dictionary, the results are similar to the counterpart with the default setting.

References

[1] H. Barlow. Possible principles underlying the transformations of sensory messages. In W. (Ed.), editor, *Sensory Communication*, chapter 13, pages 217–234. M.I.T. Press, 1961.

[2] H. Barlow. Unsupervised learning. *Neural Computation*, 1(3):295–311, 1989.

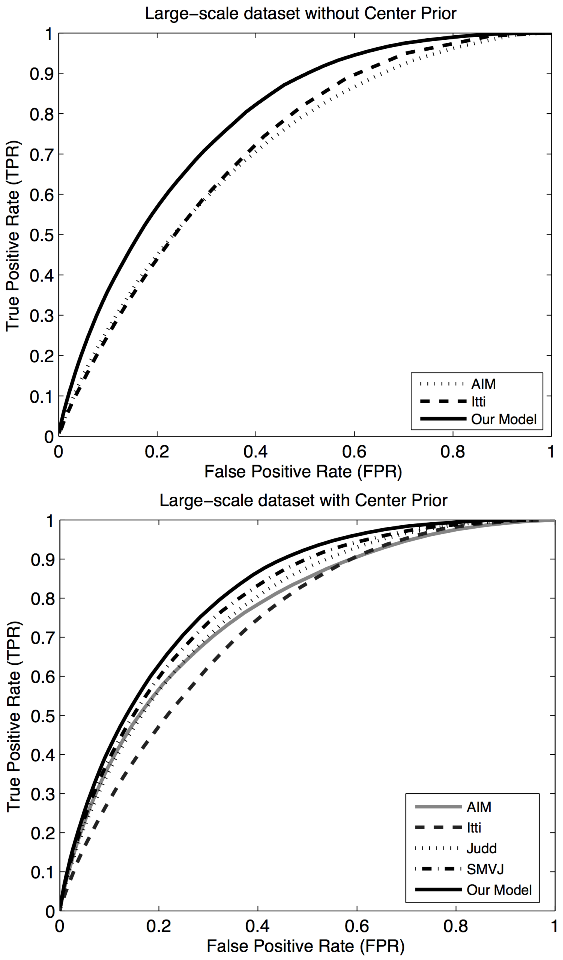

Fig. 5. The ROC curves of our model and the other state-of-the-art approaches on the Judd image dataset.

	SMWO	SMW	LGWO	LGW
AIM	0.7284	0.7734	0.7081	0.7637
Itti	0.7492	0.7612	0.7170	0.7336
Judd	-	0.7806	-	0.7734
SMVJ	-	0.8046	-	0.7910
Our Model	**0.8120**	**0.8413**	**0.7814**	**0.8105**

TABLE I

COMPARISONS OF AREA UNDER THE CURVE(AUC)

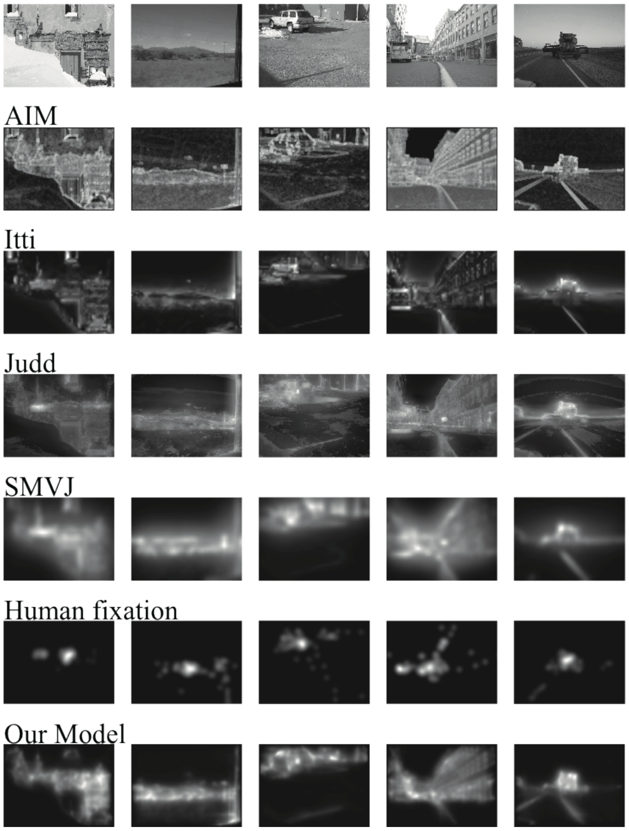

Fig. 6. The comparison with five state-of-the-art methods, human fixation density maps and the proposed model in big-scale dataset. The rows from top to down are: the original stimulus image, saliency maps generated by AIM, Itti, Judd, SMVJ, Human fixations density maps of stimulus images and saliency maps generated by Our model.

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, 2006.

[4] M. Cerf, E. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, Jan 2009.

[5] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[6] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 2007.

[7] J. Hopfinger, M. Buonocore, and G. Mangun. The neural mechanisms of top-down attentional control. *Nature neuroscience*, 3(3):284–291, 2000.

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

[9] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 1998.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *ICCV*, 2009.

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

[12] M. Lewicki. Efficient coding of natural sounds. *nature neuroscience*, 5(4):356–363, 2002.

[13] R. Linsker. Perceptual neural organization: some approaches based on network models and information theory. *Annual review of Neuroscience*, 13(1):257–281, 1990.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning*, 11:19–60, Jan 2010.

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

[17] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[18] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173, 2008.

[19] M. W. Spratling. Predictive coding as a model of the v1 saliency map hypothesis. *Neural Networks*, 26:7–28, 2012.

[20] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, Jan 1980.

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

[22] J. Wolfe. Guided search 2.0. A revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.