21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

# Visual Saliency: a Manifold Way of Perception

Hao Zhu
zhuhao_hit@hotmail.com
Beijing Normal University
Beijing, China

Biao Han
raintwoto@gmail.com
Shanghai University
Shanghai, China

Xiang Ruan
gen@omm.ncl.omron.co.jp
Omron Corporation
Kusatsu, Japan

## Abstract

*Visual saliency plays an important role in the human visual system HVS since it is indispensable for object detection and recognition. A bottom-up saliency model was proposed, following the manifold characteristic of HVS, previously developed for understanding HVS mechanism. The saliency of a given location of visual field is defined as the power of features responses after the dimensionality reduction with manifold learning for sparse representation of raw input. This saliency definition also explains the reason that HVS can suppress the response of redundant pattern and excite the response of attended pattern. Experiments show that our saliency model produces better predictions of human eye fixations on two dataset in the comparsion of four state-of-the-art methods.*

## 1. Introduction

The mechanism of visual attention plays an important role in biological vision. By recognizing certain regions of the visual field as salient points, which are more important than others, the mechanism allows a non-uniform allocation of visual attention that efficiently reduces the computational burden of HVS. It is generally acknowledged that visual attention, a process of selecting and gating visual information, includes two independent components: a bottom-up, fast, stimulus-driven mechanism, and a top-down, slower, goal-driven mechanism. In particular, bottom-up saliency, which related to the mechanism of visual coding, is more important.

There are many bottom-up computational models of visual saliency proposed in the computer and biological vision [8, 6, 5, 2]. Koch and Ullman [8] proposed the existence of a saliency map in the visual system and the relationship between saliency map and attention. Furthermore they proposed a saliency model based on feature integration theory (FIT) [15], which combined information from several abstract feature maps (e.g., local contrast, orientations, color). The well-known model [6] implemented the computational model of [8]. Bruce & Tsotsos proposed the principle that attention concerns maximizing information region of scene [2] to define saliency. In [5], authors proposed a viewpoint of information theory to the visual saliency. With the aid of machine learning method, the article [7] incorporated top-down clues, middle feature and the bottom-up model [6] to their model.

In this paper, we propose a biologically plausible bottom-up saliency model, instead of traditional model of commonly used center-surrounding mechanism, derived from the principle of the manifold ways of perception. The principle argues that HVS perceives constancy even though its raw sensory inputs are in flux, and extracts abstract, relevant features by a process of dimension reduction [13, 12, 14, 3]. Our computational model assigns a point in the image a high saliency value if it is a high value in the space of the abstract features. In addition to manifold perception theory, the proposed model is also based on following evidences of biological vision: (1)the sparse coding theory in primary visual cortex (V1) [10]. (2) the dependence of neighbour synaptic input for a neuron [9].

In order to simulate the proposed model, we establish a two-layer network, including two parts: sparse coding for simulation of early vision, and manifold learning for simulation of later stage of vision. Specifically, as shown in Fig.1,(1) the input image is divided into overlapped patches, and these patches are further convolved with Gabor-like filters, learned from a set of natural images, based on sparse coding theory. Then the image is represented as patch-wise convolved responses. (2) To project these patch-wise responses on low dimension space of abstract space, the Locality preserving projections (LPP) method is employed. (3) Finally the saliency in a point is computed by summing over

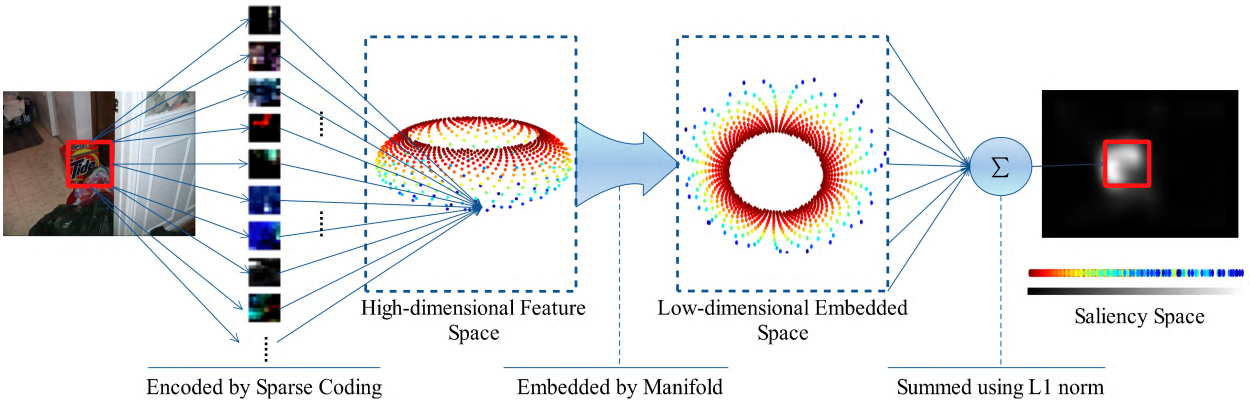



**Figure 1. The framework of the proposed method. There are three main steps in our method: encoding image patches by sparse coding, reducing dimensionality via LPP, and evaluating the saliency value. The range in color from red to blue are arranged in order of salient degree. With the mapping from high dimensional space to low dimensional space, the salient points are far from the origin of feature space.**

each component in the response in low dimensional space (e.g. L1 norm).

The rest of this paper is organized as follows: In section 2 we demonstrate a two-layer network and discuss the technical details, include how to measure saliency value from the coded representation. Section 3 shows experimental results of comparison between our model and other four state of the arts methods.

## 2. Method

### 2.1 Sparse Coding

In [10], Olshausen et al. proposed a computational model for the receptive field of simple-cell, which is able to simulate the encoding process in V1. The theory argues that the most number of neurons will be inactive while V1 receives a visual stimulus. Based on the theory, each image can be sparsely linearly represented by finite bases.

In this paper, a set of basis, employing sparse representation for natural image patches, is learned by independent component analysis (ICA) method [1]. Specifically, more than a hundred thousand RGB image patches of $8 \times 8$ size are used to train the set of basis, which include $8 \times 8 \times 3 = 192$ basis functions. Let the $A$ be the set of sparse basis, where $a_i$ is a vector of $A = [a_1, a_2, ..., a_{192}]$. Let the $W = A^{-1}$ where also includes 192 vectors $W = [w_1, w_2, ..., w_{192}]^T$ which are considered as linear filters to image patches. Vectoring the input image be $S = [s_1, s_2, ..., s_n]$, where $s_i$ is a vectorized image patch. The neuron responses of vectorized image patches $S$ on the set of basis $W$ are:

$$X = WS \tag{1}$$

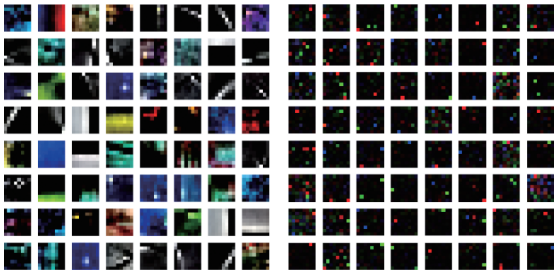
where $X = [x_1, x_2, ..., x_n]$ is sparse representation of $S$.

The set of basis, as shown in Fig.2, learned from patches looks like Gabor functions, which simulates the mechanism of primary visual cortex. Technically, this process contributes to making structure in natural signal explicit and representing complex data. However the most important of sparse coding is that the manifold of raw inputs then becomes less curvature in the higher-dimensional space defined by the neural responses, thus making it easier to learn structure in data [11].So it contributes to the following process, learning the intrinsic dimension from the high dimensional input, and makes the result more robust.

### 2.2 Dimension Reduction with Locality Preserving Projections

In this paper we choose the Locality Preserving Projections [4](LPP) to find the projection in intrinsic low-dimensional space from high dimensional data of sparse representation. The algorithmic procedure is formally stated as below:

Firstly, construct the adjacency graph and choosing the weight: for the establishing topological structure of given data, let $G = (V, E)$ be an undirected graph with vertex set $V = \{v \in X = [x_1, x_2, ...x_n]\}$ which is the sparse representation of patches $S$. In the following we assume that the graph $G$ is weighted, that is each edge



(a) (b)

**Figure 2. (a) The 64 components of 192 basis functions. (b) The 64 components of 192 filter functions.**

between two vertices $x_i$ and $x_j$ carries a non-negative weight $w_{ij} \geq 0$. The weighted adjacency matrix of the graph is the $W = (w_{ij})_{i,j=1,...n}$. Considering that a neuron's activities are driven by the total synaptic input from its neighbours [9], a graph could be defined naturally based on the spatial location of the few neighbouring pixels.

The objective of LPP can be formulated as following optimization problem:

$$P = arg\min_{P} \sum_{ij} (P^T x_i - P^T x_j)^2 W_{ij} \qquad (2)$$

Where $P$ is a projection matrix, $P = [p_1, p_2, \ldots, p_n]$ for transformation of data x in new basis. In consideration of the independence among different basis, the formula can be rewritten as:

$$p = arg\min_{p} \sum_{ij} (p^T x_i - p^T x_j)^2 W_{ij} \qquad (3)$$

Where $p$ means anyone basis in $P$. Further, let $p^T x_i$ substitute $y_i$ in equation 2:

$$\sum_{ij} (y_i - y_j)^2 W_{ij} = 2Y^T (D - W)Y = 2Y^T LY \quad (4)$$

Where $D$ is a diagonal matrix, $D_{ij} = \sum_j W_{ij}$, and the $L = D - W$ is called Laplacian matrix. One basic principle in the visual system is to suppress the response to frequently occurring input patterns, while enhancing the response to attended pattern. According to that, the constraint of $Y^T DY = 1$ is added. Because it is capable of making the similar sample close the origin of coordinates. Now, the optimization problem is given by the solution to the following generalized eigenvector problem:

$$p^T XLX^T p = \lambda p^T XDX^T p \qquad (5)$$

We can choose n eigenvectors, corresponding $n$ smallest eigenvalue to construct a projection matrix $P = [P_1^T, P_2^T, ..., P_n^T]$ and get $Y = PX$.

### 2.3 Saliency Map

The saliency of $x_i$ is supposed to be measured directly as:

$$S_i = \sum_{j=1}^{n} |y_{ij}| \qquad (6)$$

where $y_{ij}$ is $j$-th dimension of new features space in $i$-th patch of image.

## 3 Experiments

To test the performance of the proposed model, experiments on two datasets are conducted to compare the proposed model with four state-of-the-art approaches. The first dataset in common use collected by Bruce et al. [2] usually serves as the benchmark for comparing visual saliency detection results. This dataset consists of a variety of images about indoor and outdoor scenes. Eye fixations are recorded from 20 subjects on 120 color images. And the other one [7] consists of 1003 images, which mainly selected from Flicker creative commons and LableMe while the eye tracking data of 15 subjects are recorded with an eye-tracker for these images. There are some parameters in the proposed approach. Specifically, we set to 12 nearest neighbours for graph construction, and 20 eigenvectors for formula 4. In particular, we have given zero for the center prior, even it improves the value of AUC to our method, compared with the others methods ([16], [7]) incorporated into center prior.

On these datasets, we choose the area under ROC curve (AUC) to evaluate the performance of various saliency models.The AUC demonstrates the overall performance of a saliency model. Perfect prediction, corresponds to the AUC of 1, while random prediction generates an AUC of 0.5.

For evaluating with qualitative analysis, we show our saliency maps and the fixation density maps generated from the sum of all 2D Gaussians to the human fixations, and compare our saliency map on different datasets with other four state-of-the-art approaches ( [5], [6], [2], [7]). The saliency maps with Bruce dataset and Judd dataset is showed in Fig.3.

As Figure 3 shows, it is easy to get that our saliency maps, compared to other ones, are very close to the reference generated from human fixations in benchmark of dataset provided by Bruce. The proposed model is effective to predict the saliency for some warning signs.

| | Bruce Dataset | Judd Dataset |
|---|---|---|
| AIM | 0.7241 | 0.7076 |
| Itti's | 0.7455 | 0.7169 |
| ICL | 0.7705 | 0.7359 |
| Judd's | 0.7795 | 0.7732 |
| Our Method | **0.8234** | **0.8090** |

**Table 1. The AUCs comparison of all methods with different datasets.**

And in the image of dashboard (Column 5 in Fig.3), our method is unique to effectively detect the speedometer and gauges for mileage.

Table 1 presents AUC value for five approaches, and it is used for quantitative analysis. Even without center prior, the proposed model performs better than the others four method, especially 3% improvement better than Judd's method, which consists of low feature, middle feature, even top-down clue and center prior.

## 4 Conclusion

In this paper, we propose a framework that encodes the retinal input by sparse coding and maps the codes into intrinsic dimension. Saliency is defined as the distance between the patch in mainfold and the origin. Since the redundant information concentrate compactly around the origin of the coordinate, the larger the distance from original point is, the more salient the area is.

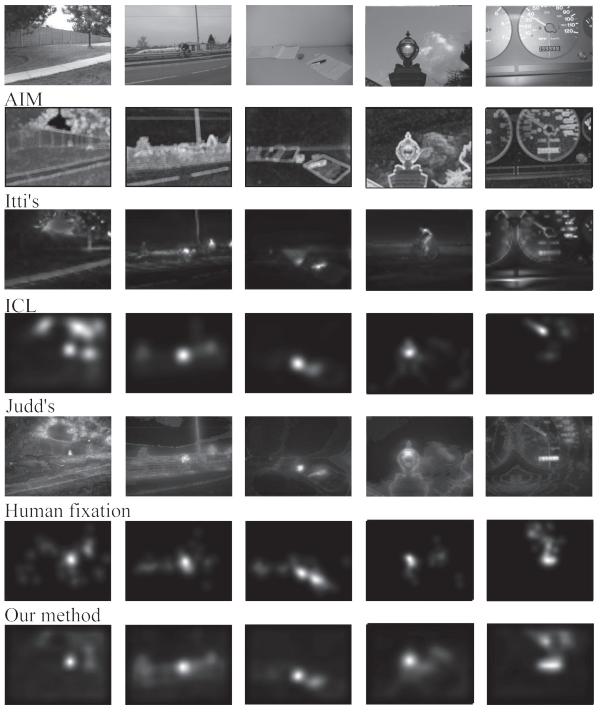



**Figure 3. Results for qualitative comparison.**

## References


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

[2] N. Bruce and J. Tsotsos. Saliency based on information maximization. *NIPS*, 2006.

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

[4] X. He and P. Niyogi. Locality preserving projections. *NIPS*, 2003.

[5] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 2008.

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

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

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

[9] W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4):115–133, 1943.

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

[11] B. Olshausen and D. Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

[12] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323, 2000.

[13] H. Seung and D. Lee. The manifold ways of perception. *Science*, 290(5500):2268, 2000.

[14] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[15] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[16] M. Wang, J. Li, T. Huang, Y. Tian, L. Duan, and G. Jia. Saliency detection based on 2D log-gabor wavelets and center bias. In *ACM MM*, pages 979–982, 2010.