Bottom-up Saliency based on Weighted Sparse Coding Residual [*]

Biao Han
Digital Media Laboratory,
Shanghai University
Shanghai China
raintwoto@gmail.com

Hao Zhu
State Key Laboratory of
Cognitive Neuroscience and
Learning, Beijing Normal
University
Beijing China
zhuhao_hit@hotmail.com

Youdong Ding
Digital Media Laboratory,
Shanghai University
Shanghai China
ydding@shu.edu.cn

ABSTRACT

The guidance of attention helps the human vision system (HVS) to detect and recognize objects rapidly. In this paper, we propose a bottom-up saliency algorithm based on sparse coding theory. Sparse coding decomposes the inputs into two parts, codes and residual. From the viewpoint of biological vision and information theory, the coding length is closely related to the local complexity while the residual is closely related to the uncertainty. The proposed algorithm defines the weighted residual using sparse coding length as saliency. By multiplying the L0 norm of sparse codes and the residual, a saliency map is obtained. The performance of the proposed method is evaluated using ROC curves with two different scale datasets and is compared with state-of-the-art models. Our algorithm outperforms all other methods and the results indicate a robust and accurate saliency.

Categories and Subject Descriptors

I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding Perceptual reasoning

General Terms

Algorithms, Experimentation

Keywords

Visual attention, Bottom-up saliency, Sparse coding

1. INTRODUCTION

What we see is determined by what we pay attention to. Humans are able to select the most relevant information from a large amount of visual data. Visual attention is used to reduce the redundancy and keep few but relevant information. Visual attention is widely studied by many researchers in the areas of psychology, neuroscience and computer vision. Koch and Ullman[8] proposed the original concept of visual saliency to represent visual attention. By calculating this visual saliency map, visual attention helps to solve computer vision problems such as thumbnail generation [11], video retargeting[16] and object recognition[13].

Bottom-up attention is a mechanism only driven by stimulus. Most of existing literature of bottom-up saliency are influenced by the idea of Feature Integration Theory (FIT)[15], such as [6], [5], [4]. These methods could lead to a good result, however, the relationship between FIT and the real neural mechanism is not clear. Other theories are also used to solve the problem of visual saliency, such as information theory[1], spectral theory[5], graph theory[4], and machine learning[7].

In this paper, we propose an algorithm based on sparse coding theory. As [12], sparse coding theory could explain what happened in the simple cells of primary visual cortex (V1). Sparse coding decomposes the inputs into two parts, codes and residual. From the viewpoint of information theory, sparse coding could reflect the flow of information. Sparse codes are the minimum entropy codes[12] which means that these codes contain the information of image parts with minimum entropy. So the residual are opposite, containing the information of the parts with maximum entropy . The higher the residual value is, the larger entropy is. The sparse coding length could be seen as the complexity when using specific sparse coding dictionary[9].

As shown in figure 1, we define the weighted sparse coding residual as saliency. The algorithm consists of following steps:

(1) Image inputs are divided into overlapping patches with the same size. For each patch, as [9], the sparse coding dictionary is calculated using the surrounding patches.

(2) Every patch is sparsely coded with its own dictionary. The patch is decomposed into two parts, the codes part and the residual part.

(3) By computing the l_0 norm of the sparse matrix, the weight is obtained. We get our saliency map by multiplying the weight and the residual. The higher the saliency value is, the more salient the area is.

The major contribution of this paper is proposing an efficient and effective algorithm based on weighted sparse coding residual. We define the multiply production of sparse coding length and residual as saliency, and get a good saliency map. This paper applies sparse coding theory to saliency

[*] Area chair: Alexander Hauptmann

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.

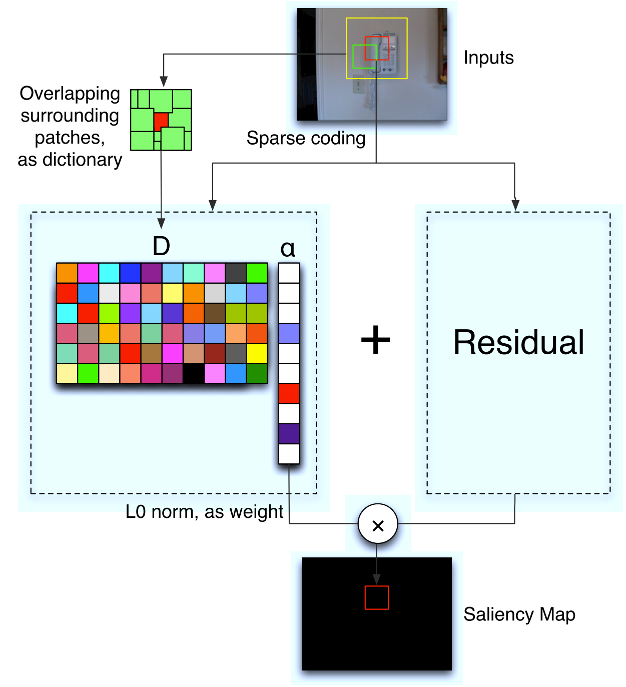

Figure 1: The framework of our method.

and tries to explain the internal relationship between sparse coding theory and visual saliency. Being compared with four state-of-the-art methods, our method outperforms all others and the results show the accuracy and robustness of our algorithm.

2. METHOD

Sparse coding theory helps to understand the response properties of simple cells in primate visual cortex and is an effective way to reduce the redundancy of input image. The basic assumption of this theory is that an image, X, can be represented by a linear superposition of bases in a dictionary, D. The sparse coding decomposes the image into two parts, the sparse codes and the residual. We calculated the saliency map by the weighted residual of sparse coding using the coding length of sparse codes. In following subsections, we will demonstrate our method.

2.1 Sparse Coding to Saliency Map

The first step of sparse coding is dividing image into patches, $X = \{x_1, x_2, \ldots x_n\}$, n is the number of patches. Each patch of the input image could be represented as following:

$$x_i = D\alpha_i + r_i \tag{1}$$

$A = \{\alpha_1, \ldots \alpha_i, \ldots \alpha_n\}$ is the sparse codes for the image and $R = \{r_1, \ldots r_i, \ldots r_n\}$ is the residual of sparse coding. D is the dictionary. The goal of sparse coding is to find the optimal balance between information loss and sparseness. It could be seen as minimizing the following cost function:

$$E = -I(X;A) + \lambda \cdot [\text{sparseness of } A] \tag{2}$$

λ is the parameter of balancing between the sparsity and distortion. The problem of minimizing information loss could be solved by maximizing mutual information $I(X;A)$, where,

$$I(X;A) = H(X) - H(X|A) \tag{3}$$

Assuming that the residual R is independent of the image X, the information relationship among X, A and R could be described as following,

$$\begin{aligned} P(X|A) &= P(X|DA) \\ &= P(DA - X) \end{aligned} \tag{4}$$

while,

$$X = DA + R \tag{5}$$

hence,

$$\begin{aligned} H(X|A) &= H(DA - X) \\ &= H(R) = H(X) - I(X;A) \end{aligned} \tag{6}$$

Therefore, the $H(R)$ implies the uncertainty when knowing A to predict X. The higher the $H(R)$ is, the more uncertainty of the patch is.

By concluding the formulas above, we can see that there is a close connection between residual R and $I(X;A)$, the minimizing residual is closely related to maximizing $I(X;A)$[17]. In our algorithm, the residual is seen as an important component of saliency.

The other important component is the sparseness of the sparse codes. Sparseness could be measured by l_0 norm of the sparse codes. For each patch, the sparseness part of equation 2 could be reformulated as following,

$$[\text{sparseness of } A] = \|\alpha_i\|_0 \tag{7}$$

The l_0 norm in this equation means the number of the non-zero elements, in other words, the coding length. In our method, the dictionary is obtained by using overlapping surrounding patches. According to [9], the sparse coding length represents the similarity and novelty between patch and its surroundings. That is to say, the coding length represents local complexity. In our algorithm, the coding length is seen as the weight.

To sum up above arguments, the saliency of our method is the result of multiplying the l_0 norm of sparse codes and the residual of sparse coding, in other words, the product of local complexity and uncertainty. In our method, the l_1 norm of residual is used. The method could be represented by following function:

$$[\text{Saliency of } x_i] = \|\alpha_i\|_0 \cdot \|x_i - D\alpha_i\|^{L1} \tag{8}$$

2.2 Solution via l_1 Norm

To solve the sparse coding problem above, we use the method of l_1 norm minimization. The cost function of sparse coding (equation 2) could be reformulated as the problem of l_0 norm optimization:

$$\min \lambda\|\alpha_i\|_0 + \|x_i - D\alpha_i\|_2^2 \tag{9}$$

However, this optimization problem is also hard to solve. Related study[2] suggests that for most systems, the problem of sparest near-solution could be represented by the minimal l_1 norm near-solution. The l_0 norm optimization problem of equation 9 could be solved by minimizing the l_1 norm:

$$\min \lambda\|\alpha_i\|_1 + \|x_i - D\alpha_i\|_2^2 \tag{10}$$

This problem is a famous linear regression problem known as Lasso[14]. By using LARS algorithm[3], the l_1 norm minimizing problem could be solved. Therefore, the residual and the coding length could be obtained. By multiplying the residual and the coding length, the saliency value of each patch is obtained. After accumulation and normalization, we get the saliency map of our algorithm. The entire method is summarized as Algorithm 1.

Algorithm 1 Bottom-up Saliency based on Weighted Sparse Coding Residual

Input: Given image X
Output: The saliency map of X.

```
for each patch $x_i$ of the image $X$ do
    Take patches from its surroundings to form dictionary $D$
    Solve the Lasso problem
        $\min \lambda \|\alpha_i\|_1 + \|x_i - D\alpha_i\|_2^2$
    and get $\alpha_i$
    Calculate the local complexity $\|\alpha_i\|_0$ as weight and the $l_1$ norm of sparse coding residual $\|x_i - D\alpha_i\|^{L1}$
    Calculate the patch saliency by
        $[\text{Saliency of } x_i] = \|\alpha_i\|_0 \cdot \|x_i - D\alpha_i\|^{L1}$
end for
Get the saliency map by accumulating and normalizing the saliency by pixels.
```

3. EXPERIMENTS

In this section, we present the results of our method on extensive experiments with two public eye-tracking datasets. The datasets used in this paper is provided by Bruce[1] and Judd[7]. The Bruce dataset is a small-scale and popularly used dataset with 120 images, and the Judd dataset is a larger dataset with 1003 images. To evaluate the performance, we compared our algorithm with four state-of-the-art methods: AIM[1], Itti's[6], Judd's[7] and Liyin's[9]. Receiver Operating Characteristic (ROC) curve is used as evaluation criterion. By the quantitative analysis with ROC curves and Area Under the Curve (AUC) , proposed method shows good performance in different datasets.

3.1 Experimental Setup

The input images are down-sampled to 0.1 of original size for Bruce dataset and 0.08 for Judd dataset. Then, the input images are divided into patches of 8 × 8 pixels with an overlapping of 4 pixels in each direction. The dictionary of each patch is obtained by using the surrounding patches. Because we use surrounding patches for dictionary, there would be a boundary effect. Therefore, we also multiply the length of dictionary to reduce the impact. To realize the LARS algorithm, we use the fast implementation in SPAMS toolbox[10].

We realize Itti's method by using the reimplementation by Harel[4], which is a bit faster and more accurate in fixation prediction. For other methods, we use the default settings of their own implementation.

To evaluate the performance, we plot the ROC curves by computing the mean value of the output from toolbox provided by Harel[4]. The AUCs of each method is also computed.

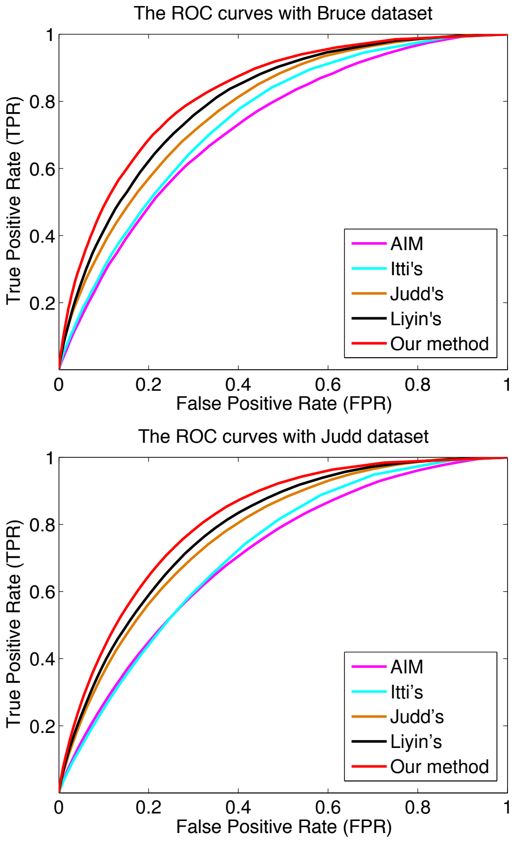

Figure 2: The ROC curves of four state-of-the-art methods and our method. The top one is with Bruce dataset and the bottom one is with Judd dataset.

3.2 Experimental Results

The results are evaluated by using quantitative and qualitative methods. On quantitative evaluation, we plot the ROC curves of different methods, by comparing the ROC curves and the AUCs, then the performance of each method is evaluated. On qualitative evaluation, we show some of the experiment results and compare them with saliency maps of other methods and the density maps of human fixation.

As shown in figure 2, we compare the ROC curves of our method with other methods. By observing the curves of the top one, we can see that our method shows competitive performance with Bruce dataset. To test the robustness of our method, we also evaluate the performance on Judd dataset, which is one of the largest public datasets. The results of our method with Judd dataset also outperform all others and the small decrease indicates the robustness of our algorithm. To know the results more directly, the AUCs for each method with two datasets are shown in table 1.

Our method is fast because we only use one scale of the input image (0.1 for Bruce dataset, 0.08 for Judd dataset) to calculate the saliency. For each image in the dataset, our method only spends about 1.5 seconds for each image in the two datasets on a Core 2 Duo 2.4GHz MAC.

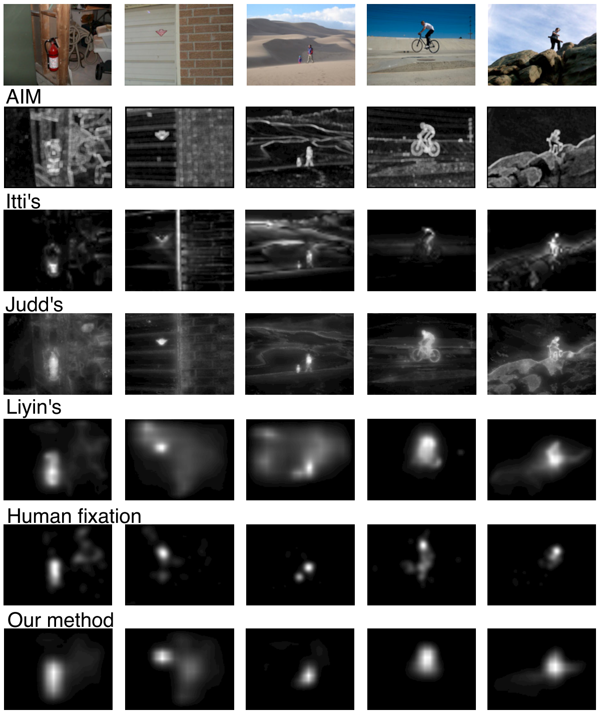

Figure 3: Some experiment results. The rows from top to down are: the original stimulus images, saliency maps generated by AIM[1], Itti's[6], Judd's[7] and Liyin's[9] methods, human fixation density maps and our saliency maps.

	Bruce Dataset	Judd Dataset
AIM[1]	0.7241	0.7076
Itti's[6]	0.7455	0.7169
Judd's[7]	0.7795	0.7732
Liyin's[9]	0.8006	0.7893
Our Method	**0.8264**	**0.8141**

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

We show some saliency map results in figure 3. By comparing with the results of AIM[1], Itti's[6], Judd's[7] and Liyin's[9] methods and human fixation, the saliency maps of our method could represent salient area accurately and show good ability in application.

4. CONCLUSION

In this paper, we propose a bottom-up saliency algorithm based on weighted sparse coding residual. The key of our method is sparse coding theory which is also regarded as a good way to understand vision. In the experiments, our method has the best performance and largest AUCs in comparison with four state-of-the-art methods. Our method is able to provide accurate and robust saliency maps for other applications. For future work, we will expand our method to temporal domain and apply our method to other computer vision problems.

5. REFERENCES

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

[2] D. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, Jan 2006.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, Jan 2004.

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

[5] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, 2007.

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

[7] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *IEEE 12th International Conference on Computer Vision, ICCV*, 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] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang. Incremental sparse saliency detection. *The International Conference on Image Processing, ICIP*, Jan 2009.

[10] 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.

[11] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. *IEEE 12th International Conference on Computer Vision, ICCV*, 2009.

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

[13] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, 2004.

[14] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.*, Jan 1996.

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

[16] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video-retargeting. *IEEE 11th International Conference on Computer Vision. ICCV*, 2007.

[17] 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.