SALIENCY DETECTION BASED ON SHORT-TERM SPARSE REPRESENTATION

Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, Pengfei Xu, Xianming Liu, Shaohui Liu

Department of Computer Science and Engineering, Harbin Institute of Technology
\{xssun,yhx,rrji,pfxu,liuxianming,shliu\}@vilab.hit.edu.cn

ABSTRACT

Representation and measurement are two important issues for saliency models. Different with previous works that learnt sparse features from large scale natural statistics, we propose to learn features from short-term statistics of single images. For saliency measurement, we define background firing rate (BFR) for each sparse feature, and then we propose to use feature activation rate (FAR) to measure the bottom-up visual saliency. The proposed FAR measure is biological plausible and easy to compute, also with satisfied performance.

Experiments on human eye fixations and psychological patterns demonstrate the effectiveness and robustness of our proposed method.

Index Terms— Saliency detection, sparse feature, feature activation rate

1. INTRODUCTION

When looking at an image or browsing a visual scene, our visual system would first focus on several important or interesting regions by saccade moments of the eyes. In psychology and computational neuroscience, eye moment is proven to be driven by visual saliency which can be either bottom-up data-driven or top-down task-driven. Thus, saliency detection is very essential to visual attention modeling and various computer vision tasks such as video compression [1] and object detection[2].

1.1. Related Works

Current works focus on proposing appropriate visual features and reasonable measurements for calculating visual saliency. Itti et al. [3] proposed a computational framework based on Koch and Ullman’s architecture of attentional selection[4]. This framework is based on the biological structure of human visual system, in which visual saliency is measured by spatial center-surround differences across different scales and feature channels. Itti’s model mainly simulates the early stage of visual attention system, in which only simple features are extracted.

Another type of methods are purely mathematical motivated. Hou et al.[5], Guo et al.[6] achieved fast and robust saliency detection by using the frequency domain signal processing. These methods utilize phase spectrum of Fourier Transform to obtain better saliency map with much less computational cost compared with previous biological inspired methods derived from Itti’s architecture.

The third type of methods is motivated both by the biological structure and mathematical optimization. Bruce[7] proposed a method based on spare coding representation and the principle of information maximization, in which self-information of the sparse coefficients is used as quantized measurement of visual saliency. Also based on spare coding, Hou[8] proposed that saliency should be dynamically measured by incremental coding length. These methods have been proven to be very effective in predicting eye fixations captured from human subjects when viewing natural images.

1.2. Main Approach and Contribution

Good representation and reasonable measurement are both critical issues in modeling visual saliency mechanism. For every input scene, we learn a group of basis functions using independent component analysis from short-term statistics instead of large scale natural statistics. The original input data can be precisely represented by linear combination of the learnt basis functions without any lost of important information. Based on this representation, each basis function provides a unique feature channel. We regard each feature channel as a group of neurons in the brain. According to neural science, we define background firing rate (BFR) to describe the average activity of the feature, and then a measurement called Feature Activation Rate (FAR) is proposed for measuring how much energy will be consumed when certain visual pattern appears. By assuming that larger energy consumption indicates larger signal saliency, bottom-up visual saliency could be estimated according to our FAR measurement.

Our contribution lies in two aspects: 1. We propose a novel image representation method based on short-term statistics which yields a more accurate (zero reconstruction error) and more sparse representation than traditional method. 2. Based on our new representation, we propose a novel saliency measurement called Feature Activation Rate (FAR), which has quite simple formulas but strong biological plausibility and satisfied performance.
2. VISUAL REPRESENTATION

In this paper, we define three types of representation method: sensory representation, short-term representation and long-term representation. Sensory representation corresponds to the original spatial representation. Short-term representation represents the image by features learnt from short-term statistics. Short-term statistics denotes the statistics of sensory information received within a short period of time. Long-term representation is based on the features learnt from long-term statistics (synonymous with large scale natural statistics).

According to our definition, most of the traditional methods [7, 8, 9] are based on long-term representation. Through our experiments, we found that long-term representation has two major drawbacks: 1. It is inaccurate; 2. It cannot capture some unique features under specific visual context. These two drawbacks make the long-term representation imperfect for saliency detection task. In our model, we use short-term sparse representation to overcome these problems.

2.1. Visual Features in Previous Models

Olshausen et al. [10] had proven that the receptive fields of simple cells in primary visual cortex yield a sparse representation which motivates lots of works to represent natural images by sparse features [7, 8, 9].

Independent Component Analysis (ICA) is an effective method for finding underlying components from statistical multi-dimensional data. The independent components learnt by ICA are both statistically independent, and non-gaussian. The coefficients of ICA basis functions yield a sparse representation of the original data, which is somehow consistent with the visual representation in primary visual cortex.

In most case, the basis functions are learnt from a large number of image patches sampled from different kinds of natural images. The resulting sparse features indeed provide better results compared with classic early visual features (intensity, color, and orientation).

Although the basis functions are learnt from large scale natural statistics, they cannot provide perfect representation for every input image with zero information lost. Fig. 1 shows some reconstruction results of natural images using a set of basis functions learned from 120,000 8×8 patches. The reconstruction error is caused by some unique features that cannot be captured by the pre-defined basis functions. As illustrated in the error maps, the un-captured features are mostly belong to the salient objects of the scene, which means they should be important cues for saliency detection.

2.2. Short-term Sparse Representation

Traditional ICA method provides a generalized long-term sparse representation for natural images without considering the adaption to the context of certain environment. This critical drawback makes the traditional representation imperfect for saliency detection task. Using sparse features learnt from short-term statistics instead of long-term statistics can perfectly solve this problem.

Given a set of basis functions A, an vectorized image patch I can be represented as the following form:

\[ I = As, \]  

where s is the coefficients of each basis function in A. For short-term representation, A is learnt by Independent Component Analysis on image patches sampled from every location of the current input image. With \( W = A^{-1} \), the coefficients can be obtained by:

\[ s = WI. \]

The coefficients of all patches yield a short-term sparse representation of the original image.

Applying ICA for every single image will cause additional computation, however, it also yields a better sparse representation for the input data which is the key factor to build a better saliency model.

3. SALIENCY MEASUREMENT

In previous models, saliency is measured by center-surround difference[3], self information[7, 9], discriminant center-surround composition[11], or incremental coding length[8]. All measurements are related to information theory except for center-surround difference which is inspired by the receptive field of retina ganglion cells. In this section, we propose a much simpler bio-inspired measurement called Feature Activation Rate.

3.1. Bottom-up Saliency as Feature Activation Rate

3.1.1. Fundamentals of Neuron Activity

According to neural science, the firing rate of specific neuron is stable at a certain value which is defined as the background firing rate. The neuron’s firing rate will increase or decrease
when certain patterns appear in its receptive field. Both increasing and decreasing activity consume energy. In purely bottom-up circumstance, there will be no influences of prior knowledge, which means the importance of the input signal is probably related to how much energy it consumes. Based on our current knowledge, we give a quite reasonable assumption that larger energy consumption causes larger bottom-up saliency.

3.1.2. Formalizing Feature Activity

According to the knowledge from neural science, we regard the features as a group of neurons so that the saliency could be measured by energy consumption of all features based on the assumption proposed in sec. 3.1.1. We first define background firing rate to describe the average activity of the feature:

$$BFR_j = \frac{\sum_{i=1}^{M} F_{ij}}{M},$$

where $BFR_j$ is the background firing rate of $j$th feature, $F_{ij}$ is the $j$th feature of $i$th patch, $M$ is the number of sampled patches. Then we define a measurement called Feature Activation Rate (FAR) for calculating how much energy will be consumed when certain visual pattern appears.

$$FAR_i = \sum_{j=1}^{N} \|F_{ij} - BFR_j\|,$$

where $FAR_i$ is the feature activation rate of $i$th patch, $N$ is the dimension of all features. To make our model easy to implement, we measure energy cost simply by the Manhattan Distance between the input features and the background firing rates.

3.2. Saliency Calculation

Algorithm 1 gives a detailed saliency detection algorithm based on the proposed temporary features and FAR measurement. We use FastICA algorithm provided by [12] as the implementation of Independent Component Analysis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Area under ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itti et al.</td>
<td>0.6146 (0.00008)</td>
</tr>
<tr>
<td>Gao et al.</td>
<td>0.6395 (0.0007)</td>
</tr>
<tr>
<td>AIM</td>
<td>0.6727 (0.00008)</td>
</tr>
<tr>
<td>SUN</td>
<td>0.6682 (0.0008)</td>
</tr>
<tr>
<td>ICL</td>
<td>0.6460 (0.00008)</td>
</tr>
<tr>
<td>Our Method</td>
<td>0.6979 (0.00007)</td>
</tr>
</tbody>
</table>

4. EXPERIMENTS

4.1. Eye fixation prediction

We tested our method on human eye fixation data provided by Bruce [7]. The database contains 11999 eye fixations captured from 20 subjects free viewing 120 natural images. As mentioned in [9], the original ROC measurement proposed in [7] is largely affected by the “edge effect” due to the central bias caused by the central composition of interesting objects. To eliminate the interference, we follow the proposal of [9], and use the same procedures to draw the ROC curves.

We compared our model against Itti et al.[3], Gao et al.[11], AIM[7], SUN[9] and ICL[8]. Our saliency maps are obtained by Algorithm 1 with the parameters $H = 60, W = 80, B = 5 \times 5 \times 3$. Saliency maps of other methods are generated using their default parameter settings. The mean and standard errors are shown in Table 1. Our method outperforms all state-of-the-art models in ROC evaluation, which demonstrates the effectiveness of the proposed short-term sparse representation and FAR measurement.

Fig. 2 shows more visual results. Due to the limitation of space, we only present the saliency maps of AIM (the previous best method in ROC evaluation) and our method for comparison. As illustrated in Fig. 2, saliency maps generated by our method are more sparse and accurate than AIM.

4.2. Response to psychological patterns

We also test our method on various psychological patterns. There is no proper evaluation measurement for this experiment. From visual observation, we can see that our method can deal with not only normal patterns such as color, orientation, intensity, curve and insertion (Fig. 3 (a)), but also patterns with noise (Fig. 3 (b)) and the conjunction of features (Fig. 3 (c)).

Experimental results on both eye fixation prediction and psychological patterns demonstrate the effectiveness of our proposed method.
Fig. 2. Examples of saliency maps. For every input image, we present (from left to right) the original color image, the density map generated using real eye fixations, and saliency maps output by AIM model and our method.

Fig. 3. Response to psychological patterns.

5. CONCLUSION

In this paper, we have proposed a saliency detection method based on short-term sparse representation and a novel bio-inspired saliency measurement called feature activation rate. Features extracted from short-term statistics could precisely represent the original image without any lost of important information. FAR measurement reflects the energy consumption of certain visual patterns, which is easier to implement and leads to better performance in predicting eye fixations.

6. ACKNOWLEDGEMENTS

This work is supported by National Basic Research Program of China (2009CB320906) and National Natural Science Foundation of China (60775024).

7. REFERENCES