

Sparse Embedding Visual Attention Systems combined with Edge Information

Cairong Zhao^{1,2}, ChuanCai Liu¹,
 1. School of Computer Science, NUST,
 Nanjing, China
 E-mail: cairong.zhao@yahoo.com

Zhihui Lai¹, Jingyu Yang¹
 2. Department of Physics and Electronics, MC
 Fuzhou, China
 E-mail: chuancailiu@yahoo.com.cn

Abstract—The general computational models of visual attention are to obtain multi-scale feature maps in terms of visual properties like intensity, color and orientation, and then combine them to get one saliency map. But due to the lack of object edge information and reasonable feature combination strategy, the visual saliency map of the image is a blur map. Being aware of these, we propose a new scheme for saliency extraction. In this paper, we firstly put forward a sparse embedding feature combination strategy, inspired by sparse representation. The strategy is used to combine the salient regions from the individual feature maps based on a novel feature sparse indicator that measures the contribution of each map to saliency. Then we combine traditional visual attention with edge information. Results on different scene images show that our method outperforms other traditional feature combination strategies.

Keywords: visual attention; sparse representation; edge information.

1. Introduction

Visual attention is an important characteristic of human visual system (HVS). A common view [1] of how attention deployed onto a given scene under bottom-up influences is as follows. Low-level feature extraction mechanisms act in a massively parallel manner over the entire visual scene to provide the bottom-up biasing cues towards salient image locations. Attention then sequentially focuses on salient image locations to be analyzed in more detail [2], [3].

Several computational models have been proposed to functionally account for many properties of visual attention in primates during the last two decades. But all of their saliency maps are blur maps of the image without any edge information. Shashua and Ullman [4] considered that not only image region has attention property but also the edge. They regard the feature contrast as the local saliency and edge as the global

saliency. Gestalt's law commonly measures the saliency property: proximity, closure and continuity.

Itti [1] proposed four feature combination strategies: the "Naive", "N(.)", "Trained", "Iterative". The four strategies studied all involve a point-wise linear combination of feature maps into the scalar saliency map. Indeed, there is mounting psychophysical evidence that different types of features do contribute additively to salience, and not, for example, through point-wise multiplication [6]. So we propose the sparse embedding feature combination strategy and define feature sparse indicator measured by sparse representation that adjusts the weights of each feature map in proportion of its contribution to the saliency map.

Being aware of these, in this paper, we propose a new scheme for saliency extraction, called sparse embedding visual attention systems with edge information. Results on different scene images show that the method outperforms the traditional visual attention models.

The proposed method relies widely on the saliency model of visual attention and sparse representation. The organization of the paper is as follows. Basics of the saliency model of visual attention and sparse representation are recalled in section 2. Then, section 3 develops the idea of the proposed method and the relevant theory and algorithm. Section 4 describes the related experiments. Section 5 offers our conclusions.

2. Outline of saliency model of visual attention and sparse representation.

2.1 Saliency model of visual attention

Itti and Koch proposed the saliency-based model of visual attention in [7]. It is based on three major steps.

First, a number of features are extracted from the scene by computing the so-called feature maps from a RGB color image and which belong to three main cues, namely intensity, color, and orientation.

Intensity feature

$$F_I = (r + g + b) / 3 \quad (1)$$

Two chromatic features based on the two color opponency filters R^+G^- and B^+Y^- .

$$F_{BY} = \frac{b - \min(r, g)}{\max(r, g, b)}, F_{RG} = \frac{r - g}{\max(r, g, b)} \quad (2)$$

For intensity and chromatic features, a Gaussian pyramid P_j is created by progressively lowpass filtering and subsampling by factor 2 the feature map F_j , using a Gaussian filter G :

$$\begin{aligned} P_j(0) &= F_j \\ P_j(i) &= (\downarrow 2)(P_j(i-1) * G) \end{aligned} \quad (3)$$

where $(*)$ refers to the spatial convolution operator and $(\downarrow 2)$ refers to the downsampling operation.

Local orientation features are obtained from I using oriented Gabor pyramids $P_o(\sigma, \theta)$, where σ represents the scale and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation.

In a second step, Center-surround is then implemented in the model as the difference between fine and coarse scales: The center is a pixel at scale $c \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{3, 4\}$.

$$F_l = \sum_{c=3}^5 \sum_{s=c+3}^{c+4} |P_l(c) - P_l(s)|, \forall l \in L = L_I \cup L_C \cup L_O \quad (4)$$

where L_I, L_C, L_O respectively indicate intensity feature set, color feature sets, and orientation feature sets. Then each feature map F_l is transformed into its conspicuity map M_l .

In the final step of the attention model, the cue conspicuity maps are integrated into a saliency map S , defined as:

$$S = \frac{1}{3} \sum_{i=1}^3 N(M_i) \quad (5)$$

where $N(\cdot)$ is a normalization operator.

2.2 A brief review of sparse representation

The role of parsimony in human perception has also been strongly supported by studies of human vision. Investigators have recently revealed that in both low-level and mid-level human vision [8], many neurons in the visual pathway are selective for a variety of specific stimuli, such as color, texture, orientation, scale.

Given a $M \times N$ matrix A containing the elements of samples in its columns, and $y \in R^m$. The problem of sparse representation is to find an $N \times 1$ coefficient vector s_i for each y through the following modified l_1 minimization problem:

$$\min \|s_i\|_1 \quad s.t. \quad y = As_i \quad (6)$$

For $y \in R^m$, let $s_i^o (i=1, 2, \dots, N)$ be the optimal solution of the above constrained optimization problem. Then, the residual of constructing y is defined as:

$$r(y) = \|y - As_i^o\|_2 \quad (7)$$

3. Sparse embedding visual attention system combined with edge information

3.1 Extraction of edge feature

In this subsection, we obtain the edge feature from intensity pyramid image using by LOG edge detector. In the field of digital image processing, the LOG detector is often replaced by DOG (Difference of Gaussians) and described as below:

$$M_E = P_I(\sigma) * DOG(\sigma_1, \sigma_2) \quad (8)$$

where $*$ is the convolution operator and σ is the pyramid scale. $P_I(\sigma)$ is computed by using Equations (1) and (3). $DOG(\sigma_1, \sigma_2)$ is obtained by

$$DOG(\sigma_1, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{x^2 + y^2}{2\sigma_1^2}\right) - \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{x^2 + y^2}{2\sigma_2^2}\right) \quad (9)$$

where σ_1, σ_2 are the variances of the DOG filter.

Combining the edge feature, the saliency map are redefined as

$$S_s = \sum_{i=1}^K w_{(i)}(f) N(M_i) \quad (10)$$

where $w_{(i)}(f)$ characterizes the conspicuity's contribution to saliency map and K represents the number of conspicuity maps.

3.2 Sparse embedding feature combination strategy.

For providing suitable weight w of each feature map (F_l) that constructs saliency map, we need a mechanism that changes the relative contribution to the final saliency map of the various features considered.

Following features extracted from the scene by computing the feature maps from a RGB color image, we obtain N feature maps (F_l), and reshape them as column vectors $\{f_i\}_{i=1}^N \in R^m$. Let matrix $F = \{f_1, f_2, \dots, f_N\}$ be the data matrix including all the feature maps in its column. To reduce the computational cost, we perform

PCA to reduce the data dimension. The projection from the raw data space to the feature space can be represented as matrix R^{pca} .

Then sparse representation firstly seeks a sparse reconstructive weight vector s_i for each f_i through the following modified l_1 minimization problem:

$$\min \|s_i\|_1, \quad (11)$$

$$s.t. x_i = R^{pca} F s_i$$

where $s_i = [s_{i,1}, \dots, s_{i,i-1}, 0, s_{i,i+1}, \dots, s_{i,N}]^T$ is a N -dimensional vector in which the i^{th} element is equal to zero (implying that the x_i is removed from X), and the elements $s_{i,j} (i \neq j)$ denote the contribution of each x_j to reconstruction x_i .

For $f_i \in R^m$, let $s_i^p (i=1,2,\dots,N)$ be the optimal solution of the above constrained optimization problem. The feature map f_i is sparsely represented by the remained $N-1$ feature maps. Based on the above sparse representation, we define the following measure of how prominent each feature is.

Definition 1. (Feature Sparse indicator): The feature saliency indicator is defined as:

$$FSI(f_i) = residuals(f_i) = \|f_i - F s_i^p\|_2, \quad (12)$$

where $FSI(f_i)$ stands for the degree of the feature f_i and s_i^p is a new vector whose only nonzero entries the greater the $FSI(f_i)$ is, the more prominent the feature is.

Definition 2. (Weight of Saliency): The weight of saliency is defined as:

$$w(f_i) = \frac{FSI(f_i)}{\sum_{i=1}^N FSI(f_i)}, \quad (13)$$

We compared the proposed feature combination strategy with other strategies. The results are showed in Fig. 1. The Naïve model, which represents the simplest solution to the problem of combining several feature maps into a unique saliency map, performed always worse than other's. The $N(\cdot)$ model yields reliable yet nonspecific detection of salient image locations. The iterative model yields much sparser maps, in which most of the noisy is strongly suppressed. The proposed sparse feature combination strategy capture more sparser maps and more reasonable saliency region than other's, attribute to more deliberate saliency weights. The proposed strategy could be refined to mimic more closely what is known of the physiology of early neurons.

4. Experiment

In this section, we use some natural color images to evaluate the proposed algorithm compared with the

Itti's methods. We extract 48 feature maps from two natural scenes and compute the each feature sparse indicator based on sparse representation. Then four conspicuity maps are normalized and combined to the sparse saliency map S_s .

Finally, the saliency locations are detected and compared with Itti's. Details are illustrated in Fig. 2. In Itti's method, the close shot target i.e. the white and red sailboats can be successfully detected, but the future white sailboat can't be recognized and taken as a prominent color region with other future regions. Obviously, in our method, we can detect the close shot i.e. the white and red sailboats as well as further sailboats, due to edge information supplement to the early visual feature and sparse embedding feature combination strategy. So the proposed method is more suitable for the human visual perception.

5. Conclusion

In this paper, we have proposed a new scheme for saliency extraction. The method is used to combine the salient regions from the individual feature maps based on a new feature sparse indicator that measures the contribution of each map to saliency. Furthermore, the edge information is supplied to the extraction of early visual features. It exploits properties of human visual system for the extraction of the feature maps. Compared to existing feature combination strategies, it improves the accuracy of salient region detection. The experimental results clearly show that the proposed method obtain more reasonable salient region for human visual perception.

Acknowledgements

This work is partially supported by the Fujian Provincial Department of Science and Technology of China under grant No.2007F5083, 2008F5045. It is also partially supported by the National Science Foundation of China under grant No. 60472061.

References

- [1] Laurent Itti, Christof Koch, Feature combination strategies for saliency-based visual attention systems, *Journal of Electronic Imaging* 10(1).161-169, 2001.
- [2] A.M. Treisman, G.Gelade, A feature-integration theory of attention, *Journal of Cognitive Psychology* 12, 97-136(1980).
- [3] C. Koch, S.Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, *Human Neurobiology* 4, 219-297,1985.
- [4] A. Shashua and S. Ullman, Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network, *Pattern Analysis Machine Intelligence* 7(1): pp90-94, 1988
- [5] Paul L. Rosin, Edges: Saliency Measures and Automatic Thresholding, *Machine Vision and Applications* 9(4) pp139-159, 1997
- [6] H.Northdurft, Saliency from feature contrast: Additively across dimensions, *Vision Research* 36, 1115-25,1996

[7] Itti,L, Koch,C, and Niebur,E, A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis Machine Intelligence* 20(11): 1254-1259, 1998.
 [8] B.Olshausen, D. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1? , *Vision Research* 37: pp. 3311–3325, 1997.

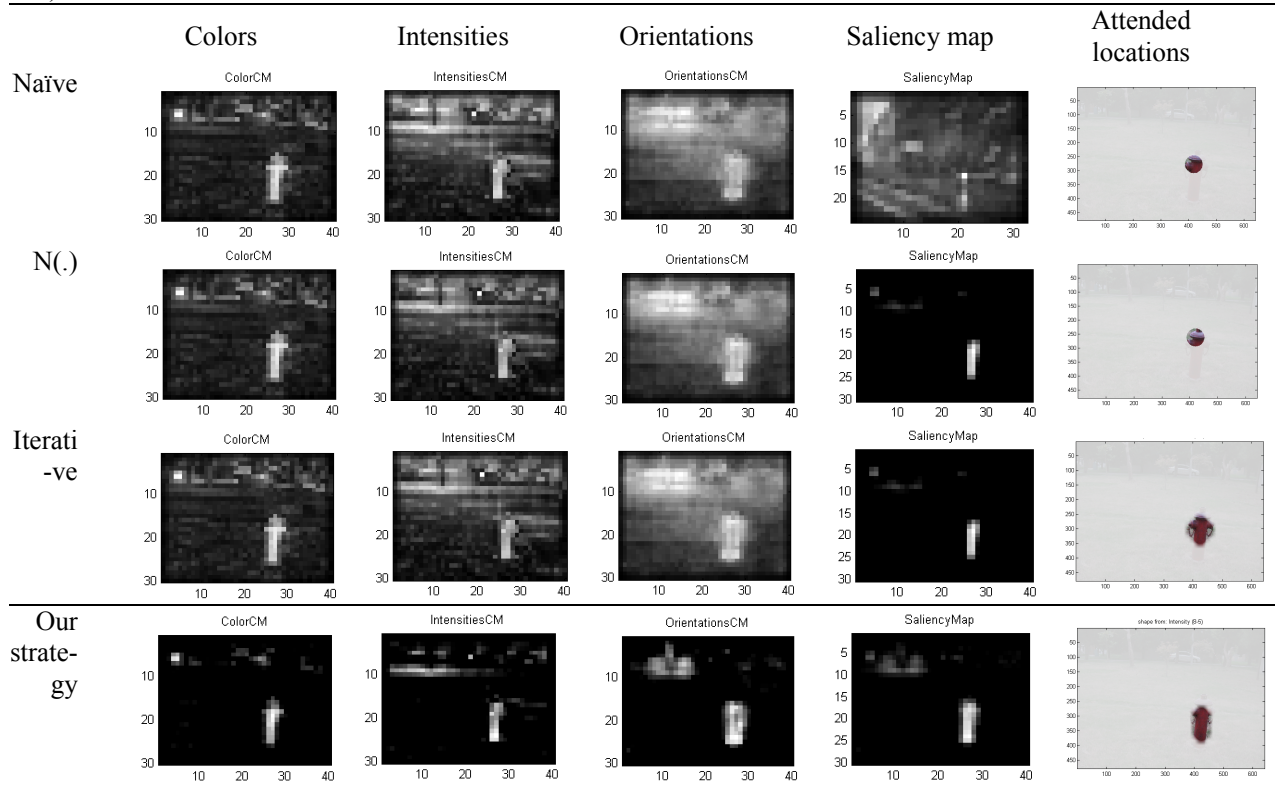


Fig. 1. Comparison with other strategies. The test image (b) is selected from the image dataset taken on campus of NJUST, in which a red fire hydrant is the most salient object.

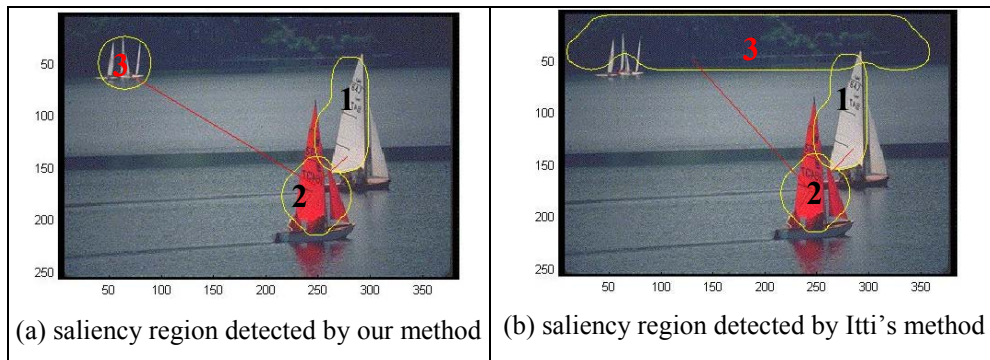


Fig. 2. The saliency locations are detected by our method and compared with Itti's. The "1", "2", "3" respectively represent the first saliency region, the second saliency region, and the third saliency region