Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

journal homepage: www.elsevier.de/aeue

Sparse embedding visual attention system combined with edge information

Cairong Zhao [a,b,*], Chuancai Liu [b], Zhihui Lai [b], Huaming Rao [b], Zuoyong Li [a,b]

[a] *Department of Physics and Electronics, Minjiang University, Fuzhou 350108, China*
[b] *Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China*

ARTICLE INFO

Article history:
Received 12 January 2010
Received in revised form 11 April 2011
Accepted 11 April 2011

Keywords:
Visual attention
Sparse representation
Feature sparse indicator
Edge information

ABSTRACT

Numerous computational models of visual attention have been suggested during the last two decades. But, there are still some challenges such as which of early visual features should be extracted and how to combine these different features into a unique "saliency" map. According to these challenges, we proposed a sparse embedding visual attention system combined with edge information, which is described as a hierarchical model in this paper. In the first stage, we extract edge information besides color, intensity and orientation as early visual features. In the second stage, we present a novel sparse embedding feature combination strategy. Results on different scene images show that our model outperforms other visual attention computational models.

1. Introduction

Visual attention is an important characteristic of human visual system (HVS). It has been a subject of research in neural science, physiology, psychology, and computer vision. A common view [1] of how attention is 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]. Visual attention hence allows for seemingly real-time performance by breaking down the complexity of scene understanding into a fast temporal sequence of localized pattern recognition problems [4].

Several computational models have been proposed to functionally account for many properties of visual attention in primates during the last two decades [5–11]. The early stage of visual feature processing appears to extract a feature subset of the available sensory information before further processing. Given an input image, the early visual feature processes consists of decomposing this input information into a set of distinct "channels", by using linear filters tuned to specific stimulus dimensions, such as luminance, color, various local orientations etc. [25]. These models typically share similar general architecture. Multi-scale topographic feature maps detect local spatial discontinuities in intensity, color, and orientation. But there are some problems such as which of early visual features should be extracted and how combine these different features into a unique "saliency" map.

In a common case, the regions with many abrupt changes or some unpredictable characteristics often attracting the human attention are considered as the salient region of images. Shashua and Ullman [12,13] hence put forward that the salient region had attention property as well as edge information. They regard the feature contrast as the local saliency and edge as the global saliency. Therefore, we bring forward to extract edge information besides color, intensity and orientation in the early visual feature processing stage. After the early visual feature extraction, it is important how to combine multi-scale feature maps, from different visual modalities with unrelated dynamic ranges, into a unique saliency map. If we merely sum up all the feature maps in a straight way, salient objects appearing strongly in only a few maps may be masked by noise or by less-salient objects present in a larger number of maps. Itti proposed four feature combination strategies in [1]: the "Naive", "$N(\cdot)$", "Trained" and "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 [14]. How to select weights of each feature map that can correctly reflect the nature contribution to saliency map is still an open problem. Sparse representation theory demonstrates that the neurons in primary visual cortex form a sparse representation of natural scenes in the viewpoint of statistics [15]. Based on this fact, we put forward 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.

* Corresponding author at: Department of Physics and Electronics, Minjiang University, Fuzhou 350108, China. Tel.: +86 15850559315; fax: +86 02584315751.
E-mail address: cairong.zhao@yahoo.com (C. Zhao).

1434-8411/$ – see front matter © 2011 Elsevier GmbH. All rights reserved.
doi:10.1016/j.aeue.2011.04.002

In this paper, we propose a sparse embedding visual attention systems combined with edge information. In this model, there are two main works: (1) Edge information is extracted in the early visual features processing stage; (2) A novel sparse embedding feature combination strategy is put forward. Results on different scene images show that the proposed model outperforms other traditional visual attention models, attributed to the more deliberate feature sparse indicator and edge information extraction in the early feature processing stage.

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 presents the idea of the proposed model 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 [6]. This model provides a massively parallel method to simulate the primate vision [16]. Visual attention acts on a multi-featured input and multiple spatial scales are created using dyadic Gaussian and Gabor pyramids. Gaussian pyramids are variable scale Gaussian filters. Gabor pyramids are variable scale and orientation Gabor filters. Gaussian filters mimic non-uniform sampling of retina. Gabor filters approximate the receptive field sensitivity profile (impulse response) of orientation-selective neurons in human visual cortex [21]. Furthermore, the response properties of the Gaussian and Gabor filters have been justified to be effective according to what is know of their neuronal equivalents in the early stage of visual processing human vision [19,20]. The early stages of visual feature processing are some kinds of visual attentions mechanism that capture a distinct subjective perceptual property which makes some stimuli stand out from among other items or locations. Then with these stimuli, our brain will rapidly compute salience in an automatic manner and in real-time over the entire visual field. Visual attention is then attracted towards salient visual locations and our attention will be attracted to visually salient stimuli. It helps our brain achieve reasonably efficient selection.

Consequently, Gabor filters are appropriate for orientation information extraction. Saliency of locations is influenced by the surrounding context. Each feature is computed in a center-surround structure akin to visual receptive fields; the saliency of locations is represented on a scalar saliency map: the saliency map. The implementation details of the model are presented in Fig. 1 and recalled below.

Given an input image, the first proceesing step consists of decomposing this input into a set of distinct feature: namely intensity, color, and orientation.

(1) Intensity feature

$$F_1 = I = \frac{(r+g+b)}{3} \tag{1}$$

where r,g,b is the red, green and blue channels of the input image and I is the intensity image.

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

$$F_2 = \frac{b-\min(r,g)}{\max(r,g,b)}, F_3 = \frac{r-g}{\max(r,g,b)} \tag{2}$$

Gaussian pyramid is produced from the convolution of a variable-scale Gaussian (G(x,y,σ)) with the input feature, $F_j(x,y)$. In the implementation, pyramids have a depth of 9 scales(where $\sigma \in [0..8]$), providing horizontal and vertical image reduction factors ranging from 1:1 (level 0; the orignial input image) to 1:256 (level 8) in consecutive power of two. $P_j(x, y, \sigma)$ is defined as:

$$P_j(x,y,\sigma) = \mathrm{Gauss}(x,y,\sigma) * F_j(x,y), \tag{3}$$

where * is the convolution operation and

$$\mathrm{Gauss}(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{4}$$

(3) For orientation features $F_{4.7}$, Gabor pyramid $P_j(x, y, \sigma, \theta)$ are obtained from I using Gabor(σ, θ), where $\sigma \in [0..8]$ represents the scale and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation.

$$P_j(x,y,\sigma,\theta) = \mathrm{Gabor}(x,y,\sigma,\theta) * I(x,y), \tag{5}$$

where * is the convolution operation and

$$\mathrm{Gabor}(x,y,\sigma,\theta) = \exp(-\frac{(x\ \cos\theta + y\ \sin\theta)^2 + r^2(-x\ \sin\theta + y\ \cos\theta)^2}{2\sigma^2}) \times \cos\left(\frac{2\pi}{\lambda}(x\ \cos\theta + y\ \sin\theta)\right) \tag{6}$$

In a second step, each feature map is computed in a center-surround structure akin to visual receptive fields.

$$F_l = \sum_{c=3}^{5}\sum_{s=c+3}^{c+4} \left|P_l(c) - P_l(s)\right|, \ \forall l \in L = L_\mathrm{I} \cup L_\mathrm{C} \cup L_\mathrm{O}, \tag{7}$$

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 $\delta \in \{3, 4\}$. L_I, L_C, L_O indicate intensity feature set, color feature sets, and orientation feature sets respectively. We hence compute six feature maps F_l(where $l \in [1..6]$, i.e., at scales 2–5,2–6,3–6,3–7,4–7,4–8) for each type of feature. Then each type of feature maps is combined into its conspicuity map M_i (where $i \in [1..3]$) at scale 4.

In the final step of the attention model, the cue conspicuity maps are integrated into a saliency map S at scale 4, defined as:

$$S = \frac{1}{3}\sum_{i=1}^{3} N(M_i) \tag{8}$$

That is, the scalar map that accounts for the final interest distribution over the image.

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 [17,18], many neurons in the visual pathway are selective for a variety of specific stimuli, such as color, texture, orientation and scale. Considering these neurons to form an over-complete dictionary of base signal elements at each visual stage, the firing of the neurons with regard to a given input image is typically highly sparse. This representation is naturally sparse, involving only small fraction of the overall training database.

Given a set of training sample $\{A_i\}_{i=1}^{N} \in R^m$, let matrix $A = [A_1, A_2, ..., A_N]$ be the data matrix including all the training samples in its columns. As any sample $y \in R^m$, the linear representation of y can be written in terms of all training samples as:

$$y = As_0 \in R^m, \tag{9}$$

Sparse representation seeks a sparse reconstructive weight vector s_i for each y through the following l_1 minimization problem:

$$\min \left\|s_i\right\|_1 \ \mathrm{s.t.}\ y = As_i, \tag{10}$$

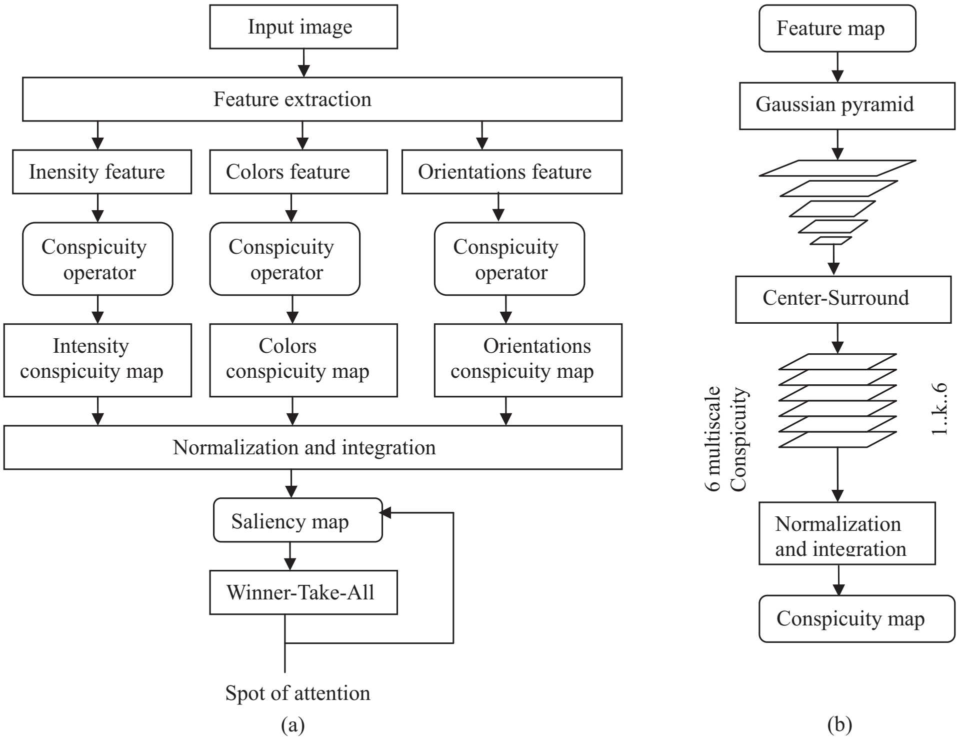


Fig. 1. Saliency-based model of visual attention. (a) Represent the three main steps of visual attention model. Feature extraction, conspicuity computation, saliency map computation by integrating all conspicuity maps and finally the diction of spots of attention by means of winner-take-all network. (b) Illustrate, with more details, the conspicuity operator, which computes six intermediate multi-scale conspicuity maps. Then, it normalizes and integrates them into the feature-related conspicuity map.

Let $s_i^o (i = 1, 2, \ldots, N)$ be the optimal solution of the above constrained optimization problem. Then, the residual of constructing y is defined as:

$$r(y) = \left\| y - As_i^o \right\|_2 \tag{11}$$

3. Sparse embedding visual attention system combined with edge information (SEVASE)

In the saliency model of visual attention, Itti et al. [6] define three main features, namely intensity, color and orientation. Shashua et al. [12,13] considered that not only image region has attention property but also the edge. Edge information is a basic feature of images since human eyes are sensitive to edge features for image perception. The related work [22,23] in this direction has proved the importance of edge information for salient region extraction. So, in the early visual feature processing stage, we extract edge information besides intensity, color, orientation. After extraction of early visual features, an important step is feature combination procedure. Itti and Koch [1] proposed four feature combination strategies: the "Naive", "N(·)", "Trained", "Iterative". Their combination strategy is based on the hypothesis that similar features

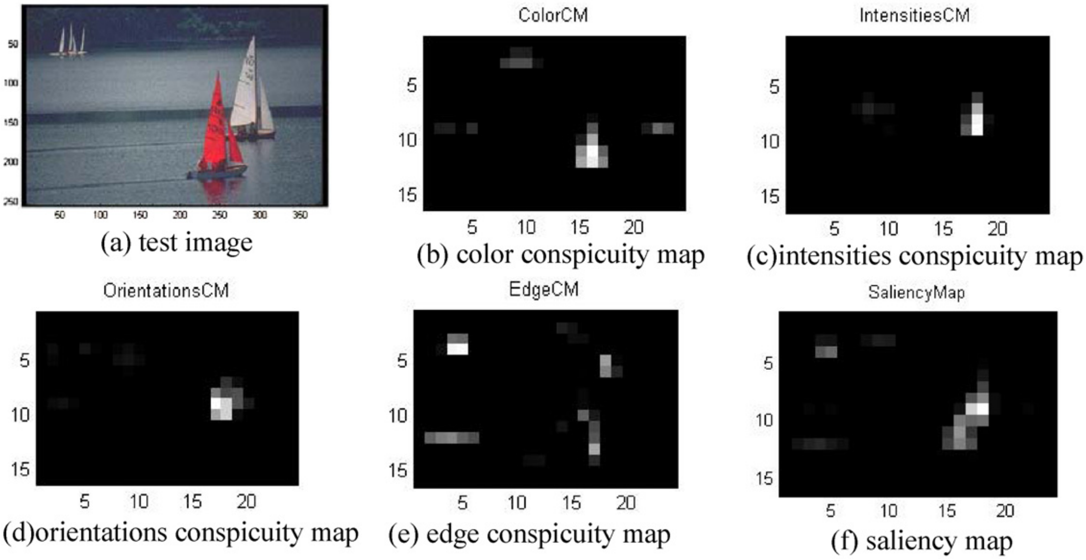
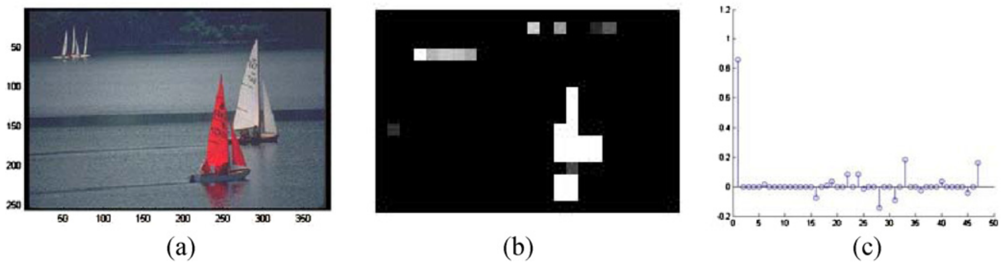


Fig. 2. Feature maps and saliency map combining with edge information.



(a) (b) (c)

Fig. 3. (a) Test image (b) Feature map f_1 of the test image (c) The sparse coefficients for representing the feature map f_1.

compete strongly for saliency, while different modalities contribute independently to the saliency map [6]. In deed, there is mounting psychophysical evidence that different types of features have different contribution to the salience, not but average contribution. Recently, the role of sparse in human perception and feature representation has been strongly supported by studies of human vision and image processing. Investigators have revealed that both in low-level and mid-level human vision, many neurons in the visual pathway are selective for a variety of specific stimuli. Inspired by the sparse property of human visual neurons, we define feature sparse indicator that measures feature's contribution to saliency map. Then we proposed a novel combination strategy, called sparse embedding feature combination strategy, which can automatically adjust the weights of each feature map in proportion of its contribution to the saliency map.

3.1. Extraction of edge feature

In this subsection, we obtain the edge feature from intensity pyramid image by using LOG edge detector. As a second-order operator, LOG edge detector can capture the acute change of the gradient direction. In the field of digital image processing, the LOG detector is often replaced by DOG (Difference of Gaussian) and described as below:

$$M_E = P_I(\sigma) * \mathrm{DOG}(\sigma_1, \sigma_2) \tag{12}$$

where * is the convolution operator and σ is the pyramid scale. $P_I(\sigma)$ is computed by using Eqs. (1) and (3). $\mathrm{DOG}(\sigma_1, \sigma_2)$ is defined as:

$$\mathrm{DOG}(\sigma_1, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp(-\frac{x^2+y^2}{2\sigma_1^2}) - \frac{1}{\sqrt{2\pi}\sigma_2}\exp(-\frac{x^2+y^2}{2\sigma_2^2}), \tag{13}$$

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

Combining the edge feature, the saliency map are redefined as

$$S_s = \sum_{i=1}^{4} w_{(i)}(f) N(M_i), \tag{14}$$

where $w_{(i)}(f)$ characterizes the feature's contribution to saliency map. The details of $w_{(i)}(f)$ will be described separately in Section

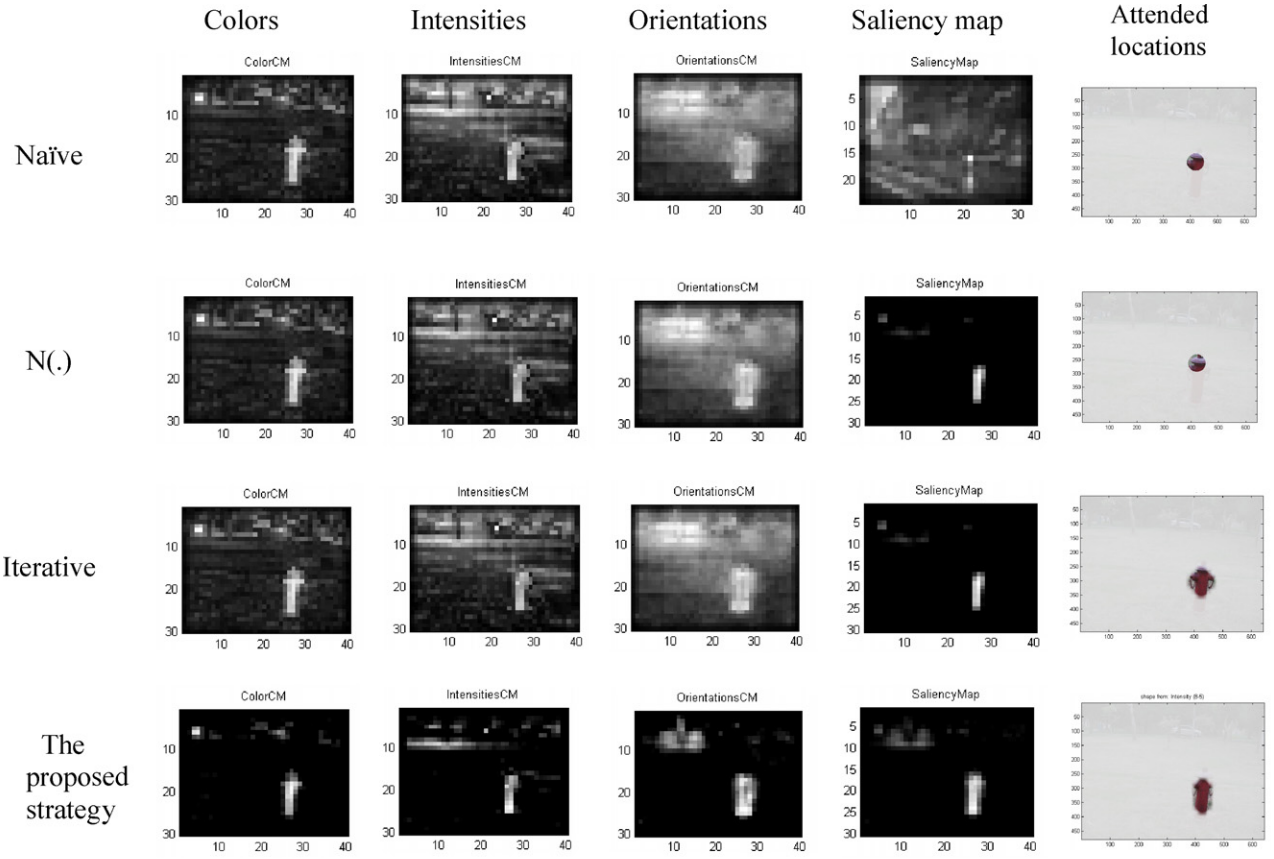


Fig. 4. Comparison with other feature combination strategies.

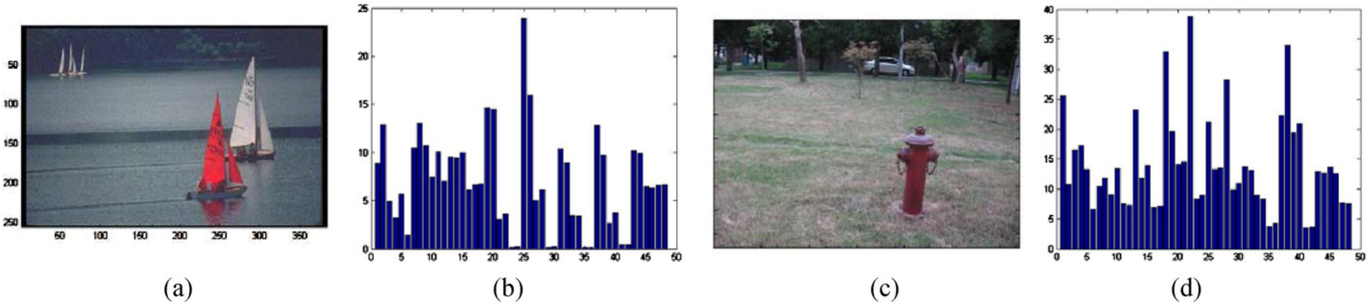


Fig. 5. (a) Test image. (b) Feature saliency indicator of test image (a). (c) Test image. (d) Feature saliency indicator of test image (b).

Table 1
User study result evaluation.

User	The proposed strategy Better	Itti's strategy Better	Both equally good
1	63.50%	9.00%	27.50%
2	61.10%	9.80%	29.10%
3	63.20%	8.80%	28.00%
4	65.20%	10.80%	24.00%
5	62.60%	10.30%	27.10%
6	65.80%	10.10%	24.10%
7	61.50%	8.40%	30.10%
8	60.70%	10.20%	29.10%
9	69.90%	7.00%	23.10%
10	66.80%	10.10%	23.10%
11	62.60%	8.30%	29.10%
12	68.90%	9.20%	21.90%
13	64.20%	11.00%	24.80%
14	62.50%	8.10%	29.40%
15	63.60%	9.40%	27.00%
16	62.90%	10.60%	26.50%
17	67.70%	8.20%	24.10%
18	70.30%	9.70%	20.00%
Average	64.61%	9.39%	26.00%

3.2 for convenience. An example of the four main features, including the edge feature, is showed in Fig. 2.

3.2. Sparse embedding feature combination strategy

Given an input image of the typical resolution, 640 × 480pixels, we extract four types of feature maps: intensity, color, orientation and edge. Subsequently, we can get $N(N = 6 \times 8)$ feature maps (F_l) at scale 4, of which the resolution is 40×30. Then we reshape them as column vectors $\{f_l\}_{l=1}^{N} \in R^M$ (where $M = 1200$). Let matrix $F = \{f_1, f_2, \ldots, f_N\}$ be the data matrix including all the feature maps in its column.

For the number of feature map, the dimension of f_l is very high, in order to avoid the singularity and accelerate the speed of obtaining the sparse solutions, dimension reduction should be performed on the feature maps. Researchers have developed many useful dimension reduction techniques. Principal Component Analysis (PCA) [24] has been a most popular linear dimensionality reduction technique in terms of the simplicity and effectiveness. Considering the algorithm complexity, we perform PCA to reduce the data dimension in our work. The projection from the raw data space to the subspace can be represented as matrix R^{pca}. Then sparse representation firstly seeks a sparse reconstructive weight vector s_l for each f_l through the modified l_1 minimization problem:

$$\begin{aligned} &\min \left\| s_l \right\|_1 \\ &\text{s.t.} f_l = R^{\text{pca}} F s_l, \end{aligned} \qquad (15)$$

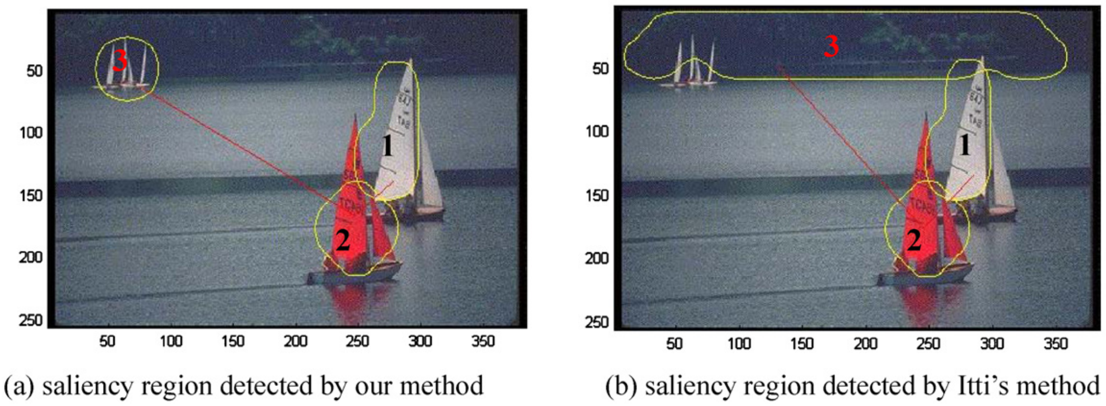
where $s_l = [s_{l,1}, \ldots, s_{l,l-1}, 0, s_{l,l+1}, \ldots, s_{l,N}]^T$ is a N-dimensional vector in which the lth element is equal to zero (implying that the x_l is removed from X), and the elements $s_{l,j} (l \neq j)$ denote the contribution of each x_j to reconstruct x_l.

For $f_l \in R^m$, let $s_l^* (l = 1, 2, \ldots, N)$ be the optimal solution of the above constrained optimization problem. The feature map f_l is sparsely represented by the remained N-1 feature maps. Specially, the sparse coefficients for representing f_1 are showed in Fig. 3.

To quantify the sparseness of each feature map, we define the following measure: feature sparse indicator.

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

$$\text{FSI}(f_l) = \text{residuals}(f_l) = \left\| f_l - F s_l^* \right\|_2, \qquad (16)$$



(a) saliency region detected by our method

(b) saliency region detected by Itti's method

Fig. 6. 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.

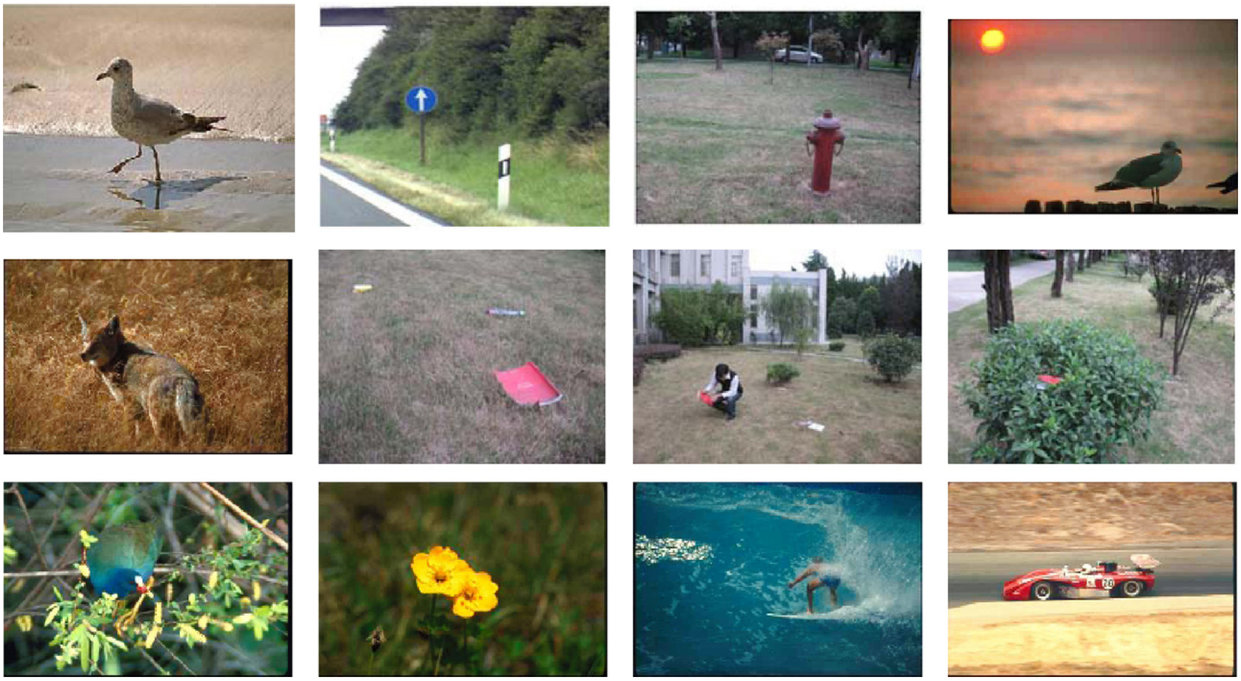


Fig. 7. Sample images of test set.

where FSI(f_l) stands for the saliency degree of the feature f_l. The greater the FIS(f_l) is, the more prominent the feature is.

To measure each type of feature's contribution to saliency map, we define the weight of saliency.

Definition 2. Weight of Saliency: The weight of each type of feature's contribution to saliency map is defined as:

$$w_{(i)}(f) = \frac{\mathrm{FSI}(f_l)}{\sum_{l=1}^{N} \mathrm{FSI}(f_l)} \tag{17}$$

We compared the proposed feature combination strategy with other strategies. The results are showed in Fig. 4. 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 captures 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.

3.3. The SEVASE algorithm

In summary of the preceding description, the following provides the complete sparse embedding visual attention system combined with edge information algorithm procedure:

Algorithm. Sparse embedding visual attention system combined with edge information (SEVASE)

```
Step 1: Extract $N$ feature maps ($F_l$) by Gaussian pyramids, Gabor pyramids and Center-surround differences.
Step 2: Convert each feature map to a column vector and form a feature matrix $F = \{f_1, f_2, \ldots, f_N\}$.
Step 3: Normalize the columns of $F$ to have unit 2-norm.
Step 4: Perform PCA on the feature matrix $F$ and obtain low-dimensional feature vectors $R^{\mathrm{pca}} * F$, where $R^{\mathrm{pca}}$ denote the transformation matrix of PCA.
Step 5: Compute the sparse reconstruction coefficients $s_i^*$
Step 6: Compute the FSI($f_l$) and $w_{(i)}(f)$
Step 7: Construct the saliency map: $S_s = \sum_{i=1}^{4} w_{(i)}(f) N(M_i)$
```

3.4. Complexity analysis

The complexity of visual attention computation model is dominated by two parts: early visual feature extraction and feature combination. Given an input image, the resolution is $n_1 \times n_2$ pixels. In the early visual feature extraction stage, the complexity of the proposed algorithm is $O(8n^2)$ and the complexity of Itti's algorithm is $O(7n^2)$, where $n = \max(n_1, n_2)$. $O\left(n^2\right)$ stands for the complexity Gaussian pyramid or Gabor pyramid. In the feature combination stage, the complexity of the proposed algorithm is $O(N^3)$, where N is the number of feature maps, which is set as 48 in the proposed method. The correspondent complexity of Itti's iterative strategy is $O(dn^2)$, where d is the number of iterative, which is set as 12 in the Itti's model. Therefore, the total complexity of the proposed algorithm is $O(8n^2 + N^3)$ and Itti's is $O((7 + d)n^2)$. Since $N \ll n$, the complexity of visual attention model is determined by the size of input image. Conclusively, the proposed algorithm is slightly more complex than Itti's.

4. Experiments

In this section, we have implemented the proposed algorithm using MATLAB 7.0 on a platform of Pentium 4 3.2 GHZ CPU and 1.5G memory. We conduct several experiments to illustrate the performance of the proposed algorithm. As an unsupervised approach, we compare the saliency location of the proposed algorithm with the

Itti's. In addition, we use 800 images selected from the standard Corel images dataset and the images taken on campus of NJUST as the test data set to evaluate the performance of the proposed algorithm. Experimental results illustrate the effectiveness of the proposed method.

4.1. Experiment 1

In this subsection, we use two natural color images to evaluate the proposed algorithm compared with the Itti's methods. In the early visual feature extraction stage, we decompose the input image into eight sets of distinct feature: a set of six intensity feature maps, a set of six edge feature maps, two sets of six color feature maps and four sets of six orientation feature maps. There are 48 feature maps totally. Then we compute the each feature sparse indicator (FSI) by using Eq. (16). The FSI are showed in Fig. 5(b) and (d). These results indicate that the f_{25} is most salient in the feature set of the test image (a). Feature maps such as f_{18}, f_{22}, f_{38} are significantly salient in the test image (c). We can get the saliency weights of each type of feature by using Eq. (17): The weights of test image (a) are 0.2863 (Color), 0.1450 (Intensity), 0.4308 (Orientation), 0.1380 (edge); the saliency weights of test image (c) are 0.2232 (Color), 0.1435 (Intensity), 0.5330 (Orientation), 0.1002 (edge). Subsequently four conspicuity maps are combined to the sparse saliency map S_s.

Finally, the saliency locations are detected and compared with Itti's. Details are illustrated in Fig. 6. 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 modeling the human visual attention system.

4.2. Experiment 2

In this subsection, we use 800 images selected from the standard Corel images dataset and the images taken on campus of NJUST as the test set to measure the performance of the proposed algorithm Fig. 7 shows several examples of the test image set.

Due to the subjectivity of human attention perception, there is not a standardized objective correctness measure for image attention analysis evaluation. So here a user study is conducted to evaluate the result of the experiments. Eighteen subjects are invited to each view any 150 of the 800 images. The subjects are asked if the saliency regions reflected the human visual attention region of the image for the proposed strategy as well as for the traditional strategies. Table 1 shows the result of user study. We can see that the proposed method outperforms the Itti's in 64.61% of the cases. About 26.00% of the responses suggest that both methods are equally good. However, 9.37% of the responses suggest that the output of Itti's combination is better.

5. Conclusions

In this paper, we have proposed a new computational model of visual attention, called sparse embedding visual attention systems combined with edge information. It can be seen as extension of our previous work [26] presented in the conference. The proposed strategy could provide complete and detailed procedures to the early visual feature processing problem. Compared with Itti's computational model of visual attention, the proposed model have two main contributions: (1) Extracting edge information besides color, intensity and orientation in the early visual features processing stage; (2) Putting forward a novel sparse embedding feature combination strategy. This strategy is based on a novel feature sparse indicator that measures the contribution of each map to saliency. Compared with existing feature combination strategies, it presents the more reasonable coefficients for feature fusion. This demonstrates the proposed strategy is an effective method to imitate human visual perception of saliency. We conduct several experiments to measure the performance of the proposed model. Results clearly show that the proposed method obtain more reasonable salient region for human visual system.

Acknowledgements

This work is partially supported by the FuJian Provincial Department of Science and Technology of China under grant No. JK2010046, JB10135. It is also partially supported by the National Science Foundation of China under grant No. 60472061, 61005005, 60973098, 60632050, 90820004 and Hi-Tech Research and Development Program of China under grant No. 2006AA04Z238. It is also partially supported by the Ministry of Industry and Information Technology of china under grant No. E0310/1112/JC01.

References

[1] Laurent Itti, Christof Koch. Feature combination strategies for saliency-based visual attention systems. Electronic Imaging 2001;10(1):161–9.

[2] Treisman AM, Gelade G. A feature-integration theory of attention. Cognitive Psychology 1980;12:97–136.

[3] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 1985;4:219–97, 916.

[4] Tsotsos JK, Culhance SM, Wai WYK, Lai YH, Davis N, Nuflo F. Modeling visual attention via selective tuning. Artifical Intelligence 1995;78:507–45.

[5] Mack M, Castelhano MS, Oliva JMA. What the visual system sees: the relationship between fixation positions and image properties during a search task in real-world scenes. In: Proceedings of Annual Object Perception Attention an Memory of Conference. 2003.

[6] Laurent Itti, Christof Koch, Niebur E. A model of saliency-based visual attention for rapid scene analysis. Pattern Analysis and Machine Intelligence 1998;20(11):1254–9.

[7] Li. S, Lee MC. An efficient spatiotemporal attention model and its application to shot matching. Circuits and Systems for Video Technology 2007;17(10):1383–7.

[8] Yu-Fei Ma, HongJiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the eleventh ACM international conference on Multimedia. 2003. p. 374–81.

[9] Liu HY, Jiang SQ, Huang QM. Region-based visual attention analysis with its application in image browsing on small displays. In: Proceedings of the fifteenth ACM international conference on Multimedia. 2007. p. 305–8.

[10] Aziz MZ, Mertsching B. Fast and robust generation of feature maps for region-based visual attention. Image Processing 2008;17(5):633–44.

[11] Meur OL, Callet PL, Dominique B, Dominique T. A coherent computational approach to model bottom-up visual attention. Pattern Analysis and Machine Intelligence 2006;28(5):803–17.

[12] Shashua A, Ullman S, Structural Saliency:. The detection of globally salient structures using a locally connected network. Pattern Analysis and Machine Intelligence 1988;17(1):90–4.

[13] Paul L, Rosin. Edges: saliency measures and automatic thresholding. Machine Vision and Applications 1997;9(4):139–59.

[14] Northdurft H. Salience from feature contrast: Additively across dimensions. Vision Research 1996;36:1115–25.

[15] Wright J, Yang A, Sastry S, Ma Y. Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence 2008;31(2):210–27.

[16] Hugli H, Bur A. Adaptive visual attention model. In: Proceeding of Image and Vision Computing. 2007. p. 233–7.

[17] Olshausen B, Field D. Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Research 1997;37:3311–25.

[18] Serre T. Learning a dictionary of shape-components in visual cortex: comparison with neurons, humans and machines, P.h.D. dissertation, MIT, 2006.

[19] Burt PJ, Adelson EH. The laplacian pyramid as a compact image code. IEEE Transactions on Communications 1983;31:532–40.

[20] Jones JP, Palmer LA. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. Neurophysiology 1987;58:1233–58.

[21] Greenspan H, Belongie S, Goodman R, Perona P, Rakshit S, Anderson CH. Overcomplete steerable pyramid filters and rotation invariance. In: CVPR 1994, IEEE Computer Vision and Pattern Recognition. 1994. p. 222–8.

[22] Shashua A, Ullman S. Structural saliency: the detection of globally salient structures using a locally connected network. IEEE Transactions on Pattern Analysis and Machine Intelligence 1988;7(1):90–4.

[23] Wang S, Kubota T, Siskind JM. Salient boundary detection using ratio contour. In: Neural information processing system conference (NIPS). 2003.

[24] Belhumeur PN, Hespanda JP, Kiregeman DJ. Eigenfaces versus fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 1997;19(7): 711–20.

[25] Laurent Itti, Models of bottom-up and top-down visual attention, dissertation (Ph.D.)(2000), California Institute of Technology Pasadena, California.

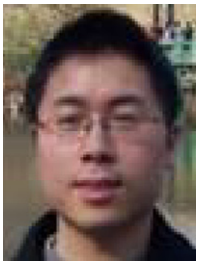
[26] Cairong Zhao, ChuanCai Liu, Zhihui Lai, Jingyu Yang. Sparse embedding visual attention systems combined with edge information. In: 20th International Conference of Pattern Recognition (ICPR). 2010.

Cairong Zhao received his B.S. degree in Electronics and Information Scientific Technology from Jilin University, China, in 2003 and M.S. degree in Electronic Circuit and System from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, respectively. He is currently pursuing his Ph.D. degree in Computer Science from Nanjing University of Science and Technology. His research interests include face recognition, building recognition and vision attention. E-mail:cairong.zhao@yahoo.com

Chuancai Liu is a full professor in the school of computer science and technology of Nanjing University of Science and Technology, China. He obtained his Ph.D. degree from the China Ship Research and Development Academy in 1997. His research interests include AI, pattern recognition and computer vision. He has published about 50 papers in international/national journals.

Zhihui Lai received the BS degree in Mathematics from South China Normal University and MS degree from Ji'nan University, China, in 2002 and 2007, respectively. He is currently pursuing the PhD degree in the School of Computer Science from Nanjing University of Science and Technology (NUST). His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research.

Huaming Rao received his B.S. degree in Computer Science from Nanjing University of Science and Technology, China, in 2009. He is currently pursuing his Ph.D. degree in Computer Science from Nanjing University of Science and Technology. His research interests include computer vision and pattern recognition. E-mail:huaming.rao@yahoo.com



Zuoyong Li received the B.S. degree in computer science and technology from Fuzhou University in 2002. He got his M.S. degree in computer science and technology from Fuzhou University in 2006. Now, Li is currently a lecturer in the department of computer science of Minjiang University, China. He received the Ph.D. degree from the School of Computer Science and Technology at Nanjing University of Science and Technology in 2010. He has published several papers in international/national journals. His current research interests include image segmentation and pattern recognition.