Bottom-Up Saliency Estimation Based on Redundancy Reduction and Global Contrast

Miao Xiaodong（缪小冬）[1], *Li Shunming*（李舜酩）[2*],
Shen Huan（沈峘）[2], *Li Aiting*（李爱婷）[3]

1. College of Mechanical and Power Engineering, Nanjing University of Technology, Nanjing, 210009, P. R. China;
2. College of Power and Energy, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, P. R. China;
3. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, P. R. China

(Received 7 November 2013; revised 11 March 2014; accepted 4 April 2014)

Abstract: A new algorithm for bottom-up saliency estimation is proposed. Based on the sparse coding model, a power spectral filter is proposed to eliminate the second-order residual correlation, which suppresses the global repeated items effectively. In addition, aiming at modeling the mechanism of the human retina prior response to high-contrast stimuli, the effect of color context is considered. Experiments on the three publicly available databases and some psychophysical images show that the proposed model is comparable with the state-of-the-art saliency models, which not only highlights the salient objects in a complex environment but also pops up them uniformly.

Key words: redundancy reduction; global contrast; saliency; bottom-up; sparse coding

CLC number: TP242.6 **Document code**: A **Article ID**: 1005-1120(2014)06-0660-08

1 Introduction

Human being receives about 80% information from vision perception. Even facing so much information, human visual system is still able to handle it quickly and efficiently. The reason lies in the fact that there is a discrepancy evaluation procedure in the early stages of human visual processing, in which some special signals (unanticipated or novel features) have been extracted and combined to salient region that occupies the limited visual resources. Since the saliency detection is broadly applied to target detection, image compression and image search, the problem of how to imitate the attention mechanism of human visual system and build effective calculation model has attracted much attention in the psychology and machine vision society[1].

The first systematic saliency calculation framework was proposed by Itti, et al[2]. The idea is first to extract some basic features such as brightness, color, orientation under different image resolution, and then to combine them together to obtain the final saliency map through a weighted center-surround processing. It is well known that the natural images obey a non-Gaussian distribution. But Itti used the difference of Gaussian as the center-surround filter, which cannot well remove the redundancy of non-Gaussian. The method was later modified by Harel, et al[3] and better results were achieved. Bruce, et al[4] proposed a framework combining sparse coding and information maximization with the inspiration of by Vinje′s neurophysiology experimental results[5], which suggested that simple cells of the visual cortex in response to external stimulation obeyed the sparse distribution. The limitation of the framework is only exploring the local contrast

Foundation items: Supported by the China Postdoctoral Science Foundation (2011M500917); the Jiangsu Innovation Program for Graduate Education (CXLX11_0180).

* **Corresponding author**: Li Shunming, Professor, E-mail: smli@nuaa.edu.cn.

information. However, the experiment indicates that human visual system also uses global contrast cue to guide attention[6]. Some saliency detection methods using global feature have been reported in recent years. For example, Hou, et al[7] found that the curve of image log spectrum presented certain statistical similarity which indicated the redundant information, while the difference corresponded to the novelty. Therefore, authors suggested that image spectrum redundancy could be used to express the visual saliency. However, whether the image spectrum redundancy can be used to express the visual saliency or not is still an open question. Guo, et al[8] argued that the phase spectrum information was more important to image saliency than spectrum, and proposed a detection model named as phase spectrum of Fourier transform and the experimental result was similar to Hou.

As a whole, although existing methods already get some successes in modeling visual saliency, there are still some limitations. The salient maps of these models are mostly concentrated on the image edges or on the positions of brightness drastic change.

To solve the above problems and obtain more perfect and uniform region of saliency object, we present an effective method. The overall structure is shown in Fig. 1. It contains two steps. The decorrelation step is based on sparse coding theory, which agrees with the human retina cells sparse response to outside stimulation, while the global contrast is used to calculate the contrast gain of the opponent features. The two steps are assembled to obtain final image saliency map.

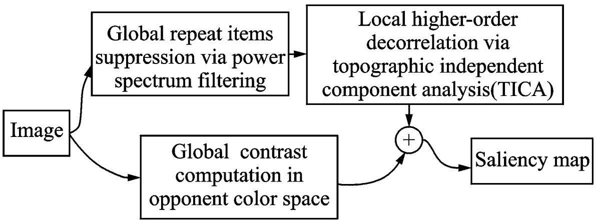

Global repeat items suppression via power spectrum filtering
Local higher-order decorrelation via topographic independent component analysis(TICA)
Image
Global contrast computation in opponent color space
+
Saliency map

Fig. 1 Overall framework of the proposed method

2 The Proposed Algorithm

From the information theory point of view, image $\boldsymbol{I}$ can be regarded as

$$\boldsymbol{I} = \boldsymbol{I}_{\text{Innovation}} + \boldsymbol{I}_{\text{Redundant}} \tag{1}$$

where $\boldsymbol{I}_{\text{Innovation}}$ corresponds to the novelty in the image, $\boldsymbol{I}_{\text{Redundant}}$ means the redundant part. In order to highlight the novel part in image, the human visual system can reduce the redundant information through sparse coding effectively. The classical sparse coding model is described as

$$x_i = \sum_{i=1}^{n} a_{ij} s_i \qquad i = 1, 2, \cdots, m \tag{2}$$

or expressed in a form of vector

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \tag{3}$$

where $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ is the observed data, $\boldsymbol{A}$ a mix matrix composed by $\{a_{ij}\}$, and $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)^{\mathrm{T}}$ the random coefficient, or called independent component. Research indicates that the mix matrix $\boldsymbol{A}$ can be applied to natural images. It has three important characteristics which are very similar to the simple cells receptive in the field of retina V1, i. e. , locality, directivity and band-pass[9]. Since neurons only have a strong response to the stimulation of the same characteristics in its receptive field, for a specifically stimulation, only a few neurons s_i can be activated simultaneously.

However, some issues will emerge when the sparse coding model is used to saliency detection directly. Firstly, the independence assumption cannot often be satisfied well. After natural images have been decomposed to independent components, there always exists weak second-order correlation between independent components[10]. In addition, the amplitude of independent component is normalized, thus the same patterns are given the same weight. An image contains a lot of repeat items, which form the spatial redundancy, but cannot be eliminated by the above models. Fortunately, the second-order correlation has some relationships with image power spectrum, and the periodicity in the spatial domain can be easily processing in frequency domain. Therefore, the above problems can be transformed to power spectrum filter in frequency domain which is easier solve. Given an image $f(x, y)$ with scale $M \times N$, the Fourier transform F and self-correlation function r can be expressed as

$$F(u,v)=\mathscr{F}[f(x,y)]=\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}f(x,y)\cdot \exp\left(-\mathrm{j}2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right)$$

$$r(x,y)=\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}f^{*}(x,y)f(m-x,n-y)\quad(4)$$

where u and v are horizontal and vertical frequency, respectively, $\mathscr{F}$ is Fourier transform, and $f^{*}(x,y)$ conjugate of $f(x,y)$. According to Wiener-Khinchin theorem, the power spectrum density P is equivalent to the Fourier transformation of self-correlation function r

$$P(u,v)=\mathscr{F}[r(x,y)]=F(u,v)F^{*}(u,v)=|F(u,v)|^{2}\quad(5)$$

This transformation demonstrates that the image power spectrum function describes the square of the amplitude of various periodic items of the whole image. The law is the more repeating patterns in one image, the bigger of corresponding frequency power spectrum magnitude. Therefore, we filter power spectrum in low-pass mode, and then inverse Fourier transform to get the image $f'(x,y)$ and to realize repeat items suppression

$$f'(x,y)=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\{\exp[g\cdot\log(P(u,v))]^{\frac{1}{2}}\}\exp\left[\mathrm{j}2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right]\quad(6)$$

where g is Gauss filter.

Although the above process can suppress the global repeated components effectively, there still exists high-order redundancy that cannot be ignored[10]. We use the sparse coding model to eliminate the residual high-order correlation in image $f'(x,y)$ as follows. From the natural images, we randomly select 10 000 image blocks (16 pixel ×16 pixel) as training samples, the TICA algorithm[9] is employed to calculate the basis functions of the image. Then we can get the hybrid matrix $\boldsymbol{A}$ in Eq. (3). The pseudo-inverse matrix $\boldsymbol{W}$ of $\boldsymbol{A}$ is equivalent to the receptive field in retina V1 and very sensitive to special spatial frequency. Through convolve operation of $\boldsymbol{W}$ and $f'(x,y)$, we can obtain the independent components of the image and eliminate the image redundancy.

The result for Fig. 2(a) by low-pass filter is shown in Fig. 2(b), and redundancy elimination is shown in Fig. 2(c). It is obvious that the coarse background in image is eliminated by repeatability inhibit process, and because of strong local correlation, the color uniform area in the image is eliminated by the sparse coding model. Only the contour profile with larger gradient gets higher saliency value. Actually, the salient value of pixel has a close relationship with its surround context. In order to get more uniform salient area and add the contribution of the color to saliency, we use color information to calculate the salient value based on the global contrast. Given two random points p_1 and p_2 in image $\boldsymbol{I}$, we define $\delta_{\text{Color}}(p_1, p_2)$ as the color difference measurement between p_1 and p_2, and $\delta_{\text{Position}}(p_1, p_2)$ as the spatial distance between p_1 and p_2. Therefore, any point p_k, the accumulated contribution of its surround pixels to it can be defined as

$$s(p_k)=\sum_{l\neq k\cap l=1}^{M\times N}\frac{\delta_{\text{Color}}(p_k,p_l)}{\delta_{\text{Position}}(p_k,p_l)}\quad(7)$$

Eq. (7) has considered the factor of spatial distance. The greater the spatial distance is, the lower the color difference contributionis, vice versa. Moreover, the color quantity that human eyes

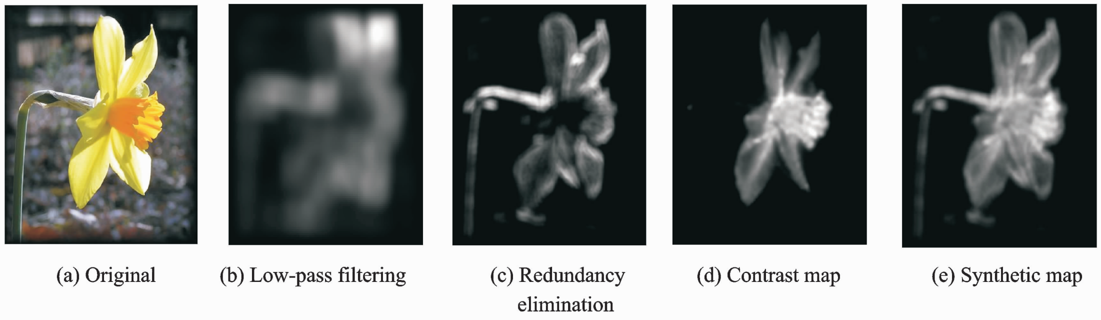

(a) Original (b) Low-pass filtering (c) Redundancy elimination (d) Contrast map (e) Synthetic map

Fig. 2 Results of processing

can distinguish is limited. Reflecting to a given image, its containing color quantity is limited. In addition, according to the different image content, the color quantity and pixel number of each color are also different. Therefore, we need to further consider the contribution of different color for the spatial contrast. We divide color into n categories according to the image hue in hue saturation intersity(HSI) color space, then Eq. (7) can be rewritten as

$$s'(p_k)=\frac{1}{N\times M-1}\sum_{l\neq k\cap l=1}^{M\times N}P[\pi(p_l)]\frac{\delta_{\mathrm{Color}}(p_k,p_l)}{\delta_{\mathrm{Position}}(p_k,p_l)} \tag{8}$$

where $\pi(\cdot)$ is color classification function, $P(\cdot)$ posterior probability of certain color.

Since the antagonistic color perceived by human eyes is similar to the description of $L^* a^* b^*$ space, we compute chromatic aberration in the $L^* a^* b^*$ space. However, when the image is converted to $L^* a^* b^*$ space from RGB space, we need to filter three channels separately by the method in Ref. [11] for preprocessing, whose effect as the spatial smoothing of the contrast sensitive function is similar to human vision system. After then, we get both color and space distance as

$$\delta_{\mathrm{Color}}(p_k,p_l)=\|\boldsymbol{v}_k-\boldsymbol{v}_l\|_2,\boldsymbol{v}_k=(L_k,a_k,b_k)^{\mathrm{T}}$$
$$\boldsymbol{v}_l=(L_l,a_l,b_l)^{\mathrm{T}}$$
$$\delta_{\mathrm{Position}}(p_k,p_l)=\|\boldsymbol{d}_k-\boldsymbol{d}_l\|_2,\boldsymbol{d}_k=(x_k,y_k)^{\mathrm{T}} \tag{9}$$
$$\boldsymbol{d}_l=(x_l,y_l)^{\mathrm{T}}$$

where x and y mean the spatial position coordinates of pixel, L, a, b are brightness, R-G and B-Y opponent color after filtering respectively.

The result of global contrast is shown in Fig. 2(c). Fig. 2(d) is the combination result of the decorrelation and global contrast. Fig. 2(e) the synthetic map by normalization. It is obvious that the object in image has been detected uniformly and completely.

3 Experiments and Analyses

We apply the proposed model on three public databases, DB1[4], DB2[12], and DB3[13]. Then we compare the proposed model with the classical IT[2] and two state-of-the-art models (i. e., GB[3], SR[7]). To make this a fair comparison among the four bottom-up based computational models, we equally add up all information channels and no special combination procedure is adopted.

3.1 Saliency object detection

Some results of saliency estimation on DB1[4] are illustrated in Fig. 3. In the figure, the rows from left to right are the original images, the saliency maps produced by IT[2], GB[3], SR[7], and the proposed model, respectively.

As shown in the first original image in Fig. 3(row(1)), the scene is with one salient object and simple background. All these tested approaches produce good saliency maps that highlight the building.

The backgrounds of in Fig. 3(row(3)) are very complex, as they are composed of several regions, including grass, bush, and a building. It is easy for the human vision system (HVS) to find out the salient objects in the two images since all of these backgrounds are with self-similar textures and less informative than the saliency objects. However, it is a challenge for the computational models. As can be seen, the IT, GB and SR models are seriously disturbed by the complex backgrounds and highlight some background regions. Since the background regions of the two images are with self-similar textures, these regions are very redundant and each pixel in them provides very little information. With redundancy reduction, the proposed model effectively estimates the saliency of the two images.

Furthermore, Fig. 3(row(4)) contains multiple saliency objects. Since the background of the image is very simple, all of the models can accurately highlight the multiple saliency regions. However, the content of Fig. 3(row(7)) is complex, which consists of a pan, a desk and albumen.

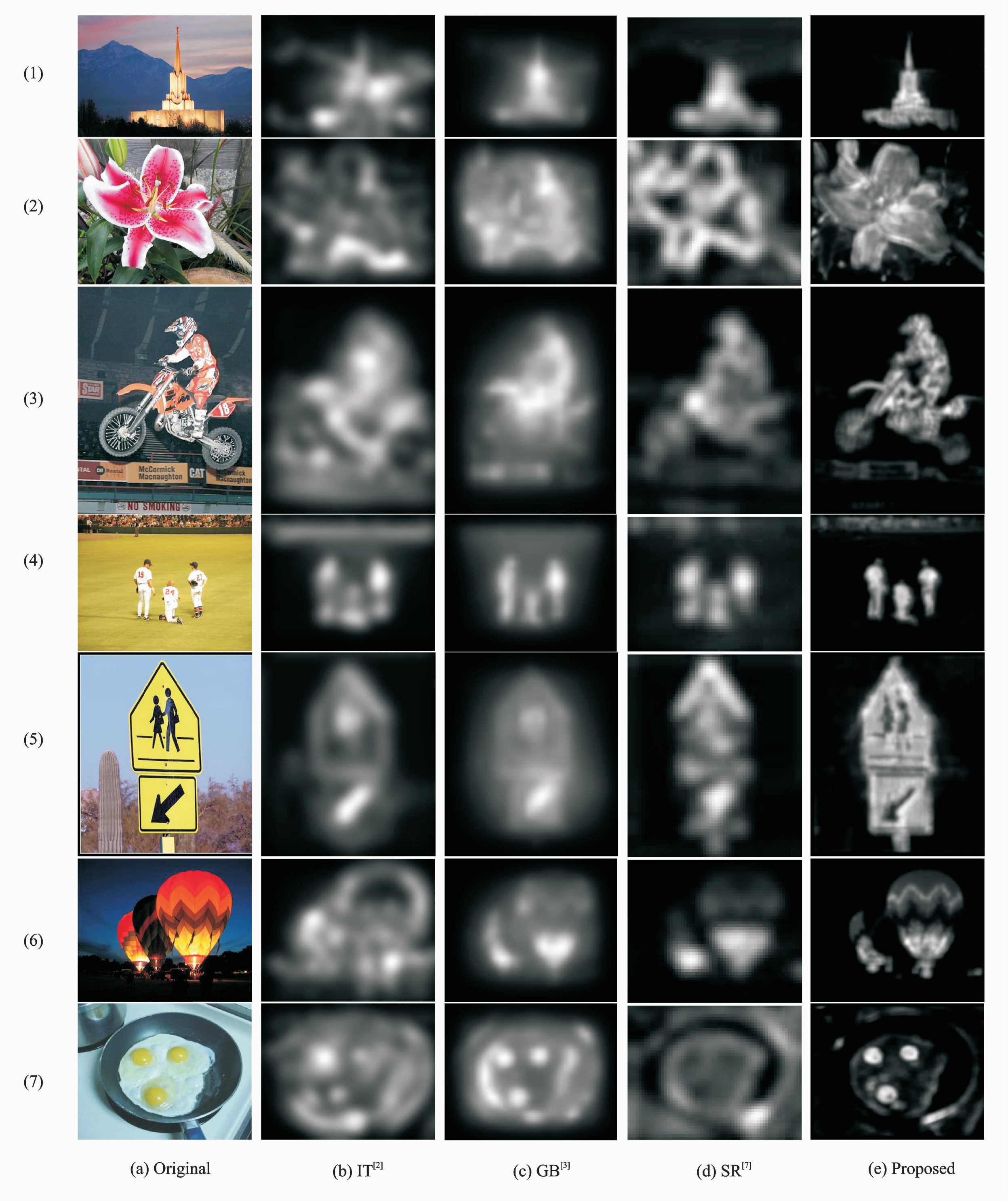


Fig. 3 Some examples from database

The IT model is totally failed in the image, which highlights most part of the desk. The GB model plays no better than the IT model, which also highlights almost all of the image. The SR model mainly highlights the edge of the pot which is not salient. Since the desk, the pan and the albumen have highly self-similar structures, the proposed model can effectively filter out these backgrounds and accurately highlight the yolk. Therefore, with the redundancy reduction and global contrast procedure, the redundancy from the image can be effectively removed. The proposed model can accurately find out saliency objects from both simple and complex backgrounds.

3.2 Overall performance

In order to make a comprehensive analysis, we verify the proposed model on three publicly available databases. These databases consist of a variety of indoor and outdoor scenes. The characteristics of the three databases are summarized in Table 1.

Table 1 Three publicly available databases for saliency estimation

Character	DB1[4]	DB2[12]	DB3[13]
Image number	120	1 000	1 003
Data achieve	Eye track	Human marked	Eye track
Ground truth	Gray map	Binary mask	Gray map

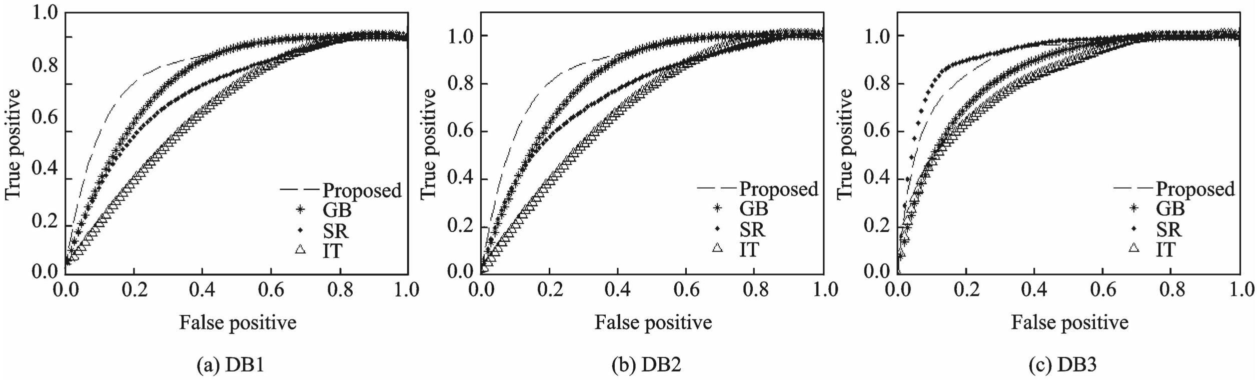
We compare the proposed model with four saliency estimation models and adopt the receiver operating characteristic (ROC) metric to assess their performances. The ROC metric measures the area under the ROC curve[14]. To calculate the measurement, the saliency map is treated as a binary classifier, where a pixel with a greater saliency value than a threshold is classified as fixation and the rest of the pixels as no fixated pixels. By varying the threshold, the ROC curve is acquired. The larger the area under the curve is, the better the saliency estimation method performs.

The ROC curves of these saliency models on the three public databases are shown in Fig. 4, and their corresponding ROC areas (A) are listed in Table 2.

Table 2 ROC areas of saliency models on three public databases

Algorithm	DB1[4]	DB2[12]	DB3[13]
IT[2]	0.713	0.814	0.714
GB[3]	0.823	0.840	0.780
SR[7]	0.767	0.919	0.600
Proposed	0.878	0.896	0.873

The best model has been highlighted in boldface for each database. As can be seen, the proposed model (with $A=0.878$) performs better than the other three bottom-up based computational models (IT with $A=0.713$, GB with $A=0.823$, and SR with $A=0.767$) on DB1. On DB2, as shown in Fig. 4(b), the proposed model (with $A=0.896$) outperforms IT model (with $A=0.814$) and GB model (with $A=0.840$), and approximates to SR (with $A=0.919$, the state-of-the-art performance on this database). And on DB3, the proposed model (with $A=0.873$) also outperforms the three bottom-up based computational models (IT with $A=0.714$, GB with $A=0.780$, and SR with $A=0.600$).



(a) DB1 (b) DB2 (c) DB3

Fig. 4 ROC curves of performance for these saliency models on three public databases

In summary, the proposed model is comparable with the state-of-the-art models and is highly consistent with the subjective visual attention.

3.3 Visual saliency of concept images

To further validate the effectiveness of the proposed model, we demonstrate our procedure on some concept images, whose saliency regions are merely determined by low-level features.

Fig. 5 shows three concept images with salience objects according to single factor. Figs. 5(a,

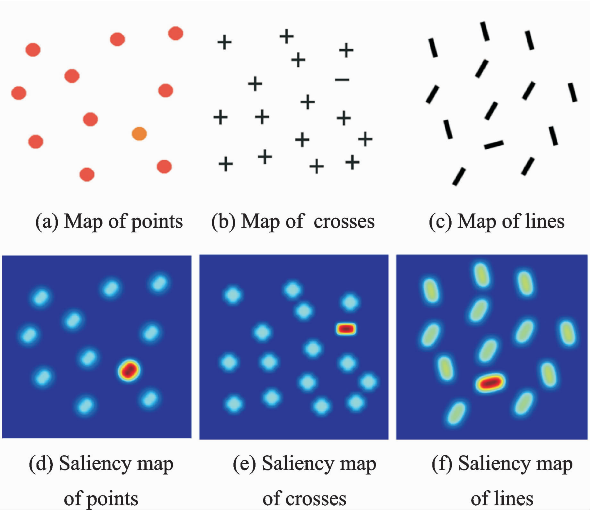

(a) Map of points
(b) Map of crosses
(c) Map of lines
(d) Saliency map of points
(e) Saliency map of crosses
(f) Saliency map of lines

Fig. 5 Visual saliency estimation on single factor

d) show an image with some colored points, and they are differing in intensity. The light point is unique and informative. While the other points are homogeneous, they are much more redundant than the lighted one. Therefore, the lighted colored point is with the highest saliency on the estimated map. Figs. 5(b,e) show salience objects under different shapes. Since the sign "—" is distinct among signs "+", the sign "—" is less redundant than that "+" in this image, and this is also well located in its corresponding saliency map. Figs. 5(c,d) show a salient case caused by the change of object orientation. There are three orientations of the objects, two orientations are with multiple objects, and the third is with a single object. As the unique object contains much information, it is the most salient as shown by its saliency map. Other objects are with nearly same saliency because the first two orientations are with six objects.

Besides these decoupled factor cases, we also test our approach with coupled factors. As Fig. 6 shows, the objects are with two colors and two orientations, and these coupled factors affect the saliency of the image. In addition, in the original image, ignoring the color, the two orientations have the same number of objects. The saliency map performs well, in which the green object is with the highest saliency, following are the red objects with the orientation same to the green one, and the lowest one is other red objects.

Fig. 7 shows a complex case whose saliency is affected by some coupled factors. It is challenging for most existing bottom-up saliency models. Objects in the original images are with different colors, sizes, orientation, and even a mirrored object. Applying the proposed approach, all these factors are successfully detected. As the saliency map shows, the most salient object is the one with different orientation, and the objects with saliency on size, color, and mirrored, also pop out. Though the orange object is distinct in color, it has the same shape to most of the objects in the image. Since these objects provide shape information jointly, the redundant coefficient is large and each one shares a very little part of shape information. Therefore, the orange one is not so salient. Meanwhile, the differences in orientation and size bring new information, and these objects gain high saliency values. As the mirrored object shares the same color and three horizontal lines to most objects, it is hard to be detected at the early stage of the visual perception.

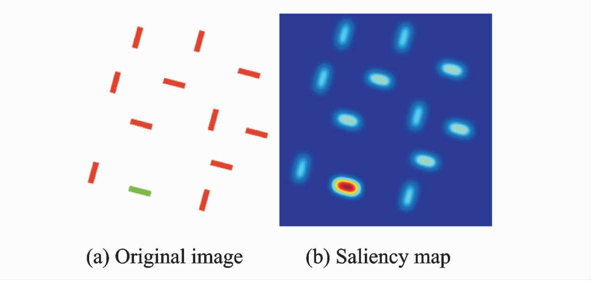


Fig. 6 Visual saliency estimation with two independent factors

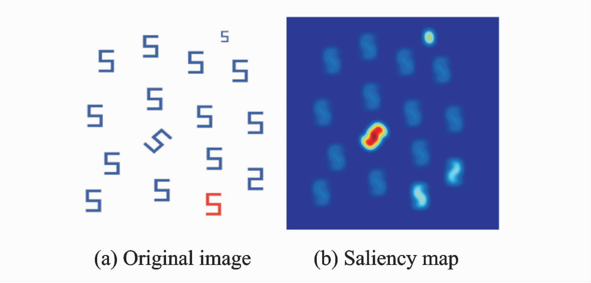


Fig. 7 Visual saliency estimation on multiple factors

4 Conclusions

A visual saliency estimation model is proposed based on redundancy reduction and global contrast, independent from the prior knowledge, which is purely data driven. The model focuses on the reduction of spatial structural redundancy in images, and considers the relationship between local and global. According to the sparse coding theory, a power spectral filter is designed to eliminate the second-order residual correlation, which suppresses the global repeated items effectively. In addition, aiming at modeling the mechanism of the human retina prior response to high-contrast stimuli, the effect of color context is considered.

Experiments on the three publicly available databases show that the proposed model is comparable with the state-of-the-art saliency models, and performs better, which is useful in image communication robort vision, code, and so on.

References:

[1] Alexander T. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(11): 2131-2146.

[2] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998,20(11): 1254-1259.

[3] Harel J, Koch C, Perona P. Graph-based visual saliency[C]// Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: Advances in Neural Information Processing Systems, 2006: 545-552.

[4] Bruce N D B, Tsotsos J K. Saliency, attention, and visual search: An information theoretic approach[J]. Journal of Vision, 2009, 9(3): 1-24.

[5] Vinje W E, Gallant J L. Sparse coding and decorrelation in primary visual cortex during natural vision [J]. Science, 2000, 287(5456): 1273-1276.

[6] Chen L. The topological approach to perceptual organization[J]. Visual Cognition, 2005,12(4): 553-627.

[7] Hou Xiaodi, Zhang Liqing. Saliency detection: A spectral residual approach[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA:IEEE, 2007:1-8.

[8] Guo Chenlei, Zhang Liming. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression[J]. IEEE Transactions on Image Processing, 2010, 19(1): 185-189.

[9] Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. Nature, 1996, 381(6583): 607-609.

[10] Hyvärinen A, Hoyer P O, Inki M. Topographic independent component analysis[J]. Neural Computation, 2001, 13(7): 1527-1558.

[11] Johnson G M, Fairchild M D. A top down description of S-CIELAB and CIEDE2000[J]. Color Research and Application, 2003, 28(6): 425-435.

[12] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection[C]//IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009). Miami, FL: IEEE, 2009: 1597-1604.

[13] Judd T, Ehinger K, Durand F, et al. Learning to predict where humans look[C]// IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009: 2106-2113.

[14] Wu Jinjian, Qi Fei, Shi Guangming, et al. Non-local spatial redundancy reduction for bottom-up saliency estimation[J]. Journal of Visual Communication and Image Representation, 2012, 23(7): 1158-1166.

(Executive editor: Xu Chengting)