Multimed Tools Appl (2016) 75:2667–2679
DOI 10.1007/s11042-015-2570-0

Hierarchical salient object detection model using contrast-based saliency and color spatial distribution

Xin Xu[1,2] · Nan Mu[1,2] · Li Chen[1,2] · Xiaolong Zhang[1,2]

Received: 30 October 2014 / Revised: 10 February 2015 / Accepted: 18 March 2015 /
Published online: 10 April 2015

Abstract Visual saliency is an important cue in human visual system to detect salient objects in natural scenes. It has attracted a lot of research focus in computer vision, and has been widely used in many applications including image retrieval, object recognition, image segmentation, and etc. However, the accuracy of salient object detection model remains a challenge. Accordingly, a hierarchical salient object detection model is presented in this paper. In order to accurately interpret object saliency in image, we propose to investigate distinctive features from a global perspective. Image contrast and color distribution are calculated to generate saliency maps respectively, which are then fused using the principal component analysis. Compared with state-of-the-art models, the proposed model can accurately detect the salient object which conform with the human visual principle. The experimental results from the MSRA database validate the effectiveness of our proposed model.

Keywords Hierarchical model · Salient object detection · Contrast measure · Color distribution

1 Introduction

Visual saliency reflects how much an image region or object stands out from its surrounding. Generally, it can be defined as what captures human perceptual attention. Salient object

✉ Xin Xu
xuxin0336@gmail.com

[1] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan, China

detection aims to build a saliency map for natural scene, and has a wide range of applications. For example, the detecting of salient object can be effectively used to automatically zoom the "interesting" areas [5] or automatically crop the "important" areas in an image [16]. Object recognition algorithms can use the results of saliency detection to quickly locate the position of visual salient objects. Salient object detection can also be used to reduce the interference of cluttered background to further improve the performance of image segmentation algorithm and image retrieval system [19].

A number of computational models for salient object detection have been developed in recent years. Typical model is the one proposed by Itti et al. [10], which depends only on low-level image features to determine salient object in image. This model is based on the Treisman theory [20] and the Koch neurobiological framework [12]. Following this model, many state-of-the-art saliency detection methods focus on low level features. Among these models, image color and contrast have been utilized as the most important and effective features to detect salient object [14, 17].

For example, the computational models proposed in [3, 13] and [6] built saliency map using image segmentation algorithms. These models divided the image into blocks according to the homogeneity of contrast and color features. However, the segmentation preprocessing is time consuming. Duan et al. [7] performed the uniform sampling on the RGB color feature maps without pre-running the complex segmentation algorithm. They used the principal component analysis (PCA) to extract the effective features of the image blocks. The image saliency can be calculated according to the global contrast between the image blocks and their spatial location. Borji et al. [4] put forward a prediction model to reflect the saliency discrimination towards eye tracking data. This model measured the scarcity of each block in both RGB and LAB color space, and then combined the local and global saliency of each color space to generate the saliency map. Achanta et al. [1, 2] used the luminance and color features to detect salient object. They calculated the contrast between local image region and its surrounding. The saliency map can be obtained by calculating the average color vector difference. Vu et al. [21] exploited image contrast in both spectral and spatial domain to measure the local perceived sharpness. The resulting sharpness maps generated by this model can well represent the saliency of the input images.

As stated above, many state-of-the-art salient object detection models compute image saliency primarily by measuring the contrast and color features of image. However, these models fail to fully consider the inter-relationship between the salient objects and the background from a global perspective. When the salient object and the background are similar in color and contrast, these models may fail to reflect image saliency. Aiming to address this problem, this paper proposes to fuse contrast based saliency with global color distribution. Because the global color distribution can compensate for the loss of saliency information caused by the similarity in color and contrast. Then, PCA was used to fuse the color saliency map and the contrast saliency map to retain the most significant information of object.

The remainder of this paper are organized as follows. Section 2 presents the proposed hierarchical salient object detection model. Experimental results obtained from the MSRA database are presented in Section 3. The conclusions are given in Section 4.

2 Proposed hierarchical salient object detection model

The visual properties of an object tend to be various, which will inevitably lead to its image features diversity. In order to fully and accurately describe an object, a variety of image features are chosen according to the visual properties of all aspects of the object. By simulating the process of people's attention, it can be found that the image color and contrast play the key role. Therefore, we extracts these two critical features to detect image saliency.

2.1 Contrast based saliency

The conventional contrast measure [21] only considers the local characteristics of the image, while the global characteristics are neglected. Aiming to address this problem, we propose to extract image contrast by exploiting both local and global characteristics of the image region.

The proposed approach converts each image block to the spectral-domain. The Euclidean distance between the mean amplitude value $m(b)$ of each image block b and the mean spectral value $m(I)$ of the whole image I is used to obtain a new spectrum value $F_1(b)$.

$$F_1(b) = \begin{cases} |m(b) - m(I)| & \text{if } \triangle\,(b) \leq T1 \text{ or } \mu(b) \leq T2 \\ 0 & \text{else} \end{cases}$$

where

$$\triangle\,(b) = max(L(b)) - min(L(b)) \tag{1}$$

$$\mu(b) = mean(L(b)) \tag{2}$$

where $\triangle(b)$ represents the difference between the maximum luminance value and the minimum luminance value, and $\mu(b)$ represents the average luminance value, $L(b) = (h + kb)^\eta$ denotes the luminance-valued block, $h = 0.7656$, $k = 0.0364$, and $\eta = 2.2$ is the Adobe RGB color space display conditions. For the thresholds T_1 and T_2, we can empirically set $T_1 = 5$ and $T_2 = 2$, assuming that image pixel values are between 0 and 255. By this means the low contrast image blocks are set to zero. Then the saliency value $F_1(b)$ of all the image blocks are combined to form the contrast saliency map.

The resulting contrast saliency map can well represent the image saliency. However, it may face difficulties when the contrast difference between the salient object and the background is quite small. In order to extract image saliency, we perform the Fourier transformation to store both saliency information and non-saliency information in the forms of statistics. The Fourier spectrum (denoted as $F(u, v)$) of the image can be decomposed into amplitude spectrum and phase spectrum which can be expressed as:

$$F(u, v) = |A(x, y)|e^{-jP(x,y)} \tag{3}$$

where $A(x, y)$ and $P(x, y)$ represent the amplitude spectrum and phase spectrum respectively, and can be given by:

$$A(x, y) = \sqrt{R(x, y)^2 + I(x, y)^2} \tag{4}$$

$$P(x, y) = arctan\left(\frac{I(x, y)}{R(x, y)}\right) \tag{5}$$

where $R(x, y)$ is the real component of Fourier spectrum and $I(x, y)$ is the imaginary component.

As can be seen from the above formulas, the image signals can be characterized as the combination of amplitudes and the sine wave of phases. For the phase spectrum, it contains the texture information of the original image and can preserve important characteristics of the signal; while the amplitude spectrum can contain the chiaroscuro information of the original image.

Because the amplitude spectrum represents the signal magnitude and the specific gravity of sinusoidal components; while the phase spectrum information represents the location of these sinusoidal components which are the important parts of saliency information. As a result, the phase spectrum should remain intact when constructing the Fourier spectrum to largely preserve the integrity of the signal.

Then we analyze the amplitude spectrum of the image. The image can be viewed as the accumulation of object shadows in a homogeneous background. After the Fourier transformation, the projection information is broken down into a series of weighted sum of the plural fundamental waves. The fundamental waves of common characteristics (non-saliency features) in the original image usually account for a large proportion, while the fundamental waves of the fresh characteristics (saliency features) only account for a small proportion. Thus, we can inhibit the common characteristics and enhance the fresh characteristics of the original image by adjusting the amplitude spectrum.

The amplitude spectrum is the weighted sum of fundamental waves for various features. The larger weighted amplitude representing non-saliency features in the amplitude spectrum should be inhibited; while the smaller weighted amplitude representing saliency features in the amplitude spectrum should be enhanced. Accordingly, we can adjust the amplitude spectrum in order to inhibit the non-saliency information while retaining the saliency information at maximum extent. Therefore, we propose to adjust the amplitude spectrum A(x, y) via:

$$A(x, y) = |A(x, y) - mean(A(x, y))| \tag{6}$$

For the image block B of scale r, the proposed method is further performed in a multi-scale strategy with three scales: R = r, $\frac{1}{2}$r, $\frac{1}{4}$r to improve the contrast between the salient region and the non-salient region. The multi-scale processing is more accurate for locating the salient object because it can well suppress the features which occur frequently, and retain the features that deviate from the norm at the same time. Our multi-scale processing in the spectral-domain will not make significant increase in calculation, but will make significant contribution to the accuracy to detect the salient object. The resulting contrast maps using our method and [21] are shown in Fig. 1.

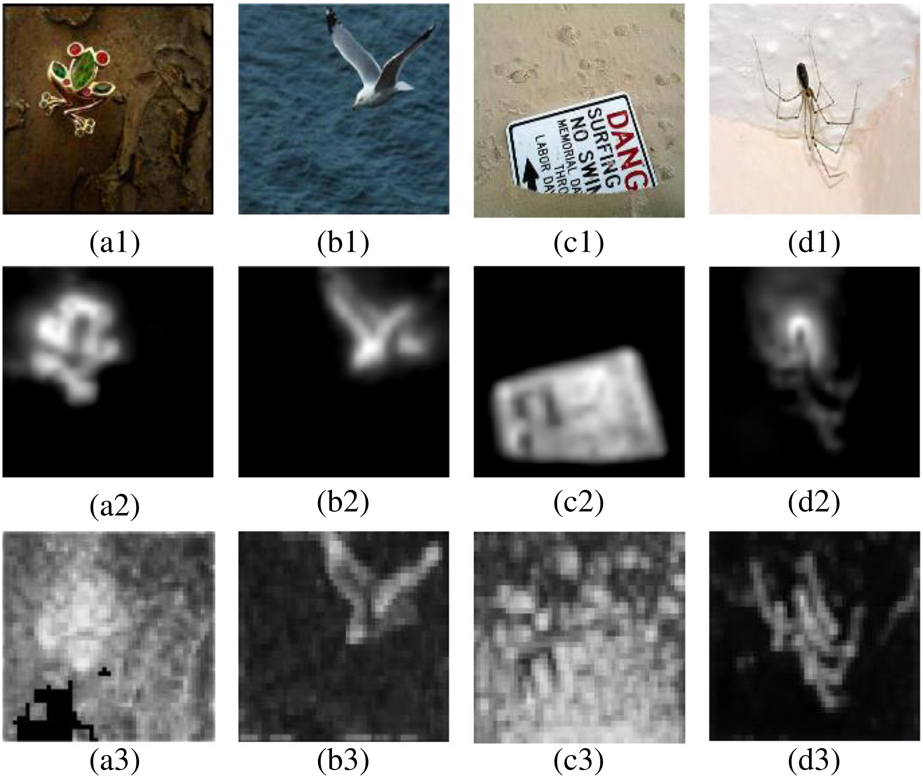
2.2 Color spatial distribution

The salient region is the area with strong contrast or variation. Accordingly, the color value in the salient region is relatively far from the average color value of the whole image; while the background region is the smooth area where the color value is closed to the average color value. Therefore, the salient object can be extracted by calculating the Euclidean distance of the average color value between each image block and the background region, in which the difference of the average color value is greater.

The LAB color space can provide a representation of color that corresponds to how human perceive chromatic differences. Thus, we convert the color space to a uniform LAB color space, and then divide the image (denoted as I) into blocks (denoted as b) with %50 overlap. Then the color saliency value $F_2(b)$ of all image blocks are combined to form the color saliency map. The color feature of each image block can be computed via:

$$F_2(b) = \sqrt{\left(\overline{A}(b) - \overline{A}(I)\right)^2 + \left(\overline{B}(b) - \overline{B}(I)\right)^2} \tag{7}$$

where $\overline{A}$ and $\overline{B}$ represent the average A and B color component in LAB color space, respectively.



(a1) (b1) (c1) (d1)

(a2) (b2) (c2) (d2)

(a3) (b3) (c3) (d3)

Fig. 1 **a1–a4** Input images, **b1–b4** contrast saliency maps obtained by the proposed method, **c1-c4** contrast saliency maps obtained by [21]

2.3 Center prior

When humans watch a picture, they will naturally gaze on the objects next to the center of it [11]. That is because the photographer usually centers the object of interest. As a result, human fixation will therefore unconsciously start at the central part of image. Thus, in order to obtain the salient objects conform with the human visual principle, more weights need to be added to the center of image region. Here we use a feature (denoted as $f_c(b)$) to indicate the distance between each image block and image center. For each image block, the contrast and color saliency value $F_n(b), n = 1, 2$ can be recalculated via:

$$F_n^*(b) = F_n(b) f_c(b) \tag{8}$$

$$f_c(b) = 1 - \frac{\sqrt{(r(b) - C^*)^2 + (c(b) - R^*)^2}}{\sqrt{(M/2)^2 + (N/2)^2}} \tag{9}$$

where $r(b)$ and $c(b)$ represent the upper-left coordinate of image block b, C^* and R^* represent the center coordinate of the estimated image region, M and N denote the width and height of the image region, respectively. The color and contrast saliency maps after center prior are shown in Fig. 2.

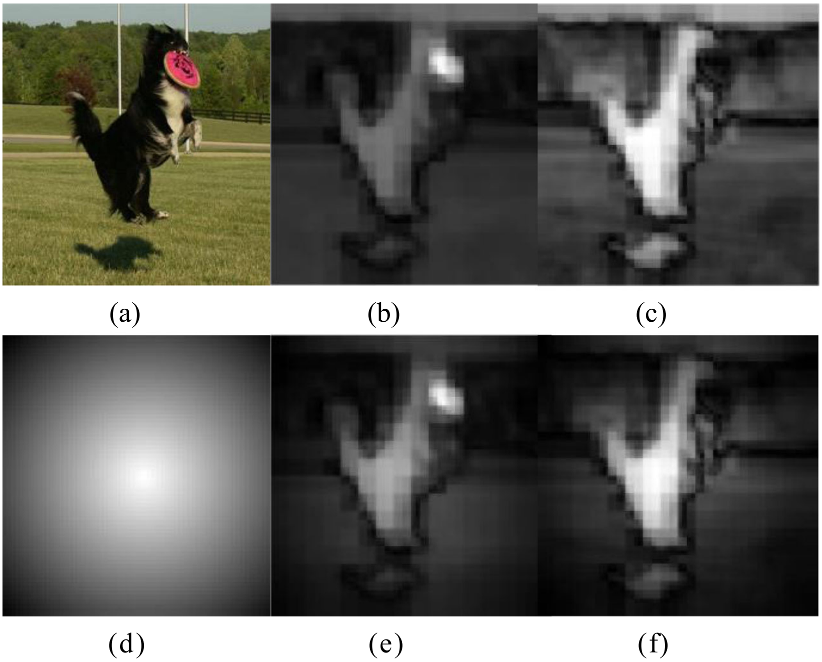


Fig. 2 Resulting feature maps. **a** Input image, **b** color saliency map, **c** contrast saliency map, **d** center weight map, **e** color saliency map after center prior, **f** contrast saliency map after center prior

Finally, the feature map $F_n(I)$ of the whole image I is generated by combined the image feature $F_n^*(b)$ of all the blocks. And the normalized feature map can be calculated by:

$$\overline{F}_n(I) = \frac{F_n(I) - min(F_n(I))}{max(F_n(I) - min(F_n(I)))} \tag{10}$$

The extracted feature maps are shown as follows:

2.4 PCA based feature fusion

Since PCA can remove the redundant information, and can well combine the most important information of two different components. Thus, we uses PCA to fuse the color feature map and contrast feature map. The proposed hierarchical salient object detection model is shown in Fig. 3.

Firstly, we construct the color feature map $F_1(I)$ and contrast feature map $F_2(I)$ into data matrixes (denoted as M_n, n $= 1, 2$), respectively. Next, calculating the covariance matrix (denoted as C) of the two data matrixes M_1 and M_2. Then, computing the eigenvalues (denoted as λ_1 and λ_2) and the corresponding eigenvectors (denoted as ξ_i and ζ_i, i $= 1, 2$) of the covariancematrix C. And then, determining the weighting coefficients (denoted as ω_i, i $= 1, 2$) via:

$$\omega_i = \begin{cases} \xi_i / \sum_{i=1}^{2} \xi_i & \text{if } \lambda_1 > \lambda_2 \\ \zeta_i / \sum_{i=1}^{2} \zeta_i & \text{else} \end{cases}$$

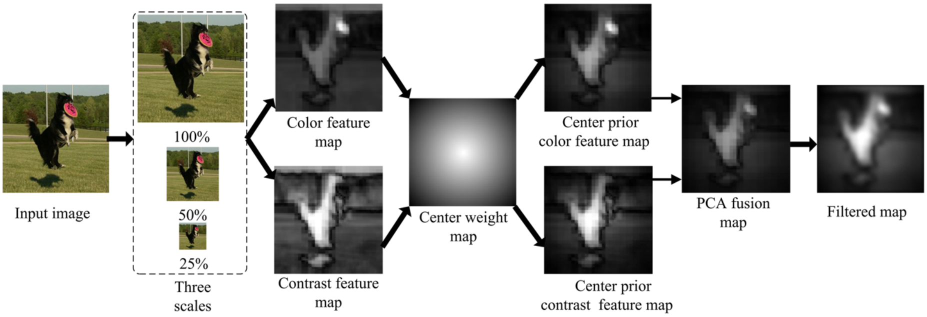


Fig. 3 The hierarchical salient object detection model

Finally, the resulting saliency map (denoted as $F(I)$) can be calculated by:

$$F(I) = \sum_{i=1}^{2} \omega_i M_i \tag{11}$$

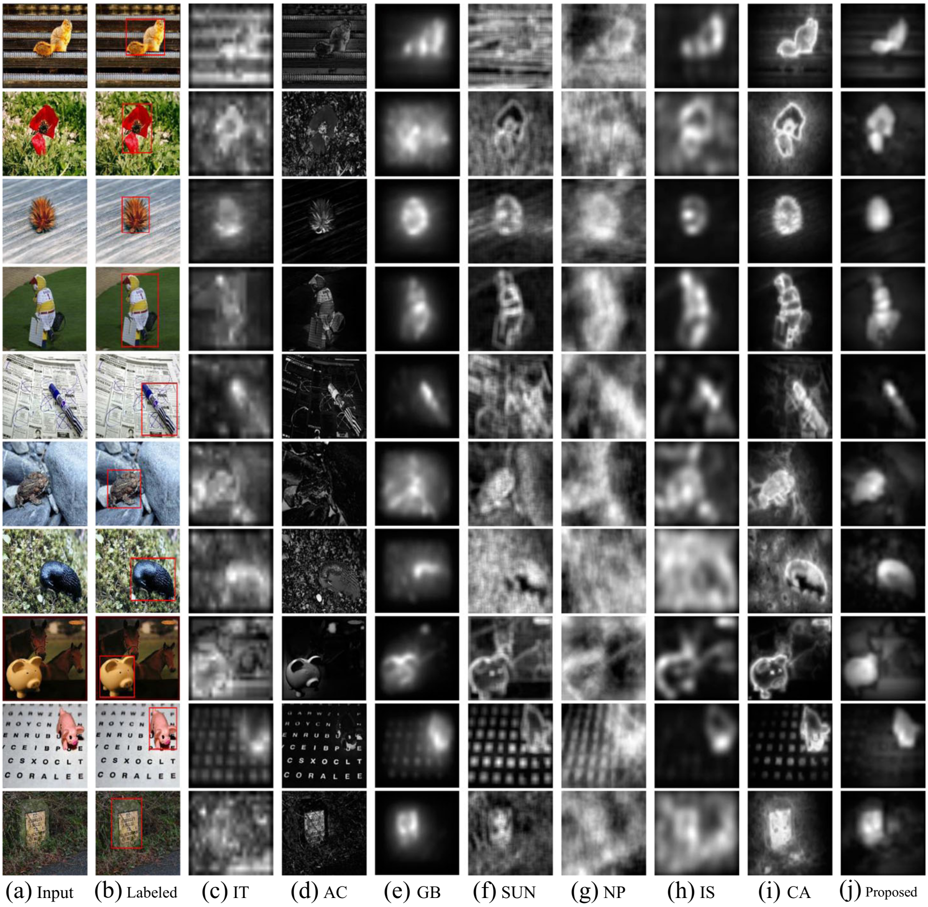
In order to achieve a preferable visual effect, the proposed method is performed in three scales: %100, %50, %25, which can better suppress the background information. Ultimately, the generated fusion map is smoothed by a Gaussian filter (the template size is $10 * 10$, and σ is 2.5). The final filtered map is shown as follows:

3 Experimental results

The experiment was implemented on the MSRA database [13] which includes two parts: (i) Image set A, contains 20,000 images and the principle salient objects are labeled by three users, and (ii) Image set B, contains 5,000 images and the principle salient objects are labeled by nine users. The images used in the experiments are representative for a certain class of image types. Some of these images show fairly typical conditions, such as objects located at different image locations. Others are more intricate, containing complex scene, such as low-contrast object, clutter backgrounds, and low lighting condition.

The proposed model (FCC) was compared with the other seven state-of-the-art models, including Itti's model (IT) [10], salient region detection and segmentation (AC) model [1], graph-based visual saliency (GB) model [8], saliency using natural statistics (SUN) model [22], non-parametric (NP) model [15], image signature (IS) model [9], and context-aware (CA) model [18]. The performance of these salient object detection models is shown in Fig. 4.

As illustrated in Fig. 4, the AC [1] and SUN [22] models fail to detect the salient object from the complex background. The saliency maps generated by IT [10] and NP [15] models appear rather blurry, and are difficult to clearly identify the salient object. The GB model [8] can not prominently reflect the outline of the salient objects, and face difficulty in detecting the texture images. The IS model [9] can only obtain the low resolution results, and it can hardly distinguish the saliency information from the non-saliency information. The CA model [18] has pretty good performance; however the highlighted salient region contains too much background information. The saliency maps obtained by our proposed model have a uniform salient region similar to the labeled rectangle, and can achieve good performance in complex background.



(a) Input (b) Labeled (c) IT (d) AC (e) GB (f) SUN (g) NP (h) IS (i) CA (j) Proposed

Fig. 4 Saliency maps obtained from different saliency computational models. **a** testing images, **b** ground-truth labeled rectangles, **c–i** saliency maps obtained by the seven state-of-the-art models, respectively, **j** saliency maps obtained by the proposed method

Given the generated saliency map S(x,y), we set a threshold T to segment the salient objects.

$$F(x, y) = \begin{cases} 1 & S(x, y) > T \\ 0 & S(x, y) \leq T \end{cases}$$

$$T = \alpha * \frac{1}{M * N} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} S(x, y) \tag{12}$$

where N and M are the height and width of the saliency map, respectively. The parameter α is empirically set to 1.7 to achieve higher registration probability.

Objective performance evaluation is conducted by calculating the True Positive Rate (TPR) and the False Positive Rate (FPR). For the obtained saliency map $F(x, y)$, a threshold

Table 1 The TPR and FPR results obtained by different salient object detection approaches

	IT	AC	GB	SUN	NP	IS	CA	FCC
TPR (t = 0.2)	0.9969	0.5357	0.9531	0.9387	1.0000	0.9036	0.9382	0.9371
FPR (t = 0.2)	0.5957	0.1229	0.2181	0.5251	0.8095	0.2845	0.2827	0.1380
TPR (t = 0.5)	0.6772	0.1211	0.5839	0.5561	0.9099	0.5654	0.7207	0.6046
FPR (t = 0.5)	0.1039	0.0228	0.0304	0.1307	0.3263	0.0447	0.0474	0.0049
TPR (t = 0.7)	0.2435	0.0324	0.2829	0.2208	0.5907	0.2645	0.4768	0.3215
FPR (t = 0.7)	0.0186	0.0076	0.0059	0.0156	0.0840	0.0106	0.0158	0.0010

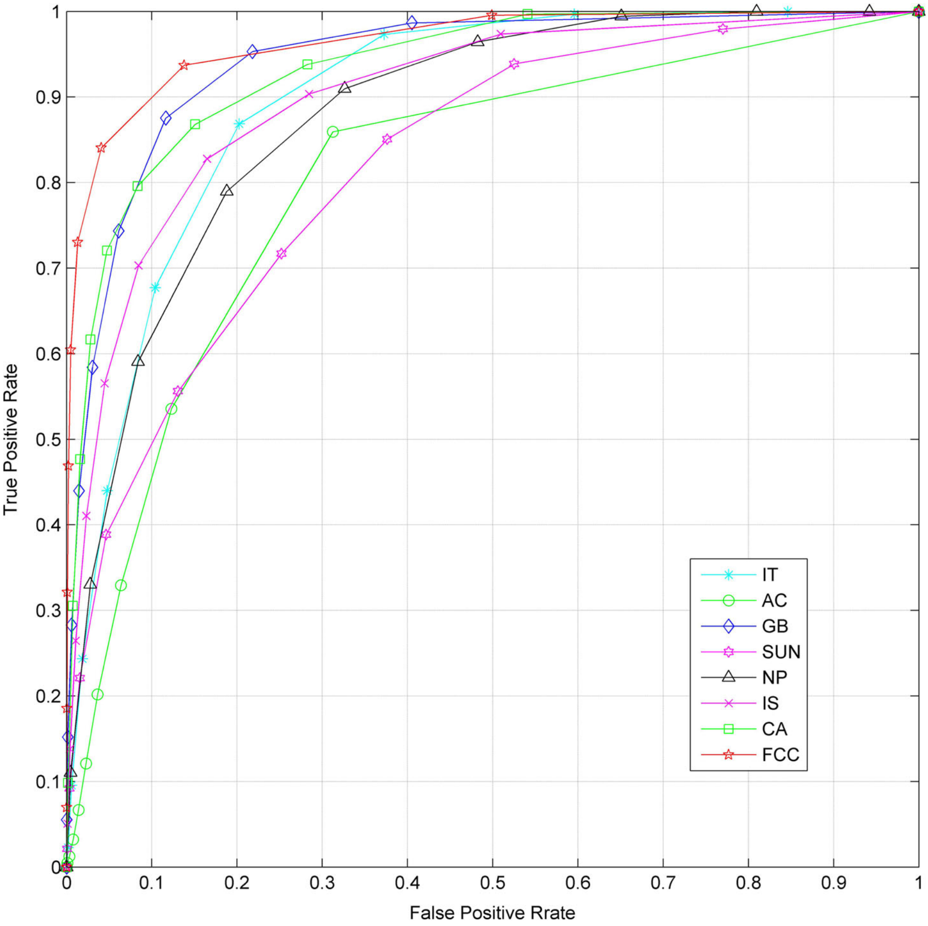


Fig. 5 ROC curves of salient object detection results obtained by different salient object detection models

Table 2 The run-time performance (in seconds) comparison of different salient object detection models

	IT	AC	GB	SUN	NP	IS	CA	FCC
Time	0.31	118.99	1.36	1.68	1.39	0.03	47.46	0.17

t $(0 \leq t \leq 1)$ is used to obtain the binary masks $B_t(x, y)$, in which 0 denote the background and 1 denote the salient objects. The TPR and FPR can be computed via:

$$TPR = E\left(\prod B_t(x, y) * G(x, y)\right) \tag{13}$$

$$FPR = E\left(\prod (1 - B_t(x, y)) * G(x, y)\right) \tag{14}$$

Table 1 shows the results of TPR and FPR obtained from these models. As illustrated in Table 1 and Fig. 5, the proposed model can achieve the best performance over the other seven models.

Finally, we compare the computational complexity of the different salient object detection models discussed. For this purpose, these models are implemented using the Matlab programming language and run on a PC with a G2020 CPU and a 4GB RAM. Each of the models is applied to the MSRA database, and then their respective average execution times can be obtained (given in seconds in Table 2).

4 Conclusions

The proposed algorithm is based on PCA method to fuse the color feature and contrast feature of the original image. The generated saliency map can highlight the salient object in different images. Experiments are conducted on the MSRA dataset to compare the performance of the proposed model with other seven state-of-the-art salient object detection models. The proposed model can achieve fairly good performance against the state-of-the-art saliency computational models, as verified in extensive experiments.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (61440016, 61375017, 61273225), the Natural Science Foundation of Hubei Provincial of China (2014CFB247), the open foundation of the key laboratory for metallurgical equipment and control of ministry of education (2013B08)

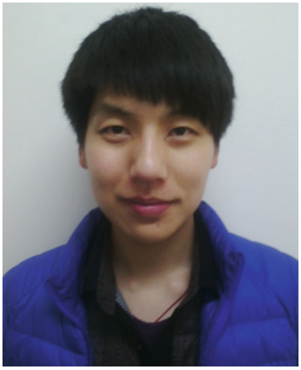
References

1. Achanta R, Estrada F, Wils P, Susstrunk S (2008) Salient region detection and segmentation. In: The 6th International conference on computer vision systems, pp 66–75
2. Achanta R, Susstrunk S (2010) Saliency detection using maximum symmetric surround. In: The IEEE International conference on image processing, pp 2653–2656
3. Aziz MZ, Mertsching B (2008) Fast and robust generation of feature maps for region-based visual attention. IEEE Trans Image Process 17:633–644
4. Borji A, Itti L (2012) Exploiting local and global patch rarities for saliency detection. In: IEEE conference on computer vision and pattern recognition, pp 478–485

5. Chen LQ, Xie X, Fan X, Ma WY, Zhang HJ, Zhou HQ (2003) A visual attention model for adapting images on small displays. Multimedia Systems 9:353–364
6. Cheng MM, Zhang GX, Mitra NJ, Huang X, Hu SM (2011) Global contrast based salient region detection. In: IEEE conference on computer vision and pattern recognition, pp 409–416
7. Duan L, Wu C, Miao J, Qing L, Fu Y (2011) Visual saliency detection by spatially weighted dissimilarity. In: IEEE conference on computer vision and pattern recognition, pp 473–480
8. Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: Advances in neural information processing systems, pp 545–552
9. Hou X, Harel J, Koch C (2012) Image signature: highlighting sparse salient regions. IEEE Trans Pattern Anal Mach Intell 34:194–201
10. Itti L, Koch C, Niebur E (1998) A model of saliency based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20:1254–1259
11. Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: The IEEE International conference on computer vision, pp 2106–2113
12. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4:219–227
13. Liu T, Sun J, Zheng NN, Tang X, Shum HY (2007) Learning to detect a salient object. In: IEEE conference on computer vision and pattern recognition, pp 1–8
14. Lu Y, Zhang W, Lu H, Xue X (2011) Salient object detection using concavity context. In: IEEE International conference on computer vision, pp 233–240
15. Murray N, Vanrell M, Otazu X, Parraga CA (2011) Saliency estimation using a non-parametric low-level vision model. In: IEEE conference on computer vision and pattern recognition, pp 433–440
16. Santella A, Agrawala M, DeCarlo D, Salesin D, Cohen M (2006) Gaze-based interaction for semi-automatic photo cropping. In: Proceedings of SIGCHI conference on human factors in computing systems, pp 771–780
17. Shi K, Wang K, Lu J, Lin L (2013) PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In: IEEE conference on computer vision and pattern recognition, pp 2115–2122
18. Stas G, Lihi ZM, Ayellet T (2012) Context-aware saliency detection. IEEE Trans Pattern Anal Mach Intell 34:1915–1926
19. Stentiford F (2007) Attention based auto image cropping. In: ICVS workshop on computational attention and application, pp 1–9
20. Treisman A, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12:97–136
21. Vu CT, Chandler DM (2012) S3: A Spectral and spatial measure of local perceived sharpness in natural images. IEEE Trans Image Process 21:934–945
22. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: a Bayesian framework for saliency using natural statistics. J Vis 8:1–20

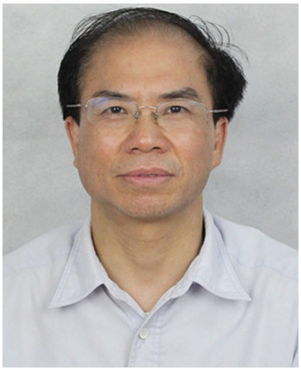
Xin Xu received the B.Sc. and Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University, China, in 2004 and 2012 respectively; and the M.Sc. degree in computer science and technology from Wuhan University of Science and Technology in 2006.

He is an associate professor in the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China. His current research interests include computer vision, pattern recognition, and visual surveillance.



Nan Mu is currently working toward the master's degree at the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China. His research interests include image saliency detection, image feature extraction, and visual surveillance.

Li Chen received the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. Currently he is an associate professor in the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China. His research interests include image processing, computer vision, and visual surveillance.



Xiaolong Zhang received the Ph.D. degree in computer science and engineering from Tokyo Institute of Technology, Japan, in 1998. During 1998 to 2002, he joined IBM in Japan and was responsible for the research and development of data warehouse & data mining. Currently he is a professor in the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China. His research interests include data mining, protein folding, and image processing.