EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

A unified framework for spatiotemporal salient region detection

Bo Wu[1,2*], Linfeng Xu[1], Liaoyuan Zeng[1], Zhengning Wang[1] and Yan Wang[2]

Abstract

This article presents a new bottom-up framework for spatiotemporal salient region detection. The generated saliency map can uniformly highlight the salient regions. In the proposed framework, the spatial visual saliency and the temporal visual saliency are first computed, respectively, then they are fused with a dynamic scheme to generate the final spatiotemporal saliency map. In the spatial attention model, the approach of joint embedding of spatial and color cues is adopted to compute the spatial saliency map. In the temporal attention model, we propose a novel histogram of average optical flow to measure the motion contrast of the different pixels. The method can suppress the motion noise efficiently because the statistical distribution of optical flow in a patch is comparatively stable. Furthermore, we combine the spatial and the temporal saliency maps through an adaptive fusion method, in which a novel motion entropy is proposed to evaluate the motion contrast of the input video. Extensive experiments demonstrate that our method can obtain higher quality saliency map compared with state-of-the-art methods.

Keywords: Visual saliency, Spatial attention model, Temporal attention model, Adaptive fusion

1 Introduction

Human visual system has an excellent ability to quickly catch salient information from complex scenes. The mechanism in the brain that determines which part of the visual data is currently of the most interest is called selective attention [1]. The mechanism is critical for human to understand scenes. In recent years, many computational models have been proposed to mimic the mechanism of selective visual attention. The models can compute saliency maps from image or video inputs. The pixels with higher intensity values in saliency map denote that the corresponding pixels are visually important. The saliency map can be used for applications, such as object-of-attention segmentation [2-4], object detection [5,6], image and video summarization [7], video surveillance [8], and image and video compression [9].

The existing methods of saliency detection can roughly be categorized into local and global methods. The local contrast-based methods compute the saliency of a specific image region based on it's local neighborhoods [10-13]. The global contrast-based methods estimate the saliency of an image region by taking the contrast relations to the entire image into account [14-19]. In contrast to the weakness of the local methods which usually highlight the object boundary instead of the entire object, the global methods can generate saliency maps with full resolution and uniformly highlighted regions. However, only simple features are considered in calculating the motion saliency, such as flicker, which limits the models to the static background. If the salient object and background change simultaneously, the quality of the estimated saliency map is degraded rapidly.

In this article, we propose a novel unified framework which can detect salient regions in both images and videos flexibly. We consider the definition of visual saliency in the global approach. Thus, we assign more visual saliency to the features which are less frequent. Our approach which is extended from our previous work [20,21] computes spatiotemporal saliency map based on a global scheme in spatial and temporal domain. Given a video, we first compute the spatial saliency map which adopts joint embedding of spatial and color cues. As for the temporal saliency, we compute the global motion contrast of dense optical flow.

*Correspondence: bo.wu.cv@gmail.com
[1]School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2]College of Physics and Information Engineering, Henan Normal University, Xinxiang, China

To suppress the motion noise, a new histogram of average optical flow (HOAOF) is proposed to compute the motion contrast of different pixels. Finally, a novel adaptive fusion technique is proposed to combine the spatial and the temporal saliency maps.

The main contributions of the article are as follows:

(1) We propose a powerful unified framework for spatiotemporal salient region detection, which can obtain higher quality salient region detection results than existing methods no matter whether the camera is fixed or not.
(2) A new HOAOF is proposed to compute the motion contrast in pixel-level. The descriptor can suppress the motion noise effectively because the statistical distribution of optical flow in a patch is comparatively stable.
(3) We propose a novel adaptive fusion scheme to combine the spatial and the temporal saliency maps. The motion contrast of video sequence is measured by motion entropy. If a video has strong motion contrast, then the motion entropy of the video will be small. Correspondingly, the temporal attention model can be assigned a high weight, and vice versa.

The remainder of the article is organized as follows. In Section 2, the related work is discussed. The spatiotemporal salient region detection method is proposed and elaborated in Section 3. The experimental results and comparisons with other methods are provided in Section 4. In Section 5, we first discuss the connections between our approach and the related methods. Then, the limitation of the proposed method is also analyzed. Finally, the conclusion is presented in Section 6.

2 Related work

Visual attention can be determined by two categories of factors: bottom-up factors and top-down factors. The idea of bottom-up attention is to seek for the "visual pop-out" saliency. The salient signals are driven solely from the visual scene. On the contrary, both the cognitive factors and the high-level stimulus (e.g., face [22], person and car [23]) are considered in top-down attention. The approach expects latent correlations between visual attributes and saliency values and aims to mine such correlations from the training data [24]. Since the data-driven stimuli are easier to control than the cognitive factors, and the exact interaction between bottom-up process and top-down process still remains elusive [25], bottom-up attention mechanisms are investigated more than top-down mechanisms.

The bottom-up models driven by low-level features can be classified into two schemes: local and global. The local contrast-based methods explore a salient feature depending on its neighborhoods. In the well-known bottom-up attention model [10], three basic low-level features (i.e., color, intensity, and orientation) are used to generate three conspicuity maps by computing the center-surround contrast. Then, the conspicuity maps are combined into a single saliency map. Based on [10], Itti et al. [12] extend the saliency detection model from the static scenes to the dynamic video clips by introducing two simple features: flicker and motion. The flicker is computed from the absolute difference between the luminance of consecutive frames. The motion is computed from spatially shifted differences between Gabor pyramids from the consecutive frames. Kim et al. [13] propose a novel method for spatiotemporal salient region detection. The approach combines the spatial saliency and the temporal saliency with fixed weight. For calculating temporal saliency, the authors simply compute the sum of absolute difference between the temporal gradients of the center and the surrounding regions.

A common limitation of the local scheme is that the generated saliency map usually produces high saliency values in the object boundary instead of the entire salient objects when only one scale is considered. In general, a multi-scale fusion scheme is used to alleviate the boundary-emphasize effect. The methods can obtain high equality motion saliency map when the camera is static. Once the camera is moving, the quality of the produced motion saliency map may be degraded rapidly. To overcome the problem, Le Meur et al. [26] apply motion contrast to compute the temporal saliency. You et al. [27] estimate the global motion to compensate camera's motion and determine the video attention regions.

The global contrast-based methods integrate the entire information features all over the visual field. The approaches in [14,17] calculate the saliency map based on the Fourier frequency spectrum. In [17], the difference between the original signal and the smooth one in the log amplitude spectrum is calculated, and then the saliency map is obtained by transforming the difference to the spatial domain. Guo and Zhang [14] use image's phase spectrum of Fourier transform instead of amplitude spectrum to calculate the saliency map. Furthermore, the phase spectrum of quaternion Fourier transform (PQFT) is applied to detect the spatiotemporal saliency in the dynamic scenes. In Guo and Zhang's model, intensity, color, and motion features are comprised into a quaternion image as an individual channel for taking phase spectrum. These features' contribution is equivalent to each other in the final saliency map. However, the psychological studies reveal that the motion contrast usually attracts more human attention than other external signals [28].

3 The proposed method

In this section, we give a detailed description of our spatiotemporal salient region detection approach. We introduce the spatial attention model in Section 3.1. Then, we show how to calculate temporal saliency in Section 3.2. In Section 3.3, we show how to fuse spatial saliency map and temporal saliency map adaptively.

3.1 Spatial attention model

The spatial saliency map is computed from joint embedding of spatial and color cues. Three factors are considered to compute the individual saliency maps, respectively. The final spatial saliency map is generated by combing these maps in a two-layer saliency structure. Please refer to [20,21] for more details about the spatial attention model.

3.1.1 Spatial constraint

The first factor is spatial constraint (SC). It is based on the observation that a pixel is salient when the adjacent pixels have strong contrast with respect to it, while a pixel is less salient when the strong contrast pixels are far away from it. Moreover, according to the "center-surround" inhibition mechanism, the surrounding pixels should make a greater contribution when calculating global contrast-based saliency. The SC-based saliency of a pixel can be formulated as

$$\mathrm{Sal}_{\mathrm{SC}}(p) = \sum_{\forall q \in I} \alpha_{p,q} \|I_p - I_q\|, \tag{1}$$

where I_p is the CIELAB color value of pixel p, $\|I_p - I_q\|$ is the Euclidean distance between I_p and I_q. The SC factor $\alpha_{p,q}$ is defined as

$$\alpha_{p,q} = \frac{1}{Z} \exp\left(-\frac{\|p-q\|^2}{\pi_1^2}\right) \exp\left(\frac{\Sigma(q)}{\pi_2^2}\right) \tag{2}$$

where Z denotes the normalization factor, $\|p - q\|$ is the spatial distance between pixels p and q, and $\Sigma(q)$ is the sum of distance to all other pixels. The surround pixels have larger $\Sigma(q)$ than the center ones. The parameters π_1^2 and π_2^2 are set to 300 and 0.06, respectively, in our experiments. We use fixed parameters for all datasets in order to perform fair comparison. The same principle is employed for all of the parameters discussed in the following sections. The obtained saliency map for Figure 1a using SC saliency is shown in Figure 1b.

3.1.2 Color double-opponent

The second factor is color double-opponent (CD), which is the color channel representation in cortex. The physiological study shows that the red–green (RG) and blue–yellow (BY) contrast have major impact on human attention [10]. We use $G_{\mathrm{RG}}(p)$ and $G_{\mathrm{BY}}(p)$ to represent the global contrasts of RG and BY, e.g., $G_{\mathrm{RG}}(p) = \frac{1}{N}\sum_{\forall q \in I} |\mathrm{RG}(p) - \mathrm{RG}(q)|$, supposing the image has N pixels. Then, the CD-based saliency of a pixel p is expressed as

$$\mathrm{Sal}_{\mathrm{CD}}(p) = \frac{G_{\mathrm{RG}}(p) + G_{\mathrm{BY}}(p)}{\beta(p)}, \tag{3}$$

where the normalization factor $\beta(p) = \max_q\{|\mathrm{RG}(p) - \mathrm{RG}(q)|, |\mathrm{BY}(p) - \mathrm{BY}(q)|\}, \forall q \in I$. The obtained saliency map for Figure 1a using CD saliency is shown in Figure 1c.

3.1.3 Similarity distribution

The third factor is similarity distribution (SD). In general, the background can be distributed over the entire image exhibiting a high spatial variance, whereas the foreground objects are generally more compact. Based on the observation, the SD-based saliency for a pixel p is defined as

$$\mathrm{Sal}_{\mathrm{SD}}(p) = \exp\left(-\frac{\Pi(p)}{\pi_3^2}\right), \tag{4}$$

where the parameter π_3^2 is set to 0.2, $\Pi(p)$ is the SD.

$$\Pi(p) = \frac{1}{N}\sum_{\forall q \in I} \frac{1}{Z'} \gamma_{p,q} \|p - q\|^2, \tag{5}$$

where Z' denotes the normalization factor, $\gamma_{p,q} \in [0, 1]$ measures the similarity between two pixels [20].

For the pixel p inside an object, $\Pi(p)$ can be approximated as the sum of distance to other pixels in the same object which is smaller. So it is more likely to assign pixels which belong to the same salient object large SD saliency values and vice versa. The obtained saliency map for Figure 1a using the SD saliency is shown in Figure 1d.

3.1.4 Two-layer fusion scheme

After the three saliency components are computed, the final saliency can be constructed from two layers [29], i.e., basic layer and enhancement layer, which are defined as follows:

(1) The SC saliency is employed as the basic layer.
(2) The enhancement layer is designed based on the CD and SD saliency.

According to the two-layer fusion scheme [29], we can obtain the final saliency map

$$S^S(p) = \mathrm{Sal}_{\mathrm{SC}}(p)(1 + w_1 \mathrm{Sal}_{\mathrm{CD}}(p) + w_2 \mathrm{Sal}_{\mathrm{SD}}(p)), \tag{6}$$

where the weight factors w_1 and w_2 regulate the extent of importance for the CD and the SD saliency. In our experiments, we set $w_1 = w_2 = 1$.

In the two-layer saliency fusion scheme, the basic layer (SC) always works when the CD or the SD is either high or low. The enhancement layer (CD and SD) aims to attract more human attention when the CD contrast is strong or

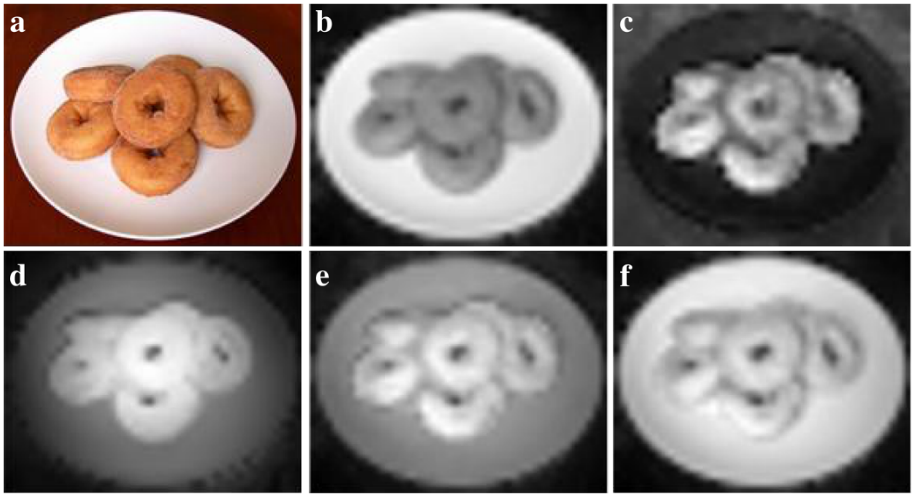

Figure 1 Example of image saliency detection. (**a**) Original image. (**b**) SC saliency map. (**c**) CD saliency map. (**d**) SD saliency. (**e**) The final saliency map is obtained through pooling mechanism [10]. (**f**) The final saliency map is synthesized through the two-layer structure.

the SD is compact. As shown in Figure 1e,f, the saliency map constructed from two-layer structure highlights the two salient objects (i.e., pastry and plate) more uniformly than the pooling mechanism used in [10].

3.2 Temporal attention model

In the temporal attention model, temporal saliency maps are often calculated by temporal gradient which is computed by using the intensity difference between successive frames. The models work well when the camera is static. Once the camera moves, the evaluated saliency maps will incorporate much noise. In this study, we find that the object in video sequences exhibiting high motion saliency usually has the following properties: (a) there are clear motion patterns in the scene; (b) the motion of object exhibits difference from the global motion of the scene; (c) compared with the size of the scene, the object is relatively small.

Based on the observations, we define the saliency of a pixel as its motion contrast to all the other pixels in the frame. In this study, we select dense optical flow[a] [30] which is a modified version based on [31,32], to compute the motion field, because the computational complexity is low. The pixel's saliency can be formulated as

$$S^T(p) = \sum_{\forall q \in I} |D(V_p, V_q)|, \tag{7}$$

where V_p and V_q are the optical flow of pixels p and q in frame I, respectively, $D(V_p, V_q)$ is the vector difference between the optical flow of pixels p and q, and $|.|$ represents magnitude of vector.

Due to the changing illumination conditions or the fixed camera noise, there is a considerable noise in the estimated optical flow. Moreover, if multiple motion layers exist in the scene, inaccurate estimation of optical flow may yield at the edge pixels in different motion regions. If we use formula (7) to compute the saliency map directly, much noise will be generated. Some examples present in the third column of Figure 2. In contrast, the statistical distribution of optical flow in a patch is comparatively stable. To suppress the background noise, we propose a novel HOAOF to measure the motion contrast of the different pixels. Specifically, the optical flow is first computed at every frame of video sequence. Then, a smooth procedure is applied to the optical flow. Finally, the histogram of optical flow belonging to the local patch centered at the pth pixel is generated. Flow orientations are quantized into N levels according to its primary angle from the horizontal axis and weighted according to its magnitude. The HOAOF can be defined as follows:

$$\begin{aligned} &H_p = (h_{p,1}, h_{p,2}, \ldots, h_{p,N}) \\ &\text{with } h_{p,n} = \sum_{\substack{(x,y)\in w_p \\ \theta(x,y)\in n}} m(x,y), \end{aligned} \tag{8}$$

where $m(x,y)$ and $\theta(x,y)$ denote the flow magnitude and the quantized orientation at the pixel position (x,y) of a frame, respectively. The parameter w_p is a local patch centered at the pth pixel. The number of bins N is set to 4 and w_p is set to 7×7 pixels in our experiments. Figure 3 illustrates the procedure.

So, the formula (7) can be rewritten as

$$S^T(p) = \sum_{\forall q \in I} D(H_p, H_q), \tag{9}$$

where H_p and H_q represent the HOAOF of local patch centered at the pixel p and q in frame I,

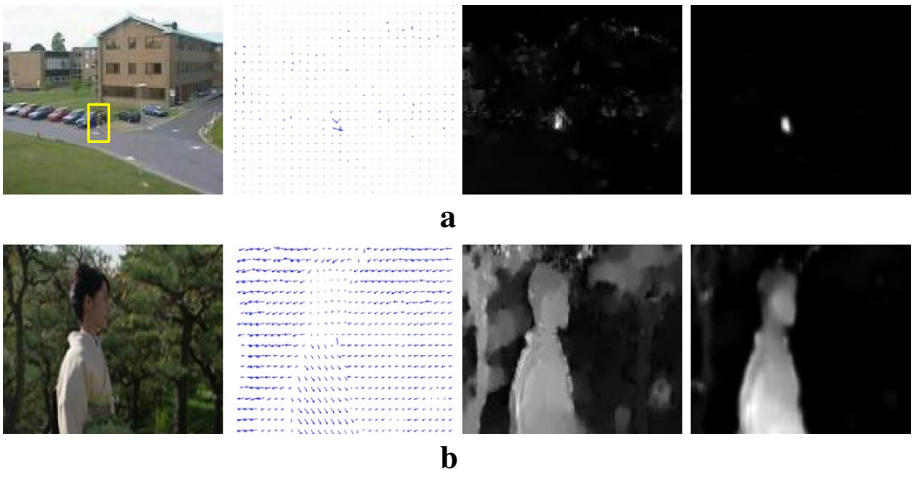

Figure 2 Examples of motion saliency detection. (**a**) A pedestrian under a static camera attracts attentions. (**b**) The camera is tracking a walking person, who attracts more attention. First column: sample frames of two videos. Second column: the corresponding optical flow. Third column: motion saliency map calculated with optical flow directly. Fourth column: motion saliency map calculated with HOAOF.

respectively, $D(H_p, H_q)$ is χ^2 distance between the two histograms:

$$D(H_p, H_q) = \frac{1}{2} \sum_k \frac{(H_p(k) - H_q(k))^2}{H_p(k) + H_q(k)} \tag{10}$$

Finally, the temporal saliency map is normalized to a fixed range [0,1].

An example is shown in Figure 2. It is clear that the temporal attention model can suppress the background noise efficiently. In Figure 2a, the camera is fixed and the global motion is nearly static. Compared with the background, the moving pedestrian produces a high-salient region in the frame. In Figure 2b, the camera tracks a pedestrian such that the person has small optical flow, while the background has large motion. In this case, the direction of global motion is opposite to that person. A clearly saliency map can still be obtained by using our model. If a dynamic scene has strong motion contrast, the main motion will be gathered in few directions and the motion object will pop out explicitly.

3.3 Adaptive fusion

We have obtained the spatial and the temporal saliency maps separately. The two maps need to be fused in a meaningful way to generate the final spatiotemporal saliency map. It is shown in [28] that the human vision system is more sensitive to motion information compared with the static signals. In a dynamic scene, the camera is tracking a pedestrian, while the motion direction of background is opposite to the camera's movement. In general, people are more interested in the followed person instead of his surrounding regions. In surveillance video, the camera is fixed and the moving objects in video attract more human attention than the static background. In these examples, motion contrast is the prominent feature for the saliency detection compared with other features, such as intensity, texture, and color. In contrast, if the motion of the video is cluttered or the motion contrast is insignificant, human attention is attracted more to the contrasts caused by the static visual stimuli. Thus, simple linear combination with fixed weights between the spatial saliency map and the temporal saliency map may lead to unsatisfactory result. Instead, we adopt an adaptive fusion scheme, which is consistent with the above considerations. The adaptive fusion scheme can give higher weight to the temporal saliency map when strong motion contrast is present in the dynamic scene. In contrast, a higher weight is assigned to the spatial saliency map when the motion contrast is weak.

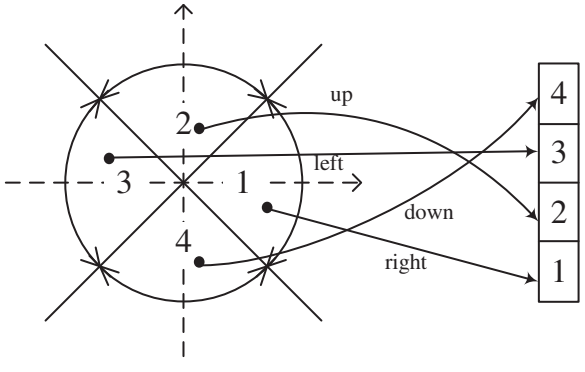

Figure 3 Histogram formation with four bins, N = 4. The optical flow is quantized into one of four cardinal directions (up, down, left, and right).

In this article, the motion entropy is proposed to evaluate how strong the motion contrast is in the video sequence. First, the HOAOF of the frame is calculated. Second, according to the HOAOF, the motion entropy is computed as follows:

$$E = -\sum_{i=0}^{L} h_i \log h_i, \tag{11}$$

where L is the number of bins. The parameter h_i is the value of ith bin in HOAOF. The more cluttered the distribution of motion direction in video frame is, the larger the entropy is, and vice versa.

It is important to note the differences from the aforementioned HOAOF. First, we use one additional bin with $i = 0$ which incorporate all pixels that the flow magnitude is lower than a preset threshold. For instance, in the surveillance video, there is considerable motion noise in the static background. To weaken the effect of flow noise on the motion entropy computing, we collect the flow into an individual bin. Second, the number of bins L is set to 16. A relative fine quantization is beneficial for the estimated entropy to reflect the motion distribution correctly. Two examples of the spatiotemporal saliency detection are shown in Figures 4 and 5. The first columns show sample frames of two different videos, respectively. In the first video, the moving car and person have a strong relative motion with respect to the static background. Thus, a higher weight is assigned to the temporal attention model. On the contrary, the second video has a cluttered background motion because the grass and branches present irregular motion. In this case, our attention is attracted more to the static visual stimuli (e.g., color) than motion. Hence, our algorithm allocates high weight to spatial saliency map. Attribute to the adaptive fusion scheme, the fused spatiotemporal saliency map, successfully detects the pedestrian, car, and bird as salient region.

4 Experimental results

In this section, we first introduce the datasets used for performance evaluation. Then, we compare the proposed method with three state-of-the-art methods [11,13,14] and provide the qualitative and quantitative results, respectively.

4.1 Video sequences datasets

The performance of the proposed algorithm is evaluated extensively on two types of videos, named Video Set 1 and Video Set 2, respectively. Video Set 1 contains surveillance videos, which are collected from PETS2001.[b] There are 6,000+ images totally. In this dataset, the camera is fixed and the background is still. People's attention is mainly attracted to the moving objects [28], such as the pedestrian and the moving car. The examples of frames are shown in Figure 4. Since the size of the moving object is small, we use the bounding boxes of the moving objects as the ground truth. We collect Video Set 2 with 60 video clips from the Internet and the video segmentation datasets [33]. Each video clip contains about 60–200 frames with the same salient objects. There are 6,000+ images totally. Different from Video Set 1, the camera in this dataset is moving or the background presents clutter motion when the camera is still. It means that the objects and the background of the scenes are moving. Since the size of salient objects in Video Set 2 is large, the annotated ground truth masks are object-contour based.

4.2 Performance evaluation

In Figure 4a, we show the representative frames of Video Set 1, as well as the individual saliency detection results of the proposed method in different stages. Figure 4b is the

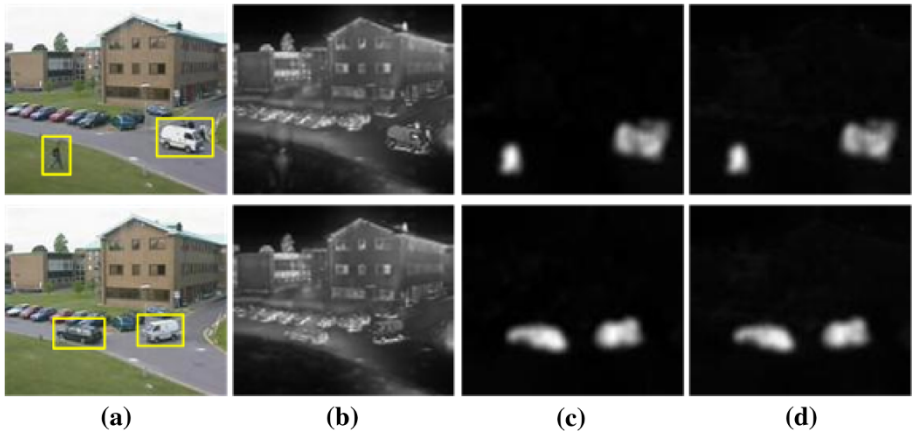

Figure 4 The example saliency detection in Video Set 1. Column (a) shows two example frames of Video Set 1 (PETS2001). Yellow boxes represent the moving objects; Column (**b**) presents spatial saliency maps; Column (**c**) is temporal saliency maps; Column (**d**) is the fused spatiotemporal saliency maps.

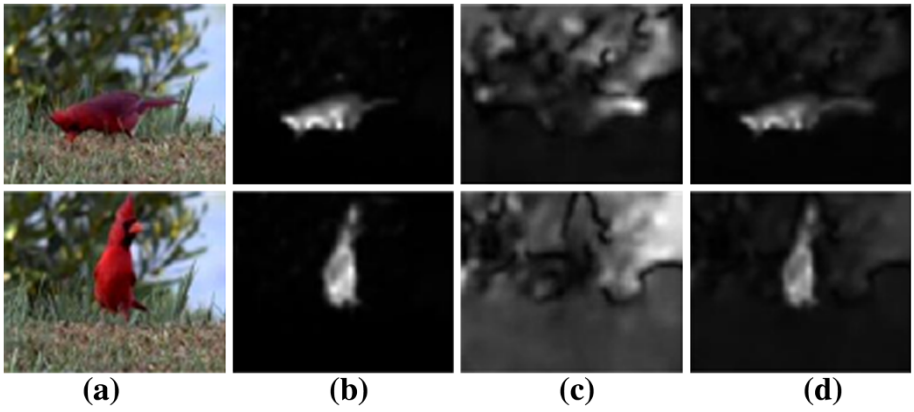

Figure 5 The example saliency detection in Video Set 2. Column (a) shows two example frames of bird; Column (**b**) presents spatial saliency maps; Column (**c**) is temporal saliency maps; Column (**d**) is the fused spatiotemporal saliency maps.

computed spatial saliency map. Figure 4c is the temporal saliency map. The fused spatiotemporal saliency map is presented in Figure 4d. It is seen from the figure that the spatial saliency map does not highlight the salient object successfully. The main reason is that the scene has the highly texture background and the static features of the small foreground objects are not significantly distinctive. However, compared with the still background, the moving foreground objects have strong motion contrast. This leads to a temporal saliency map which can detect the moving salient objects clearly. In our adaptive fusion scheme, the strong motion contrast results in the dominant contribution of temporal attention model in the final spatiotemporal saliency map. As shown in Figure 4d, the effect of spatial saliency map is negligible. Another example in Video Set 2 is presented in Figure 5. The video records a bird by a fixed camera in the wild where the branches of the background present clutter motion. The motion contrast in the scene is weak, so that the spatial saliency is dominant over the temporal saliency in the adaptive fusion scheme. The spatial, the temporal, and the spatiotemporal saliency maps are shown in Figure 5b–d, respectively.

Recently, two spatiotemporal saliency models were presented in [13,14]. To justify the effectiveness of proposed model, we compare the proposed method with PQFT model [14] and Kim et al.'s model [13] in Figures 6 and 7. The pedestrian and the moving car in Figure 6 are captured as salient region by all models. Compared with PQFT [14], Kim's method assigns the moving objects much higher saliency value. Nevertheless, the highly texture backgrounds are not suppressed by these models. In contrast to that, our method not only detects the pedestrian and the car as the most salient regions but also suppress the background area effectively. This is attributed to the adaptive fusion scheme of our framework. The strong motion contrast can lead to the dominant contribution of temporal attention model in the final spatiotemporal saliency map. Figure 7 shows the example images from the different video segments of Video Set 2 and the computed saliency maps using different models. The methods in [13,14] cannot detect the salient region successfully due to the changing background, while our model can detect the salient region clearly.

Finally, the spatiotemporal saliency map of the frame in the video sequence is formulated as follows:

$$S^{\mathrm{ST}}(p) = (1 - w_t)S^S(p) + w_t S^T(p) \tag{12}$$

with $w_t = e^{-\alpha E}$, α is a constant factor which adjust the weight. In our experiments, we set $\alpha = 0.15$. Our model can also deal with static images easily by setting $w_t = 0$.

Furthermore, the proposed model can be employed to extract the salient objects from the video sequences by thresholding the spatiotemporal saliency map via a moderate threshold. To this end, a non-parametric significance testing is adopted [34].We compute the empirical PDF from all the saliency values and set a threshold to achieve 95% confidence level in deciding whether the given values are in the extremely right tails of the estimated distribution. In addition to the comparison between the methods in [13,14], we also compare the proposed method with Liu et al.'s model [11], which is a salient object detection method.In [11], a group of static and dynamic saliency features are computed and the optimal linear weights are learned through CRF learning method. Given an image pair, the model outputs a binary label map, which is further transformed to a bounding rectangle representing the salient object. In order to facilitate comparison, we take the binary label map of [11] as the detection result.In Video Set 1, the background is static, which is different from Video Set 2. Training in two video sets together can degrade the overall performance of salient object detection. So, we train CRFs in two datasets separately. In Video

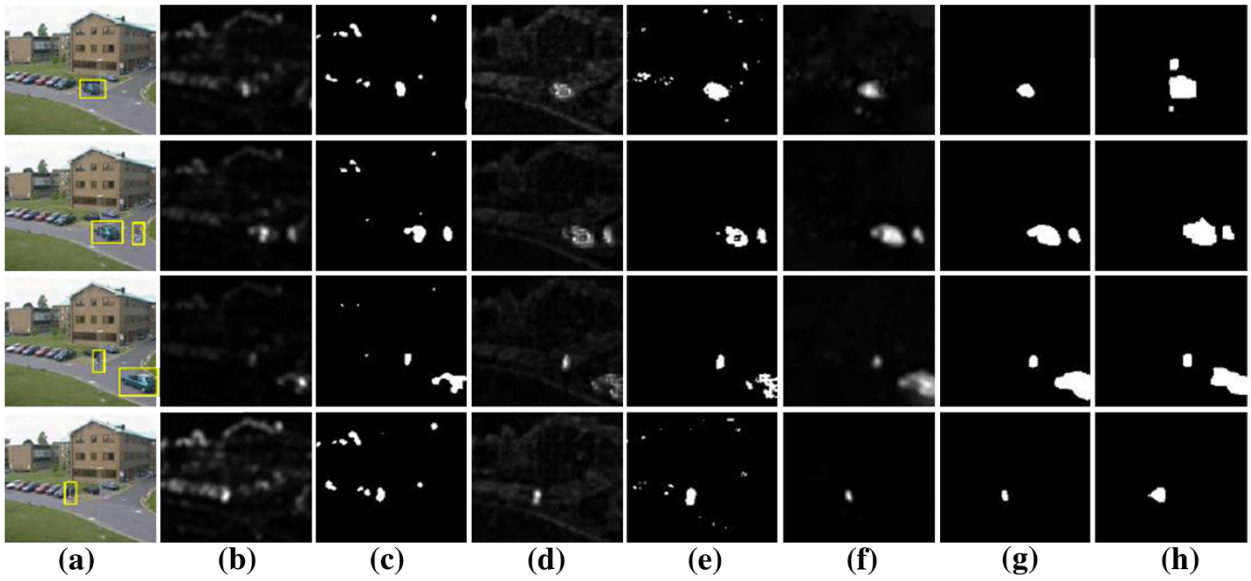

Figure 6 Spatiotemporal saliency in video sequences and salient object extraction. Column (**a**) shows the representative frames of Video Set 1. Yellow boxes represent the ground truth; Columns (**b**), (**d**), and (**f**) show the spatiotemporal saliency map of PQFT [14], Kim's [13], and proposed method, respectively; Column (**c**), (**e**), and (**g**) show the extracted objects by corresponding models; (**h**) is the generated binary mark of [11].

Set 1, the surveillance video is divided into 60 video segments. We randomly select 40 video segments with 2,000+ image pairs to construct a training set, and use the others for testing. In Video Set 2, we randomly select 40 video segments with 2,000+ image pairs to construct a training set, and use the others for testing.

The subjective results are shown in Figures 6 and 7. For the quantitative comparison, precision, recall, and F-beta, which are defined in [18], are computed by comparing the segmented region and the ground truth. To perform comparison experiment in the same settings, we use the method in [13] to calculate the performance index from 15 frames which are taken from every test video segment randomly, and then averaged in the test set of each Video Set. Because of the small change of scene and motion patterns in each short video segment, the variance of computed performance indexes in each video segment is small. The results are shown in Tables 1 and 2. It is clear that Kim et al.'s model [13] outperforms PQFT [14] in Video Set 1, but it is lower than Liu et al.'s model [11] and our method. Compared with the results of Kim et al.'s model, our method yields 3, 13, and 6% gain with regard to recall, precision, and F-beta, respectively. The performance of the salient object detection in [11] is superior to our method, which is mainly attributed to the static background and the singular motion pattern. It is beneficial for CRF learning. In Video Set 2, our method presents the optimal performance. The methods in [13,14] fail to extract salient objects mainly because the methods cannot deal with the scene with background moving. Due to the diverse scene and motion patterns, learning the optimal linear weights of various saliency features to satisfy all situations is difficult. The performance of the salient object detection of [11] in Video Set 2 is not as good as in Video Set 1. Compared with the results of [11], our method yields 8, 20, and 7% gain with regard to recall, precision, and F-beta, respectively. To further verify whether the differences between these methods are statistically significant, we use approximate randomization [35] for statistical significance testing on F-beta. The test results (Tables 1 and 2) show that our model outperforms [13,14] in all evaluations with strong statistical significance. In Video Set 2 our model outperforms [11] significantly, while [11] outperforms our model significantly in Video Set 1. The main reason is the background in Video Set 1 is still, which is beneficial for CRF learning.

Table 1 Performance evaluation for salient object extraction

Method	Dataset		
	Video set 1		
	Recall	Precision	F-beta
PQFT [14]	0.33	0.56	0.39[a]
Kim et al. [13]	0.35	0.72	0.46[a]
Liu et al. [11]	0.62	0.83	0.64[b]
Our	0.38	0.85	0.52

Our model has significantly better results than the methods which are indicated with [a] for $p \leq 0.05$. The methods which perform significantly better than our model are indicated with [b] for $p \leq 0.1$.

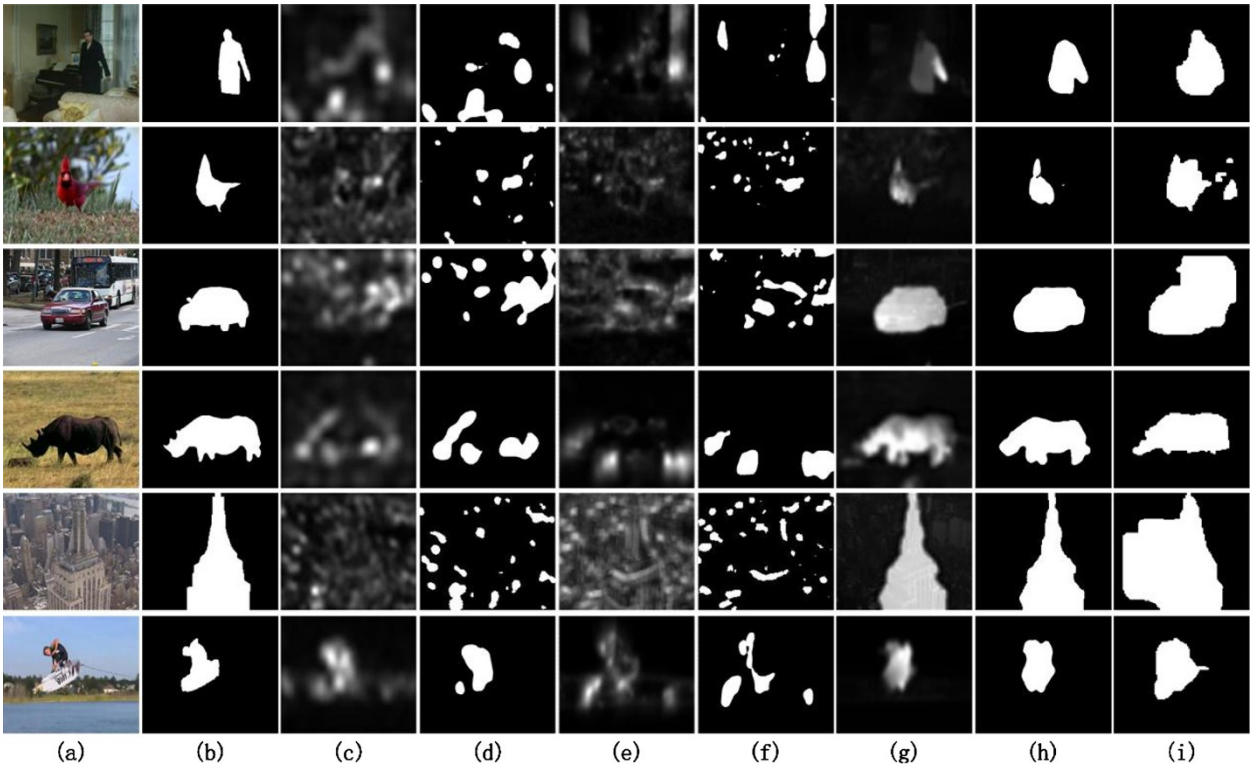

Figure 7 Spatiotemporal saliency in video sequences and salient object extraction. Column (**a**) shows the representative frames of Video Set 2; Column (**b**) shows the pixel-wise ground truth annotation; Column (**c**), (**e**), and (**g**) show the spatiotemporal saliency map of PQFT, Kim, and proposed method; Column (**d**), (**f**), and (**h**) show the extracted objects by corresponding models; Column (**i**) shows the generated binary mask of [11].

5 Discussion

In this section, we first discuss the difference between our approach for salient region detection and other saliency detection models similar with Itti et al.'s model [10]. Furthermore, we discuss the limitations and implement failure analysis of the proposed method.

5.1 Salient region versus visual saliency

Salient region detection is different from the visual saliency computation in [10,36] or other based on the biologically plausible computational models of attention. Itti et al.'s model [10] and those similar to it usually focus on mimicking the properties of vision and predicting eye fixations. The resulting saliency maps are often overemphasize small, purely local features, and fail to detect the internal part of the target, which makes the approach less useful for applications, such as segmentation and detection. This kind of model is usually evaluated by comparing the saliency map with the real human attention density map. Salient region detection method is part of the computational approach which is inspired by the biological theory, but is closely related to the typical applications in computer vision, such as adaptive content delivery, adaptive region-of-interest-based image compression, salient object segmentation [37], and object recognition. The resulted saliency map can uniformly highlight the entire salient regions in scenes. This kind of model is usually evaluated by comparing the resulted saliency map with the manually labeled binary ground-truth mask, such as [18,38].

Table 2 Performance evaluation for salient object extraction

Method	Dataset		
	Video set 1		
	Recall	Precision	F-beta
PQFT [14]	0.20	0.32	0.23[a]
Kim et al. [13]	0.17	0.26	0.19[a]
Liu et al. [11]	0.62	0.65	0.60[a]
Our	0.70	0.85	0.67

Our model has significantly better results than those methods which are indicated with [a] for $p \leq 0.05$.

5.2 Limitations

Since the proposed salient region detection method is based on global scheme, the computation cost is high. Suppose there are N pixels in an image, the computational

complexity is proportional to $O(N^2)$. In Table 3, we give the average running time of our approach and the others on the benchmark videos.

For the proposed method, the motion saliency is computed based on the assumptions given in Section 3.2. If the videos with strong motion contrast do not comply with the assumptions, the computed saliency map will be incorrect. For example, when the salient motion object accounts for a large proportion of the scene, the resulting saliency map will highlight the background instead of the salient object. An example is shown in Figure 8a. In addition, according to the fusion scheme, the final saliency detection result is mainly determined by static saliency if the scene's motion contrast is lower. The static saliency detection method itself cannot produce good result when the scene has the highly texture background. An example is shown in Figure 8b.

Table 3 Comparison of running times

Method	PQFT [14]	Kim et al. [13]	Liu et al. [11]	Ours
Times (s)	0.28	2.46	3.30	48.26
Code	Matlab	Matlab	Matlab	Matlab

Image's resolution is resized to **96 × 96**. Algorithms are tested using a Dual-Core 3.2-GHz machine with 4-GB RAM.

6 Conclusion

In this article, we propose a novel spatiotemporal salient region detection framework based on global scheme. The saliency maps are calculated separately by using the static and motion information of the videos. In the spatial attention model, we adopt joint embedding of spatial and color cues. The pixel-level saliency map is computed by using three components which are SC, CD, and SD. In the temporal attention model, the dense optical flow is used to calculate the global motion contrast of object in dynamic scene. To suppress the produced noise while estimating optical flow, a novel HOAOF is proposed to measure the motion contrast. To achieve the final spatiotemporal saliency map, an adaptive fusion scheme is adopted to combine the spatial and the temporal saliency. The dynamic weights of the two individual components are controlled by the motion entropy of the video frames. Extensive experiments show that the proposed method can obtain higher quality salient region detection results than existing methods no matter whether the camera is fixed or not.

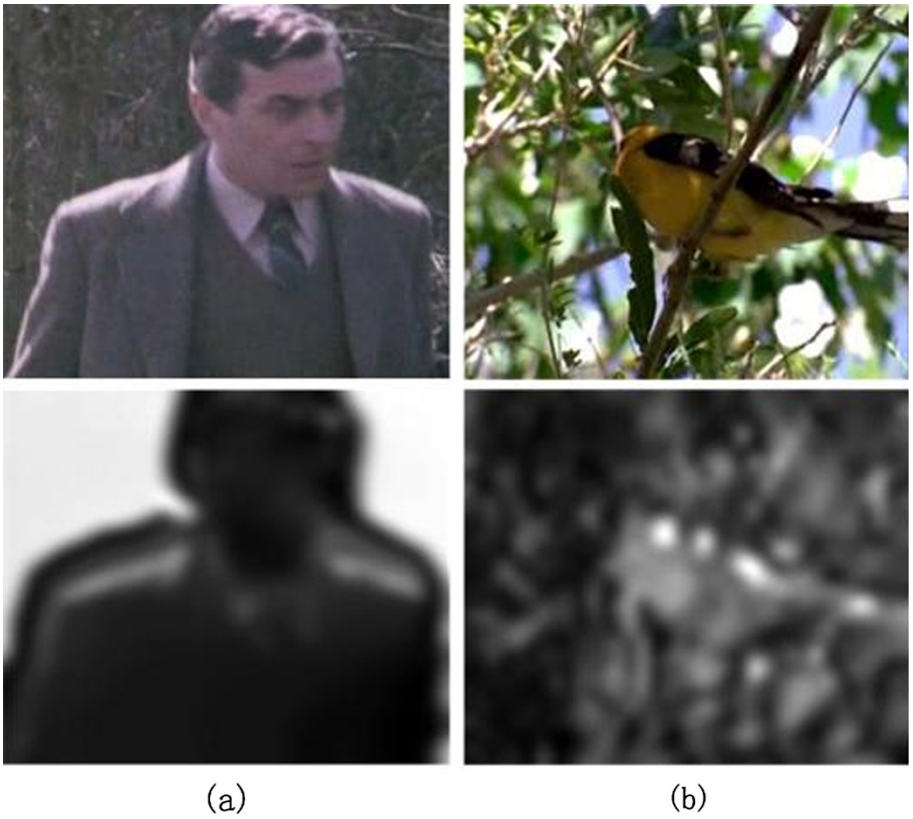

Figure 8 Failure cases. Original images are shown in row 1. The corresponding saliency maps are shown in row 2.One failure case is shown in (**a**), and another failure case is shown in (**b**)" in the end of figure caption.

Endnotes

[a]Code is available from http://people.csail.mit.edu/celiu/OpticalFlow/.

[b]ftp://ftp.pets.rdg.ac.uk/pub/PETS2001.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This study was supported by the National Natural Science Foundation of China (No. 61179060) and by the grants from the Fundamental Research Funds for the Central Universities (No. ZYGX2012J019).

Received: 1 June 2012 Accepted: 12 March 2013
Published: 15 April 2013

References

1. S Frintrop, E Rome, H Christensen, Computational visual attention systems and their cognitive foundations: a survey. ACM Trans. Appl. Perception (TAP). **7**(1), 1–39 (2010)
2. F Meng, H Li, G Liu, K Ngan, Object co-segmentation based on shortest path algorithm and saliency model. IEEE Trans. Multimed. **14**(5), 1429–1441 (2012)
3. C Jung, C Kim, A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. IEEE Trans. Image Process. **21**(3), 1272–1283 (2012)
4. H Li, K Ngan, Saliency model-based face segmentation and tracking in head-and-shoulder video sequences. J. Visual Commun. Image Represent. **19**(5), 320–333 (2008)
5. D Gao, N Vasconcelos, in *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, Vol. 2. Integrated learning of saliency, complex features, and object detectors from cluttered scenes (Los Alamitos, CA, USA, 2005), pp. 282–287
6. H Li, K Ngan, A co-saliency model of image pairs. IEEE Trans. Image Process. **20**(12), 3365–3375 (2011)
7. W Cheng, C Wang, J Wu, Video adaptation for small display based on content recomposition. IEEE Trans. Circuits Syst. Video Technol. **17**(1), 43–58 (2007)
8. V Mahadevan, N Vasconcelos, in *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, vol. 5. Background subtraction in highly dynamic scenes (Piscataway, NJ, USA, 2008), pp. 1–6
9. K Liu, Prediction error preprocessing for perceptual color image compression. EURASIP J. Image Video Process. **2012**(1), 1–14 (2012)
10. L Itti, C Koch, E Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
11. T Liu, Z Yuan, J Sun, J Wang, N Zheng, X Tang, H Shum, Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2011)
12. L Itti, N Dhavale, F Pighin, in *SPIE*, vol. 5200. Realistic avatar eye and head animation using a neurobiological model of visual attention (San Diego, CA, USA, 2003), pp. 64–78
13. W Kim, C Jung, C Kim, Spatiotemporal saliency detection and its applications in static and dynamic scenes. IEEE Trans. Circuits Syst. Video Technol. **21**(4), 446–456 (2011)
14. C Guo, L Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process. **19**(1), 185–198 (2010)
15. W Luo, H Li, G Liu, K Ngi Ngan, Global salient information maximization for saliency detection. Signal Process.: Image Commun. **27**(3), 238–248 (2011)
16. Y Zhai, M Shah, in *Proceedings of the 14th annual ACM international conference on Multimedia*, vol. 1. Visual attention detection in video sequences using spatiotemporal cues (New York, NY, USA, 2006), pp. 815–824
17. X Hou, L Zhang, in *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, vol. 3. Saliency detection: a spectral residual approach (Piscataway, NJ, USA, 2007), pp. 1–8
18. R Achanta, S Hemami, F Estrada, S Susstrunk, in *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, vol. 2. Frequency-tuned salient region detection (Piscataway, NJ, USA, 2009), pp. 1597–1604
19. H Li, K Ngan, Unsupervised video segmentation with low depth of field. IEEE Trans. Circuits Syst. Video Technol. **17**(12), 1742–1751 (2007)
20. L Xu, H Li, Z Wang, in *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 1. Saliency detection from joint embedding of spatial and color cues (Piscataway, NJ, USA, 2012), pp. 2673–2676
21. L Xu, H Li, L Zeng, KN Ngan, Saliency detection using joint spatial-color constraint and multi-scale segmentation. J. Visual Commun. Image Represent. **24**(4), 465–476 (2013)
22. M Cerf, J Harel, W Einhäuser, C Koch, in *Advances in Neural Information Processing Systems*, vol. 20. Predicting human gaze using low-level saliency combined with face detection (New York, NY, USA, 2008), pp. 241–248
23. T Judd, K Ehinger, F Durand, A Torralba, in *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 5. Learning to predict where humans look (Piscataway, NJ, USA, 2009), pp. 2106–2113
24. J Li, D Xu, W Gao, Removing label ambiguity in learning-based visual saliency estimation. IEEE Trans. Image Process. **21**(4), 1513–1525 (2012)
25. R Kountchev, K Nakamatsu, *Advances in Reasoning-Based Image Processing Intelligent Systems: Conventional and Intelligent Paradigms*, vol. 29 (Springer, New York, 2012)
26. O Le Meur, D Thoreau, P Le Callet, D Barba, in *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 3. A spatio-temporal model of the selective human visual attention (Piscataway, NJ, USA, 2005), pp. 1–4
27. J You, G Liu, H Li, A novel attention model and its application in video analysis. Appl. Math. Comput. **185**(2), 963–975 (2007)
28. A Bur, P Wurtz, R Miiri, H Hugli, in *Proceedings of the International Conference on Computer Vision Systems*, vol. 1. Dynamic visual attention: competitive versus motion priority scheme (Bielefeld, Germany, 2007), pp. 1–10
29. H Li, L Xu, G Liu, Two-layer average-to-peak ratio based saliency detection. Signal Process.: Image Commun. **28**(1), 55–68 (2013)
30. W Freeman, E Adelson, C Liu, et al., Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009)
31. T Brox, A Bruhn, N Papenberg, J Weickert, in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 4. High accuracy optical flow estimation based on a theory for warping (Prague, Czech Republic, 2004), pp. 25–36
32. A Bruhn, J Weickert, C Schnörr, Lucas/kanade meets horn/schunk: combining local and global optical flow methods. Int. J. Comput. Vision. **61**(3), 211–231 (2005)
33. K Fukuchi, K Miyazato, A Kimura, S Takagi, J Yamato, in *Proceedings of the IEEE Conference on Multimedia and Expo (ICME)*, vol. 4. Saliency-based video segmentation with graph cuts and sequentially updated priors (Piscataway, NJ, USA, 2009), pp. 638–641
34. H Seo, P Milanfar, Static and space-time visual saliency detection by self-resemblance. J. Vis. **9**(12), 1–27 (2009)
35. A Yeh, in *Proceedings of the 18th Conference on Computational Linguistics*, vol. 2. More accurate tests for the statistical significance of result differences (Saarbrücken, Germany, 2000), pp. 947–953
36. V Courboulay, MPD Silva, in *Optics, Photonics, and Digital Technologies for Multimedia Applications*, vol. 8436. Real-time computational attention model for dynamic scenes analysis: from implementation to evaluation (Brussels, France, 2012), pp. 1–15

37. H Li, KN Ngan, Q Liu, Faceseg: automatic face segmentation for real-time video. IEEE Trans. Multimed. **11**(1), 77–88 (2009)
38. M Cheng, G Zhang, N Mitra, X Huang, S Hu, in *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, vol. 2. Global contrast based salient region detection (Piscataway, NJ, USA, 2011), pp. 409–416

doi:10.1186/1687-5281-2013-16
Cite this article as: Wu *et al.*: A unified framework for spatiotemporal salient region detection. *EURASIP Journal on Image and Video Processing* 2013 **2013**:16.