Superpixel-Based Spatiotemporal Saliency Detection

Zhi Liu, *Member, IEEE*, Xiang Zhang, Shuhua Luo, and Olivier Le Meur

***Abstract*— This paper proposes a superpixel-based spatiotemporal saliency model for saliency detection in videos. Based on the superpixel representation of video frames, motion histograms and color histograms are extracted at the superpixel level as local features and frame level as global features. Then, superpixel-level temporal saliency is measured by integrating motion distinctiveness of superpixels with a scheme of temporal saliency prediction and adjustment, and superpixel-level spatial saliency is measured by evaluating global contrast and spatial sparsity of superpixels. Finally, a pixel-level saliency derivation method is used to generate pixel-level temporal and spatial saliency maps, and an adaptive fusion method is exploited to integrate them into the spatiotemporal saliency map. Experimental results on two public datasets demonstrate that the proposed model outperforms six state-of-the-art spatiotemporal saliency models in terms of both saliency detection and human fixation prediction.**

***Index Terms*— Motion vector field, spatial saliency, spatiotemporal saliency detection, superpixel, temporal saliency.**

I. Introduction

HUMANS can effortlessly identify salient objects even in a complex dynamic scene by exploiting the inherent visual attention mechanism. The research on computational models for saliency detection is originally motivated by simulating human visual attention mechanism, and a number of saliency models based on different schemes and theories have been proposed in past decades. Saliency maps generated using saliency models are used for human fixation prediction and also have been widely exploited in a variety of applications including salient object detection [1], [2] and segmentation [3], [4] from image/video, content-aware image/video retargeting [5], [6], content-based image/video compression [7], [8], and image/video quality assessment [9], [10].

A good deal of research effort has already been devoted to saliency models for images, while the research on spatiotemporal saliency models for videos has increasingly received attention in the recent years. Although image saliency models can be used for saliency detection independently in each video frame, their saliency detection performance is generally lower than spatiotemporal saliency models, which also considers the temporal information between the video frames. This paper focuses on saliency detection in videos and the following will mainly review the state-of-the-art spatiotemporal saliency models proposed in recent years.

As a pioneer work, Itti *et al.* [11] proposed a well-known saliency model based on the biologically plausible visual attention architecture [12] and the feature integration theory [13]. This model first computes feature maps of luminance, color, and orientation using a center-surround operator across different scales, and then performs normalization and summation to generate the saliency map, which highlights regions showing high local contrast with their surrounding regions in terms of any of the three features. The center-surround scheme exploited in Itti's model [11] has a clear interpretation of visual attention mechanism and a concise computation form; thus, a number of spatiotemporal saliency models implement this scheme using different features and formulations. Itti *et al.* [14] further introduced the features of flicker and motion energy in their surprise model for saliency detection in videos. Akin to Itti's model, a multiscale background (MB) model [10] represented using Gaussian pyramid is exploited to identify salient pixels, whose changes differ significantly from the changes undergone by most of the pixels in the video frame. As a general paradigm, the difference between each patch/volume and its spatiotemporal surroundings is evaluated using the Kullback-Leibler divergence on dynamic texture feature [15] under the discriminant center-surround hypothesis [16], the local regression kernel-based self-resemblance (SR) [17] and the earth mover's distance between the weighted histograms [18], and then is assigned as the spatiotemporal saliency measure to the center pixel of the patch/volume.

Recently, various formulations for measuring spatiotemporal saliency have been proposed based on different theories and methods such as information theory, control theory, frequency domain analysis, machine learning, and low-rank decomposition. Based on information theory,

Manuscript received July 30, 2013; revised November 1, 2013 and January 10, 2014; accepted February 13, 2014. Date of publication February 26, 2014; date of current version August 31, 2014. This work was supported in part by the National Natural Science Foundation of China under Grants 61171144 and 61105001, in part by the Shanghai Natural Science Foundation under Grant 11ZR1413000, in part by the Innovation Program of Shanghai Municipal Education Commission under Grant 12ZZ086, in part by the Key (Key Grant) Project of Chinese Ministry of Education under Grant 212053, and in part by a Marie Curie International Incoming Fellowship within the 7th European Community Framework Programme under Grant 299202. This paper was recommended by Associate Editor Y. Fu.

Z. Liu is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with IRISA/INRIA-Rennes, Rennes 35042, France (e-mail: liuzhisjtu@163.com).

X. Zhang is with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: uesthero@gmail.com).

S. Luo is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: taokenyizhong@163.com).

O. Le Meur is with the University of Rennes 1, Rennes 35042, France (e-mail: olemeur@irisa.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2014.2308642

self-information [19], minimum conditional entropy (CE) [20] and incremental coding length with different formulations [21], [22] are calculated on the basis of patch/volume and used as spatiotemporal saliency measures. In the control theory-based models, the video sequence is first represented using the state space model of linear system, and then either observability [23] or controllability [24] of the linear system is exploited to define spatiotemporal saliency measure for discrimination between the salient object and background motions. Using frequency domain analysis methods, the phase spectrum of quaternion Fourier transform (QFT) on the 4-D feature space composed of luminance, two chrominance components and frame difference [7], and the spectral residual of the amplitude spectrum of Fourier transform [25] on temporal slices at the horizontal/vertical direction [26] is exploited to generate the spatiotemporal saliency map. Using eye tracking data as the training set, machine learning methods such as probabilistic multitask learning [27], support vector regression [28] and support vector machine with Gaussian kernels [29] are exploited to build spatiotemporal saliency models. Using a low-rank decomposition method, the matrix composed by temporally aligned video frames [30] or temporal slices [31] is decomposed into a low-rank matrix for background and a sparse one for salient objects.

Due to the difference of nature between temporal and spatial domains, some models first measure temporal saliency and spatial saliency, respectively, and then combine them using linear or nonlinear fusion schemes to generate a spatiotemporal saliency map. In [32], spatial saliency is measured using the center-surround difference on ordinal signatures of edge and color orientation histograms, and temporal saliency is measured using temporal gradient difference. Based on sparse representation of each video frame, spatial saliency is evaluated using entropy gain, and temporal saliency is evaluated based on temporal difference and temporal consistency in [33]. Similarly, in [34], using sparse representation method, sparse reconstruction at the patch level in both temporal and spatial domains is exploited to measure temporal and spatial saliency, respectively. Since motion in most videos plays an important role on drawing human visual attention, the motion vector field is exploited as an explicit temporal feature in some saliency models. In [35], based on pixel correspondences estimated using motion vectors, the saliency map of each video frame is predicted by the result of its previous frame. By analysis of motion vector field, directionally consistent motion vectors are accumulated to identify salient motion in [36], and global motion compensation is performed to obtain residual motion vectors, whose amplitudes [37] or the weighted amplitudes by global median filtering [38] are used as temporal saliency measures. Besides, global motion parameters are exploited to generate position-based temporal saliency map in [39], and distributions of motion orientation and amplitude are combined to measure temporal saliency in [40]. In the above models, interactions between the cortical-like filters [37], contrast sensitivity function in the frequency domain [38], and center-surround scheme on spatial features [39], [40] are exploited to measure spatial saliency, accordingly. In an effort to obtain motion vector fields suitable for measuring saliency [41], a dynamic consistent (DC) optical flow estimation method is proposed to pop out the consistent motion of the prominent object with enough amplitude for generating the temporal saliency map, on which a pixel-wise maximum operation is performed with the spatial saliency map generated using the graph-based saliency model [42] to obtain the spatiotemporal saliency map.

Although a number of spatiotemporal saliency models as mentioned above have been proposed for saliency detection in videos, we observed that the existing models are insufficient to effectively highlight the complete salient objects with well-defined boundaries and suppress background regions. We also found that the existing spatiotemporal saliency models measure saliency at either regular-shaped patch/volume level, temporal slice level or pixel level, which neglects the factor of video representation for reasonable saliency measurement. Therefore, motivated by some recent works on image saliency models, in which saliency detection performance is effectively elevated by measuring saliency on the basis of over-segmented regions [4], [43]–[45] or superpixels [46]–[52], we propose a superpixel-based (SP) spatiotemporal saliency model aiming to improve saliency detection performance for videos. Our main contributions are threefold. First, we propose to jointly use local superpixel-level and global frame-level features in both temporal and spatial domain as the basis of our spatiotemporal saliency model. Second, at superpixel level, we propose to measure temporal saliency based on motion distinctiveness of superpixels and a scheme of temporal saliency prediction and adjustment, and integrate global contrast with spatial sparsity of superpixels to measure spatial saliency. Finally, we propose a pixel-level saliency derivation method and an adaptive fusion method to generate the pixel-level spatiotemporal saliency map. Subjective observations and objective evaluations demonstrate that the proposed model achieves the better performance on both saliency detection and human fixation prediction compared with six state-of-the-art spatiotemporal saliency models.

The rest of this paper is organized as follows. The proposed model is described in Section II, experimental results and analysis are presented in Section III, and the conclusion is given in Section IV.

II. SP Spatiotemporal Saliency Model

The proposed SP spatiotemporal saliency model is shown in Fig. 1, which consists of five components marked using the dash-dotted lines, i.e., feature extraction, superpixel-level temporal saliency, superpixel-level spatial saliency, pixel-level temporal/spatial saliency derivation, and pixel-level spatiotemporal saliency generation. The following five sections from Section II-A to E will elaborate the five components in turn.

In this paper, we build the proposed model on the superpixel representation of input video frame, because the features extracted at superpixels, which are over-segmented regions and fit well to the boundaries between the salient objects and background regions, are more meaningful than block-level and pixel-level features. In addition, the superpixel representation facilitates to preserve the well-defined object boundaries in the

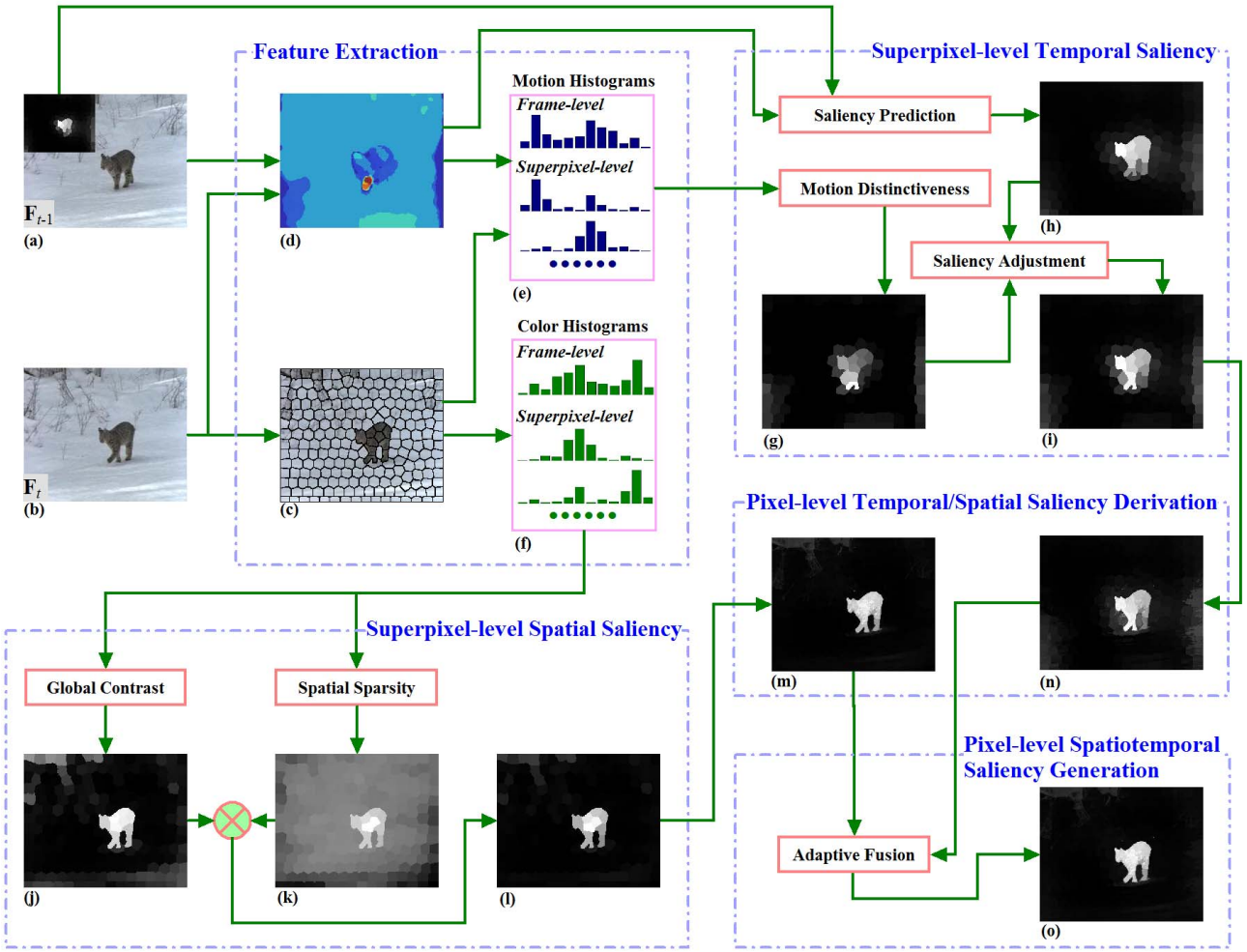


Fig. 1. Pictorial illustration of SP spatiotemporal saliency model. (a) Previous video frame $\mathbf{F}_{t-1}$ with its superpixel-level temporal saliency map at the top-left corner. (b) Current video frame $\mathbf{F}_t$. (c) Superpixel representation for $\mathbf{F}_t$. (d) Quantized motion vector field for $\mathbf{F}_t$. (e) and (f) Motion and color histograms, respectively, for $\mathbf{F}_t$. (g)–(i) Superpixel-level motion distinctiveness map, temporal saliency prediction map and temporal saliency map, respectively, for $\mathbf{F}_t$. (j)–(l) Superpixel-level global contrast map, spatial sparsity map and spatial saliency map, respectively, for $\mathbf{F}_t$. (m) and (n) Pixel-level spatial saliency map and temporal saliency map, respectively, for $\mathbf{F}_t$. (o) Pixel-level spatiotemporal saliency map for $\mathbf{F}_t$.

generated saliency map as well. In Section II-A, we extract motion and color histograms at superpixel level, which can effectively represent the local distribution of motion and color, and also extract both histograms at frame level to represent the global distribution of motion and color. In a comparative sense, superpixel-level and frame-level motion/color histograms are used as local features and global features, respectively, for saliency measurement.

Using the extracted motion and color histograms, we measure the temporal saliency and spatial saliency for each superpixel by taking into account the difference between its superpixel-level and the frame-level histogram as well as its similarities with other superpixels in both temporal and spatial domain. Specifically, we propose to evaluate the motion distinctiveness of superpixels and exploit a scheme of temporal saliency prediction and adjustment to measure the superpixel-level temporal saliency in Section II-B, and jointly evaluate the global contrast and spatial sparsity of superpixels to measure the superpixel-level spatial saliency in Section II-C.

To generate the saliency map with pixel-level accuracy, a pixel-level saliency derivation method is used to transform the superpixel-level temporal/spatial saliency measures into the pixel-level temporal/spatial saliency map in Section II-D. Finally, an adaptive fusion method is exploited to integrate the temporal saliency map with the spatial saliency map at pixel-level for generating the spatiotemporal saliency map in Section II-E.

A. Feature Extraction

Each input video frame is partitioned into a set of superpixels, which usually have more regular and compact shape with better boundary adherence compared with the commonly segmented regions generated using conventional image segmentation approaches. Superpixels are used as the primitives for the following feature extraction of superpixel-level motion/color histogram and effective temporal/spatial saliency measurement.

Specifically, each video frame $\mathbf{F}_t$ is first transformed into the *Lab* color space, in which luminance channel and

two chrominance channels are separated. Then the simple linear iterative clustering (SLIC) algorithm [53] is used to partition the video frame into a number of superpixels $\mathrm{sp}_t^i (i = 1, \ldots, n_t)$, where n_t is the number of the generated superpixels. If not otherwise stated, in the following related mathematical notations, the temporal index is put at the subscript such as t in sp_t^i, and the spatial index is put at the superscript such as i in sp_t^i. Note that, the notations may omit the superscript representing the spatial index, such as sp_t, which denotes all superpixels in the frame $\mathbf{F}_t$.

The parameter of the starting size of superpixels in SLIC is set to $\sqrt{w \cdot h/200}$, where w and h denote frame's width and height, respectively, and thus n_t is approximately 200, which is generally sufficient to preserve different boundaries well. Two consecutive video frames $\mathbf{F}_{t-1}$ and $\mathbf{F}_t$ are shown in Fig. 1(a) and (b), and the superpixel representation result of $\mathbf{F}_t$ is shown in Fig. 1(c), in which the boundaries between the adjacent superpixels are delineated using black lines.

1) Motion Histogram: Motion histograms are extracted at superpixel level as local features and frame level as global features for temporal saliency measurement. For each frame $\mathbf{F}_t$, using its previous frame $\mathbf{F}_{t-1}$ as the reference frame, the pixel-level motion vector field $\mathbf{MVF}_t$ is calculated using the optical flow estimation method in [54]. Then, amplitudes and orientations of all motion vectors in $\mathbf{MVF}_t$ are uniformly quantized into 10 and 16 intervals, respectively, which are sufficient for quantization of motion vector fields estimated between the adjacent frames, to generate a motion quantization table $\mathbf{MQ}_t$ with $q_M = 160$ entries. For the video frames in Fig. 1(a) and (b), the quantized motion vector field $\mathbf{MVF}_t$ is visualized using Fig. 1(d), in which a hotter color denotes a higher motion amplitude. Based on $\mathbf{MQ}_t$, the frame-level motion histogram $\mathbf{MH}_t^0$ for $\mathbf{F}_t$ is calculated using all motion vectors in $\mathbf{MVF}_t$, and then normalized to have $\sum_{k=1}^{q_M} \mathbf{MH}_t^0(k) = 1$. The quantized motion vector for each bin, $\mathbf{qmv}_t(k)$, is calculated as the mean of motion vectors that fall into the kth bin of $\mathbf{MH}_t^0$. Similarly, on the basis of each superpixel $\mathrm{sp}_t^i (i = 1, \ldots, n_t)$, the superpixel-level motion histogram $\mathbf{MH}_t^i (i = 1, \ldots, n_t)$ is calculated and normalized to have $\sum_{k=1}^{q_M} \mathbf{MH}_t^i(k) = 1$. Note that, the superscript 0 in $\mathbf{MH}_t^0$ is not a spatial index. Since the superscript i in superpixel-level motion histograms $\mathbf{MH}_t^i$ represents the spatial index and starts from 1, we use the superscript 0 to denote the frame-level motion histogram as $\mathbf{MH}_t^0$ so as to keep the consistent format of notations for motion histograms. The similar explanation is applicable to the notation for frame-level color histogram defined in Section II-A2.

2) Color Histogram: Color histograms are also extracted at superpixel level as local features and frame level as global features. For each frame $\mathbf{F}_t$ in the *Lab* color space, each color channel is uniformly quantized into q_b bins to generate a color quantization table $\mathbf{CQ}_t$ with $q_C = q_b \times q_b \times q_b$ bins. The parameter q_b is set to a moderate value, 16, which is generally sufficient for color quantization of video frames. Based on $\mathbf{CQ}_t$, the frame-level color histogram $\mathbf{CH}_t^0$ is calculated using all pixels in $\mathbf{F}_t$ and normalized to have $\sum_{k=1}^{q_C} \mathbf{CH}_t^0(k) = 1$. The quantized color of each bin, $\mathbf{qc}_t(k)$, is calculated as the mean color of those pixels that fall into the k^{th} bin of $\mathbf{CH}_t^0$. Similarly, for each superpixel $\mathrm{sp}_t^i (i = 1, \ldots, n_t)$, the superpixel-level color histogram $\mathbf{CH}_t^i$ is calculated and normalized to have $\sum_{k=1}^{q_C} \mathbf{CH}_t^i(k) = 1$.

B. Superpixel-Level Temporal Saliency

We evaluate the temporal saliency of each video frame based on the following two observations: 1) the motion of salient object is generally distinctive from background regions in videos, and we can exploit the motion vector field of each frame to evaluate the motion distinctiveness of superpixels; and 2) the motion of salient object as well as the motion induced by the camera movement are usually coherent during a period of frames, and thus it is reasonable that the temporal saliency measures of superpixels should also keep the temporal coherence.

Therefore, for each superpixel in the current frame, we exploit the temporal saliency measures of some correlated superpixels in the previous frame to predict its temporal saliency, so as to keep the coherence of temporal saliency between adjacent frames, and then use the motion distinctiveness of the current superpixel to adjust such a temporal saliency prediction for generating the temporal saliency measure of the current superpixel. The following describes the definition of motion distinctiveness, temporal saliency prediction and temporal saliency adjustment in Sections II-B1–3, respectively, and finally illustrates the overall superpixel-level temporal saliency measurement process in Section II-B4.

1) Motion Distinctiveness: If motion exhibits at each frame $\mathbf{F}_t$, those superpixel-level motion histograms associated with salient object regions generally show a higher difference with the frame-level motion histogram. Therefore, the motion distinctiveness for each superpixel sp_t^i is defined by evaluating the difference between $\mathbf{MH}_t^i$ and $\mathbf{MH}_t^0$ as

$$\mathbf{S_{MD}}(\mathrm{sp}_t^i) = \sum_{j=1}^{q_M} \left[\mathbf{MH}_t^i(j) \sum_{k=1}^{q_M} \left\| \mathbf{qmv}_t(j) - \mathbf{qmv}_t(k) \right\|_2 \cdot \mathbf{MH}_t^0(k) \right]. \quad (1)$$

Using (1), the difference between $\mathbf{MH}_t^i$ and $\mathbf{MH}_t^0$ is measured as a sum of distances between different quantized motion vectors weighted by their occurrence probabilities. Based on superpixel-level motion distinctiveness measures, the frame-level motion significance for each frame $\mathbf{F}_t$ is then defined as

$$\mathrm{MS}_t = \sum_{i=1}^{n_t} \mathbf{S_{MD}}\left(\mathrm{sp}_t^i\right) \cdot \left|\mathrm{sp}_t^i\right| \quad (2)$$

where $|\mathrm{sp}_t^i|$ denotes the number of pixels in sp_t^i.

2) Temporal Saliency Prediction: Given the superpixel-level temporal saliency measures $\mathbf{S_T}(\mathrm{sp}_{t-1})$ of the frame $\mathbf{F}_{t-1}$, the motion vector field $\mathbf{MVF}_t$, which represents the motion between $\mathbf{F}_{t-1}$ and $\mathbf{F}_t$ is used to find for each current superpixel sp_t^i its correlated superpixels in $\mathbf{F}_{t-1}$, and then the temporal saliency measures of these correlated superpixels are exploited to predict the temporal saliency of sp_t^i.

As shown in Fig. 2, for each superpixel sp_t^i in the frame $\mathbf{F}_t$, the projected region for sp_t^i in the frame $\mathbf{F}_{t-1}$ is determined

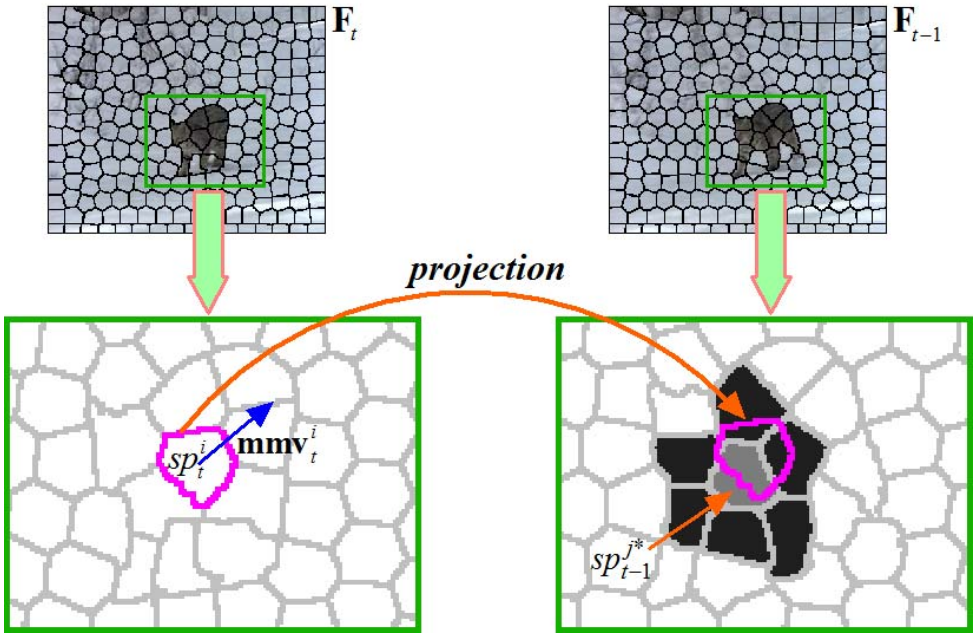


Fig. 2. Superpixel projection and selection of correlated superpixels.

using $\mathbf{mmv}_t^i$, which is the mean of pixel-level motion vectors in sp_t^i, as the displacement. Then in $\mathbf{F}_{t-1}$, the most matched superpixel for sp_t^i is selected as the one with the shortest distance to the projected region for sp_t^i using the following criterion:

$$\mathrm{sp}_{t-1}^{j*} = \arg \min_{\mathrm{sp}_{t-1}^j \in \mathbf{F}_{t-1}} \left\| \boldsymbol{\mu}_t^i + \mathbf{mmv}_t^i - \boldsymbol{\mu}_{t-1}^j \right\|_2 \tag{3}$$

where μ_t^i and μ_{t-1}^j denote the spatial center position of sp_t^i and sp_{t-1}^j, respectively. As shown in the bottom-right part of Fig. 2, the most matched superpixel sp_{t-1}^{j*} (filled with the light color) and its adjacent superpixels (filled with the dark color) are selected as the correlated superpixels for sp_t^i, to constitute a superpixel set denoted by Ψ_t^i.

The interframe similarity between the current superpixel sp_t^i and each of its correlated superpixel sp_{t-1}^j in Ψ_t^i is defined as

$$\lambda_{\mathrm{inter}}(\mathrm{sp}_t^i, \mathrm{sp}_{t-1}^j) = \sum_{k=1}^{qc} \sqrt{\mathbf{CH}_t^i(k) \cdot \mathbf{CH}_{t-1}^j(k)} \cdot \left[\mathbf{1} - \frac{\left\| \boldsymbol{\mu}_t^i + \mathbf{mmv}_t^i - \boldsymbol{\mu}_{t-1}^j \right\|_2}{d} \right] \tag{4}$$

where d denotes the diagonal length of video frame. In (4), the former term is the Bhattacharyya coefficient between the corresponding superpixel-level color histograms, $\mathbf{CH}_t^i$ and $\mathbf{CH}_{t-1}^j$, and the latter term is the distance factor, which weights more to the correlated superpixel nearer to the projected region for sp_t^i. Based on the temporal saliency measures of the correlated superpixels in Ψ_t^i and their interframe similarities with the current superpixel sp_t^i, the temporal saliency prediction for sp_t^i is defined as

$$\mathbf{S_{TP}}(\mathrm{sp}_t^i) = \frac{\sum_{\mathrm{sp}_{t-1}^j \in \Psi_t^i} \lambda_{\mathrm{inter}}(\mathrm{sp}_t^i, \mathrm{sp}_{t-1}^j) \cdot \mathbf{S_T}(\mathrm{sp}_{t-1}^j)}{\sum_{\mathrm{sp}_{t-1}^j \in \Psi_t^i} \lambda_{\mathrm{inter}}(\mathrm{sp}_t^i, \mathrm{sp}_{t-1}^j)}. \tag{5}$$

3) Temporal Saliency Adjustment: The motion distinctiveness measures of superpixels are used to adjust the temporal saliency predictions for superpixels in the frame $\mathbf{F}_t$, by considering the change of frame-level motion significance measures in a period of frames, to generate the superpixel-level temporal saliency measures in $\mathbf{F}_t$ as

$$\mathbf{S_T}(\mathrm{sp}_t) = \omega_t \cdot \mathbf{S_{TP}}(\mathrm{sp}_t) + (1 - \omega_t) \cdot \mathbf{S_{MD}}(\mathrm{sp}_t) \tag{6}$$

where the weight ω_t controls the adjustment effect. We exploit the comparison between the current motion significance measure MS_t and an equivalent motion significance measure for $\mathbf{S_{TP}}(\mathrm{sp}_t)$ to set the weight ω_t. A reasonable estimate of motion significance for $\mathbf{S_{TP}}(\mathrm{sp}_t)$ can be set as the median value of motion significance measures in several previous frames, and thus the weight ω_t is defined as

$$\omega_t = \max \left[\frac{\mathrm{med}(\mathrm{MS}_{t-f}, \ldots, \mathrm{MS}_{t-1})}{\mathrm{med}(\mathrm{MS}_{t-f}, \ldots, \mathrm{MS}_{t-1}) + \mathrm{MS}_t}, 0.5 \right] \tag{7}$$

where med(.) denotes the median operator. The parameter f determines the number of previous frames involved in the estimation of motion significance for $\mathbf{S_{TP}}(\mathrm{sp}_t)$. A small value of f will increase the fluctuation of such an estimate between the adjacent frames, but a large value of f cannot appropriately adapt to the frames with high motion. Therefore, in our implementation, f is set to a moderate value, 5, by the experiments, with the clamp factor, 0.5, which is used to limit the adjustment effect of $\mathbf{S_{MD}}(\mathrm{sp}_t)$ due to suddenly high motion in some frames, for the maintenance of temporal saliency coherence.

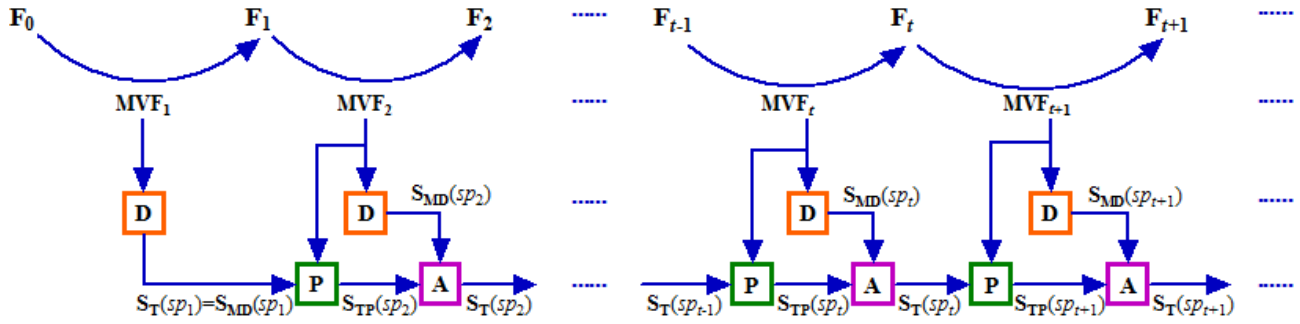
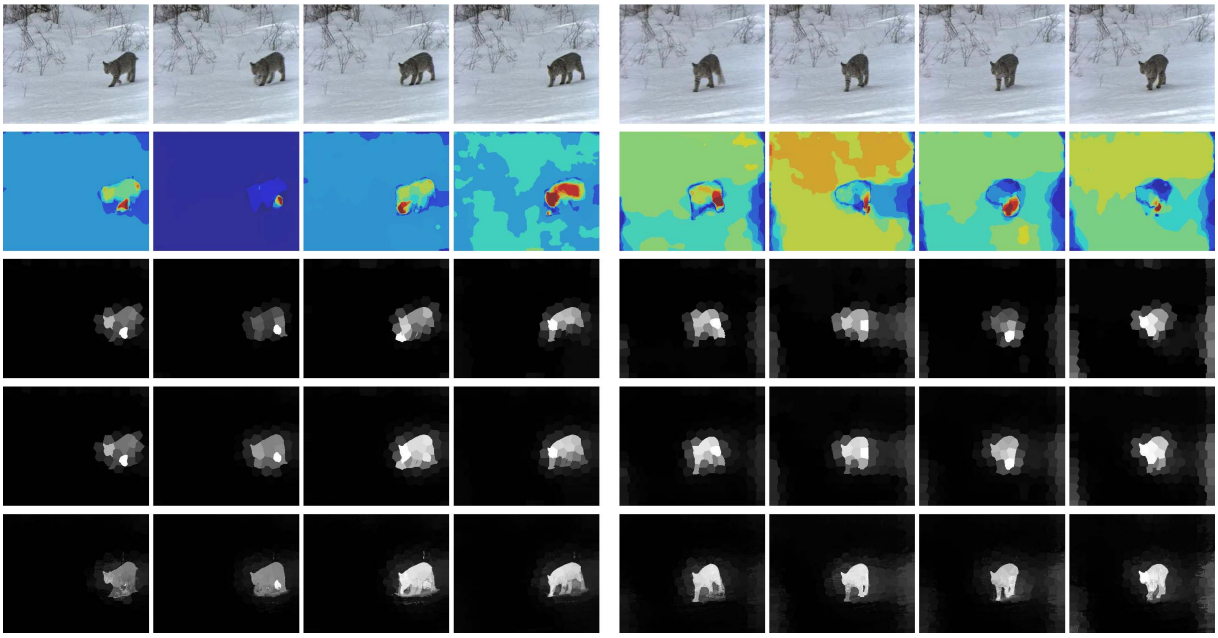


Fig. 3. Overall superpixel-level temporal saliency measurement process.



(a) (b)

Fig. 4. Examples of temporal saliency measurement process for (a) frames 1–4 and (b) frames 27–30 in the video sequence VWC102T. From top to bottom: video frames, the quantized motion vector fields, motion distinctiveness maps, superpixel-level temporal saliency maps, and pixel-level temporal saliency maps.

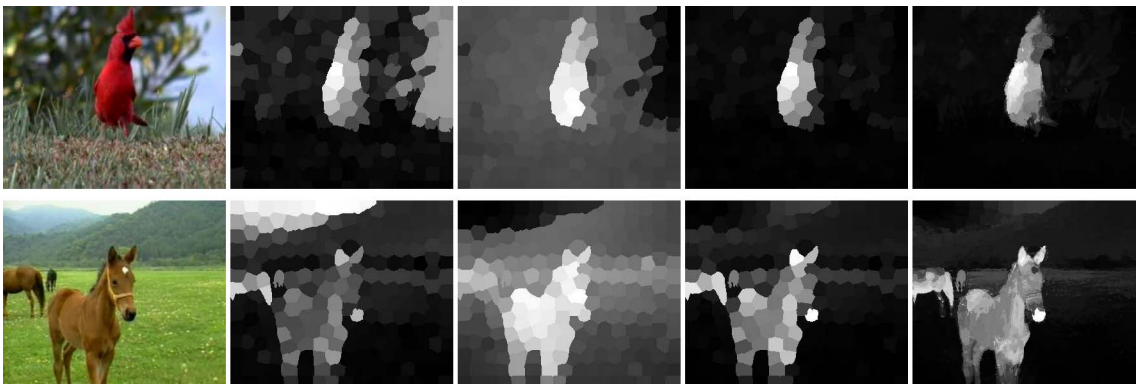
4) Overall Process: The pictorial examples of superpixel-level temporal saliency measurement for adjacent frames are shown in Fig. 1(g)–(i), and the scheme of the overall measurement process is shown in Fig. 3, in which the three types of functional blocks, i.e., D, P, and A denote the motion distinctiveness estimation, temporal saliency prediction and temporal saliency adjustment, respectively. It should be noted that at the beginning of the video sequence, the motion distinctiveness measure at the frame $\mathbf{F}_1$ is used as its temporal saliency measure, i.e., $\mathbf{S_T}(\mathrm{sp}_1) = \mathbf{S_{MD}}(\mathrm{sp}_1)$, to initiate the process of temporal saliency measurement.

Fig. 4 shows the results on several consecutive frames generated by the temporal saliency measurement process. Using the video frames and the quantized motion vector fields in Fig. 4 as reference, we can observe that the motion distinctiveness maps (the third row in Fig. 4) highlight some parts of the salient object, which show distinctive motion from the remaining regions. In contrast, the superpixel-level temporal saliency maps (the fourth row in Fig. 4) highlight the complete salient object more effectively and also show the better temporal coherence due to the use of temporal saliency prediction and adjustment.

C. Superpixel-Level Spatial Saliency

Besides temporal saliency induced by motion of salient objects, spatial saliency is evaluated on the basis of each video frame. In a variety of video scenes, salient objects usually show contrast with the surrounding background regions, and the spatial distribution of salient object colors is sparser than background colors, which usually scatter over the scene. Based on the above two observations, we evaluate the global contrast and spatial sparsity of superpixels using the frame-level and superpixel-level color histograms, as the basis for measuring the spatial saliency of superpixels.

Referring to the motion distinctiveness measure in (1), which is evaluated by comparing superpixel-level and frame-level motion histogram, the global contrast of each superpixel sp_t^i is defined by comparing the superpixel-level color his-



(a) (b) (c) (d) (e)

Fig. 5. Examples of spatial saliency generation process. (a) Video frames. (b) Global contrast maps. (c) Spatial sparsity maps. (d) Superpixel-level spatial saliency maps. (e) Pixel-level spatial saliency maps.

togram with the frame-level color histogram as

$$\mathbf{S_{GC}}(\mathrm{sp}_t^i) = \sum_{j=1}^{qC} \left[\mathbf{CH}_t^i(j) \sum_{k=1}^{qC} \left\| \mathbf{qc}_t(j) - \mathbf{qc}_t(k) \right\|_2 \cdot \mathbf{CH}_t^0(k) \right]. \tag{8}$$

It can be observed from (8) that the contrast between $\mathbf{CH}_t^i$ and $\mathbf{CH}_t^0$ is evaluated as a sum of distances between different quantized colors weighted by their occurrence probabilities.

Referring to the interframe similarity in (4), the intraframe similarity between each pair of superpixels, sp_t^i and sp_t^j, is defined as

$$\lambda_{\mathrm{intra}}(\mathrm{sp}_t^i, \mathrm{sp}_t^j) = \sum_{k=1}^{qC} \sqrt{\mathbf{CH}_t^i(k) \cdot \mathbf{CH}_t^j(k)} \cdot \left[1 - \frac{\left\| \mu_t^i - \mu_t^j \right\|_2}{d} \right]. \tag{9}$$

Using (9), $\lambda_{\mathrm{intra}}(\mathrm{sp}_t^i, \mathrm{sp}_t^j)$ is evaluated higher when the color distributions of sp_t^i and sp_t^j are similar and the spatial distance between them is shorter. For each superpixel sp_t^i, different colors in its color histogram $\mathbf{CH}_t^i$ may have different spatial distributions over the whole scene. Therefore, in terms of color histogram, the spatial spread of color distribution for sp_t^i is defined as

$$\mathbf{SD}(\mathrm{sp}_t^i) = \frac{\sum_{j=1}^{n_t} \lambda_{\mathrm{intra}}(\mathrm{sp}_t^i, \mathrm{sp}_t^j) \cdot \mathbf{D}(\mathrm{sp}_t^j)}{\sum_{j=1}^{n_t} \lambda_{\mathrm{intra}}(\mathrm{sp}_t^i, \mathrm{sp}_t^j)} \tag{10}$$

where $\mathbf{D}(\mathrm{sp}_t^j)$ denotes the Euclidean distance from the spatial center of sp_t^j to the spatial center of video frame. Then, an inverse normalization operation is performed on the above spatial spread measures to obtain the spatial sparsity measure for each superpixel as

$$\mathbf{S_{SS}}(\mathrm{sp}_t^i) = \frac{\max\left[\mathbf{SD}(\mathrm{sp}_t)\right] - \mathbf{SD}(\mathrm{sp}_t^i)}{\max\left[\mathbf{SD}(\mathrm{sp}_t)\right] - \min\left[\mathbf{SD}(\mathrm{sp}_t)\right]}. \tag{11}$$

Finally, by performing a superpixel-wise multiplication operation between the global contrast measure and spatial sparsity measure, the spatial saliency measure for each superpixel sp_t^i is defined as

$$\mathbf{S_S}(\mathrm{sp}_t^i) = \mathbf{S_{GC}}(\mathrm{sp}_t^i) \cdot \mathbf{S_{SS}}(\mathrm{sp}_t^i). \tag{12}$$

As shown in Figs. 1(j)–(l) and 5(a)–(d), global contrast map and spatial sparsity map can complement each other to generate superpixel-level spatial saliency map, which can highlight salient object regions and suppress background regions more effectively compared with either global contrast map or spatial sparsity map.

D. Pixel-Level Temporal/Spatial Saliency Derivation

To finally generate spatiotemporal saliency maps with pixel-level accuracy, pixel-level temporal/spatial saliency maps are derived from superpixel-level temporal/spatial saliency measures. For each pixel p_t^i at each frame $\mathbf{F}_t$, its temporal/spatial saliency measure $\mathbf{S_{T/S}}(p_t^i)$ (the subscript $\mathbf{T/S}$ actually denotes $\mathbf{T}$ for temporal saliency or $\mathbf{S}$ for spatial saliency) is defined as a sum of temporal/spatial saliency measures of those superpixels in its local neighborhood weighted by the pixel color's probabilities in the corresponding superpixel-level color histograms

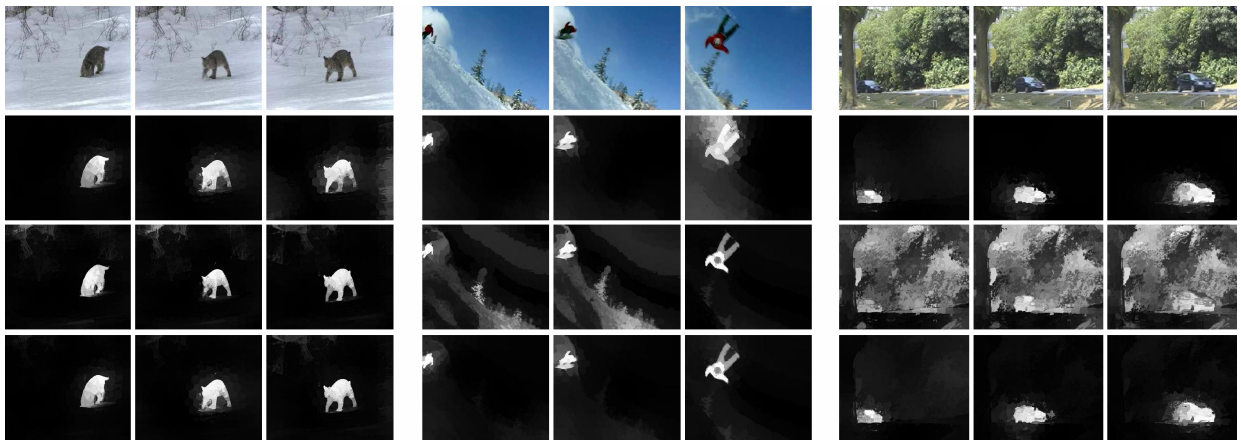
$$\mathbf{S_{T/S}}(p_t^i) = \frac{\sum_{\mathrm{sp}_t^j \in \mathrm{N}(p_t^i)} \mathbf{S_{T/S}}(\mathrm{sp}_t^j) \cdot \mathbf{CH}_t^j\left[\mathrm{bin}(p_t^i)\right]}{\sum_{\mathrm{sp}_t^j \in \mathrm{N}(p_t^i)} \mathbf{CH}_t^j\left[\mathrm{bin}(p_t^i)\right]} \tag{13}$$

where $\mathrm{N}(p_t^i)$ denotes the local neighborhood of p_t^i, and $\mathrm{bin}(p_t^i)$ denotes the entry number for the quantized color of p_t^i in the color quantization table $\mathbf{CQ}_t$. The local neighborhood $\mathrm{N}(p_t^i)$ includes the superpixel containing p_t^i and its adjacent superpixels.

As shown in Figs. 1(m)-(n), 5(e) and the bottom row of Fig. 4, the pixel-level temporal/spatial saliency maps can highlight salient object regions more uniformly with well-defined boundaries compared with the corresponding superpixel-level temporal/spatial saliency maps.

E. Pixel-Level Spatiotemporal Saliency Generation

The pixel-level spatiotemporal saliency map $\mathbf{S}(p_t)$ for each frame $\mathbf{F}_t$ is generated by adaptively fusing pixel-level temporal saliency map $\mathbf{S_T}(p_t)$ and spatial saliency map $\mathbf{S_S}(p_t)$.



(a) (b) (c)

Fig. 6. Examples of spatiotemporal saliency generation for three video examples with the frame interval of (a) and (b) 10 frames and (c) 5 frames. From top to bottom: original video frames, pixel-level temporal, spatial, and spatiotemporal saliency maps.

Based on $\mathbf{S_T}(p_t)$ and $\mathbf{S_S}(p_t)$, we first measure their mutual consistency as

$$\mathrm{MCT}_t = \frac{\sum_i \left[\mathbf{S_T}(p_t^i) \cdot \mathbf{S_S}(p_t^i)\right]}{\sum_i \mathbf{S_T}(p_t^i)}$$
$$\mathrm{MCS}_t = \frac{\sum_i \left[\mathbf{S_T}(p_t^i) \cdot \mathbf{S_S}(p_t^i)\right]}{\sum_i \mathbf{S_S}(p_t^i)}. \quad (14)$$

Specifically, MCT_t denotes the consistency of $\mathbf{S_T}(p_t)$ relative to $\mathbf{S_S}(p_t)$, and MCS_t denotes the consistency of $\mathbf{S_S}(p_t)$ relative to $\mathbf{S_T}(p_t)$. Both MCT_t and MCS_t fall into the range of [0, 1].

The proposed adaptive fusion scheme is based on collaborative interaction and selection between the temporal and spatial saliency maps, and thus the spatiotemporal saliency map $\mathbf{S}(p_t)$ is defined as

$$\mathbf{S}(p_t) = \gamma_t \cdot \mathbf{S_{int}}(p_t) + (1 - \gamma_t) \cdot \mathbf{S_{sel}}(p_t) \quad (15)$$

where the interaction term $\mathbf{S_{int}}(p_t)$ denotes the interaction between $\mathbf{S_T}(p_t)$ and $\mathbf{S_S}(p_t)$, and the selection term $\mathbf{S_{sel}}(p_t)$ denotes the selection between $\mathbf{S_T}(p_t)$ and $\mathbf{S_S}(p_t)$. The weight γ_t denotes the interaction confidence between $\mathbf{S_T}(p_t)$ and $\mathbf{S_S}(p_t)$, and is set to $\max(\mathrm{MCT}_t, \mathrm{MCS}_t)$, i.e., if either MCT_t or MCS_t indicates a higher consistency, the interaction term is given a higher weight. However, if both MCT_t and MCS_t show a lower consistency, a higher weight is then given to the selection term.

The interaction term $\mathbf{S_{int}}(p_t)$ is a combination of $\mathbf{S_T}(p_t)$ and $\mathbf{S_S}(p_t)$ proportional to their consistency measures

$$\mathbf{S_{int}}(p_t) = \frac{\mathrm{MCT}_t \cdot \mathbf{S_T}(p_t) + \mathrm{MCS}_t \cdot \mathbf{S_S}(p_t)}{\mathrm{MCT}_t + \mathrm{MCS}_t}. \quad (16)$$

The selection term $\mathbf{S_{sel}}(p_t)$ is determined based on a prior that salient objects are usually rarer than background regions and are surrounded by background regions. Therefore, for the same video frame, a saliency map with compact saliency distribution is generally more reasonable than one with wide spread of saliency distribution, in view of saliency detection. Let $\mathbf{D}(p_t^i)$ denote the Euclidean distance from the pixel p_t^i to the center position of saliency map, the saliency spread measures for $\mathbf{S_T}(p_t)$ and $\mathbf{S_S}(p_t)$ are defined as

$$\mathrm{SDT}_t = \sum_i \left[\mathbf{S_T}(p_t^i) \cdot \mathbf{D}(p_t^i)\right]$$
$$\mathrm{SDS}_t = \sum_i \left[\mathbf{S_S}(p_t^i) \cdot \mathbf{D}(p_t^i)\right]. \quad (17)$$

The selection term $\mathbf{S_{sel}}(p_t)$ is then defined as

$$\mathbf{S_{sel}}(p_t) = \begin{cases} \mathbf{S_T}(p_t), & \text{if } \mathrm{SDT}_t < \mathrm{SDS}_t \\ \mathbf{S_S}(p_t), & \text{otherwise.} \end{cases} \quad (18)$$

Fig. 6 shows the fusion of temporal and spatial saliency maps on three video examples. Compared with either temporal saliency maps or spatial saliency maps, in which some background regions are falsely highlighted and/or salient objects are not highlighted well for some frames [see the examples in Fig. 6(b) and (c)], the proposed adaptive fusion method can reasonably integrate temporal saliency maps with spatial saliency maps to generate spatiotemporal saliency maps, which can more effectively highlight salient objects and suppress background regions.

III. Experimental Results

A. Data Sets and Experimental Settings

We evaluated our SP spatiotemporal saliency model on two public video datasets, DS1 [55] and DS2 [56]. DS1 contains 10 uncompressed video clips of natural scenes with 12 fps, and the video length varies from 5 to 10 seconds. Besides, DS1 provides the pixel-level binary ground truths of salient objects for objectively evaluating the saliency detection performance of spatiotemporal saliency models. DS2 is known as abnormal surveillance crowd moving noise (ASCMN) dataset, which contains 24 videos with eye tracking data from 13 viewers. ASCMN dataset covers a wider spectrum of video types classified into five categories: abnormal motion, surveillance video, crowd motion, moving camera, and motion noise with sudden salient motion, which make it suitable for comprehensively evaluating the performance on human fixation prediction of spatiotemporal saliency models.

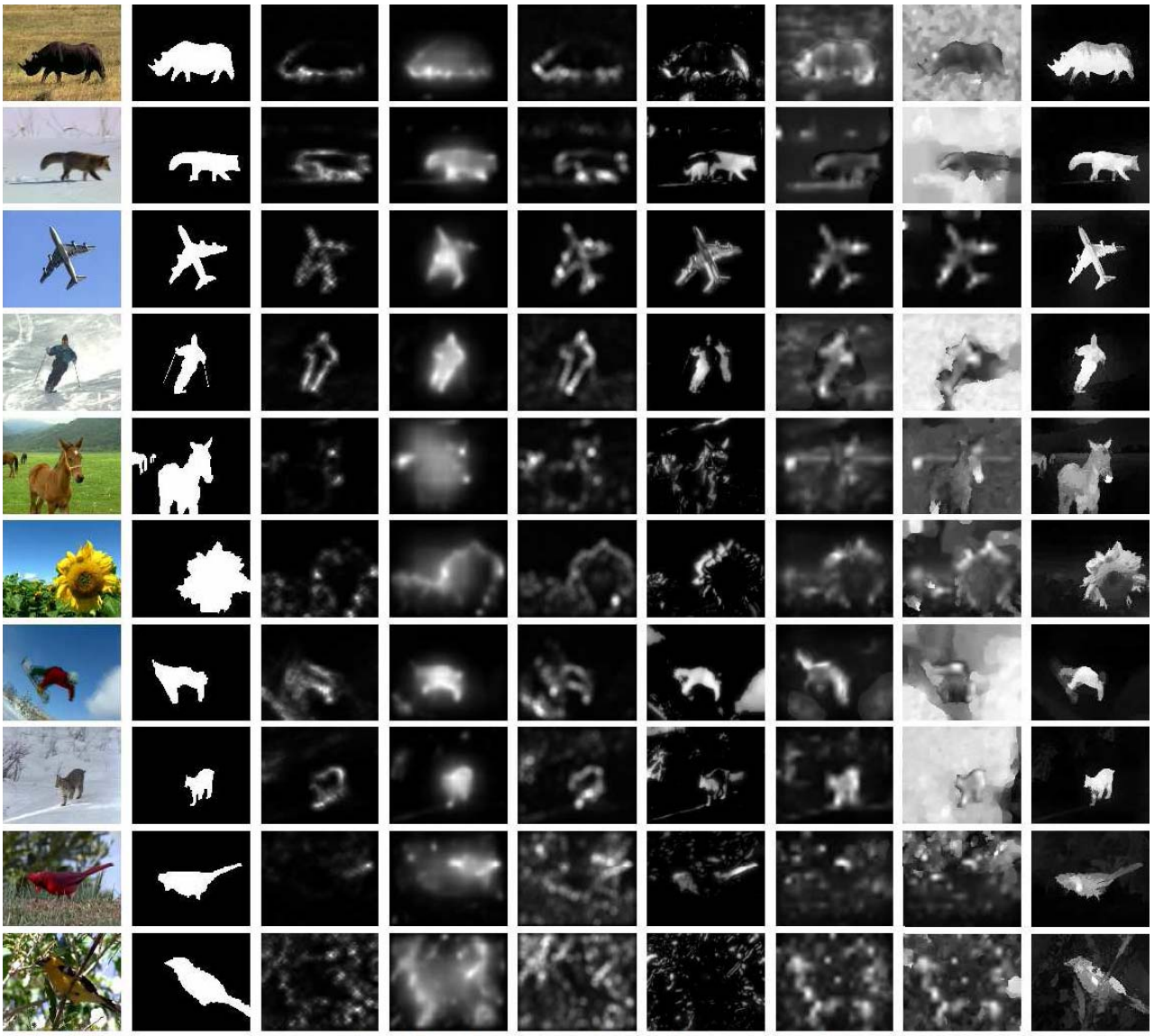


Fig. 7. Spatiotemporal saliency maps for sample video frames in DS1. (a) video frame, (b) ground truths, and spatiotemporal saliency maps generated using, (c) SR [17], (d) CE [20], (e) QFT [7], (f) MB [10], (g) PS [35], (h) DC [41], and (i) our SP model.

We compared our SP model with six state-of-the-art spatiotemporal saliency models, including SR [17], CE [20], QFT [7], MB [10] based saliency models, predictive saliency (PS) model [35] and DC saliency model [41]. We implemented the PS model and tuned its parameter setting to achieve the optimal performance considering the results on both DS1 and DS2 comprehensively. The optimal parameters are set as follows: $\alpha = 0.9$, $\beta = 0.1$, $num = 5$ and $Mthresholds$ is set to 70% of the maximum amplitude of motion vectors in the current frame (please refer to [35] for definitions of these parameters). For the other five models, we used the source codes with default parameter settings or executables provided by the authors. For a fair comparison, all saliency maps generated using different models are normalized into the same range of [0, 255] with the full resolution of original videos.

Experimental results on DS1 and DS2 are, respectively, presented in Section III-B and C, which compare the performance of our model with the other six models both subjectively and objectively, and a brief analysis on the complexity of different models is given in Section III-D.

B. Performance Evaluation on Saliency Detection

Some saliency maps for sample frames from the 10 video sequences in DS1 are shown in Fig. 7 for a glance. Fig. 7 shows that our SP model highlights the complete salient objects with well-defined boundaries and suppresses background regions more effectively compared with other models. In contrast, SR and QFT only highlight regions around salient object boundaries, CE and MB cannot effectively highlight salient objects in some frames, and PS and DC cannot effectively suppress background regions in some frames.

To objectively evaluate saliency detection performance of different models, we adopted the commonly used receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) and

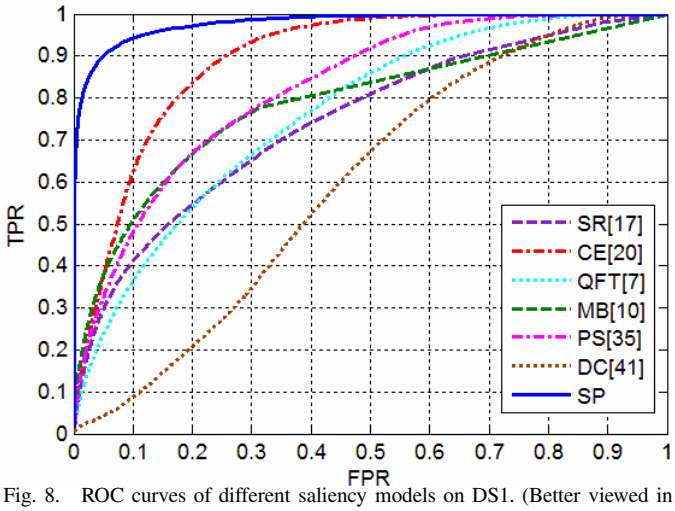


Fig. 8. ROC curves of different saliency models on DS1. (Better viewed in color; see the color online version.)

presents a robust evaluation of saliency detection performance. Specifically, the thresholding operations using a series of fixed integers from 0 to 255 are first performed on each saliency map to obtain 256 binary salient object masks, and a set of TPR and FPR values are calculated by comparing with the corresponding binary ground truth. Then for each model, at each threshold, the TPR/FPR values of all saliency maps are averaged, and as shown in Fig. 8, the ROC curve for each model plots the 256 average TPR values against the 256 average FPR values. Fig. 8 shows that the ROC curve of our SP model is higher than the other six ROC curves, and thus objectively demonstrates that our SP model outperforms other models on saliency detection performance. Besides, as a quantitative metric, the area under the ROC curve (AUC) for each model, on the basis of each video sequence and the overall dataset, is also calculated and shown in Fig. 9 for a more intuitive comparison. We can see from Fig. 9 that our SP model consistently outperforms other models with the highest AUC values on all 10 video sequences.

To evaluate the applicability of saliency maps for salient object detection more explicitly, an adaptive thresholding operation is performed on each saliency map using the well-known Otsu's method [57], which is simple yet effective, to obtain the binary salient object mask. We calculate the measures of precision and recall by comparing each binary salient object mask with the corresponding binary ground truth, and then calculate F-measure, which is the harmonic mean of precision and recall, to evaluate the overall performance as

$$F_\beta = \frac{(1+\beta)\cdot \text{precision}\cdot \text{recall}}{\beta\cdot \text{precision} + \text{recall}} \tag{19}$$

where the coefficient β is set to 0.3. Similarly as Fig. 9, for each model, the average F-measure on the basis of each video sequence and the overall dataset is calculated and shown in Fig. 10, which demonstrates the consistently higher applicability of our SP model for salient object detection in all video sequences. Furthermore, using AUC/F-measure values calculated for each video frame in DS1, we performed the paired t-test between our SP model and each of the other six models, and the statistical results show that our SP model significantly outperforms the other six models with p-value $\ll 0.001$.

To show the dynamics of spatiotemporal saliency maps, a series of frames with a fixed interval from three videos, on which our SP model achieves the high, medium and low performance in terms of F-measure, are shown with spatiotemporal saliency maps generated using all models in Fig. 11. In the top example, the camera pans to track the rhinoceros, which moves from right to left and then stops to graze. We can observe that our SP model maintains the coherence of saliency maps, which steadily highlight the salient object with well-defined boundaries. In the middle example, the posture and the scale of the player change quickly due to the fast and complex motion of the player, and the camera also moves to track the player. We can see that our SP model adapts to such fast and complex motion and generates the saliency maps with better quality. In the bottom example, the leaves swing moderately and the bird is with several fast intermittent movements, and thus the motion vector fields are too cluttered to identify the motion of the bird. Nonetheless, due to the spatial saliency measurement and the adaptive fusion method exploited in our SP model, our spatiotemporal saliency maps also show relatively better quality compared with other spatiotemporal saliency maps.

In principle, the use of superpixels for measuring saliency at superpixel level and then the pixel-level saliency derivation method enable our SP model to preserve salient object boundaries more accurately and to highlight the complete salient object more effectively than other models. The superpixel-level temporal saliency measurement can better adapt to different motions to highlight moving objects with temporal coherence, and the superpixel-level spatial saliency measurement can highlight the visually salient regions in individual frames. On the basis of pixel-level temporal and spatial saliency maps, the adaptive fusion method is effective to generate spatiotemporal saliency maps with reasonable quality, and thus ensures a better saliency detection performance of our SP model. In contrast, the common drawback of other models is that they cannot highlight the complete salient objects with well-defined boundaries. Besides, for videos with fast and complex motion such as the middle example in Fig. 11, other models cannot effectively suppress background regions, which are visually salient in view of individual frames, and their performances are also severely degraded on videos with complex scenes such as the bottom example in Fig. 11.

In summary, as the overall advantage over other models, our SP model can highlight the complete salient objects with well-defined boundaries more effectively, and can better handle videos with fast and complex motion as well as with complex scenes. However, due to the difference on the underlying scheme for spatiotemporal saliency measurement, it should be noted that each model has its inherent limitation, which will be summarized with more detailed analysis in Section III-C.

C. Performance Evaluation on Human Fixation Prediction

Based on the data of human fixation positions associated with each video frame in DS2, we evaluated the performance

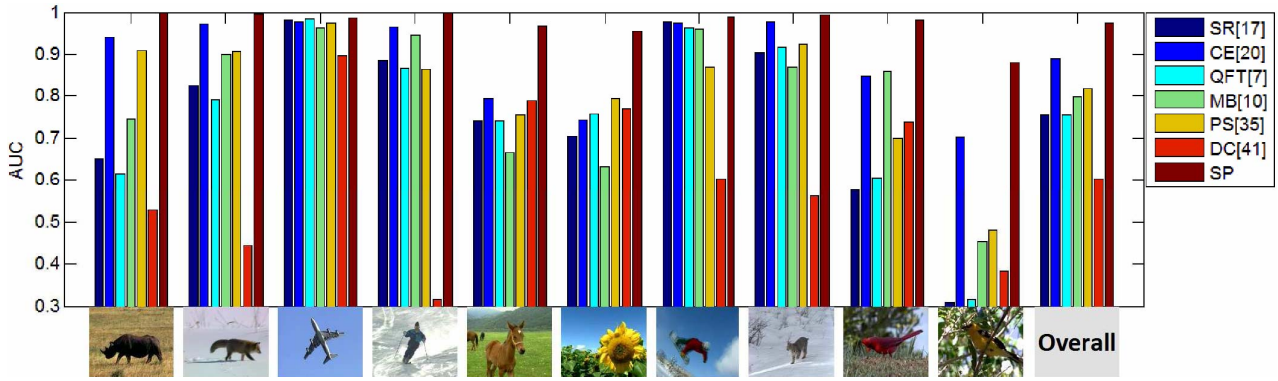


Fig. 9. AUC achieved using different saliency models on the basis of each video sequence and the overall dataset DS1.

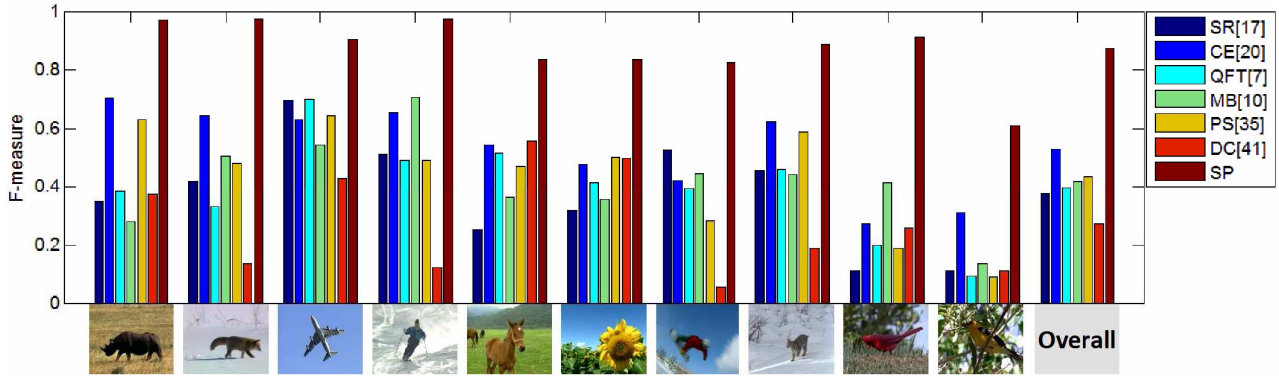


Fig. 10. F-measure achieved using different saliency models on the basis of each video sequence and the overall dataset DS1.

on human fixation prediction of different spatiotemporal saliency models. Figs. 12 and 13 show a series of video frames marked with human fixation positions, the heatmaps generated by performing Gaussian filtering around each fixation point, and spatiotemporal saliency maps generated using different models for a subjective comparison. Using human fixation points and heatmaps in Figs. 12 and 13 as reference, we can observe that most of saliency models can highlight salient regions, which locate around fixation points, either partially or completely, while cannot effectively suppress some type of background regions without fixation points. Specifically, such background regions are classified into four types, as shown in Table I. For a clear description, Table II qualitatively summarizes the performance of each model and analyzes why each model has its inherent limitation.

For objective performance evaluation on human fixation prediction, we adopted three quantitative metrics, i.e., AUC, the Pearson correlation coefficient (CC) and the normalized scanpath saliency (NSS) [58]. Since each metric has its own strength and weakness [59], the use of multiple metrics not only ensures that the performance evaluation is independent of the choice of metric, but also uses their complementary nature to adequately measure the similarity between the spatiotemporal saliency map and human fixation data. CC is calculated between the heatmap and the spatiotemporal saliency map, and falls into the range of $[-1, 1]$, in which a value of 0 indicates no linear correlation between the two maps, 1 indicates a perfect correlation, and -1 also indicates a perfect correlation, but in the opposite direction. NSS is the average of saliency values at human fixation positions in the spatiotemporal saliency map, which has been normalized to have zero mean and unit standard deviation.

The average values of AUC, CC, and NSS are calculated on the basis of each video category and the overall dataset, and shown in Figs. 14–16, respectively, for an objective performance comparison. We can observe from Figs. 14–16 that for the three video categories, i.e., abnormal, crowd, and noise, our SP model outperforms other models in terms of all metrics. For video category moving, our SP model outperforms other models in terms of CC and NSS, while CE slightly outperforms our SP model in terms of AUC. For video category surveillance, our SP model outperforms other models in terms of AUC, while MB and QFT outperform our SP model in terms of CC and NSS. The performance of our SP model on this category is degraded due to the surveillance video containing escalators (the bottom example in Fig. 13), which are not suppressed by our SP model. From another perspective, in terms of either AUC, CC, or NSS, our SP model outperforms other models on four out of five video categories. As an overall performance evaluation, Figs. 14–16 show that for the overall dataset DS2, our SP model outperforms other models in terms of all metrics. Furthermore, using AUC/CC/NSS values calculated for each video frame in DS2, we also performed the paired t-test between our SP model and each of the other six models, and the statistical results show that our SP model significantly outperforms the other six models with p-value $\ll 0.001$. In conclusion, our SP model achieves an overall better performance on human fixation prediction than the other six models.

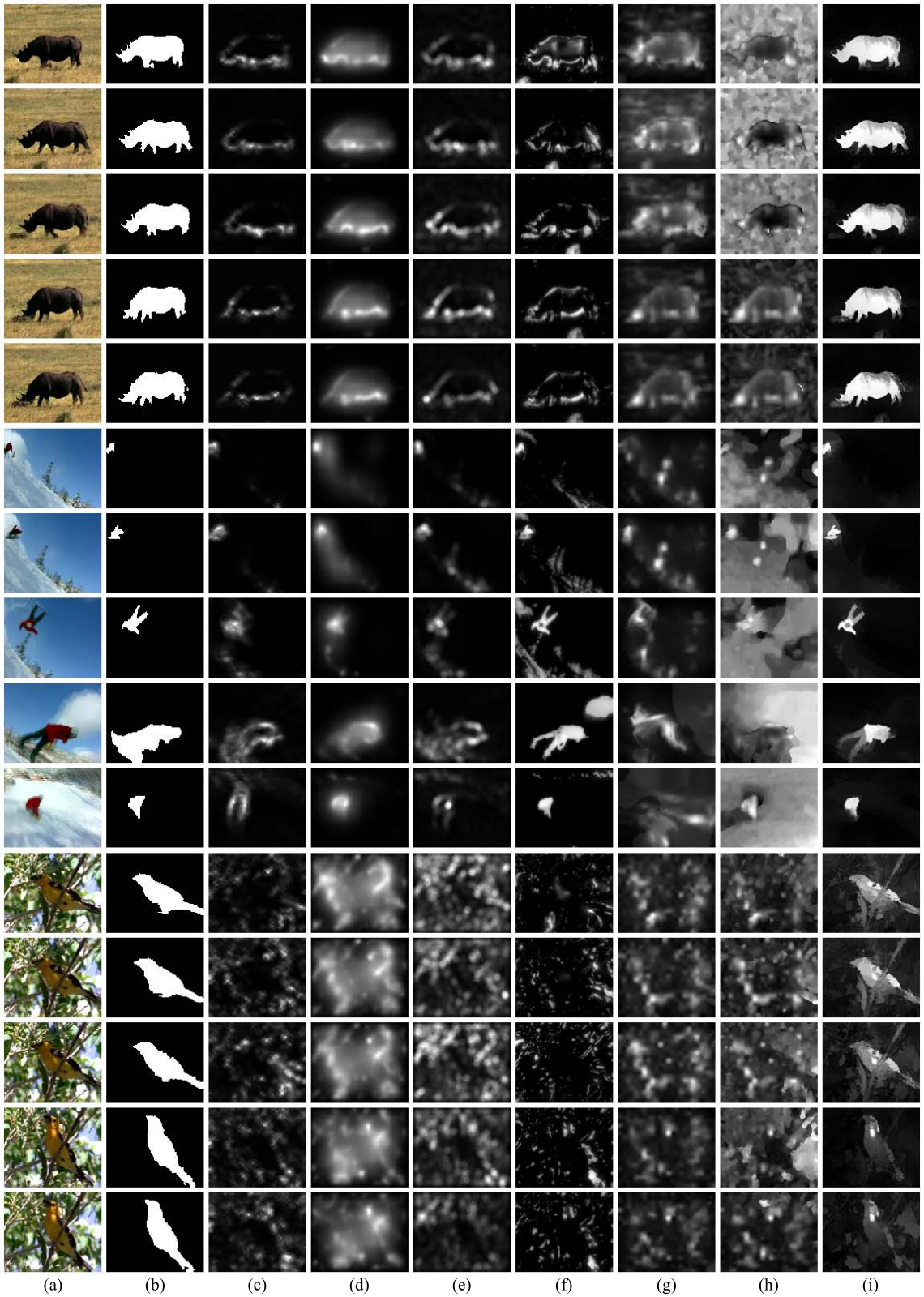
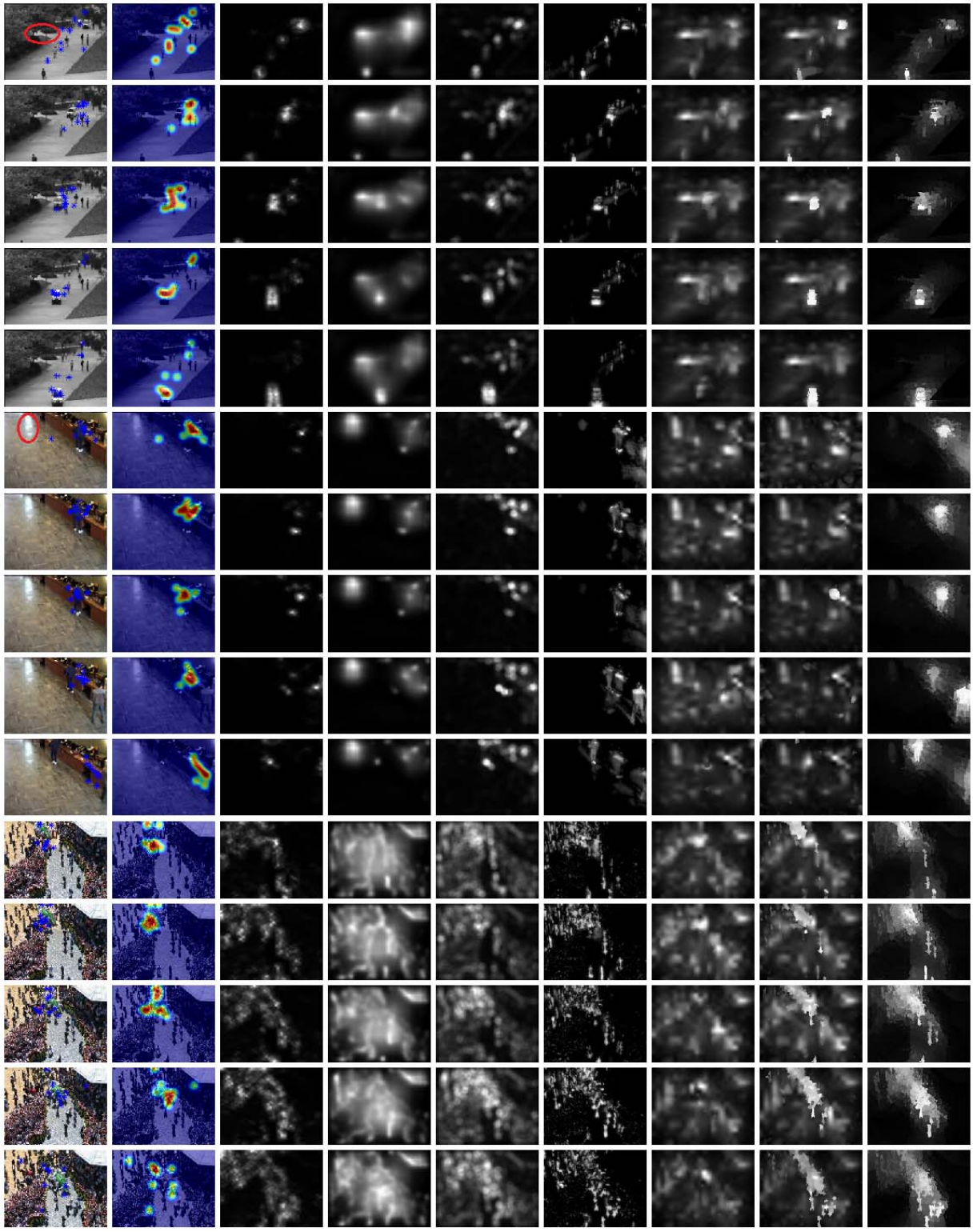


Fig. 11. Spatiotemporal saliency maps for three video clips (shown with an interval of 15, 10, and 10 frames, respectively, from top to bottom) in DS1. (a) video frames, (b) ground truths, and spatiotemporal saliency maps generated using (c) SR [17], (d) CE [20], (e) QFT [7], (f) MB [10], (g) PS [35], (h) DC [41], and (i) our SP model.

D. Complexity Analysis and Discussion

Besides the performance comparison of different spatiotemporal saliency models in terms of saliency detection and human fixation prediction, the following will briefly analyze the space complexity and time complexity of different models.

Regarding the space complexity, SR model has the largest demand for memory space due to batch processing, where a bunch of frames are always maintained in memory at run time. In contrast, other models process videos basically in a frame-by-frame manner, and at most the results of several previous

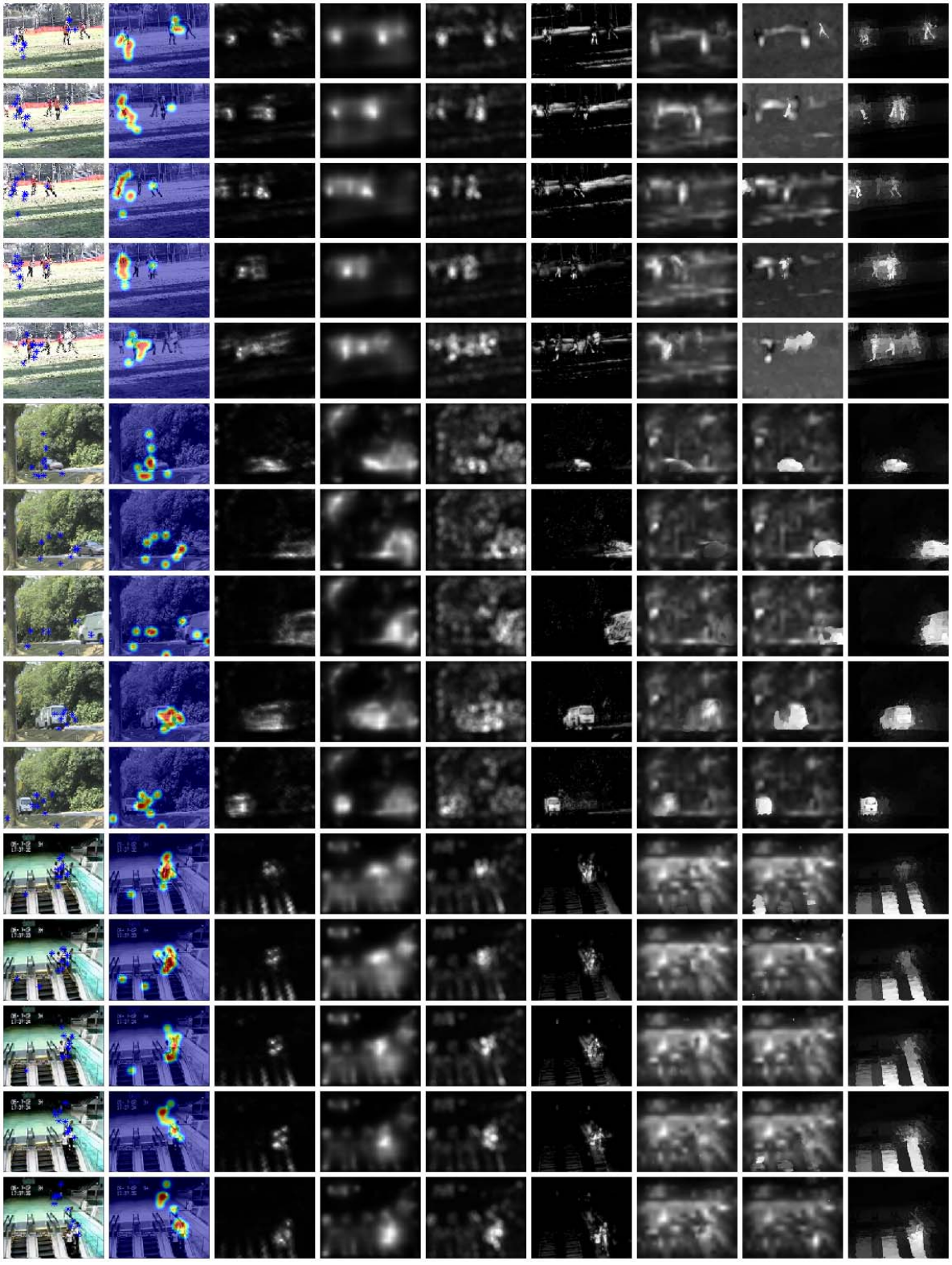


(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 12. Spatiotemporal saliency maps for video clips (shown with an interval of 10 frames) from the category abnormal (top), surveillance (middle), and crowd (bottom) in DS2. (a) Video frames with fixations, (b) fixation heatmaps superimposed on video frames, and spatiotemporal saliency maps generated using (c) SR [17], (d) CE [20], (e) QFT [7], (f) MB [10], (g) PS [35], (h) DC [41], and (i) our SP model. In (a), the blue asterisks are used to mark fixation points, and the red ovals overlaid on the two video frames are used to show the examples of background regions of *Type A* (see Table I). (Better viewed in color; see the color online version.)

frames are required for processing each current frame. Due to a relatively higher space complexity of SR model, other models are more suitable to process high-resolution and long-duration videos than SR model.

Regarding the time complexity, Table III shows the average processing time per frame for videos with resolution of 352 $\times$ 288 on a PC with Intel Core i7-2600 3.4 GHz CPU and 4 GB RAM, for an intuitive comparison. It is obvious



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 13. Spatiotemporal saliency maps for video clips (shown with an interval of 10 frames) from the category moving (top), noise (middle), and surveillance (bottom) in DS2. (a) Video frames with fixations, (b) fixation heatmaps superimposed on video frames, and spatiotemporal saliency maps generated using (c) SR [17], (d) CE [20], (e) QFT [7], (f) MB [10], (g) PS [35], (h) DC [41], and (i) our SP model. In (a), the blue asterisks are used to mark fixation points. (Better viewed in color; see the color online version.)

that PS, DC and our SP model have a higher time complexity than the other four models, because these models use the time-consuming optical flow estimation methods to obtain motion vector fields. Among all models, QFT model has the lowest time complexity, since its main operation, i.e., QFT, is very fast. Table IV shows the average processing time per frame taken by each component in the current MATLAB implementation of our SP model. It can be observed from

TABLE I
FOUR TYPES OF BACKGROUND REGIONS WITH EXAMPLES

Type	Description	Example
A	Background regions showing high contrast with the surroundings.	In the top and middle examples of Fig. 12, such a background region is marked using the red oval in the column (a).
B	Background regions with complex and dynamic scenes in videos captured using static cameras.	The crowd in the bottom example of Fig. 12 and swaying trees in the middle example of Fig. 13.
C	Background regions with global motion in videos captured using moving cameras.	The background regions in the top example of Fig. 13 (also in the 1st, 2nd, 4th, 7th and 8th video example of Fig. 7).
D	Background regions with coherent motion in videos captured using static cameras.	The escalators in the bottom example of Fig. 13.

TABLE II
QUALITATIVE DESCRIPTION AND ANALYSIS ON THE PERFORMANCE OF EACH SPATIOTEMPORAL SALIENCY MODEL

Model	Qualitative description (the top paragraph) and analysis (the bottom paragraph)
SR[17]	SR sparsely highlights high-contrast regions, which are usually fixated by human, but cannot effectively highlight the fixated homogenous regions (such as regions inside the object in some frames of the middle example in Fig. 12).
	Due to the use of center-surround scheme in SR, the fixated homogenous regions which have low contrast with the surroundings cannot be highlighted in the saliency maps.
CE[20]	CE highlights salient regions more completely, but cannot suppress background regions of *Type A*, *B* and *D*.
	CE uses the minimum conditional entropy which represents the minimum uncertainty of region to define the saliency. Therefore, background regions with high contrast and dynamic changes are assigned with high entropy values and falsely highlighted in the saliency maps.
QFT[7]	QFT can suppress background regions of *Type A* and *C*, but mainly suffers from the cluttered saliency maps due to background regions of *Type B*.
	QFT jointly uses the features of luminance, chrominance and frame difference. Background regions with complex and dynamic scenes are different from the spatial surrounding regions or/and the temporally co-located regions in terms of these features, and thus they are falsely highlighted in the saliency maps.
MB[10]	MB can suppress background regions of *Type A*, *B* and *D*, but cannot suppress background regions of *Type C*.
	The background modeling scheme used in MB can suppress motion-coherent background regions such as escalator, whose appearance changes slightly and repeatedly. However, it cannot build a reliable background model for videos captured using moving cameras, and thus cannot suppress background regions with global motion.
PS[35]	PS can highlight salient regions around fixation points, but suffers from the cluttered saliency maps due to the four types of background regions.
	PS exploits the linear combination to integrate spatial saliency with temporal saliency. However, in the PS model, the spatial saliency map generated using [42] falsely highlights background regions of *Type A* and *B*, and the temporal saliency map predicted using motion vectors cannot suppress background regions of *Type C* and *D* by itself. Finally, the linear combination using manually set weights cannot adapt well to different videos.
DC[41]	DC can uniformly highlight salient regions with motion in videos captured using static cameras, but suffers from the cluttered saliency maps due to the four types of background regions.
	DC exploits the pixel-wise maximum operation to integrate spatial saliency with temporal saliency. However, in the DC model, the spatial saliency map generated using [42] falsely highlights background regions of *Type A* and *B*, and the temporal saliency map, which is designed to highlight regions with consistent motion proportional to motion amplitude, falsely highlights background regions of *Type C* and *D* with enough motion amplitude. Finally, the maximum operation results in that the above falsely highlighted background regions are still preserved in the spatiotemporal saliency map.
SP	Compared to the above six models, SP can highlight salient regions and suppress background regions of *Type A*, *B* and *C* more effectively to generate spatiotemporal saliency maps with better quality.
	However, due to the use of motion vector field, SP cannot suppress background regions of *Type D*, which are with coherent motion different from the major background.

Table IV that all the other components of our SP model excluding optical flow estimation take a total of 1.678 s per frame, while the two highest time-consuming components, i.e., optical flow estimation and superpixel segmentation, respectively, occupy 85.8% and 7.3% of the total processing time.

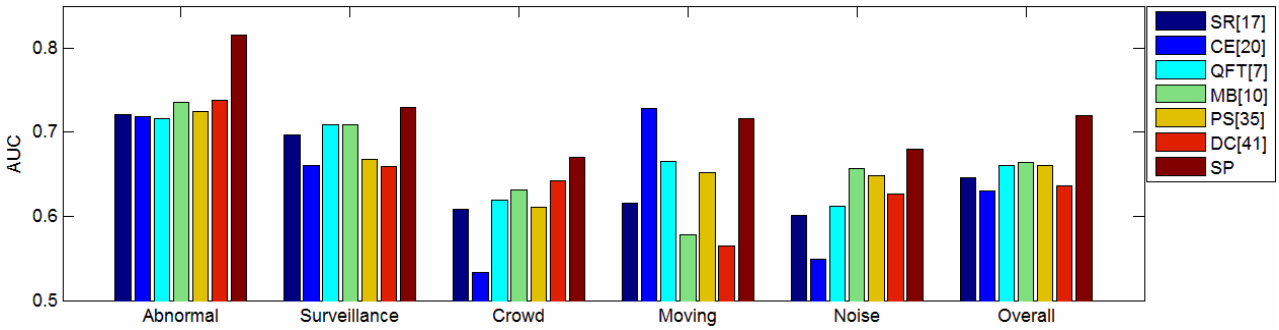


Fig. 14. Average AUC achieved using different saliency models on the basis of each video category and the overall dataset DS2.

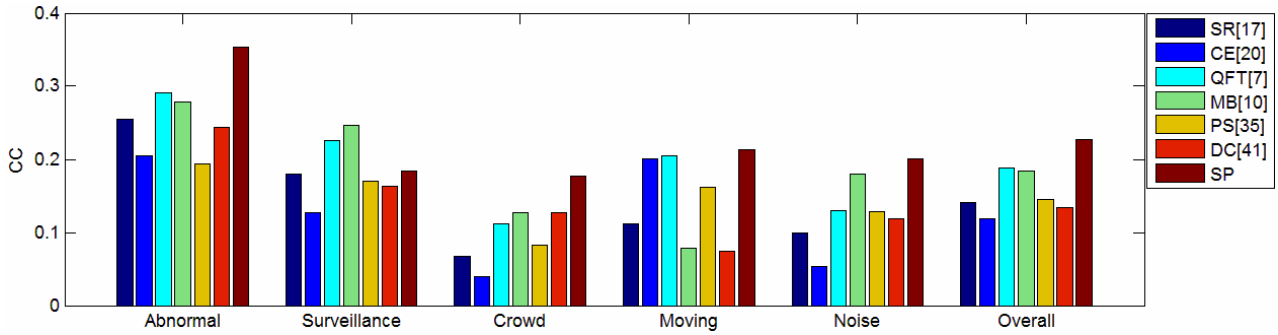


Fig. 15. Average CC achieved using different saliency models on the basis of each video category and the overall dataset DS2.

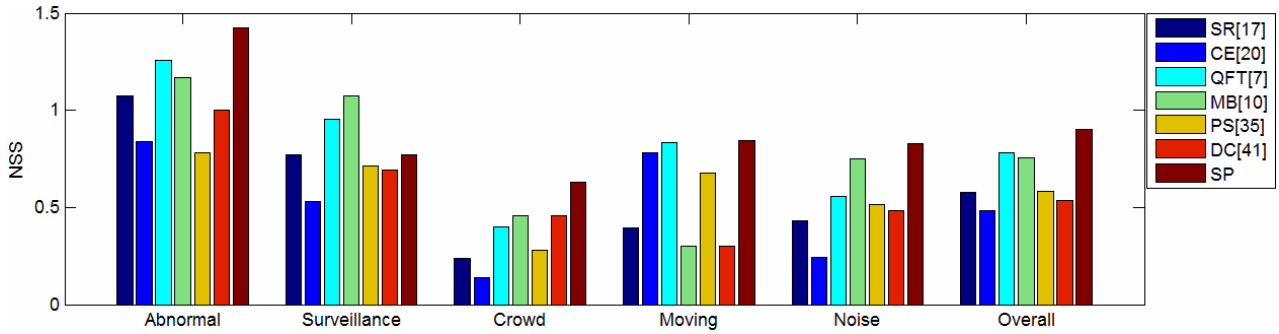


Fig. 16. Average NSS achieved using different saliency models on the basis of each video category and the overall dataset DS2.

TABLE III
COMPARISON OF AVERAGE PROCESSING TIME PER FRAME USING DIFFERENT SPATIOTEMPORAL SALIENCY MODELS

Model	SR[17]	CE[20]	QFT[7]	MB[10]	PS[35]	DC[41]	SP
Time (second)	0.427	4.377	0.117	0.174	10.472	21.577	11.850
Code	Matlab	Matlab	Matlab	C++	Matlab	Matlab	Matlab

To alleviate the time complexity of our SP model as well as those models that use optical flow estimation, one solution is to use block-level motion vectors in the video bitstream instead of optical flow. Specifically, block-level motion vectors are obtained during the video decoding process so as to bypass the optical flow estimation, and then an edge-preserving upsampling algorithm can be exploited to generate the pixel-level motion vector field with reasonable motion boundaries. Another solution is to resize the input video to a low resolution for generating saliency maps, which are then resized to the original resolution for the final use. Depending on the specific requirements of different applications, either or both solutions can be adopted to improve the computational efficiency of our model with a compromise of performance. Last but not least, a C++ implementation of our model or even a parallel GPU implementation, on which different components

TABLE IV
AVERAGE PROCESSING TIME PER FRAME TAKEN BY EACH COMPONENT OF OUR SP MODEL

Component		Time (second)
Feature extraction	Optical flow estimation	10.172
	Superpixel segmentation	0.866
	Motion histograms and color histograms	0.070
Superpixel-level temporal saliency		0.168
Superpixel-level spatial saliency		0.500
Pixel-level temporal/spatial saliency derivation		0.058
Pixel-level spatiotemporal saliency generation		0.016
Total		11.850 (1.678*)

*Excluding optical flow estimation.

of our model can be parallelized on the basis of superpixel or pixel, will also substantially improve the computational efficiency.

IV. CONCLUSION

In this paper, we have presented a SP spatiotemporal saliency model, which uses the effectiveness of measuring saliency at superpixel level to generate the pixel-level spatiotemporal saliency map with better quality. Superpixels have been used for saliency detection in images, while this paper explores its use for spatiotemporal saliency detection in videos. Based on motion and color histograms, which are extracted at superpixel level as local features and frame level as global features, temporal saliency is measured by exploiting motion distinctiveness estimation of superpixels and a scheme of temporal saliency prediction and adjustment, and spatial saliency is measured by integrating global contrast with spatial sparsity of superpixels. Then, based on temporal and spatial saliency measures of superpixels, the pixel-level saliency derivation method and adaptive fusion method are jointly exploited to generate the pixel-level spatiotemporal saliency map effectively. Extensive experimental results on two public video datasets demonstrate the better performance of the proposed model on both saliency detection and human fixation prediction.

In our future work, we will adapt the proposed model into the framework of high efficiency video coding (HEVC), to design a saliency-aware rate control algorithm to improve coding efficiency without noticeable degradation of subjective visual quality. For this purpose, the proposed model can directly use the motion vectors of variable-sized blocks from the motion estimation in HEVC, and will be extended with an edge-preserving upsampling algorithm to handle the motion vector field with variable-sized blocks. The proposed model with the above adaptation and extension will generate the block-level and pixel-level spatiotemporal saliency maps more efficiently.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the associate editor for their valuable comments, which have greatly helped us to make improvements.

REFERENCES

[1] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[2] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, "Region diversity maximization for salient object detection," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 215–218, Apr. 2012.

[3] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *Proc. ECCV*, Sep. 2010, pp. 366–379.

[4] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[5] A. Shamir and S. Avidan, "Seam carving for media retargeting," *Commun. ACM*, vol. 52, no. 1, pp. 77–85, Jan. 2009.

[6] Z. Yuan, T. Lu, Y. Huang, D. Wu, and H. Yu, "Addressing visual consistency in video retargeting: A refined homogeneous approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 890–903, Jun. 2012.

[7] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[8] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.

[9] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, Jul. 2011.

[10] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokric, V. Crnojević, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 948–958, Apr. 2011.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[12] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[13] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[14] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE CVPR*, Jun. 2005, pp. 631–637.

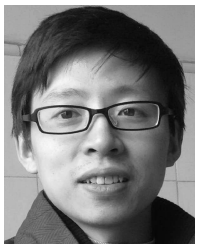
[15] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[16] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Proc. NIPS*, Dec. 2007, pp. 497–504.
[17] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, Nov. 2009.
[18] Y. Lin, Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth mover's distance based saliency measurement and nonlinear feature combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 314–328, Feb. 2013.
[19] C. Liu, P. C. Yuen, and G. Qiu, "Object motion detection using information theoretic spatio-temporal saliency," *Pattern Recognit.*, vol. 42, no. 11, pp. 2897–2906, Nov. 2009.
[20] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," in *Proc. ACCV*, Sep. 2009, pp. 246–257.
[21] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. NIPS*, Dec. 2008, pp. 681–688.
[22] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, "Incremental sparse saliency detection," in *Proc. IEEE ICIP*, Nov. 2009, pp. 3093–3096.
[23] V. Gopalakrishnan, D. Rajan, and Y. Hu, "A linear dynamical system framework for salient motion detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 683–692, May 2012.
[24] K. Muthuswamy and D. Rajan, "Salient motion detection through state controllability," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 1465–1468.
[25] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.
[26] X. Cui, Q. Liu, and D. N. Metaxas, "Temporal spectral residual: Fast motion saliency detection," in *Proc. 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp. 617–620.
[27] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, Nov. 2010.
[28] W. F. Lee, T. H. Huang, S. L. Yeh, and H. H. Chen, "Learning-based prediction of visual attention for video signals," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3028–3038, Nov. 2011.
[29] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "Intrinsic dimensionality predicts the saliency of natural dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1080–1091, Jun. 2012.
[30] Z. Ren, L.-T. Chia, and D. Rajan, "Video saliency detection with robust temporal alignment and local-global spatial contrast," in *Proc. 2nd ACM ICMR*, Jun. 2012, article 47.
[31] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 1485–1488.
[32] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 446–456, Apr. 2011.
[33] Y. Luo and Q. Tian, "Spatio-temporal enhanced sparse feature selection for video saliency estimation," in *Proc. IEEE CVPR Workshops*, Jun. 2012, pp. 33–38.
[34] Z. Ren, S. Gao, D. Rajan, L.-T. Chia, and Y. Huang, "Spatiotemporal saliency detection via sparse representation," in *Proc. IEEE ICME*, Jul. 2012, pp. 158–163.
[35] Q. Li, S. Chen, and B. Zhang, "Predictive video saliency detection," *Commun. Comput. Inf. Sci.*, vol. 321, pp. 178–185, Sep. 2012.
[36] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 774–780, Aug. 2000.
[37] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, May 2009.
[38] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, Sep. 2007.
[39] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp, "Camera motion-based analysis of user generated video," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 28–41, Jan. 2010.
[40] Y. Tong, F. A. Cheikh, F. F. E. Guraya, H. Konik, and A. Trémeau, "A spatiotemporal saliency model for video surveillance," *Cognit. Comput.*, vol. 3, no. 1, pp. 241–263, Mar. 2011.
[41] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modeling," in *Proc. 27th AAAI Conf.*, Jul. 2013, pp. 1063–1069.
[42] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, Dec. 2006, pp. 545–552.
[43] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Global contrast based salient region detection," in *Proc. IEEE CVPR*, Jun. 2011, pp. 409–416.
[44] Z. Liu, O. Le Meur, S. Luo, and L. Shen, "Saliency detection using regional histograms," *Opt. Lett.*, vol. 38, no. 5, pp. 700–702, Mar. 2013.
[45] X. Zhang, Z. Ren, D. Rajan, and Y. Hu, "Salient object detection through over-segmentation," in *Proc. IEEE ICME*, Jul. 2012, pp. 1033–1038.
[46] Z. Liu, O. Le Meur, and S. Luo, "Superpixel-based saliency detection," in *Proc. IEEE WIAMIS*, Jul. 2013, pp. 1–4.
[47] F. Perazzi1, P. Krähenbül, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE CVPR*, Jun. 2012, pp. 733–740.
[48] Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan, "Improved saliency detection based on superpixel clustering and saliency propagation," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1099–1102.
[49] J. Wang, C. Zhang, Y. Zhou, Y. Wei, and Y. Liu, "Global contrast of superpixels based salient region detection," in *Proc. 1st Conf. CVM*, Nov. 2012, pp. 130–137.
[50] Z. Tan, L. Wan, W. Feng, and C.-M. Pun, "Image co-saliency detection by propagating superpixel affinities," in *Proc. IEEE ICASSP*, May 2013, pp. 2114–2117.
[51] S. Sahli, D. A. Lavigne, and Y. Sheng, "Saliency detection in aerial imagery using multi-scale SLIC segmentation," in *Proc. Int. Conf. Image Process., Comput. Vis., Pattern Recognit.*, Jul. 2012, pp. 647–653.
[52] Y. Xie, H. Lu, and M. H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.
[53] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[54] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.
[55] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE ICME*, Jun. 2009, pp. 638–641.
[56] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit, "Dynamic saliency models and human attention: A comparative study on videos," in *Proc. ACCV*, Nov. 2012, pp. 586–598.
[57] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[58] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, Aug. 2005.
[59] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behavior Res. Meth.*, vol. 45, no. 1, pp. 251–266, Mar. 2013.

Zhi Liu (M'07) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 1999, 2002, and 2005, respectively.

He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From Aug. 2012 to Aug. 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 100 refereed technical papers in international journals and conferences. His research interests include saliency models, image/video segmentation, image/video retargeting, video coding, and multimedia communication.

Dr. Liu was a TPC member in ICME 2014, WIAMIS 2013, IWVP 2011, PCM 2010, and ISPACS 2010. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014, and currently serves as a guest editor for a special issue on *Recent Advances in Saliency Models, Applications and Evaluations*, which will appear in the journal of *Signal Processing: Image Communication*.



Xiang Zhang received the B.E. and M.E. degrees from University of Electronic Science and Technology of China, Chengdu, China, and the Ph.D. degree from Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 2003, 2006, and 2009, respectively.

He is an Associate Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image/video processing and video coding.

Shuhua Luo received the B.E. degree from Hunan Institute of Science and Technology, Yueyang, China, in 2011, and the M.E. degree from Shanghai University, Shanghai, China, in 2014.

Her research interests include saliency models and image/video retargeting.

Olivier Le Meur received the Ph.D. degree from University of Nantes, Nantes, France, in 2005.

He was with the Media and Broadcasting Industry from 1999 to 2009. In 2003, he joined the Research Center of Thomson-Technicolor, Rennes, France, where he supervised a Research Project concerning the modeling of human visual attention. He has been an Associate Professor of image processing with the University of Rennes 1 since 2009. In the SIROCCO Team of IRISA/INRIA-Rennes, his current research interests include human visual attention, computational modeling of visual attention, and saliency-based applications, such as video compression, objective assessment of video quality, and retargeting.