This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

Saliency Detection for Unconstrained Videos Using Superpixel-level Graph and Spatiotemporal Propagation

Zhi Liu, *Senior Member, IEEE*, Junhao Li, Linwei Ye, Guangling Sun, and Liquan Shen

***Abstract*—This paper proposes an effective spatiotemporal saliency model for unconstrained videos with complicated motion and complex scenes. First, superpixel-level motion and color histograms as well as global motion histogram are extracted as the features for saliency measurement. Then a superpixel-level graph with the addition of a virtual background node representing the global motion is constructed, and an iterative motion saliency measurement method which utilizes the shortest path algorithm on the graph is exploited to reasonably generate motion saliency maps. Temporal propagation of saliency in both forward and backward directions is performed by using efficient operations on inter-frame similarity matrices to obtain the integrated temporal saliency maps with the better coherence. Finally, spatial propagation of saliency both locally and globally is performed via the use of intra-frame similarity matrices to obtain the spatiotemporal saliency maps with the even better quality. Experimental results on two video datasets with various unconstrained videos demonstrate that the proposed model consistently outperforms the state-of-the-art spatiotemporal saliency models on saliency detection performance.**

***Index Terms*—Spatiotemporal saliency detection, saliency model, superpixel-level graph, temporal propagation, spatial propagation, motion saliency, unconstrained video.**

I. Introduction

Visual attention mechanism enables human observers to capture the visually salient objects in a complex scene effortlessly. The research on computational models for saliency detection was motivated by simulating the visual attention mechanism based on the biologically plausible architecture [1] and the feature integration theory [2], and was originally used for human fixation prediction [3]. In the past decade, saliency detection has been a booming research topic, and has benefited a broad range of applications such as salient object detection [4-6], salient object segmentation [7-11], content-aware image/video retargeting [12-14], content-based image/video compression [15-17], image/video quality assessment [18, 19], and visual scanpath prediction [20], etc.

Recent years have witnessed a lot of research efforts on saliency models for images, and some recent benchmarks have been reported in [21, 22]. In contrast, the research on spatiotemporal saliency models for videos is not as abundant as that for images, but it has received increasing attention from the research community. Saliency detection in videos is different from static images due to the introduction of temporal dimension, and the motion information in video is definitely important for measuring saliency. This paper focuses on saliency detection in videos, and the following will mainly review the related spatiotemporal saliency models for videos.

Due to the clear interpretation of human visual attention mechanism and the concise computation form, the center-surround scheme has been widely implemented with different features and formulations in a number of spatiotemporal saliency models, which measure the saliency of a pixel/patch/volume as the feature difference with its surroundings. For example, the surprise model [23] exploits the difference on a set of features including luminance, color, orientation, flicker and motion energy to generate the spatiotemporal saliency map. The feature difference between each patch/volume and its spatiotemporal surroundings can be measured by using the local regression kernel based self-resemblance [24] and the earth mover's distance between the weighted histograms [25], and then is assigned as the spatiotemporal saliency of the center pixel in the patch/volume. Based on the discriminant center-surround hypothesis [26], the Kullback-Leibler (KL) divergence on dynamic texture feature is used to measure spatiotemporal saliency [27]. In [28], the contrast of directional coherence, which is calculated on the distribution of spatiotemporal gradients, is evaluated under the multiscale framework to measure the spatiotemporal saliency. From the view of feature change in the temporal direction, salient object pixels significantly differ from most of pixels in the video frames, and thus a multiscale background model [19] represented using Gaussian pyramid is exploited for saliency measurement.

Besides the center-surround scheme, recent spatiotemporal saliency models are also formulated based on a number of theories, methods and schemes including information theory, control theory, frequency domain analysis method, machine

Manuscript received January 28, 2016. This work was supported by the National Natural Science Foundation of China under Grants 61471230 and 61171144, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

Z. Liu (corresponding author), J. Li, L. Ye, G. Sun and L. Shen are with the School of Communication and Information Engineering, Shanghai University, Shanghai, China (email: liuzhisjtu@163.com; jxlijunhao@163.com; yelinweimail@163.com; sunguangling@shu.edu.cn; jsslq@163.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

learning method, sparse representation method and the fusion scheme on spatial saliency and temporal saliency.

Some concepts in the information theory can be naturally adapted to measure spatiotemporal saliency in video. For example, self-information [29], minimum conditional entropy [30] and incremental coding length [31, 32] are calculated on the basis of patch/volume and used as spatiotemporal saliency measures. From the view of modern control theory, the video sequence can be represented using the state space model of linear system, and then either the observability of output [33] or the controllability of states [34] is exploited to measure spatiotemporal saliency in order to discriminate salient object motion from background motion.

The frequency domain analysis method is first introduced into image saliency detection [35], in which the spectral residual of the amplitude spectrum of Fourier transform is used as saliency measures. Then as an extension, the spectral residual along the temporal slices at the horizontal/vertical direction is exploited to generate spatiotemporal saliency maps for video [36]. Besides, the phase spectrum of quaternion Fourier transform, which incorporates color, orientation and intensity in each frame and difference between frames in parallel, is exploited to efficiently perform multiresolution spatiotemporal saliency detection in [15].

Machine learning methods such as probabilistic multi-task learning [37], support vector machine [38] and support vector regression [39] have been exploited to build spatiotemporal saliency models by using eye tracking data as the training set. In [4], the conditional random field (CRF) learning is used to integrate multiscale contrast, center-surround histogram and spatial distribution of color and motion vector field for saliency measurement. In [40], with the aim of dominant camera motion removal, the one-class support vector machine is used to classify the trajectories, and the saliency of each trajectory is diffused to its surrounding regions for generating the spatiotemporal saliency map.

Sparse representation method, which has been effectively utilized for saliency detection in images [41], is introduced into spatiotemporal saliency models. Sparse feature selection [42] and patch-level reconstruction error with sparse regularization term [43] are utilized to measure spatiotemporal saliency. The matrix which is composed by temporally aligned video frames [44] or temporal slices [45] can be decomposed into a low-rank matrix for background and a sparse matrix for salient objects to measure spatiotemporal saliency.

The fusion scheme has become a commonly used paradigm in a number of spatiotemporal saliency models. Considering the difference of nature between temporal domain and spatial domain, temporal saliency and spatial saliency are measured respectively and then combined using different fusion schemes to generate spatiotemporal saliency map. Since motion in most videos plays an important role on drawing human visual attention, motion vector field is generally used as the most important cue for temporal saliency measurement. The directional consistency of motion [46], temporal coherence of motion amplitude and orientation [47], distribution of motion amplitude and orientation [48], amplitudes of residual motion vectors after global motion compensation [49, 50], global motion parameters [51] and temporal gradient difference [52] are exploited to measure temporal saliency. Generally, spatial saliency measurement is performed on the basis of each video frame. For example, the center-surround difference on spatial features [48, 51, 52], contrast sensitivity function [49], interactions between cortical-like filters [50] and entropy of HOG features [47] are utilized to measure spatial saliency. Besides, some models [53, 54] directly use block-level motion vectors and DCT coefficients decompressed from the video bitstream to calculate spatial saliency map and temporal saliency map, respectively. To integrate temporal saliency map with spatial saliency map, fusion schemes such as normalization, linear addition, max, multiplication and their combination are widely adopted in the above mentioned models to generate spatiotemporal saliency map.

Recently, saliency models that are built on the basis of segmented regions/superpixels such as [7-9] have significantly elevated the saliency detection performance on images. Inspired by the improvement on image saliency models, a superpixel-based spatiotemporal saliency model is proposed in [55], which exploits superpixel-level motion distinctiveness, global contrast and spatial sparsity to measure temporal saliency and spatial saliency, respectively, and adaptively fuses them based on their interaction and selection to generate spatiotemporal saliency map. Furthermore, a superpixel-level trajectory based spatiotemporal saliency model is proposed in [56], which utilizes the long-term motion cues induced by superpixel-level trajectories to effectively enhance the coherence of temporal saliency through the video, and exploits a quality-guided fusion scheme to integrate with spatial saliency for spatiotemporal saliency generation. In [11], the geodesic distance from each superpixel to the frame boundaries is adopted in an intra-frame graph for computing the object probability of each superpixel, and an inter-frame graph is constructed to calculate the geodesic distance from each superpixel to the estimated background regions, which are obtained by thresholding the object probabilities of superpixels, for generating superpixel-level spatiotemporal saliency maps. The joint utilization of intra-frame graph and inter-frame graph enables [11] to be an effective spatiotemporal saliency model, which is distinctive from previous models. Compared to other models, the above three models [11, 55, 56] can better highlight salient objects with well-defined boundaries and suppress background regions.

However, the existing spatiotemporal saliency models are still insufficient to effectively handle a variety of unconstrained videos, which cause the difficulty to highlight salient objects and suppress background due to the unconstrained condition of capturing videos. There are no constraints on motion patterns of salient objects and background regions in unconstrained videos. Specifically, salient objects may have rigid motion or deformation, consistent motion or intermittent motion, and background regions may exhibit the complicated motion due to the mixed camera movements including pan, tilt, zoom and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

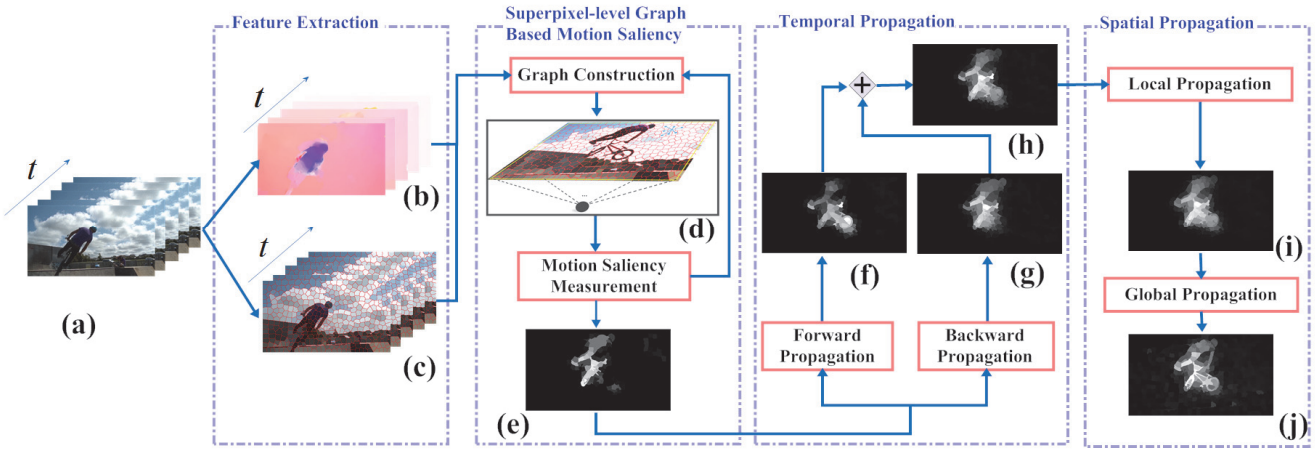

Fig. 1. Illustration of the proposed SGSP model. (a) Original video frames; (b) motion vector fields; (c) superpixel segmentation results; (d) superrpixel-level graph; (e) motion saliency map; (f) forward temporal saliency map; (g) backward temporal saliency map; (h) integrated temporal saliency map; (i) initial spatiotemporal saliency map; (j) final spatiotemporal saliency map.

camera jitter. Besides, the cluttered background and low contrast between salient objects and background in some unconstrained videos even increase the difficulty of saliency detection.

Therefore, in order to effectively improve the saliency detection performance on unconstrained videos with complicated motion and complex scenes, this paper proposes a superpixel-level graph and spatiotemporal propagation (SGSP) based saliency model for unconstrained videos. Specifically, our main contributions are fourfold. First, inspired by the previous works in [11, 55, 56] as well as those graph based image saliency models [57-59], we propose a superpixel-level graph based motion saliency measurement method, which can effectively estimate and refine the motion saliency of each superpixel in an iterative manner. Second, we propose an effective method for temporal propagation of saliency in both forward and backward directions by using efficient operations on inter-frame similarity matrices. Third, we propose a two-phase spatial propagation method to locally and globally propagate saliency via the use of intra-frame similarity matrices. With the introduction of bidirectional temporal propagation, two-phase spatial propagation and the unified operations on inter-frame/intra-frame similarity matrices, the above two contributions enable the spatiotemporal saliency propagation in video frames to be more systematic and more effective for the better saliency detection performance. Last, we created a dataset containing 18 unconstrained videos for saliency detection (UVSD), to introduce more challenging videos with complicated motion and complex scenes for comprehensive evaluation of spatiotemporal saliency detection performance.

Note that the recent spatiotemporal saliency models use superpixel-level graph in [11] and temporal propagation of saliency in [55, 56], the related components in the proposed SGSP model are considerably different from those in the above three models. Specifically, the proposed motion saliency measurement method realizes an iteratively updated superpixel-level graph with the introduction of a virtual background node, which is associated with the iteratively refined global motion histogram for each frame. Both construction and update nature of the superpixel-level graph in our method are considerably different from the superpixel-level intra-frame graph and inter-frame graph used in [11]. Our model enables an effective bidirectional temporal propagation of saliency, while in [55] saliency is only propagated from the previous frame to the current frame and in [56] saliency propagation is only performed for each superpixel-level trajectory in the temporal direction.

Thanks to the above mentioned main contributions, the proposed SGSP model shows new prospects of spatiotemporal saliency modeling, and extensive experimental results on UVSD and another video dataset SegTrackV2 [60] demonstrate that the proposed SGSP model consistently outperforms the state-of-the-art spatiotemporal saliency models including [11, 55, 56]. The UVSD dataset and the code of the proposed SGSP model will be made publicly available at http://www.ivp.shu.edu.cn/ upon the acceptance.

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

II. Proposed Spatiotemporal Saliency Model

The proposed SGSP model is illustrated in Fig. 1, which consists of four components marked using the dash-dotted lines, and will be detailed from Section II-A to II-D in turn. Feature extraction is first performed on the superpixel segmentation results of video frames to obtain motion histograms and color histograms in Section II-A, and then a superpixel-level graph is built for measuring motion saliency of superpixels in Section II-B. The spatiotemporal propagation of saliency is actually performed in two constitutive steps, i.e., temporal propagation in Section II-C is exploited to generate temporal saliency maps, and then spatial propagation in Section II-D is exploited to generate spatiotemporal saliency maps. It should be noted that only motion features are utilized in motion saliency measurement (Section II-B), while motion features and color

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

features are jointly utilized in temporal propagation (Section II-C) and spatial propagation (Section II-D).

A. Feature Extraction

For each input video frame $\mathbf{F}_t$, its pixel-level motion vector field $\mathbf{MVF}_{t,t-1}$ with respect to the previous frame $\mathbf{F}_{t-1}$, is calculated using the large displacement optical flow (LDOF) method in [61], which allows capturing large displacements between frames. Each frame $\mathbf{F}_t$ is transformed into the *Lab* color space, in which luminance channel and two chrominance channels are separated, and then the simple linear iterative clustering (SLIC) algorithm [62] is used to segment $\mathbf{F}_t$ into a set of perceptually homogenous superpixels, $sp_{t,i}(i=1,...,n_t)$, where n_t is the number of the generated superpixels and should be set to preserve different boundaries between objects and background in the superpixel segmentation results. The analysis of this parameter on saliency detection performance is presented in Section III-B. Note that the notation sp_t, which omits the subscript i representing the spatial index of superpixel, denotes all superpixels in $\mathbf{F}_t$. For the example video in Fig. 1(a), the motion vector fields are visualized in Fig. 1(b) and the superpixel segmentation results are shown in Fig. 1(c), which delineates the boundaries between adjacent superpixels using red lines.

Superpixel is the primitive processing unit in the proposed SGSP model. Specifically, superpixel-level motion/color histograms are extracted as the features for saliency measurement. Superpixel-level graph is built for measuring motion saliency in Section II-B, and temporal propagation in Section II-C and spatial propagation in Section II-D are also performed on the basis of superpixels.

Based on the superpixel segmentation result of each frame $\mathbf{F}_t$ and the corresponding motion vector field $\mathbf{MVF}_{t,t-1}$, motion histograms are extracted at superpixel level to represent the local motion pattern and also extracted at frame level to represent the global motion pattern. The orientation space of motion vectors in the complete range of $[-\pi,\pi]$ is uniformly quantized into $b_{\mathrm{M}}=8$ intervals, each covers $\pi/4$ radians. For each superpixel $sp_{t,i}(i=1,...,n_t)$, its superpixel-level motion histogram $\mathbf{H}_{t,i}^{\mathrm{M}}$ with b_{M} bins is calculated using the motion vectors of all pixels in $sp_{t,i}$. For the k^{th} bin of $\mathbf{H}_{t,i}^{\mathrm{M}}$, those motion vectors that fall into the k^{th} orientation space are accumulated to obtain $\mathbf{H}_{t,i}^{\mathrm{M}}(k)$. To represent the local motion pattern, $\mathbf{H}_{t,i}^{\mathrm{M}}$ is normalized to have $\sum_{k=1}^{b_{\mathrm{M}}}\mathbf{H}_{t,i}^{\mathrm{M}}(k)=1$.

Similarly, the global motion histogram $\mathbf{H}_{t,0}^{\mathrm{M}}$, which is exploited to represent the motion pattern of background in $\mathbf{F}_t$, is initialized using the motion vectors of all pixels in $\mathbf{F}_t$ and normalized to have $\sum_{k=1}^{b_{\mathrm{M}}}\mathbf{H}_{t,0}^{\mathrm{M}}(k)=1$. Please note that the subscript '0' in $\mathbf{H}_{t,0}^{\mathrm{M}}$ is not actually an index. Since the subscript i in the superpixel-level motion histogram $\mathbf{H}_{t,i}^{\mathrm{M}}$ represents the index of superpixel and starts from 1, the subscript '0' is used here to denote the global motion histogram as $\mathbf{H}_{t,0}^{\mathrm{M}}$, so as to keep the consistent format of notations for motion histograms.

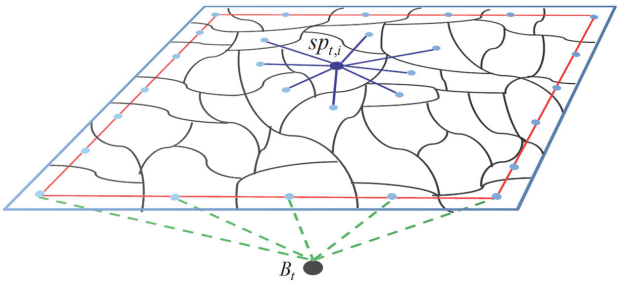

Fig. 2. Schematic example of superpixel-level graph.

Color histogram is also extracted for each superpixel to represent the color distribution of superpixel. For each frame $\mathbf{F}_t$ in the *Lab* color space, each color channel is uniformly quantized into $b_{\mathrm{C}}=16$ bins for a moderate color quantization of video frames. For each superpixel $sp_{t,i}$, the color histogram at each color channel, i.e., $\mathbf{H}_{t,i}^{\mathrm{C,L}}$, $\mathbf{H}_{t,i}^{\mathrm{C,a}}$ and $\mathbf{H}_{t,i}^{\mathrm{C,b}}$, respectively, is calculated with the normalization operation similarly as $\mathbf{H}_{t,i}^{\mathrm{M}}$. Then the above three color histograms are concatenated to form the superpixel-level color histogram $\mathbf{H}_{t,i}^{\mathrm{C}}=\left[\mathbf{H}_{t,i}^{\mathrm{C,L}},\mathbf{H}_{t,i}^{\mathrm{C,a}},\mathbf{H}_{t,i}^{\mathrm{C,b}}\right]$ with $3b_{\mathrm{C}}$ bins for $sp_{t,i}$.

B. Superpixel-level Graph Based Motion Saliency

Based on the superpixel segmentation result of each frame $\mathbf{F}_t$, an undirected weighted graph $\mathbf{G}_t=(\mathbf{V}_t,\mathbf{E}_t)$ is built at superpixel-level, and an schematic example is shown in Fig. 2. The node set $\mathbf{V}_t$ contains regular nodes $\{v_{t,i}\}_{i=1}^{n_t}$, which correspond to all superpixels $\{sp_{t,i}\}_{i=1}^{n_t}$ in $\mathbf{F}_t$, and an additional virtual background node B_t, which represents the motion of background regions in $\mathbf{F}_t$. As shown in Fig. 2, the edge set $\mathbf{E}_t$ contains three types of edges: a) blue edge between each pair of adjacent regular nodes; b) red edge between any two regular nodes, which correspond to two superpixels at the border of video frame; c) green edge between the virtual background node B_t and each regular node, which corresponds to a superpixel at the border of video frame.

The weight for the blue/red edge, which connects two regular nodes, $v_{t,i}$ and $v_{t,j}$, is defined as the motion difference between the corresponding superpixels, $sp_{t,i}$ and $sp_{t,j}$, with the following form:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

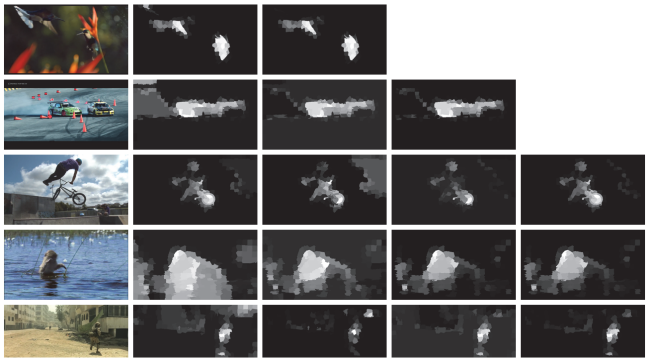

Fig. 3. Examples of iterative motion saliency measurement process. From left to right: original video frames, motion saliency maps after the 1st, 2nd, 3rd and 4th iteration, respectively.

$$\omega_a\left(v_{t,i}, v_{t,j}\right) = \exp\left[\lambda \cdot \chi^2\left(\mathbf{H}^{\mathrm{M}}_{t,i}, \mathbf{H}^{\mathrm{M}}_{t,j}\right)\right], \quad (1)$$

where $\chi^2(.)$ is the chi-squared distance between the two motion histograms, and the coefficient λ is set to 0.1 for a moderate exponential effect.

The weight for the green edge, which connects the virtual background node B_t with a regular node $v_{t,i}$, is similarly defined as follows:

$$\omega_b\left(v_{t,i}, B_t\right) = \exp\left[\lambda \cdot \chi^2\left(\mathbf{H}^{\mathrm{M}}_{t,i}, \mathbf{H}^{\mathrm{M}}_{t,0}\right)\right], \quad (2)$$

where B_t is associated with the global motion histogram $\mathbf{H}^{\mathrm{M}}_{t,0}$.

The motion saliency of each superpixel $sp_{t,i}$ can be measured as the difference between its motion histogram $\mathbf{H}^{\mathrm{M}}_{t,i}$ and the global motion histogram $\mathbf{H}^{\mathrm{M}}_{t,0}$. However, $\mathbf{H}^{\mathrm{M}}_{t,0}$ needs to be refined by excluding those superpixels possibly belonging to salient objects for a more accurate estimation. For each node $v_{t,i}$ on the superpixel-level graph $\mathbf{G}_t$, the sum of edge weights along the shortest path from $v_{t,i}$ to B_t can be naturally used to measure the difference between $\mathbf{H}^{\mathrm{M}}_{t,i}$ and $\mathbf{H}^{\mathrm{M}}_{t,0}$. Therefore, based on the above considerations, an iterative motion saliency measurement method which utilizes the shortest path algorithm on the graph is proposed as follows:

1) Calculate superpixel-level and global motion histograms $\left\{\mathbf{H}^{\mathrm{M}}_{t,i}\right\}_{i=0}^{n_t}$ and initialize $\mathbf{G}_t$ using Eqs. (1) and (2).

2) Find the shortest path from each regular node $v_{t,i}$ to the virtual background node B_t using Dijkstra's algorithm, and the summation of edge weights along the shortest path is defined as the motion saliency of $sp_{t,i}$ with the following form:

$$\mathbf{M}_t(i) = \min_{u_1 = v_{t,i}, u_2, \ldots, u_{m-1}, u_m = B_t} \left[\sum_{k=1}^{m-2} \omega_a\left(u_k, u_{k+1}\right) + \omega_b\left(u_{m-1}, u_m\right)\right], \quad (3)$$

where $\mathbf{M}_t(i)$ denotes the motion saliency of $sp_{t,i}$ compared to the global motion associated with B_t. A higher value of $\mathbf{M}_t(i)$ indicates that the motion of $sp_{t,i}$ is more distinctive from the global motion.

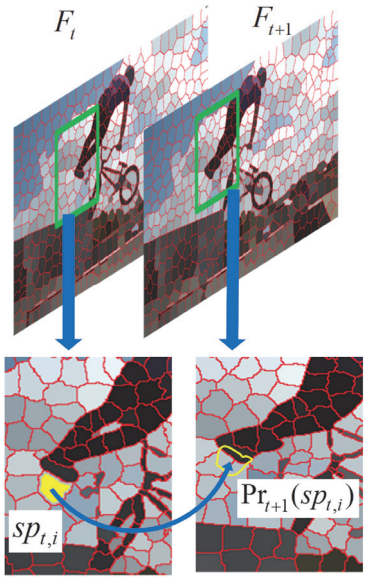

Fig. 4. Pictorial illustration of projection operation on superpixel.

3) In order to more accurately estimate the global motion histogram $\mathbf{H}^{\mathrm{M}}_{t,0}$, the superpixel-level motion saliency map $\mathbf{M}_t$ is thresholded using the Otsu's method [63] to estimate background regions, which have the relatively lower motion saliency. Then $\mathbf{H}^{\mathrm{M}}_{t,0}$ is updated by using the motion vectors in the estimated background regions.

4) Calculate the L_1-norm distance between the two versions of $\mathbf{H}^{\mathrm{M}}_{t,0}$ (after update and before update). If this distance is smaller than a pre-determined value, which is set to a rather smaller value, 0.01, for a stable estimation of $\mathbf{H}^{\mathrm{M}}_{t,0}$, the whole iteration process is terminated. Otherwise, based on the current $\mathbf{H}^{\mathrm{M}}_{t,0}$, the graph $\mathbf{G}_t$ is updated by recalculating the weights of green edges using Eq. (2), and then go to step 2) for the next round of iteration.

Some examples of the iterative superpixel-level motion saliency measurement process are shown in Fig. 3, which contains video frames with various motion activities. It can be seen from Fig. 3 that the visual quality of motion saliency maps improves with the iteration process, i.e., background regions are more effectively suppressed and more accurate object regions are highlighted with the increase of the iteration times. Besides, Fig. 3 also shows that the whole iteration process for each video frame adaptively terminates as the motion saliency map becomes visually stable.

C. Temporal Propagation

It should be noted that the motion saliency maps only utilize the motion information between adjacent frames. In order to enhance the temporal coherence of saliency measurement through consecutive frames, motion saliency measures of superpixels are temporally propagated over video frames to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

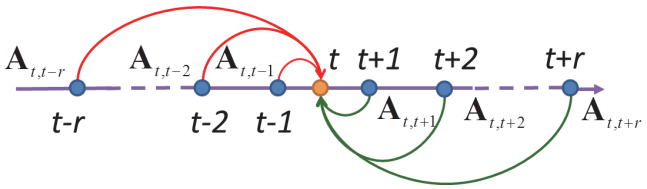

Fig. 5. Illustration of temporal propagation in both forward and backward directions.

obtain for superpixels the temporal saliency measures, which are more temporally coherent and more reliable for saliency measurement on video. The following will describe the temporal propagation of saliency in detail.

First, the similarity between any pair of superpixels in the adjacent frames, $\mathbf{F}_t$ and $\mathbf{F}_{t+1}$, is evaluated based on their appearance similarity and overlap degree. All similarity measures between each superpixel in $\mathbf{F}_t$ and each superpixel in $\mathbf{F}_{t+1}$ form an inter-frame similarity matrix $\mathbf{A}_{t,t+1}$ with a dimension of $n_t \times n_{t+1}$. Each element $\mathbf{A}_{t,t+1}(i,j)$ which represents the similarity between $sp_{t,i}$ and $sp_{t+1,j}$ is defined as follows:

$$\mathbf{A}_{t,t+1}(i,j) = \exp\left[-\frac{1}{\gamma}\cdot\chi^2\left(\mathbf{H}^{\mathrm{C}}_{t,i},\mathbf{H}^{\mathrm{C}}_{t+1,j}\right)\right] \cdot \frac{|\Pr_{t+1}(sp_{t,i})\bigcap sp_{t+1,j}|}{|sp_{t+1,j}|}. \quad (4)$$

The former term in Eq. (4) measures the appearance similarity between $sp_{t+1,j}$ and $sp_{t,i}$ using the chi-squared distance between their color histograms. The coefficient γ is set to the mean of all chi-squared distances calculated for superpixel pairs, to serve as the normalization factor for calculating the appearance similarity. The overlap degree between $sp_{t+1,j}$ and the projected region of $sp_{t,i}$ in the frame $\mathbf{F}_{t+1}$ is measured by the latter term in Eq. (4), where $|.|$ denotes the number of pixels in the superpixel or the overlapped region. As shown in Fig. 4, the projection operation on the superpixel $sp_{t,i}$ projects each pixel in $sp_{t,i}$ with a displacement of the pixel's motion vector in $\mathbf{MVF}_{t,t+1}$ into the adjacent frame $\mathbf{F}_{t+1}$ to obtain the projected region $\Pr_{t+1}(sp_{t,i})$.

Using the similarity matrices between adjacent frames such as $\mathbf{A}_{t,t+1}$ or $\mathbf{A}_{t,t-1}$, the similarity matrix between any two frames, $\mathbf{F}_t$ and $\mathbf{F}_r$, can be calculated based on matrix multiplication operations in the forward or backward direction as follows:

$$\mathbf{A}_{t,r} = \begin{cases} \mathbf{A}_{t,t+1}\cdot\mathbf{A}_{t+1,t+2}\cdots\cdots\mathbf{A}_{r-1,r}, & r > t \\ \mathbf{A}_{t,t-1}\cdot\mathbf{A}_{t-1,t-2}\cdots\cdots\mathbf{A}_{r+1,r}, & r < t \end{cases}. \quad (5)$$

With the similarity matrix $\mathbf{A}_{t,r}$ for any pair of frames, the temporal saliency of each superpixel $sp_{t,i}$ is measured through the temporal propagation of motion saliency measures in its preceding frames as well as subsequent frames. Specifically, the temporal propagation in the forward direction is performed to obtain the forward temporal saliency measure of $sp_{t,i}$ as follows:

$$\mathbf{T}^{\mathrm{F}}_t(i) = \mathbf{M}_t(i) + \frac{\sum_{r=t+1}^{t+\tau}\sum_{j=1}^{n_r}\mathbf{A}_{t,r}(i,j)\cdot\mathbf{M}_r(j)}{\sum_{r=t+1}^{t+\tau}\sum_{j=1}^{n_r}\mathbf{A}_{t,r}(i,j)}, \quad (6)$$

where the number of frames involved in the propagation, τ, is set to 5, a moderate value for a coherent saliency estimation. Similarly, the temporal propagation in the backward direction is performed to obtain the backward temporal saliency measure of $sp_{t,i}$ as follows:

$$\mathbf{T}^{\mathrm{B}}_t(i) = \mathbf{M}_t(i) + \frac{\sum_{r=t-\tau}^{t-1}\sum_{j=1}^{n_r}\mathbf{A}_{t,r}(i,j)\cdot\mathbf{M}_r(j)}{\sum_{r=t-\tau}^{t-1}\sum_{j=1}^{n_r}\mathbf{A}_{t,r}(i,j)}. \quad (7)$$

The temporal propagation in the forward and backward direction is illustrated in Fig. 5. The temporal propagation in both directions actually exploits the motion saliency measures of superpixels, which have a higher similarity with the current superpixel $sp_{t,i}$, in the preceding frames and subsequent frames, to obtain a more reliable and temporally coherent estimation of saliency for $sp_{t,i}$. Using a combination of forward and backward propagation, the integrated temporal saliency measure of $sp_{t,i}$ is defined as follows:

$$\mathbf{T}_t(i) = \mathbf{T}^{\mathrm{F}}_t(i) + \mathbf{T}^{\mathrm{B}}_t(i). \quad (8)$$

As shown in Fig. 1, both forward and backward temporal saliency maps in Fig. 1(f) and (g) can highlight some salient object regions more effectively than the motion saliency map in Fig. 1(e), and their combination, i.e., the integrated temporal saliency map in Fig. 1(h), obviously highlights the salient object more completely than the previous three saliency maps in Figs. 1(e)-(g). Besides, the performance analysis of different saliency maps generated using the proposed model will be presented in Section III-B with Fig. 7, which demonstrates the saliency detection performance is progressively improved from motion saliency maps to forward/backward temporal saliency maps, and to integrated saliency maps. Therefore, the proposed temporal propagation of saliency in this subsection effectively enhances the reliability and coherence of saliency measurement, and substantially improves the saliency detection performance.

D. Spatial Propagation

Based on the integrated temporal saliency map, which has been enhanced on the temporal coherence, we also need to enhance the spatial coherence of saliency estimation in each frame for the even better saliency detection performance. Therefore, the spatial propagation of saliency, which is performed on the basis of each frame, is proposed to for this purpose.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

For any pair of superpixels in each frame $\mathbf{F}_t$, their similarity is evaluated based on the difference between their color histograms. All similarity measures constitute an intra-frame similarity matrix $\mathbf{A}_{t,t}$ with a dimension of $n_t \times n_t$. Following Eq. (4), each element $\mathbf{A}_{t,t}(i,j)$ which represents the similarity between $sp_{t,i}$ and $sp_{t,j}$ is defined as follows:

$$\mathbf{A}_{t,t}(i,j) = \exp\left[-\frac{1}{\gamma}\cdot\chi^2\left(\mathbf{H}^{\mathrm{C}}_{t,i},\mathbf{H}^{\mathrm{C}}_{t,j}\right)\right]. \quad (9)$$

Based on the integrated temporal saliency map $\mathbf{T}_t$, the local spatial propagation of saliency for each superpixel $sp_{t,i}$ is performed to obtain its initial spatiotemporal saliency measure as follows:

$$\mathbf{ST}^{\mathrm{L}}_t(i) = \mathbf{T}_t(i) + \frac{\sum_{sp_{t,j}\in \mathrm{N}(sp_{t,i})}\mathbf{A}_{t,t}(i,j)\cdot\mathbf{T}_t(j)}{\sum_{sp_{t,j}\in \mathrm{N}(sp_{t,i})}\mathbf{A}_{t,t}(i,j)}, \quad (10)$$

where the local neighborhood $\mathrm{N}(sp_{t,i})$ contains all superpixels spatially adjacent to $sp_{t,i}$.

Then the global spatial propagation of saliency exploits those superpixels, which could be outside of the local neighborhood of $sp_{t,i}$ but have a higher similarity with $sp_{t,i}$ over the whole frame $\mathbf{F}_t$, to propagate their saliency measures to $sp_{t,i}$, and obtain the final spatiotemporal saliency measure of $sp_{t,i}$ as follows:

$$\mathbf{ST}^{\mathrm{G}}_t(i) = \mathbf{ST}^{\mathrm{L}}_t(i) + \frac{\sum_{sp_{t,j}\in \mathrm{K}(sp_{t,i})}\mathbf{A}_{t,t}(i,j)\cdot\mathbf{D}_t(i,j)\cdot\mathbf{ST}^{\mathrm{L}}_t(j)}{\sum_{sp_{t,j}\in \mathrm{K}(sp_{t,i})}\mathbf{A}_{t,t}(i,j)\cdot\mathbf{D}_t(i,j)}, \quad (11)$$

where $\mathrm{K}(sp_{t,i})$ is the global neighborhood for $sp_{t,i}$ and $\mathbf{D}_t(i,j)$ is the factor of spatial distance between $sp_{t,i}$ and any superpixel $sp_{t,j}$ in $\mathrm{K}(sp_{t,i})$. Specifically, $\mathrm{K}(sp_{t,i})$ contains the superpixels, which are most similar with $sp_{t,i}$, found in the frame $\mathbf{F}_t$ by using the KD tree. $\mathbf{D}_t(i,j)$ is defined as follows:

$$\mathbf{D}_t(i,j) = \exp\left[-d_t(i,j)\right], \quad (12)$$

where $d_t(i,j)$ is the Euclidian distance between the centroids of $sp_{t,i}$ and $sp_{t,j}$, which are normalized using the diagonal length of video frame to the range of [0, 1].

Using the above spatial propagation, for each superpixel $sp_{t,i}$, the saliency measures of superpixels which have a similar appearance with $sp_{t,i}$ and a short distance to $sp_{t,i}$ will contribute higher to the spatiotemporal saliency of $sp_{t,i}$. As shown in Fig. 1(i) and (j), the local and global spatial propagation of saliency can enhance the spatial coherence of saliency and highlight the complete salient object more accurately than the integrated temporal saliency map in Fig. 1(h). Besides, the performance analysis presented in Section III-B with Fig. 7 will demonstrate the effectiveness of spatial propagation, which further improves the saliency detection performance compared to integrated temporal saliency maps.

III. Experimental Results

We performed a comprehensive performance evaluation of the proposed SGSP model based on extensive experimental results on two video datasets. The video datasets and experimental settings are introduced in Section III-A, and the performance analysis of our SGSP model is shown in Section III-B. Then objective evaluation and subjective evaluation including comparisons with eight state-of-the-art spatiotemporal saliency models are presented in Section III-C and III-D, respectively, and some failure cases are analyzed in Section III-E. Finally, the computation issue of spatiotemporal saliency models is discussed in Section III-F, and the effect of motion vector field on performance and computational efficiency is discussed in Section III-G.

A. Datasets and Experimental Settings

We performed extensive experiments on two video datasets with manually annotated binary ground truths for salient objects. The first dataset SegTrackV2 [60] contains 14 videos with various scenes and diverse motion activities. Note that for videos containing multiple salient objects, we combined the original ground truth maps for individual objects into a unified ground truth map. Nonetheless, for the video *penguins* (see the bottom example in Fig. 12), since only several penguins in the center are annotated as salient objects in the original ground truths, we annotated all penguins as salient objects to generate more suitable ground truths for the purpose of evaluating saliency detection performance. The second dataset UVSD is created by us to introduce more unconstrained videos with complicated motion and complex scenes. The UVSD dataset contains a total of 18 videos, and we accurately segmented salient objects in each video frame using the Adobe Photoshop CC software, to generate the binary ground truths.

We compared our SGSP model with eight state-of-the-art spatiotemporal saliency models including SR [24], CE [30], QFT [15], MB [19], DCMR [40], SP [55], SLT [56] and GD [11]. For all the eight models, the source codes with default parameter settings or executables provided by the authors were used for all videos in the two datasets. 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.

B. Performance Analysis

Our SGSP model first generates motion saliency (MS) map in Section II-B, then generates forward temporal saliency (FTS) map, backward temporal saliency (BTS) map and integrated temporal saliency (ITS) map in Section II-C, and finally generates initial spatiotemporal saliency (IStS) map and final spatiotemporal saliency (FStS) map in Section II-D. Since our SGSP model is built on the basis of superpixels, the number of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

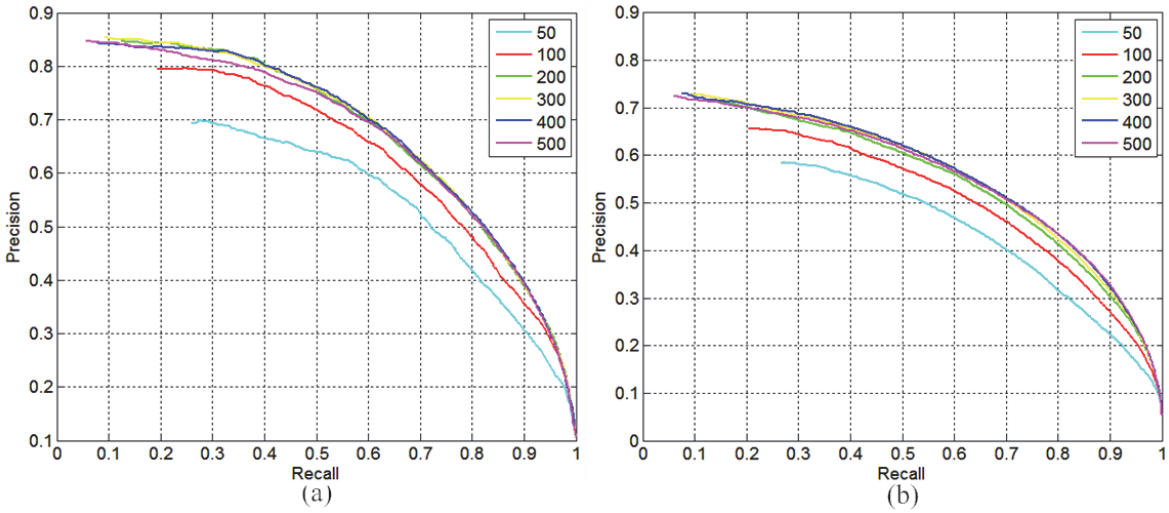

Fig. 6. (better viewed in color) PR curves of final spatiotemporal saliency (FStS) maps generated using our SGSP model with different number of superpixels on (a) SegTrackV2 and (b) UVSD.

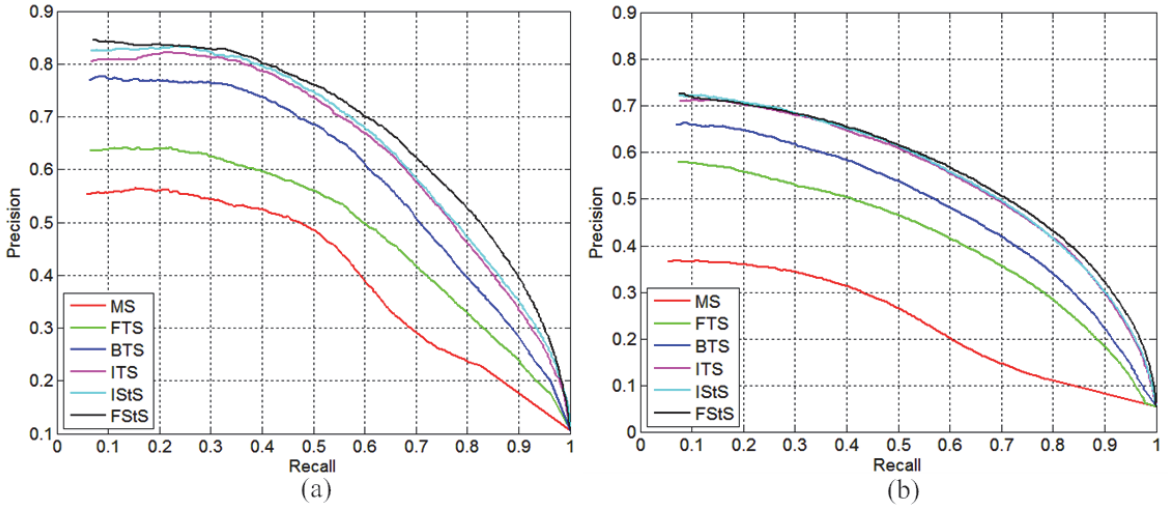

Fig. 7. (better viewed in color) PR curves of different saliency maps generated using our SGSP model, i.e., MS, FTS, BTS, ITS, IStS and FStS, respectively, on (a) SegTrackV2 and (b) UVSD.

the generated superpixels for each frame, i.e., n_t, is important to the saliency detection performance. Therefore, we first analyzed the effect of n_t with a series of values from 50 to 500, and we adopted the commonly used precision-recall (PR) curve, which plots the precision measure against the recall measure to characterize the saliency detection performance.

Specifically, the thresholding operations using a series of fixed integers from 0 to 255 are performed on each saliency map, and thus a total of 256 binary salient object masks are obtained for each saliency map and a set of precision and recall values are calculated by comparing with the corresponding binary ground truth. Then for each class of saliency maps (for example, all the FStS maps generated for all video frames in a dataset using our SGSP model with a specific value of n_t), at each threshold, the precision and recall values of all saliency maps are averaged, and the 256 average precision values against the 256 average recall values are plotted to generate the PR curve.

The PR curves of our FStS maps generated with different values of n_t are shown in Fig. 6. We can observe from the PR curves on the two datasets in Fig. 6 that the saliency detection performance is close with n_t not less than 200, while the performance obviously degrades with the decrease of n_t less than 200. Therefore, we can conclude that the performance of our SGSP model is rather stable with moderate segmentation and over-segmentation of superpixels, while the performance degrades with the under-segmentation of superpixels. Based on Fig. 6, n_t is set to 400, which achieves the overall better performance on both datasets, and is fixed for all the following experiments including the comparisons with all the other spatiotemporal saliency models.

To objectively evaluate the contribution of different parts in our SGSP model to the saliency detection performance, the PR curves for different classes of saliency maps, i.e., MS, FTS, BTS, ITS, IStS and FStS, are shown in Fig. 7. We can observe from the PR curves in Fig. 7 the better saliency detection performance of ITS over FTS/BTS and FTS/BTS over MS on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

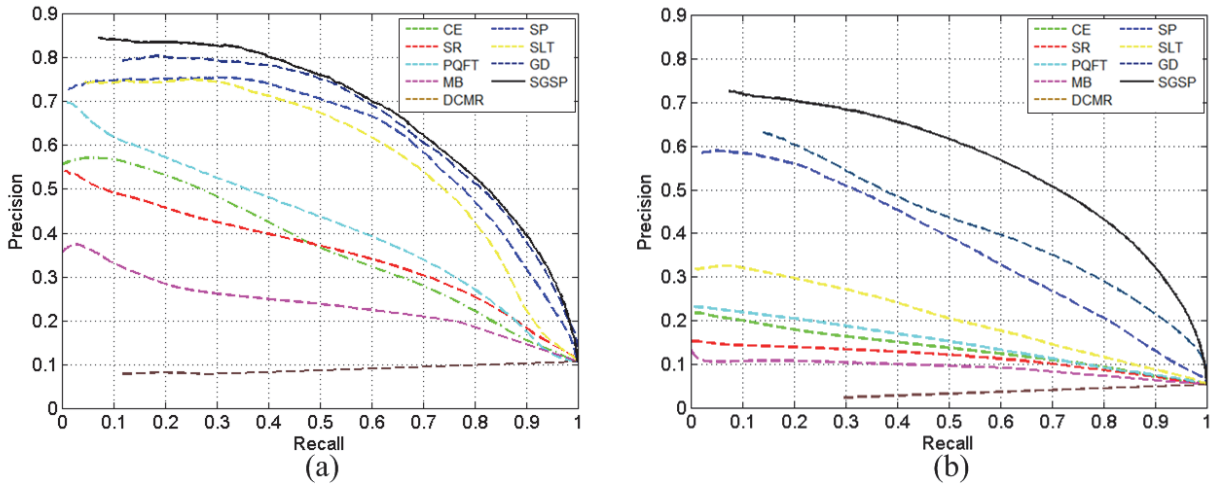

Fig. 8. (better viewed in color) PR curves of different saliency models on (a) SegTrackV2 and (b) UVSD.

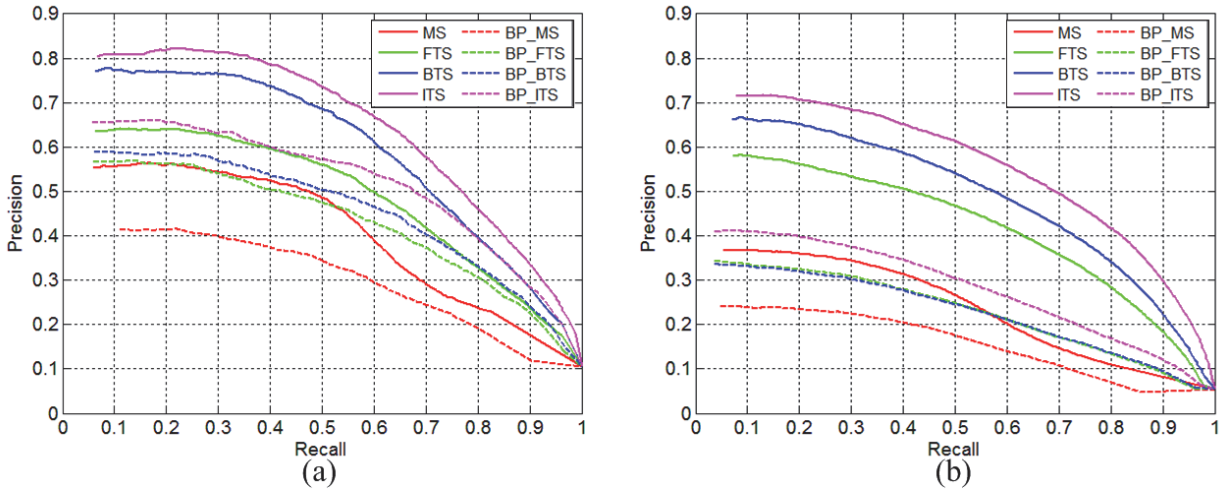

Fig. 9. (better viewed in color) Comparison with PR curves of saliency maps generated using the belief propagation algorithm and the proposed temporal propagation method on (a) SegTrackV2 and (b) UVSD.

both datasets. This clearly demonstrates the effectiveness of the proposed temporal propagation of saliency on the reasonable estimate of motion saliency. Furthermore, we can also observe from Fig. 7 the better saliency detection performance of FStS over IStS and IStS over ITS on both datasets. This clearly shows that the proposed spatial propagation of saliency can further improve the saliency detection performance. In summary, these PR curves in Fig. 7 objectively demonstrate the contribution of each part in our SGSP model to saliency detection performance.

C. Objective Evaluation

(a) Performance comparison with other models: In order to objectively evaluate the saliency detection performance of different models, the PR curves of all the other eight models and our SGSP model on the two video datasets are plotted in Fig. 8 for a comparison. It can be seen from Fig. 8 that our SGSP model consistently outperforms all the other eight models on both datasets. As shown in Fig. 8(b), the performance gain of our SGSP model over all the other eight models is more significant on the UVSD dataset. This demonstrates that our SGSP model effectively improves the saliency detection performance on more unconstrained videos with complicated motion and complex scenes. Besides, the PR curves of the four models including SP, SLT, GD and SGSP are substantially higher than those of the other five models. The common characteristic of SP, SLT, GD and SGSP is that they all use superpixel segmentation and the optical flow estimation method LDOF, which serve as the base for the relatively higher performance. In this sense, the better performance of our SGSP model over SP, SLT and GD is persuasive.

(b) Comparison with the global optimization method on motion saliency measurement: On the basis of the superpixel-level graph $\mathbf{G}_t$ initialized in Section II-B, some global optimization method such as belief propagation [64, 65] can be used to replace the proposed iterative motion saliency measurement method for estimating the superpixel-level motion saliency. Inspired by the work in [66], which uses the belief propagation algorithm to optimize the saliency measures of superpixels, the optimal motion saliency $\mathbf{M}_t(i)$ for each superpixel $sp_{t,i}$ is obtained by minimizing the energy function as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

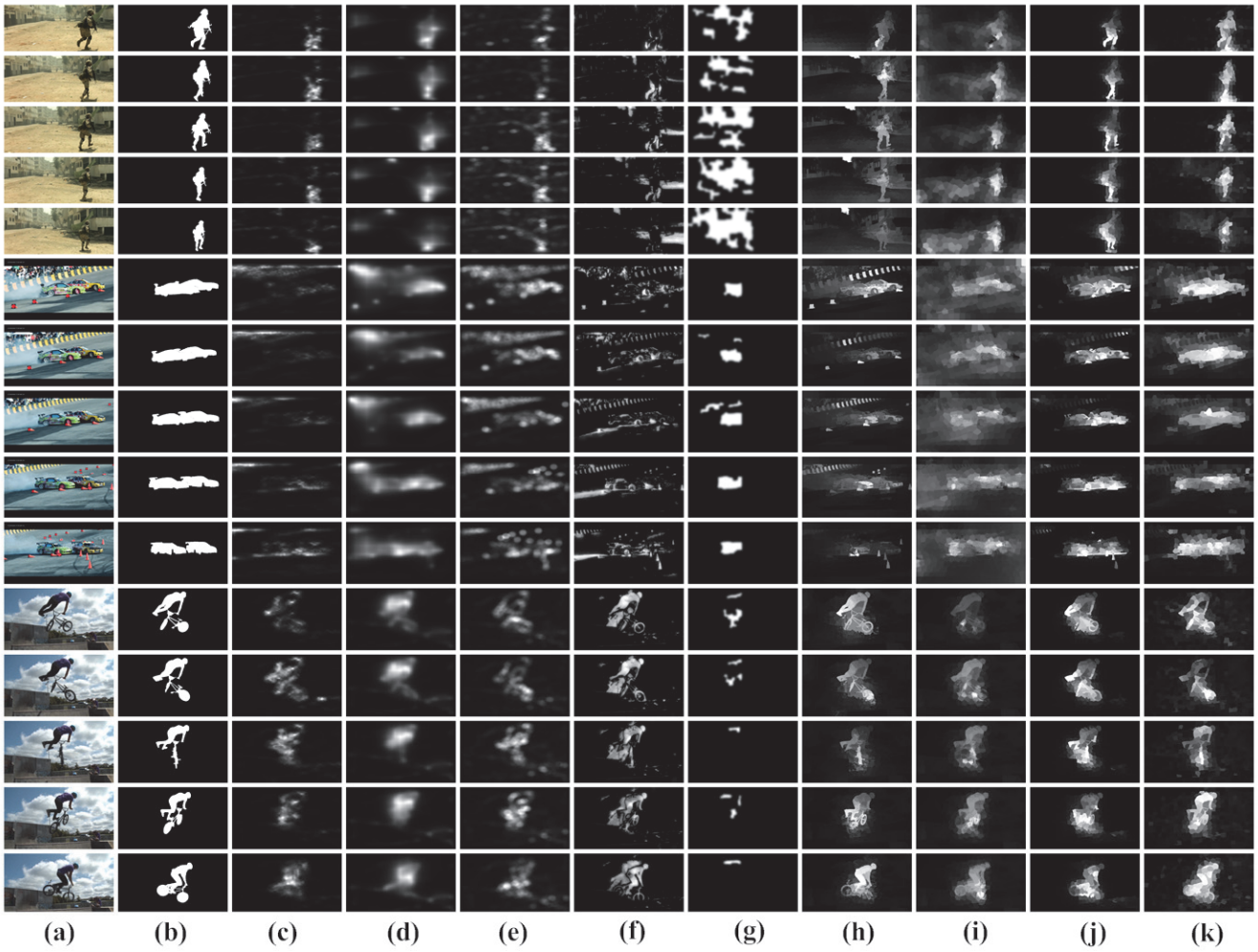

Fig. 10. Examples of spatiotemporal saliency maps for some videos in SegTrackV2 (shown with an interval of 6, 10, and 5 frames, respectively, from top to bottom). (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) SR, (d) CE, (e) QFT, (f) MB, (g) DCMR, (h) SP, (i) SLT, (j) GD and (k) our SGSP model.

$$\sum_i E_d\left(\mathbf{M}_t(i)\right)+\eta\sum_{sp_{t,j}\in\mathrm{N}(sp_{t,i})} E_s\left(\mathbf{M}_t(i),\mathbf{M}_t(j)\right), \tag{13}$$

where the data term is defined as

$$E_d\left(\mathbf{M}_t(i)\right)=\left\|\mathbf{M}_t(i)-\mathbf{M}_t^O(i)\right\|^2, \tag{14}$$

the smoothness term is defined as

$$E_s\left(\mathbf{M}_t(i),\mathbf{M}_t(j)\right)=\frac{\left\|\mathbf{M}_t(i)-\mathbf{M}_t(j)\right\|^2}{\omega_a\left(v_{t,i},v_{t,j}\right)}, \tag{15}$$

and η is the weight for balancing the two terms. Using the similar formulation as Eq. (1) and (2), the observed motion saliency for each superpixel $sp_{t,i}$ is defined as

$$\mathbf{M}_t^O(i)=\exp\left[\lambda\cdot\chi^2\left(\mathbf{H}_{t,i}^{\mathrm{M}},\mathbf{H}_{t,0}^{\mathrm{M}}\right)\right]. \tag{16}$$

The belief propagation algorithm is used to minimize the above energy function and to obtain the optimal motion saliency for each superpixel, and the corresponding motion saliency maps are abbreviated as BP_MS maps. The PR curves of BP_MS maps with the weight η equal to 0.1, which achieves the overall better saliency detection performance on both datasets, are shown in Fig. 9 for a comparison with our MS maps. It can be seen from Fig. 9 that the PR curves of our MS maps are consistently higher than those of BP_MS maps on both datasets. Therefore, compared to the above belief propagation based method, the proposed iterative motion saliency measurement method achieves the better saliency detection performance. The advantage of the proposed method is mainly due to the reasonable use of shortest path algorithm for calculating motion saliency and the update of superpixel-level graph with the iteratively refined global motion histogram.

(c) Evaluation of the proposed temporal propagation method: Our SGSP model enables an effective bidirectional temporal propagation of saliency by using the proposed method in Section II-C, which exploits efficient operations on inter-frame similarity matrices to propagate saliency in both forward and backward directions. To further evaluate the effectiveness of bidirectional temporal propagation, we used another class of motion saliency maps, for example, the above BP_MS maps, as the input, to correspondingly generate the forward temporal saliency (BP_FTS) maps, backward temporal saliency (BP_BTS) maps and integrated temporal saliency (BP_ITS)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

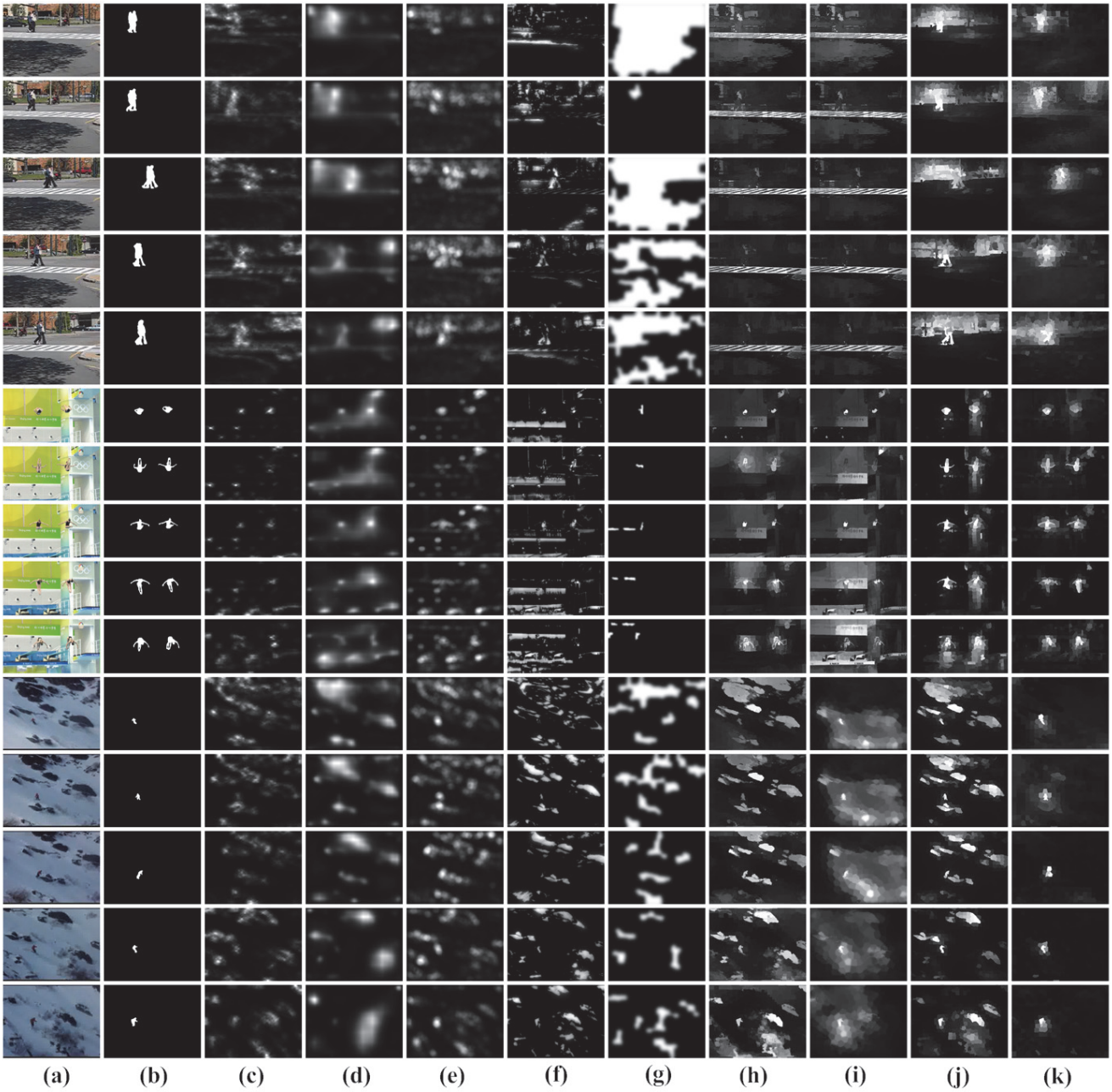

Fig. 11. Examples of spatiotemporal saliency maps for some videos in UVSD (shown with an interval of 10, 15, and 40 frames, respectively, from top to bottom). (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) SR, (d) CE, (e) QFT, (f) MB, (g) DCMR, (h) SP, (i) SLT, (j) GD and (k) our SGSP model.

maps in turn by using the proposed temporal propagation method in Section II-C. The PR curves of these classes of saliency maps are also shown in Fig. 9. With different classes of motion saliency maps, i.e., either MS maps or BP_MS maps, as the input, we can observe from the PR curves in Fig. 9 the better saliency detection performance of integrated temporal saliency maps over forward/backward temporal saliency maps and forward/backward temporal saliency maps over motion saliency maps on both datasets. This clearly demonstrates the robustness of the proposed temporal propagation method to different input saliency maps, and the advantage of bidirectional temporal propagation over unidirectional temporal propagation. Besides, by comparing each pair of PR curves for MS and BP_MS, FTS and BP_FTS, BTS and BP_BTS, ITS and BP_ITS, respectively, it can be seen from Fig. 9 that with the better motion saliency maps, the correspondingly generated forward/backward/integrated saliency maps achieve the better saliency detection performance.

D. Subjective Evaluation

Spatiotemporal saliency maps generated using our SGSP model and all the other eight spatiotemporal models for some videos in SegTrackV2 and UVSD are shown in Fig. 10 and Fig. 11, respectively, for a subjective comparison. These videos exhibit various motions of salient objects such as fast motion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

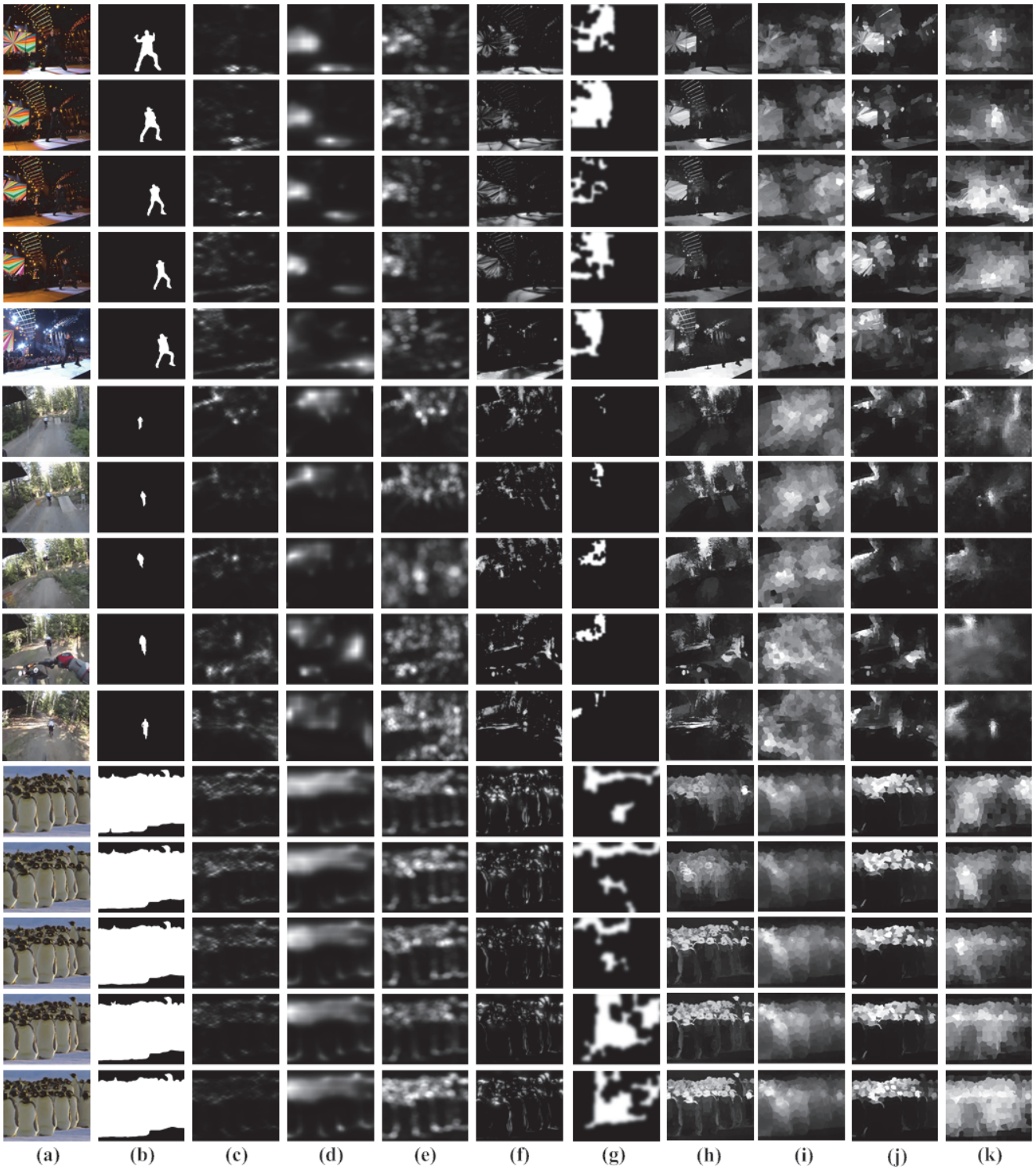

Fig. 12. Failure examples of spatiotemporal saliency maps for some challenging videos (shown with an interval of 30, 15 and 5 frames, respectively, from top to bottom). (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) SR, (d) CE, (e) QFT, (f) MB, (g) DCMR, (h) SP, (i) SLT, (j) GD and (k) our SGSP model.

with cluttered scene (top example in Fig. 10), fast change of motion (middle example in Fig. 10), complicated non-rigid motion (bottom example in Fig. 10), multiple object motions (top and middle examples in Fig. 11) and fast motion of tiny object (bottom example in Fig. 11). Besides, these videos also show various global motions induced by the mixed camera operations of pan, tilt and zoom (all videos in Figs. 10-11) and camera jitter (top example in Fig. 11).

Compared to all the other eight models, we can observe from Figs. 10-11 that the saliency maps generated using our SGSP model can better highlight the complete salient objects and suppress background regions more effectively. SR, CE and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

Table I. Comparison of average processing time per frame taken by different spatiotemporal saliency models.

Model	SR	CE	QFT	MB	DCMR	SP	SLT	GD	SGSP
Time (second)	0.391	3.825	0.098	0.132	0.032	9.432	9.326	9.107	8.766
Code	Matlab	Matlab	Matlab	C++	C++	Matlab	Matlab	Matlab	Matlab

Table II. Average processing time per frame and percentage taken by each component of our SGSP model.

Component	Optical flow estimation (LDOF)	Motion/color histograms	Motion saliency	Temporal propagation	Spatial propagation
Time (second)	8.026	0.338	0.139	0.007	0.256
Percentage (%)	91.6	3.8	1.6	0.1	2.9

QFT only highlight regions around salient object boundaries or a part of salient object regions, but fail to highlight the complete salient object regions and/or falsely highlight some background regions with high contrast. The reason why the above three models generate low-quality saliency maps is that they exploit center-surround differences of features on the basis of patch/volume, which in essence cannot highlight salient object regions completely and cannot adapt well to complicated motions in videos. MB employs background modeling for detecting salient regions, but cannot effectively handle videos with global motion such as these videos in Figs. 10-11. DCMR exploits key point tracking method, which is not robust to form reliable point-level trajectories on salient objects with fast motion and deformation, and the complicated motion further decreases the reliability of point-level trajectories. Therefore, DCMR fails to highlight the complete salient objects in these example videos or even totally misses salient objects in some video frames.

SP, SLT and GD are built on the basis of superpixels, and generally can highlight salient objects and suppress background better than the former five models. However, SP, SLT and GD cannot effectively handle videos with camera jitter and fast object motion (see the top and bottom examples in Fig. 11). In contrast, our SGSP model can better handle such unconstrained videos to improve the quality of saliency maps, because the iterative refinement of global motion histogram and the update of superpixel-level graph in Section II-B enhance the reliability of motion saliency measurement. However, both SP and SLT only exploit the frame-level motion histogram as the representation of global motion without any refinement, and GD exploits the superpixel-level graphs for each frame without an update scheme. Besides, the use of temporal propagation and spatial propagation in Section II-C and II-D, respectively, also effectively enhances the spatiotemporal coherency and improves the quality of our saliency maps.

E. Failure Cases and Analysis

As shown in the previous two subsections, our SGSP model outperforms the state-of-the-art spatiotemporal saliency models on both objective and subjective evaluations. However, it is still difficult for our SGSP model to effectively handle some challenging videos such as the examples shown in Fig. 12. In the top example, besides the salient object (the singer) with body motion, the colorful screen content is changing and the audiences are clapping hands. Therefore the regions covering the screen and audiences also show distinctive motions in contrast with the remaining large background regions, and with the shifting light, such regions are falsely highlighted in our spatiotemporal saliency maps. In the middle example, the salient object (the cyclist) is captured by a camera mounted on the rear bicycle (see the 4th row), so the camera with a severe jitter tries to capture an inward motion of salient object. Although the salient object is highlighted in some frames, it is difficult for our SGSP model to effectively discriminate such unstable inward motion of object from the cluttered motion of background regions. In the bottom example, the area of salient objects (a flock of slowly moving penguins) is rather large, so the global motion histogram is insufficient to effectively represent the motion of background regions, which occupy a small area of the scene. Therefore, the complete salient object regions cannot be consistently highlighted through the video using our SGSP model. As shown in Fig. 12, all the other eight saliency models also cannot handle such challenging videos. Nonetheless, due to the effectiveness of superpixel-level graph based motion saliency measurement and bidirectional temporal propagation, our SGSP model still handles complicated motions and highlights salient objects better than the other saliency models.

F. Computation Cost

We performed all experiments on a PC with Intel Core i7 2600 3.4GHz CPU and 8GB RAM, and compared the computation cost of all saliency models. Table I reports for each model the average processing time per frame for videos with a resolution of 320×240. The average processing time per frame taken by the MATLAB implementation of our SGSP model is 8.766 seconds. It can be seen from Table I that the four models including SGSP, GD, SLT and SP have the higher computation cost than the other five models, since the four models use the time-consuming LDOF method for optical flow estimation to obtain motion vector fields. The average

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

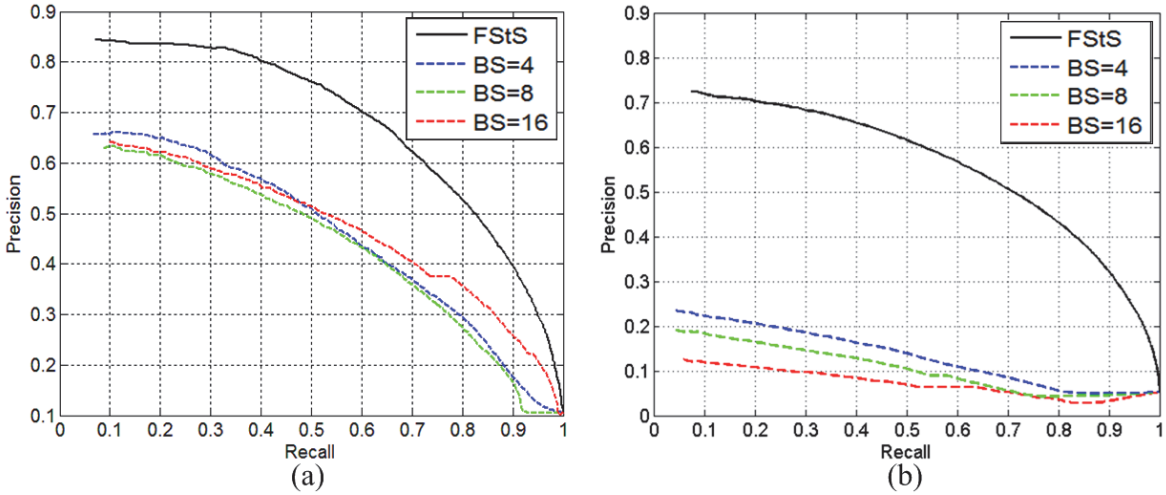

Fig. 13. (better viewed in color) Comparison with PR curves of final spatiotemporal saliency maps generated with block-level motion vector fields on (a) SegTrackV2 and (b) UVSD.

processing time taken by each component of our SGSP model is shown in Table II, and it can be seen that the optical flow estimation process occupies 91.6% of the total processing time, while the saliency measurement process from Section II-B to II-D (the latter three components in Table II) only takes 4.6% of the total processing time.

Therefore, in order to make our SGSP model more practical for applications with runtime requirements, the computational efficiency of the LDOF method, which is the bottleneck of runtime, should be elevated with the highest priority. Fortunately, as reported in [67], using the GPU-accelerated implementation of the LDOF method, the processing time has come down from 143 seconds for the serial implementation on CPU to 1.84 seconds for the parallel implementation on GPU, i.e., with a speedup of 78× for the resolution of 640×480 on an Intel Core2 Quad Q9550 processor running at 2.83GHz in conjunction with a NVIDIA GTX 480 GPU. Therefore, with the GPU-accelerated implementation of the LDOF method and an optimized C/C++ implementation of other components in our SGSP model, we believe that the computation efficiency of our SGSP model can be substantially accelerated.

G. *Effect of Motion Vector Field*

The quality of motion vector field is important to the saliency detection performance of our SGSP model. We performed the following experiments to analyze the effect of motion vector field on the saliency detection performance. By using the block matching method [68], which is commonly used in the motion estimation module of video encoder, the motion vector of each block in each frame $\mathbf{F}_t$ is estimated with the previous frame $\mathbf{F}_{t-1}$ as the reference frame. We used different block sizes including 4×4, 8×8 and 16×16, respectively, and set the search range to 64×64, which is large enough to cover the range of motions between adjacent frames. The obtained block-level motion vector field is finally upsampled to generate the pixel-level motion vector field $\mathbf{MVF}_{t,t-1}$, for the use in our SGSP model. Therefore, for each dataset, we generated a total of three groups of final spatiotemporal saliency maps with three block sizes, which are denoted as BS=4, BS=8 and BS=16, respectively, and their PR curves and the PR curve of our FStS maps are shown in Fig. 13 for a comparison.

We can see from Fig. 13 that the saliency detection performance of our SGSP model degrades on both datasets with the use of block-level motion vector fields. Note that the LDOF method is designed for better capturing the large and complicated motion between frames, while the block matching method is designed for finding the matched block with the lowest error such as the minimum sum of absolute difference between blocks. The block-level motion vector fields are with the lower accuracy to represent the complicated motions of object and background regions, and thus degrade the saliency detection performance. Compared to the SegTrackV2 dataset, on the UVSD dataset, which contains more challenging videos, the block-level motion vector fields are with the even lower accuracy, and thus the saliency detection performance decreases more significantly. Using the MATLAB implementation, the average processing time for estimating the block-level motion vector field and upsampling to the pixel-level motion vector field per frame with a resolution of 320×240 is 1.256, 0.399 and 0.136 seconds for the block size of 4×4, 8×8 and 16×16, respectively. It can be seen that the use of block-level motion estimation can substantially reduce the computation cost of our SGSP model with the degradation on saliency detection performance.

In summary, the quality of motion vector field obviously affects the saliency detection performance, and thus the optical flow estimation methods such as the LDOF method, which can better handle the large and complicated motion, is the better choice for spatiotemporal saliency models. Indeed, those spatiotemporal saliency models with the higher performance, including SP, SLT, GD and our SGSP model, all use the LDOF method for estimating motion vector fields. However, as shown by those challenging videos in Fig. 12, even using the motion vector fields estimated by the LDOF method, it is still difficult for SP, SLT, GD and our SGSP model to effectively highlight the complete salient objects. In such cases that motion vector

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595324, IEEE Transactions on Circuits and Systems for Video Technology

fields are insufficient for effective saliency detection, we need to resort to other sources of high-performing saliency maps if available. In our recent work [69], an adaptive fusion method via learning a quality prediction model for different saliency maps has shown the effectiveness to improve the saliency detection performance on images. The quality prediction model is learned by using various quality measures [70] for saliency maps, and is used to generate for a given saliency map its quality score, which serves as the weight for adaptive fusion. Therefore, with the tailored quality prediction model, we believe that the adaptive fusion of our spatiotemporal saliency maps with other saliency maps generated using some high-performing saliency models for images may facilitate to further improve the saliency detection performance.

IV. Conclusions

In this paper, we have presented our SGSP model which utilizes superpixel-level graph and spatiotemporal propagation to effectively improve the saliency detection performance on unconstrained videos. Our SGSP model enables the iterative motion saliency measurement based on superpixel-level graph as well as the bidirectional (both forward and backward) temporal propagation of saliency and glocal (both global and local) spatial propagation of saliency via the efficient use of inter-frame and intra-frame similarity matrices, respectively. Extensive experimental results on two video datasets with various unconstrained videos demonstrate the consistently better saliency detection performance of our SGSP model compared to other state-of-the-art spatiotemporal saliency models. However, it is still difficult for our SGSP model to effectively handle some challenging videos, in which the area of salient objects is rather large as well as some background regions show distinctive motions in contrast with the remaining large background regions, and the low-quality motion vector fields. Considering the above limitations of our SGSP model, in our future work, we will investigate an adaptive saliency fusion method to further improve the saliency detection performance on unconstrained videos. In addition, as a meaningful application of spatiotemporal saliency model, we will explore an effective salient object segmentation framework for unconstrained videos.

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] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] Y. Luo, J. Yuan, P. Xue, and Q. Tian, "Saliency density maximization for efficient visual objects discovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1822-1834, Dec. 2011.

[7] 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.

[8] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937-1952, May 2014.

[9] M. M. Cheng, N. J. Mitra, X. L. Huang, P. Torr, S. M. Hu, Salient object detection and segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569-582, Mar. 2015.

[10] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," *Proc. IEEE ICME*, Jun. 2009, pp. 638-641.

[11] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," *Proc. IEEE CVPR*, Jun. 2015, pp. 3395-3402.

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

[13] 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.

[14] H. Du, Z. Liu, J. Jiang, and L. Shen, "Stretchability-aware block scaling for image retargeting," *J. Vis. Commun. Image Represent.*, vol. 24, no. 4, pp. 499-508, May 2013.

[15] 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.

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

[17] L. Shen, Z. Liu, and Z. Zhang, "A novel H.264 rate control algorithm with consideration of visual attention," *Multimedia Tools and Applications*, vol. 63, no. 3, pp. 709-727, Apr. 2013.

[18] 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.

[19] 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.

[20] O. Le Meur, and Z. Liu, "Saccadic model of eye movements for free-viewing condition," *Vision Res.*, vol. 116, pp. 152-164, Nov. 2015.

[21] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," *Proc. ECCV*, Oct. 2012, pp. 414-429.

[22] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55-69, Jan. 2013.

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

[24] H. J. Seo, and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vision*, vol. 9, no. 12, article 15, Nov. 2009.

[25] 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.

[26] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *J. Vision*, vol. 8, no. 7, article 13, Jun. 2008.

[27] 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.

[28] W. Kim, and C. Kim, "Spatiotemporal saliency detection using textural contrast and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 646-659, Apr. 2014.

[29] 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.

[30] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," *Proc. ACCV*, Sep. 2009, pp. 246-257.

[31] X. Hou, and L. Zhang, "Dynamic visual attention: searching for coding length increments," *Proc. NIPS*, Dec. 2008, pp. 681-688.

[32] Y. Li, Y. Zhou, L. Xu, X. Yang and J. Yang, "Incremental sparse saliency detection," *Proc. IEEE ICIP*, Nov. 2009, pp. 3093-3096.

[33] 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.

[34] K. Muthuswamy, and D. Rajan, "Salient motion detection through state controllability," *Proc. IEEE ICASSP*, Mar. 2012, pp. 1465-1468.

[35] X. Hou, and L. Zhang, "Saliency detection: a spectral residual approach," *Proc. IEEE CVPR*, Jun. 2007, pp. 1-8.

[36] X. Cui, Q. Liu, and D. N. Metaxas, "Temporal spectral residual: fast motion saliency detection," *Proc. ACM MM*, Oct. 2009, pp. 617-620.

[37] 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.

[38] 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.

[39] 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.

[40] C. Huang, Y. Chang, Z. Yang, and Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1336-1349, Aug. 2014.

[41] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2009-2021, Dec. 2013.

[42] Y. Luo, and Q. Tian, "Spatio-temporal enhanced sparse feature selection for video saliency estimation," *Proc. IEEE CVPR Workshops*, Jun. 2012, pp. 33-38.

[43] Z. Ren, S. Gao, L. Chia, and D. Rajan, "Regularized feature reconstruction for spatiotemporal saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3120-3132, Aug. 2013.

[44] Z. Ren, L.-T. Chia, and D. Rajan, "Video saliency detection with robust temporal alignment and local-global spatial contrast," *Proc. ACM ICMR*, Jun. 2012, article 47.

[45] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," *Proc. IEEE ICASSP*, Mar. 2012, pp. 1485-1488.

[46] 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.

[47] D. Mahapatra, S. O. Gilani, and M. K. Saini, "Coherency based spatio-temporal saliency detection for video object segmentation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 454-462, Jun. 2014.

[48] 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.

[49] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Res.*, vol. 47, no. 19, pp. 2483-2498, Sep. 2007.

[50] 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.

[51] 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.

[52] 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.

[53] K. Muthuswamy, and D. Rajan," Salient motion detection in compressed domain," *IEEE Signal Process. Lett.*, vol. 20, no. 10, pp. 996-999, Oct. 2013.

[54] Y. Fang, W. Lin, Z. Chen, C. Tsai, and C. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27-38, Jan. 2014.

[55] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522-1540, Sep. 2014.

[56] J. Li, Z. Liu, X. Zhang, O. Le Meur, and L. Shen, "Spatiotemporal saliency detection based on superpixel-level trajectory," *Signal Process.: Image Commun.*, vol. 38, pp. 100-114, Oct. 2015.

[57] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proc. NIPS*, Dec. 2006, pp. 545-552.

[58] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," *Proc. IEEE CVPR*, Jun. 2013, pp. 3166-3173.

[59] Y. Li, K. Fu, L. Zhou, Y. Qiao, J. Yang, and B. Li, "Saliency detection based on extended boundary prior with foci of attention," *Proc. IEEE ICASSP*, May 2014, pp. 2798-2802.

[60] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments", *Proc. IEEE ICCV*, Dec. 2013, pp.2192-2199.

[61] 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.

[62] 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.

[63] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62-66, Jan. 1979.

[64] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068-1080, Jun. 2008.

[65] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 25-47, Oct. 2000.

[66] S. C. Pei, W. W. Chang, and C. T. Shen, "Saliency detection using superpixel belief propagation," *Proc. IEEE ICIP,* Oct. 2014, pp. 1135-1139.

[67] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by GPU-accelerated large displacement optical flow," *Proc. ECCV*, Sep. 2010, pp. 438-451.

[68] Y. Nie, and K. K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Trans. Image Process*., vol. 11, no. 12, pp. 1442-1448, Dec. 2002.

[69] X. Zhou, Z. Liu, G. Sun, L. Ye, and X. Wang, "Improving saliency detection via multiple kernel boosting and adaptive fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 517-521, Apr. 2016.

[70] L. Mai, and F. Liu, "Comparing salient object detection results without ground truth," *Proc. ECCV*, Sep. 2014, pp. 76-91.

Zhi Liu (M'07-SM'15) 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 120 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. He was a TPC member in ICME 2014, WIAMIS 2013, IWVP 2011, PCM 2010, ISPACS 2010, etc. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an area editor of *Signal Processing: Image Communication* and served as a guest editor for the special issue on *Recent Advances in Saliency Models, Applications*

and Evaluations in this journal.

Junhao Li received the B.E. degree from Hubei Normal University, Huangshi, China, in 2013, and the M.E. degree from Shanghai University, Shanghai, China, in 2016. His research interests include saliency model and salient object detection.

Linwei Ye received the B.E. degree from Hangzhou Dianzi University, Hangzhou, China, in 2013, and the M.E. degree from Shanghai University, Shanghai, China, in 2016. His research interests include saliency model and salient object segmentation.

Guangling Sun received the B.S. degree in electronic engineering from Northeast Forestry University, China, in 1996 and the M.E. and Ph.D. degrees in computer application technology from Harbin Institute of Technology, China, in 1998 and 2003, respectively. Since 2006, she has been with the faculty of the School of Communication and Information Engineering, Shanghai University, where she is currently an Associate Professor. She was with the University of Maryland, College Park as a visiting scholar from December 2013 to December 2014. Her research interests include saliency detection, face recognition, and image/video processing.

Liquan Shen received the B.S. degree in automation control from Henan Polytechnic University, Shanyang, China, and the M.E. and Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China, in 2001, 2005, and 2008, respectively. He was with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA, as a Visiting Professor from 2013 to 2014. He has been with the Faculty of the School of Communication and Information Engineering, Shanghai University, since 2008, where he is currently a Professor. He has authored and co-authored more than 80 refereed technical papers in international journals and conferences in the field of video coding and image processing. He holds 10 patents in the areas of image/video coding and communications. His research interests include High Efficiency Video Coding, perceptual coding, video codec optimization, 3DTV, and video quality assessment.