Signal Processing: *Image Communication* ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Spatiotemporal saliency detection based on superpixel-level trajectory

Junhao Li [a], Zhi Liu [a,*], Xiang Zhang [b], Olivier Le Meur [c], Liquan Shen [a]

[a] *School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China*
[b] *School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[c] *IRISA, University of Rennes 1, Campus Universitaire de Beaulieu, Rennes 35042, France*

ARTICLE INFO

Keywords:
Spatiotemporal saliency detection
Trajectory
Superpixel
Temporal saliency
Spatial saliency

ABSTRACT

In this paper, we propose a novel spatiotemporal saliency model based on superpixel-level trajectories for saliency detection in videos. The input video is first decomposed into a set of temporally consistent superpixels, on which superpixel-level trajectories are directly generated, and motion histograms at superpixel level as well as frame level are extracted. Based on motion vector fields of multiple successive frames, the inside–outside maps are estimated to roughly indicate whether pixels are inside or outside objects with motion different from background. Then two descriptors, i.e. accumulated motion histogram and trajectory velocity entropy, are exploited to characterize the short-term and long-term temporal features of superpixel-level trajectories. Based on trajectory descriptors and inside–outside maps, superpixel-level trajectory distinctiveness is evaluated and trajectory classification is performed to obtain trajectory-level temporal saliency. Superpixel-level and pixel-level temporal saliency maps are generated in turn by exploiting motion similarity with neighboring superpixels around each trajectory, and color-spatial similarity with neighboring superpixels around each pixel, respectively. Finally, a quality-guided fusion method is proposed to integrate the pixel-level temporal saliency map with the pixel-level spatial saliency map, which is generated based on global contrast and spatial sparsity of superpixels, to generate the pixel-level spatiotemporal saliency map with reasonable quality. Experimental results on two public video datasets demonstrate that the proposed model outperforms the state-of-the-art spatiotemporal saliency models on saliency detection performance.

1. Introduction

Humans are able to detect salient objects in complex and dynamic scene effortlessly and rapidly. To mimic such a remarkable capability of human visual system, a number of computational models for saliency detection have been proposed in the past decades, and they harness a variety of applications for image and video, such as salient object detection [1–3], salient object segmentation [4–6], content-aware image/video retargeting [7–9], content-based image/video compression [10,11], quality assessment [12], visual scanpath prediction [13], smart editing [14], image/video memorability estimation [15,16] and geospatial target detection [17,18].

An enormous amount of research efforts has been focused on saliency models for images [19], while the research on spatiotemporal saliency models for videos has increasingly received attention in the recent years. Videos are fundamentally different from static images due that each video contains a sequence of consecutive frames with larger volume of data and richer visual evidences and covers spatial and temporal features simultaneously. Psychological studies showed that

* Corresponding author. Tel.: +86 21 66137270.
E-mail address: liuzhisjtu@163.com (Z. Liu).

http://dx.doi.org/10.1016/j.image.2015.04.014
0923-5965/

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

the temporal cue, i.e. motion, in most videos is the primary factor contributing to saliency detection [20]. However, the analysis of motion is challenging, since many factors such as camera motion (moving background), non-rigid deformation and occlusion may mix together and complicate saliency detection in videos. Intuitively, image saliency models can be used for saliency detection independently in each video frame regardless of temporal cue. However, it ignores the fact that human visual system determines salient regions in still images and video sequences in different ways [21]. Therefore, many researchers have proposed various models using spatiotemporal information among multiple frames based on different theories and methods.

Based on information theory, self-information [22], minimum conditional entropy (CE) [23] and incremental coding length with different formulations [24,25] are calculated on the basis of patch/volume and used as spatiotemporal saliency measures. Inspired by the modern control theory, video sequence is modeled as a linear dynamic system, and then the observability of output [26] and the controllability of states [27] are exploited to identify regions with salient motion. Using frequency domain analysis method, a multi-resolution spatiotemporal saliency model [10] is built in the frequency domain by using the phase spectrum of quaternion Fourier transform (QFT), which incorporates color, orientation and intensity in each frame and difference between frames in parallel. The spectral residual of the amplitude spectrum of Fourier transform is used for saliency detection in image [28], and as an extension, the temporal spectral residual along the X–T and Y–T slices is exploited to generate the spatiotemporal saliency map for video [29]. Sparse representation method, which is effectively utilized for saliency detection in images [30], is also introduced to compute both spatial and temporal saliency in [31]. Low rank matrix recovery method is used in [32] to decompose video frames into background which is represented by a low rank matrix and salient objects which are recovered by a sparse matrix.

Recently machine learning has been effectively exploited in image saliency models, such as the deep learning architecture for modelling background in [33] and the principled probabilistic formulation of object saliency as a sampling problem in [34]. While for video, spatiotemporal saliency detection is formulated as a dominant camera motion removal (DCMR) problem in [35]. To solve such a problem, the one-class support vector machine (SVM) is utilized to classify the trajectories, which are constructed by extracting a set of key points and tracking them during the video sequence, and the saliency of each trajectory is diffused to its surrounding regions for generating the spatiotemporal saliency map.

Apart from the aforementioned models, the center-surround scheme, which has been widely exploited in a number of image saliency models such as [36], has a straightforward interpretation of human visual attention mechanism and a concise computation form. Therefore, some spatiotemporal saliency models implemented this scheme with different features and formulations. For example, local regression kernel based self-resemblance (SR) [37] is exploited to measure the likeness of a pixel to its spatiotemporal surroundings and then to derive the pixel's spatiotemporal saliency. In [38], the difference between each patch/volume and its spatiotemporal surroundings is evaluated using the Kullback–Leibler divergence on the dynamic texture feature. The multiscale background (MB) model [12], which is represented using Gaussian pyramid, is exploited to identify salient object pixels, which differ significantly from the background pixels. Besides, in the compressed domain [39,40], DCT coefficients and motion vectors extracted from the video bitstream are used to calculate spatial saliency map and temporal saliency map, respectively, and then they are combined together to generate spatiotemporal saliency map. In [41], saliency is measured based on the coherency information, in which the temporal coherency map is obtained using the center-surround operation on motion magnitude and direction, while the spatial coherency map is obtained using the entropy of HOG features. In [42], the directional coherence is measured on the distribution of spatiotemporal gradients, and the contrast of directional coherence is evaluated using a multiscale framework to generate the saliency map.

Note that the above models measure saliency at either regular-shaped patch/volume level, temporal slice level or pixel level, which neglects the importance of video representation for reasonable saliency measurement and usually cannot generate saliency maps with well-defined boundaries. Inspired by the fact that a number of saliency models built on the basis of regions/superpixels significantly elevated the saliency detection performance on images [43–50,4–6], a superpixel-based (SP) spatiotemporal saliency model presented in our previous work [51] exploits motion distinctiveness, global contrast and spatial sparsity of superpixels to measure temporal saliency and spatial saliency, and then combines them using an adaptive fusion scheme to generate spatiotemporal saliency maps with the better saliency detection performance on videos.

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 still insufficient to effectively highlight the complete salient objects with well-defined boundaries and suppress irrelevant background regions in challenging videos with complicated motions. Motivated by the recent work on point based trajectory for action recognition [52] and segmentation [53] in video sequences, we propose a superpixel-level trajectory based spatiotemporal saliency model aiming to improve the saliency detection performance on challenging videos. Compared with the existing spatiotemporal saliency models, our main contributions are threefold. First, to handle motion variability of different videos, two trajectory descriptors, i.e. accumulated motion histogram and trajectory velocity entropy, are exploited to effectively utilize both short-term and long-term motion cues. Second, we propose a novel pipeline, which estimates superpixel-level trajectory distinctiveness and then systematically measures trajectory-level, superpixel-level and pixel-level temporal saliency in turn, to effectively enhance the coherence of temporal saliency through the whole video. Finally, a quality-guided fusion method is proposed to reasonably integrate temporal saliency maps with spatial saliency maps to generate pixel-level spatiotemporal saliency maps. Experimental results on two public datasets show that the proposed model achieves a better saliency detection performance than the state-of-the-art spatiotemporal saliency models.

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

The rest of this paper is organized as follows. The proposed spatiotemporal saliency model is described in Section 2, experimental results and comparisons are presented and analyzed in Section 3, and conclusions are drawn in Section 4.

2. Superpixel-level trajectory based spatiotemporal saliency model

The proposed spatiotemporal saliency model consists of four components, i.e. feature extraction, trajectory descriptors, temporal saliency measurement, and spatiotemporal saliency generation. The following four subsections from Sections 2.1–2.4 will elaborate on the four components in turn. Specifically, the major work of this paper (from Sections 2.1–2.3) focuses on how to effectively generate high-quality temporal saliency maps for videos with complicated motions. The spatial saliency maps, which are used to integrate with the temporal saliency maps in Section 2.4, can be generated using any saliency model for images, and this paper adopted the spatial saliency model presented in our previous work [51]. For clarity, only the temporal saliency measurement process from Sections 2.1–2.3 is illustrated in Fig. 1.

The proposed model is built on the basis of superpixels, which not only fit well to boundaries between salient objects and background in each frame but also show temporal consistency among frames. The superpixel representation of video frames will facilitate to preserve well-defined boundaries and inter-frame coherence of the generated saliency maps through the whole video. In Section 2.1, superpixel-level trajectories are directly generated from the superpixel representations of video frames, and motion histograms are extracted at superpixel level and frame level to effectively represent the local and global distribution of motion, respectively. Besides, inside–outside maps which roughly indicate pixels inside or outside moving objects are estimated for the latter use in trajectory classification. In Section 2.2, two trajectory descriptors, i.e. accumulated motion histogram and trajectory velocity entropy, are exploited to characterize the superpixel-level trajectories.

In Section 2.3, based on trajectory descriptors and inside–outside maps, we propose to evaluate superpixel-level trajectory distinctiveness and exploit a trajectory classification scheme for calculating the trajectory-level temporal saliency, which is adjusted by saliency prediction from neighboring superpixels around each trajectory to obtain the superpixel-level temporal saliency measures along each trajectory. Then a pixel-level saliency derivation method is used to transform superpixel-level temporal saliency measures to pixel-level temporal saliency map. Finally in Section 2.4, a quality-guided fusion method is proposed to integrate temporal saliency map with spatial saliency map at pixel level for generating spatiotemporal saliency map.

2.1. Feature extraction

As the pre-processing step, for each video frame F_t, its pixel-level motion vector field MVF_t is calculated with its previous frame F_{t-1} as the reference frame, by using the state-of-the-art optical flow estimation algorithm [54], which allows capturing large displacements between frames. Then the temporal superpixel (TSP) algorithm [55] is used to partition each video frame F_t into a set of perceptually homogenous and temporally consistent superpixels $sp_t^i(i=1,2,\ldots,r_t)$ where r_t is the number of the generated superpixels. Over the whole video, superpixels are tracked along the temporal direction to yield a set of superpixel-level trajectories, which can be directly obtained from the result of TSP algorithm. 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.

2.1.1. Motion histogram

Based on superpixels and motion vector fields, motion histograms are extracted at superpixel level as the short-term features for superpixels. In order to construct superpixel-level and frame-level motion histograms, the orientation space of motion vectors in the complete range of $[-\pi,\pi]$ is uniformly quantized into $b=8$ intervals, each with $\pi/4$ radians. For each superpixel sp_t^i in F_t, its superpixel-level motion histogram $MH_t^i(i=1,2,\ldots,r_t)$ with b bins, which correspond to the b quantized orientation space, is calculated using the motion vectors of all pixels in sp_t^i. For the kth $(k=1,2,\ldots,b)$ bin of MH_t^i, the amplitudes of those motion vectors that fall into the kth orientation space are accumulated to obtain $MH_t^i(k)$, and the mean value of orientations of those motion vectors is used as the quantized orientation $O_t^i(k)$. Similarly, the frame-level motion histogram MH_t^0 for F_t is calculated using the motion vectors of all pixels in F_t. Please note that the superscript '0' in MH_t^0 is not a spatial index. Since the superscript i in the superpixel-level motion histogram MH_t^i represents the spatial index and starts from 1, the superscript '0' is used here to denote the frame-level motion histogram as MH_t^0, so as to keep the consistent format of notations for motion histograms.

2.1.2. Inside–outside map estimation

Based on the motion vector field between each pair of adjacent frames, a computationally efficient method [56] is used to roughly estimate which pixels fall into a moving object. The estimation result for each frame F_t is represented using the inside–outside map IOM_t, which is a binary map indicating some possible pixels belonging to moving objects. For the video frames shown in the top row of Fig. 2, the inside–outside maps estimated using [56] are shown in the middle row of Fig. 2. However, we found that in some videos containing an object with intermittent motion (the object might be static in some frames), some inside–outside maps such as the latter two in the middle row of Fig. 2 miss a part of object regions, due to no motion on the motion vector fields of the corresponding frames. Therefore, we follow the method in [56] but use the motion vector fields of multiple successive frames, to better handle the intermittent motion of objects and obtain more temporally coherent inside–outside maps. For a visual comparison, the inside–outside maps generated by our scheme are shown in the bottom row of Fig. 2, which is more informative and more reliable than the middle row of Fig. 2. The inside–outside maps approximately indicate the location of moving objects, and will be used for trajectory classification in Section 2.3.2.

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

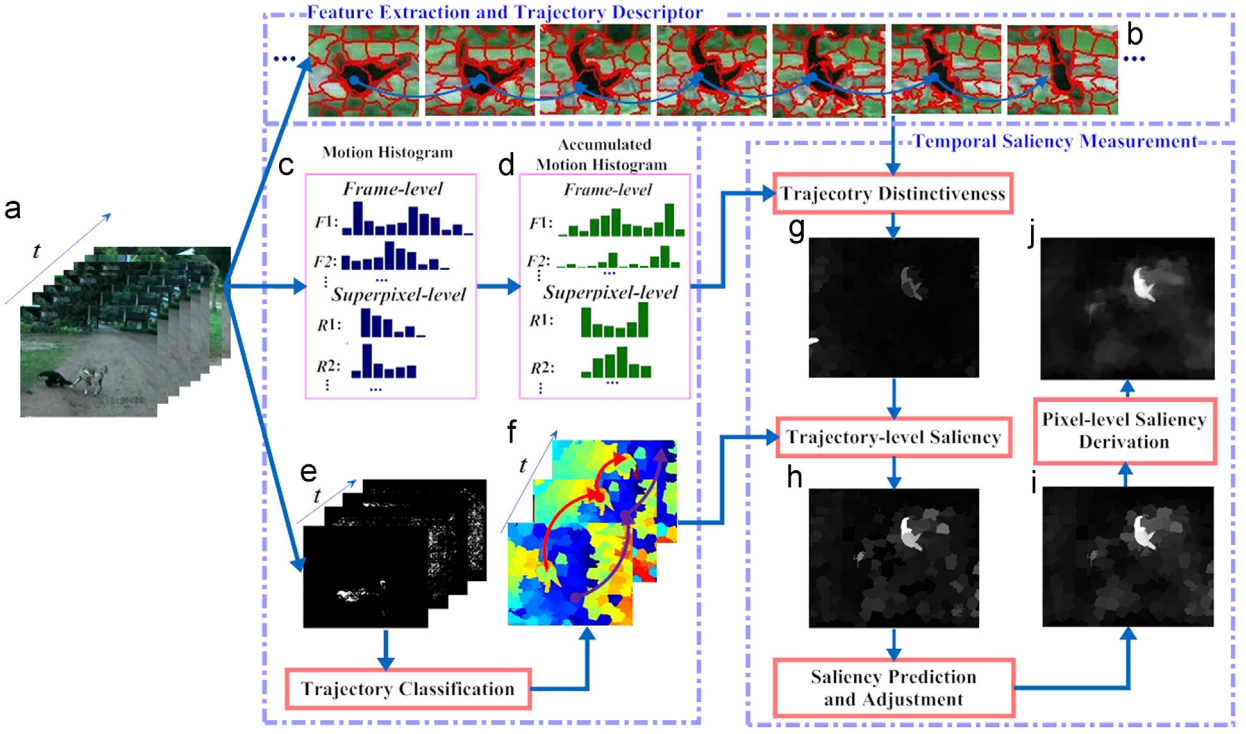

Fig. 1. Overview of temporal saliency measurement. (a) Input video frames; (b) example of superpixel-level trajectory; (c) frame-level and superpixel-level motion histograms; (d) frame-level and superpixel-level accumulated motion histograms; (e) inside–outside maps; (f) examples of salient trajectory (with red arrows) and non-salient trajectory (with purple arrows); (g) superpixel-level trajectory distinctiveness map; (h) trajectory-level temporal saliency map; (i) superpixel-level temporal saliency map; (j) pixel-level temporal saliency map.

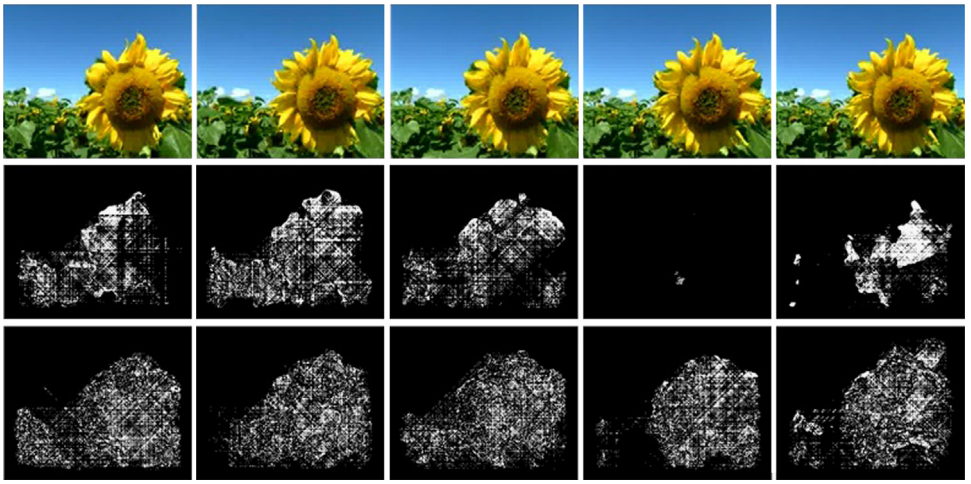

Fig. 2. Examples of inside–outside maps. From top to bottom: original video frames; inside–outside maps generated based on a pair of adjacent frames and multiple successive frames, respectively.

2.2. Trajectory descriptors

On the basis of superpixel representation of video frames, any part of object/background in different frames can be tracked by the corresponding superpixels. It is intuitive to concatenate the corresponding superpixels to construct a superpixel-level trajectory. Suppose that K trajectories $T_i (i = 1, 2, \ldots, K)$ are extracted from the whole video, where each trajectory T_i consists of a series of superpixels along the temporal direction. More specifically, each trajectory T_i can be represented by a collection of corresponding superpixels in successive frames, denoted as $T_i = \{sp^i_m, sp^i_{m+1}, \ldots, sp^i_t, \ldots, sp^i_{m+n}\}$. The indices of the first and the last frame where T_i resides in are m and $m+n$, respectively. Due that different objects may exhibit a wide range of motions, deformations and occlusions, old superpixels disappear and new

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

superpixels emerge, the extracted trajectories are usually asynchronous, i.e. they start and end in different frames, resulting in different lifespans. It is quite common that in natural videos a salient object has varying velocity and intermittent motion, and thus motions only from adjacent frames are insufficient to identify salient objects and keep the temporal coherence of saliency estimation through the entire video. To address this problem, we introduce the accumulated motion histogram and trajectory velocity entropy to describe each trajectory.

2.2.1. Accumulated motion histogram

Conceptually, the motion information of trajectory T_i can be represented by using a set of motion histograms $\{MH^i_m, MH^i_{m+1}, \ldots, MH^i_t, \ldots, MH^i_{m+n}\}$. For each motion histogram MH^i_t in the trajectory T_i as well as in the frame F_t, its temporally adjacent motion histograms are accumulated to enhance the reliability of motion information. Specifically for MH^i_t, the accumulation window is set to $[F_{t-L}, F_{t+L}]$ with a size of $2L+1$ frames (including the current frame F_t), in which the parameter L is set to 5 by experiments to suitably accommodate superpixels with fast motion, and its accumulated motion histogram is defined as follows:

$$AMH^i_t(k) = \frac{\sum_{\tau=t-L}^{t+L} MH^i_\tau(k)}{\sum_{\tau=t-L}^{t+L} |sp^i_\tau|} \tag{1}$$

$$AO^i_t(k) = \frac{\sum_{\tau=t-L}^{t+L} O^i_\tau(k)}{2L+1} \tag{2}$$

where $|sp^i_\tau|$ denoting the number of pixels within the superpixel sp^i_τ is used to normalize the motion amplitudes. $AMH^i_t(k)$ and $AO^i_t(k)$ are used to represent the normalized motion amplitude and the quantized orientation, respectively, of the kth bin in the accumulated motion histogram AMH^i_t. Similarly, the frame-level accumulated motion histogram AMH^0_t and the related quantized orientation AO^0_t are calculated by using the frame-level motion histograms in the same accumulation window.

2.2.2. Trajectory velocity entropy

Motivated by the fact that the velocity is a perceivable element in human visual system and the observation that trajectories associated with background regions are usually consistent with the camera movement, a trajectory containing background superpixels is usually smoother than a trajectory containing object superpixels, which shows more variations on velocity at different frames. The velocity of a trajectory T_i at frame F_t is defined as the scale change of corresponding superpixels in the successive frames, i.e.,

$$v^i_t = |sp^i_{t-1} \cup sp^i_t| - |sp^i_{t-1} \cap sp^i_t| \tag{3}$$

where the former and the latter term denote the number of union and intersection, respectively, of pixels between the corresponding superpixels in the successive frames F_t and F_{t-1}, in terms of pixel's location.

To describe the velocity variation over the trajectory T_i, the velocities at all frames within T_i are normalized as $p(v^i_t) = v^i_t / \sum_{t=m}^{m+n} v^i_t$, and then the trajectory velocity entropy, which represents the velocity diversity of trajectory T_i in its lifespan, is defined as follows:

$$E_v(T_i) = -\sum_{v^i_t \in T_i} p(v^i_t) \log [p(v^i_t)] \tag{4}$$

where a larger value of velocity entropy indicates for the trajectory T_i a higher probability belonging to a salient object.

2.3. Temporal saliency measurement

The proposed superpixel-level temporal saliency measurement is based on the observation that trajectories containing background superpixels are coherent with the dominant camera motion and have more consistent motion amplitude and orientation over time. In contrast, the trajectories containing salient object superpixels, which show more variations on motion amplitude and orientation, are distinctive from the trajectories containing background superpixels. Therefore, we exploit the two trajectory descriptors, i.e. accumulated motion histogram and trajectory velocity entropy as the short-term and long-term temporal feature, respectively, to evaluate the temporal saliency in the following subsections.

2.3.1. Superpixel-level trajectory distinctiveness

For each trajectory T_i at the current frame F_t, the trajectory motion contrast with respect to the global motion of F_t is measured by evaluating the difference on the accumulated motion histogram between the corresponding superpixel sp^i_t and the entire frame F_t. Specifically, such a motion contrast of T_i at F_t, denoted as $MC(sp^i_t)$, is measured based on the difference between AMH^i_t and AMH^0_t, which is defined as a sum of distances between each pair of the quantized motion orientations weighted by their accumulated motion amplitudes with the following form:

$$MC(sp^i_t) = \sum_{j=1}^{b} AMH^i_t(j) \sum_{k=1}^{b} \left[\|AO^i_t(j) - AO^0_t(k)\|_2 \cdot AMH^0_t(k) \right] \tag{5}$$

The above motion contrast measure for trajectory captures the local motion in a set of successive frames with the use of accumulated motion histograms. Besides, regarding the lifespan of trajectory, the trajectory velocity entropy is used as a global factor to weight the motion contrast measure, for a better discrimination among different trajectories. Therefore, the superpixel-level trajectory distinctiveness measure is defined as follows:

$$TD(sp^i_t) = MC(sp^i_t) \cdot E_v(T_i) \tag{6}$$

Eq. (6) combines both short-term and long-term motion cues, i.e. the motion contrast in several successive frames and the motion change in the complete lifespan of trajectory, respectively, to better discriminate object superpixels from background superpixels.

2.3.2. Trajectory-level temporal saliency

Based on trajectory distinctiveness measures of superpixels in each trajectory, the trajectory-level saliency is evaluated by taking into account the inside–outside maps, which roughly indicate whether each superpixel in the trajectory belongs to an object or not. Based on inside–outside maps, salient trajectories which are highly likely to correspond to salient objects are selected to enhance the coherence

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

of temporal saliency. For each trajectory, the number of its associated superpixels, which fall into the inside region of the corresponding inside–outside maps during its lifespan, is counted as the occurrence index. All trajectories are sorted based on the descending order of occurrence index, and the top α percentage of trajectories are labeled as salient, while the remaining trajectories are treated as non-salient. The parameter α is set to 10 by experiments, due that a majority of trajectories in most videos correspond to background, and noisy trajectories with short durations should be avoided selecting as salient.

Then for each salient trajectory, its temporal saliency is defined as the maximum trajectory distinctiveness measure of its associated superpixels, to uniformly highlight the salient trajectory over time. Similarly, for each non-salient trajectory, its temporal saliency is defined as the minimum trajectory distinctiveness measure of its associated superpixels for a uniform suppression. Specifically, the temporal saliency for each trajectory T_i is defined as follows:

$$TTS_i = \begin{cases} \max\limits_{m \le t \le m+n} TD(sp_t^i), & \text{if } T_i \text{ is salient} \\ \min\limits_{m \le t \le m+n} TD(sp_t^i), & \text{otherwise} \end{cases} \tag{7}$$

2.3.3. Superpixel-level temporal saliency

The temporal saliency for each superpixel is initially assigned with the temporal saliency of its associated trajectory to keep the temporal consistency among frames. However, as aforementioned, trajectories are usually asynchronous, and thus spatially neighboring superpixels in the same frame could show obvious difference on their initial temporal saliency measures. Therefore, to enhance the spatial consistency of neighboring superpixels in the superpixel-level temporal saliency map, the temporal saliency measures of spatially neighboring superpixels are used as a prediction to adjust the initial temporal saliency of each superpixel. Let N_t^i denote the local neighborhood which includes the superpixel sp_t^i and its spatially neighboring superpixels in F_t, the motion similarity between sp_t^i and any neighboring superpixel sp_t^j is defined based on the chi-square distance between their accumulated motion histograms AMH_t^i and AMH_t^j, i.e.,

$$MS(sp_t^i, sp_t^j) = exp\left(-\frac{1}{2} \sum_{k=1}^{b} \frac{\left[AMH_t^i(k) - AMH_t^j(k)\right]^2}{AMH_t^i(k) + AMH_t^j(k)} \right). \tag{8}$$

The temporal saliency prediction for sp_t^i is then defined as follows:

$$S_T^p(sp_t^i) = \frac{\sum_{sp_t^j \in N_t^i, j \ne i} MS(sp_t^i, sp_t^j) \cdot TTS_j}{\sum_{sp_t^j \in N_t^i, j \ne i} MS(sp_t^i, sp_t^j)}. \tag{9}$$

Finally, for each superpixel sp_t^i, its initial temporal saliency is adjusted by the above temporal saliency prediction to obtain the final temporal saliency as follows:

$$S_T(sp_t^i) = TTS_i + S_T^p(sp_t^i). \tag{10}$$

The superpixel-level temporal saliency measure using Eq. (10) keeps the temporal consistency effectively using the former term, which assigns a uniform temporal saliency to superpixels along the trajectory, and also enhances the spatial consistency between neighboring superpixels using the latter term, which propagates the temporal saliency measures of superpixels within the spatial neighborhood. Using a combination of the two terms, the discontinuities of temporal saliency measures in both temporal domain and spatial domain are effectively alleviated to better handle object superpixels with intermittent motion and also to enhance the saliency uniformity of superpixels belonging to the same object or background.

2.3.4. Pixel-level temporal saliency

Based on temporal saliency measures of superpixels, the temporal saliency measure for each pixel is derived from its neighboring superpixels with the consideration of similarity between pixel and superpixel. For each pixel p_t^k in the superpixel sp_t^i, its temporal saliency is defined as a weighted linear combination of the temporal saliency measures of superpixels in the local neighborhood N_t^i with the following form:

$$S_T(p_t^k) = \sum\nolimits_{sp_t^j \in N_t^i} \left[exp\left(-\frac{d_t^{j,k}}{2\sigma_d} \right) + exp\left(-\frac{c_t^{j,k}}{2\sigma_c} \right) \right] \cdot S_T(sp_t^j), \ \forall p_t^k \in sp_t^i \tag{11}$$

where the weight consists of two exponential terms to evaluate the spatial similarity and color similarity, respectively, between the pixel p_t^k and the superpixel sp_t^j. The Euclidean spatial distance between p_t^k and the centroid of sp_t^j is denoted as $d_t^{j,k}$, and the Euclidean color distance between p_t^k and the mean color of all pixels in sp_t^j is denoted as $c_t^{j,k}$. The two parameters σ_d and σ_c are set to 30 and 40, respectively, for a moderate weighting effect.

Two examples of the above temporal saliency measurement process is shown in Fig. 3. We can see that superpixel-level trajectory distinctiveness maps only highlight some parts of salient objects but fail to suppress background regions in some frames. After trajectory-level enhancement and temporal saliency prediction and adjustment, salient object regions are better highlighted and background regions are more effectively suppressed as shown in the trajectory-level and superpixel-level temporal saliency maps. Finally, the pixel-level temporal saliency maps achieve an even better quality with more uniform highlighting and suppression effect.

2.4. Spatiotemporal saliency generation

In the previous subsections, the generation of trajectory-level, superpixel-level and pixel-level temporal saliency map are described in details. As for the spatial saliency map of each video frame, any saliency model for images can be used for this purpose, and here a simple saliency model based on global contrast and spatial sparsity of superpixels in [51] is used to generate the pixel-level spatial saliency map $S_S(p_t)$ for each frame F_t. In this subsection, a quality-guided fusion method is proposed to integrate temporal saliency map $S_T(p_t)$ with spatial saliency map $S_S(p_t)$ for generating pixel-level spatiotemporal saliency map. To quantitatively evaluate the

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

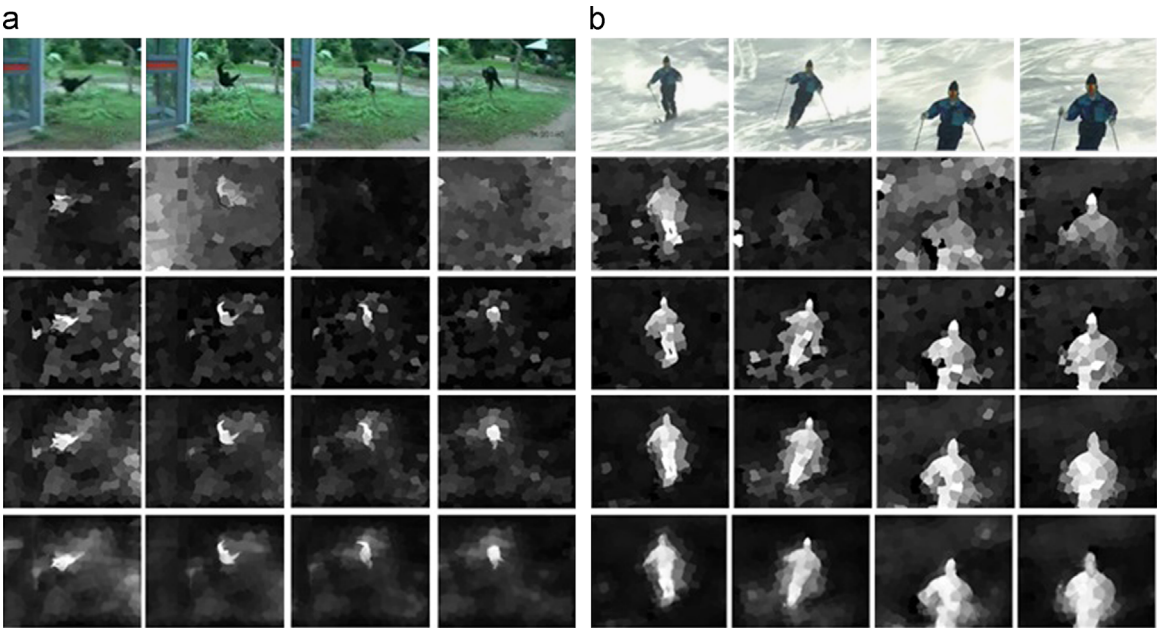

Fig. 3. Examples of temporal saliency measurement for two videos. From top to bottom: original video frames, superpixel-level trajectory distinctiveness maps, trajectory-level, superpixel-level and pixel-level temporal saliency maps, respectively.

quality of $S_T(p_t)$ and $S_S(p_t)$ of each video frame F_t, two measures, i.e. object area based confidence measure and consistency measure with respect to the inside–outside map IOM_t, are defined to evaluate the relative quality of temporal and spatial saliency maps.

Given $S_T(p_t)$ and $S_S(p_t)$ of F_t, an adaptive thresholding operation using the well-known Otsu's method [57] is performed on both saliency maps to obtain the two binary object masks denoted as $B_T(p_t)$ and $B_S(p_t)$, respectively. The number of object pixels in the two binary masks $B_T(p_t)$ and $B_S(p_t)$ is denoted as A_T and A_S, respectively (the subscript t is omitted for a concise notation due that the following description of fusion process is on the basis of each video frame F_t).

In most natural videos, background covers a majority of areas and surrounds salient objects, which generally have a compact spatial distribution. Therefore, for $B_T(p_t)$ and $B_S(p_t)$, the one with a relatively smaller object area indicates that the corresponding saliency map is generally more reliable for saliency detection than the other saliency map in a comparative sense. Based on the above consideration, the object area based confidence measures for $S_T(p_t)$ and $S_S(p_t)$ are defined as follows:

$$\lambda_T = \frac{A_S}{A_T + A_S}, \quad \lambda_S = \frac{A_T}{A_T + A_S} \tag{12}$$

where $\lambda_T + \lambda_S = 1$. Eq. (12) indicates that a higher confidence measure will be assigned to the temporal/spatial saliency map with a smaller area of salient objects in the corresponding binary mask.

As the inside–outside map IOM_t can roughly indicate the possible location of salient objects, the consistency between $B_T(p_t)/B_S(p_t)$ and IOM_t is exploited to measure the reliability of $S_T(p_t)/S_S(p_t)$ to predict salient objects, and a higher consistency value indicates a higher quality of saliency map to some degree. The inside–outside map IOM_t is dilated to fill missing pixels inside the estimated object regions, and then the consistency between $B_T(p_t)/B_S(p_t)$ and IOM_t is determined by their overlapping area of object regions as follows:

$$\gamma_T = \frac{|B_T(p_t) \otimes IOM_t|}{|B_T(p_t)|}, \quad \gamma_S = \frac{|B_S(p_t) \otimes IOM_t|}{|B_S(p_t)|} \tag{13}$$

where $\otimes$ denotes the pixel-wise multiplication operation, and $|\cdot|$ denotes the number of object pixels in the binary mask, and $\gamma_T + \gamma_S = 1$.

By multiplying the above defined two measures, the quality index for $S_T(p_t)$ and $S_S(p_t)$ is measured as $\lambda_T \cdot \gamma_T$ and $\lambda_S \cdot \gamma_S$, respectively. Then the quality-guided fusion is exploited to integrate $S_T(p_t)$ with $S_S(p_t)$ for generating the pixel-level spatiotemporal saliency map $S_{ST}(p_t)$ as follows:

$$S_{ST}(p_t) = \lambda_T \cdot \gamma_T \cdot S_T(p_t) + \lambda_S \cdot \gamma_S \cdot S_S(p_t) \tag{14}$$

where the quality index $\lambda_T \cdot \gamma_T$ and $\lambda_S \cdot \gamma_S$ control the relative contributions of temporal and spatial saliency maps to the spatiotemporal saliency map.

Some examples of integrating temporal and spatial saliency maps for generation of spatiotemporal saliency maps are shown in Fig. 4, in which all saliency maps are normalized into the same range of [0, 255]. In temporal and spatial saliency maps, some background regions such as the sunlight regions in the 3rd row of Fig. 4(a) and the bush regions in the 3rd row of Fig. 4(c) are falsely highlighted, and/or salient object regions such as the missed parachute region in the 3rd row of Fig. 4(a) and the antelope region in the 2nd and 3rd row of Fig. 4(b) are not completely highlighted in some frames. In contrast, the spatiotemporal saliency maps generated using the proposed quality-guided fusion scheme can better highlight the complete salient object and effectively suppress background regions. Fig. 4(d) shows an example with a relatively larger salient object compared to Fig. 4(a)–(c), we can observe from the spatiotemporal saliency maps that

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

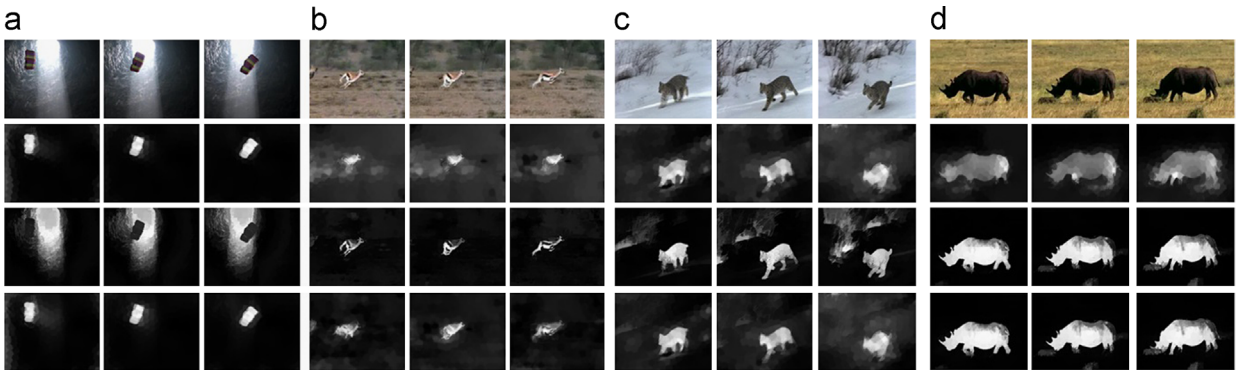

Fig. 4. Examples of spatiotemporal saliency map generation for four videos. From top to bottom: original video frames, pixel-level temporal, spatial and spatiotemporal saliency maps, respectively.

the proposed fusion scheme is effective for salient objects with different scales.

3. Experimental results

In this section, we perform a comprehensive performance evaluation of the proposed superpixel-level trajectory (SLT) based model and comparison with state-of-the-art spatiotemporal saliency models. The test datasets and experimental settings are first described in Section 3.1. Both subjective evaluation and objective evaluation on the test datasets are presented and analyzed in Sections 3.2 and 3.3, respectively. The results on some challenging videos are shown and discussed in Section 3.4, and the computation issue is addressed in Section 3.5.

3.1. Datasets and experimental settings

We performed experiments on two publicly available datasets with different characteristics, i.e. VideoSeg [58] and SegTrack [59]. VideoSeg contains 10 uncompressed video clips of natural scenes with 12 fps and varied video length from 5 to 10 s. SegTrack contains 6 videos with diverse motion patterns of objects and background with video length varying from 21 to 71 frames. Both datasets provide the manually labeled binary ground truths of salient objects. Following [60], the video 'penguin' is also discarded from the SegTrack dataset, since the ground truths provided for this video only contain the center penguin, which is just a part of moving objects.

We compared our SLT model with six state-of-the-art spatiotemporal saliency models including SR [37], CE [23], QFT [10], MB [12], SP [51] and DCMR [35]. 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. For all the compared saliency models, we used the source codes with default parameter settings or executables provided by the authors.

3.2. Subjective evaluation

Spatiotemporal saliency maps generated using our SLT model and the other six spatiotemporal models for several videos from the two datasets are shown in Figs. 5 and 6 for a visual comparison. SR, CE and QFT only highlight regions around salient object boundaries or a part of salient object regions, and fail to highlight the complete salient object regions with well-defined boundaries. MB employs background modeling for detecting salient regions, but cannot completely highlight salient objects with slow and intermittent motions such as the two examples in Fig. 5, and also falsely highlights background regions in videos with global motion such as the two examples in Fig. 6. SP can better highlight salient objects and suppress background regions than the former four models, but cannot keep the temporal coherence of saliency maps as well as SLT, and may falsely highlight high-contrast background regions such as the sunlight region in the bottom example of Fig. 6. 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 thus DCMR fails to highlight the complete salient objects in the two examples of Fig. 6. In contrast, DCMR performs slightly better on the top example of Fig. 5, in which the salient object (sunflower) is with a slow motion on the static background. Compared with the other six models, our SLT model can generate temporally coherent saliency maps, which better highlight the complete salient objects with well-defined boundaries and suppress background regions more effectively for videos with various motions.

3.3. Objective evaluation

In order to objectively evaluate the saliency detection performance of different models, we adopted the commonly used precision-recall (PR) curve, which plots the precision measure against the recall measure to characterize saliency detection performance. Specifically, the thresholding operations using a series of fixed integers in the range of [0,255] are first performed on each saliency map to obtain 256 binary salient object masks, and a set of precision and recall values are calculated by comparing with the corresponding binary ground truth. Then for each model, at each threshold, the precision and recall values of all saliency maps are averaged, and as shown in Fig. 7, the PR curve for each model plots the 256 average precision values against the 256 average recall values. Fig. 7(a) shows that our SLT model and SP significantly outperform the other five models on the VideoSeg dataset,

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

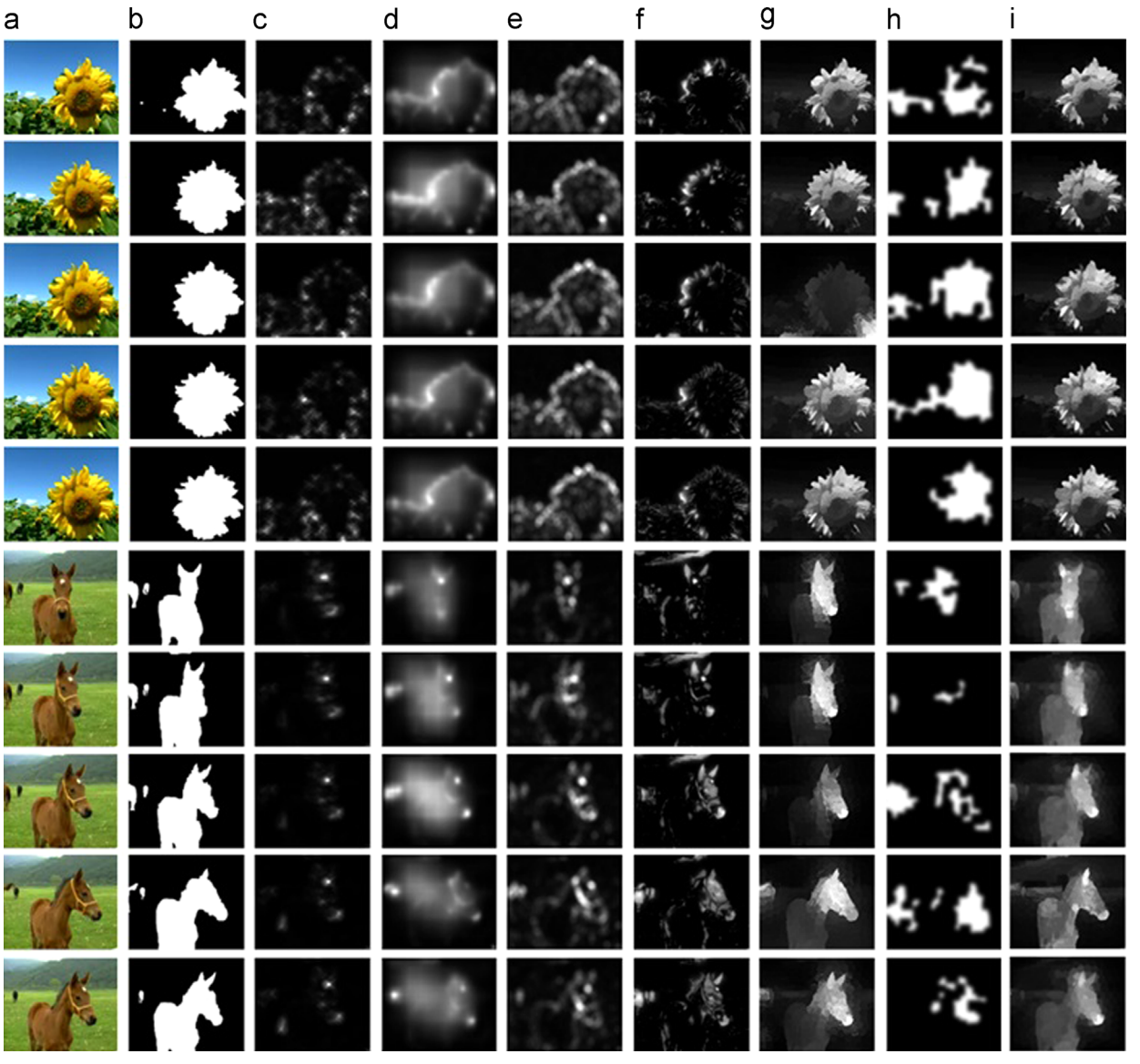

Fig. 5. Examples of spatiotemporal saliency detection on two videos from the VideoSeg dataset. (a) Videos frames (shown with an interval of 20 and 10 frames, respectively, for the top and bottom example); (b) ground truths; spatiotemporal saliency maps generated using (c) SR, (d) CE, (e) QFT, (f) MB, (g) SP, (h) DCMR and (i) SLT, respectively.

while SLT still achieves a better saliency detection performance than SP with a slightly higher PR curve. For the SegTrack dataset, which contains more challenging videos with complicated motions, Fig. 7(b) clearly shows that our SLT model significantly outperforms all the other six models. In conclusion, our SLT model shows the consistently better performance over all the other six models on the two datasets.

Besides, in order to evaluate the contributions of each component in our SLT model, we also quantitatively plot in Fig. 8 the PR curves of the intermediate maps generated by our SLT model. As shown in Fig. 8, these maps include trajectory motion contrast map (MC) using Eq. (5), superpixel-level trajectory distinctiveness map (TD) using Eq. (6), trajectory-level temporal saliency map (TL-TS) using Eq. (7), superpixel-level temporal saliency map (SPL-TS) using Eq. (9), pixel-level temporal saliency map (PL-TS) using Eq. (11), pixel-level spatial saliency map (PL-SS), and pixel-level spatiotemporal saliency map (PL-STS) using Eq. (14). We can observe from Fig. 8 the progressive improvement trend on MC, TD, TL-TS, SPL-TS and PL-TS in turn, and it is obvious that the PR curves of PL-TS are higher than those PR curves of the preceding four maps. This demonstrates the effectiveness of the proposed pipeline for measuring temporal saliency (from Sections 2.1–2.3). Furthermore, as shown in Fig. 8, the PR curves of PL-STS are higher than the PR curves of both PL-TS and PL-SS on the two datasets, and this demonstrates the

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

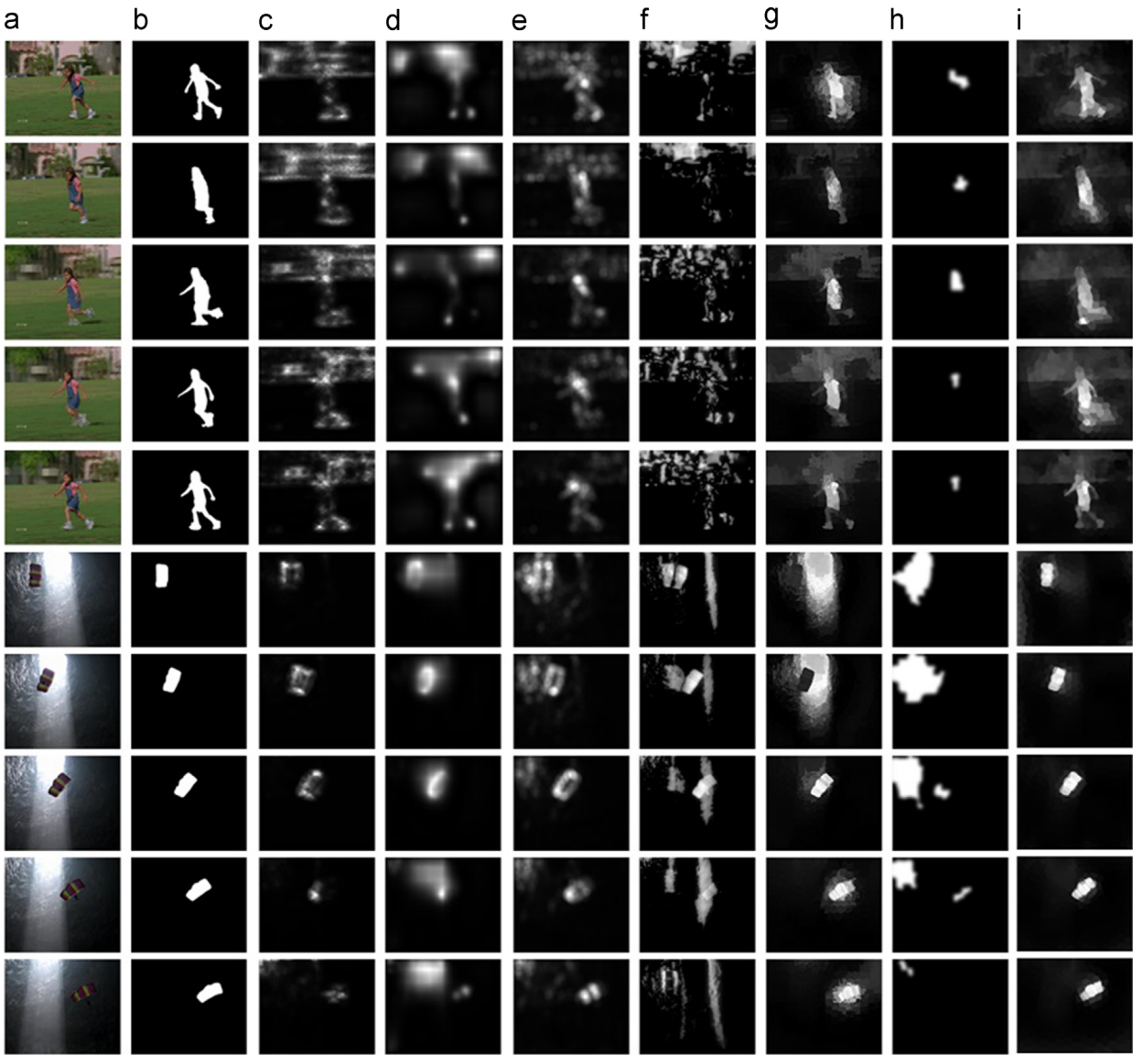

Fig. 6. Examples of spatiotemporal saliency detection on two videos from the SegTrack dataset. (a) Videos frames (shown with an interval of 4 and 10 frames, respectively, for the top and bottom example); (b) ground truths; spatiotemporal saliency maps generated using (c) SR, (d) CE, (e) QFT, (f) MB, (g) SP, (h) DCMR and (i) SLT, respectively.

effectiveness of the proposed quality-guided fusion method in Section 2.4. In conclusion, each component in our SLT model has its contribution to the finally better saliency detection performance achieved by the whole model.

3.4. Analysis

Both subjective evaluation and objective evaluation on two public datasets in the previous two subsections demonstrate that our SLT model outperforms the other state-of-the-art spatiotemporal saliency models. To better understand the characteristics and limitations of our SLT model and other models, we tested all the models on some challenging videos downloaded from the internet, and the corresponding spatiotemporal saliency maps of some video frames are shown in Figs. 9 and 10 for the following analysis.

Fig. 9 shows the results on the video *MotorRolling* with motion blur and the video *SkateBoarding* which exhibits complicated motion with 3D parallax. These two videos contain large and complicated motion of salient objects and/or background. We can observe from Fig. 9 that SR, CE, QFT and MB falsely highlight some background regions and cannot effectively highlight salient objects in these two videos. SR exploits the space-time local kernel to encode the geometric structures which could represent motion orientation in video, but it is not suitable for the complicated motion with large magnitude and may falsely highlight some boundaries of geometric structures in the background. CE measures saliency as the minimum uncertainty of a local region given

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

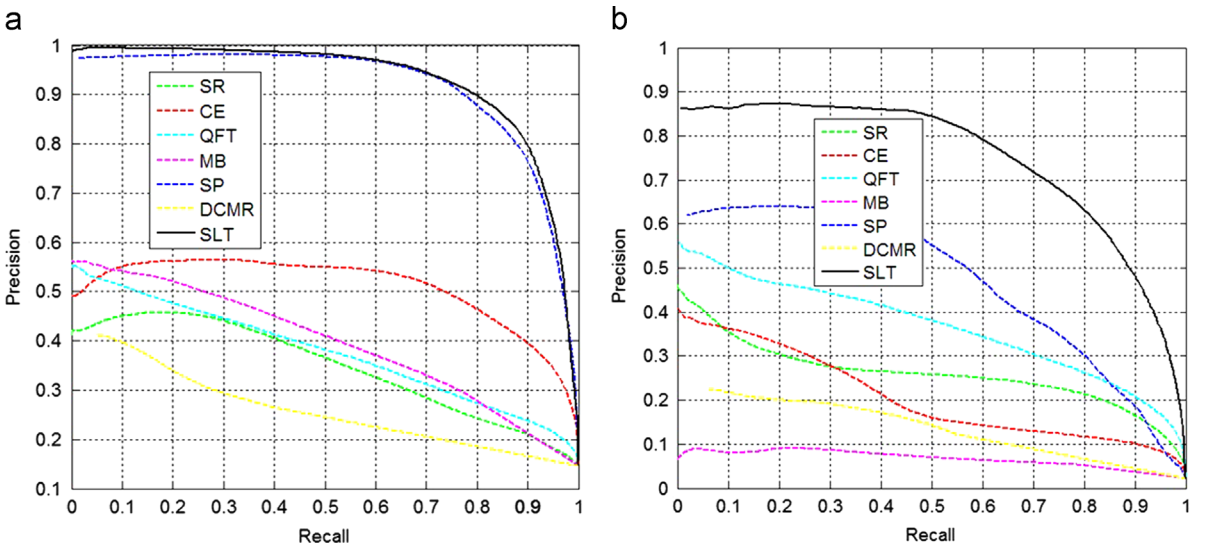

Fig. 7. (better viewed in color) Precision-recall curves of different spatiotemporal saliency models on (a) the VideoSeg dataset and (b) the SegTrack dataset, respectively.

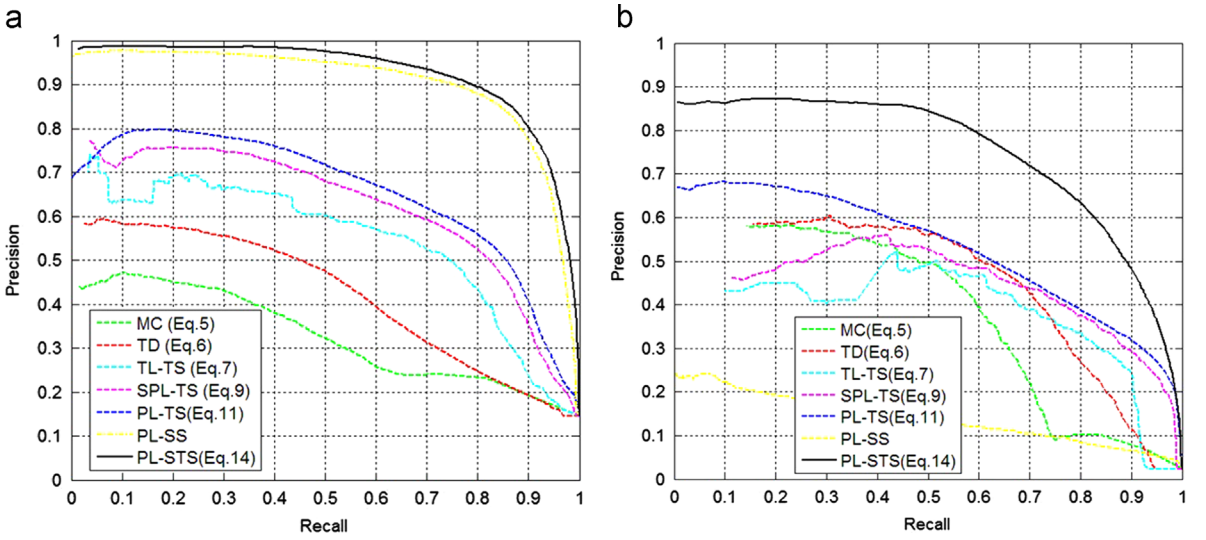

Fig. 8. (better viewed in color) Precision-recall curves of different maps obtained using our SLT model on (a) the VideoSeg dataset and (b) the SegTrack dataset, respectively.

the spatiotemporal surrounding areas in consecutive frames, without explicitly taking motion information into account. QFT exploits the frame difference as the motion feature for saliency measurement, and thus it cannot appropriately handle videos with moving background. MB is based on the background subtraction scheme, which also significantly degrades the saliency detection performance on videos with moving background. DCMR utilizes motion information implicitly by matching and tracking the key points through the video, but the performance is degraded due to the complicated background motion. Both SP and our SLT model measure saliency at superpixel level using motion vector fields, and thus highlight salient object regions better than previous models. Compared to all the other six models, our SLT model suppresses background regions and highlights the complete salient objects more effectively, because SLT exploits trajectories of superpixels to propagate motion information from faraway video frames, and this allows SLT to better handle large and complicated motion and to generate temporally coherent saliency maps.

However, it is still difficult for our SLT model to effectively handle some other challenging videos such as the two examples in Fig. 10. The top example is a video captured by a handheld device with severe camera shaking, which results in an unreliable optical flow estimation for generating motion vector field. Therefore, it is difficult for our SLT model which relies on the motion vector field to generate reliable trajectories for temporal saliency measurement. In the bottom example of Fig. 10, the bird shows fast intermittent motion with the shaking leaves, and the bird is partly occluded by the leaves. Such a case results in a cluttered motion vector field, in which the

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

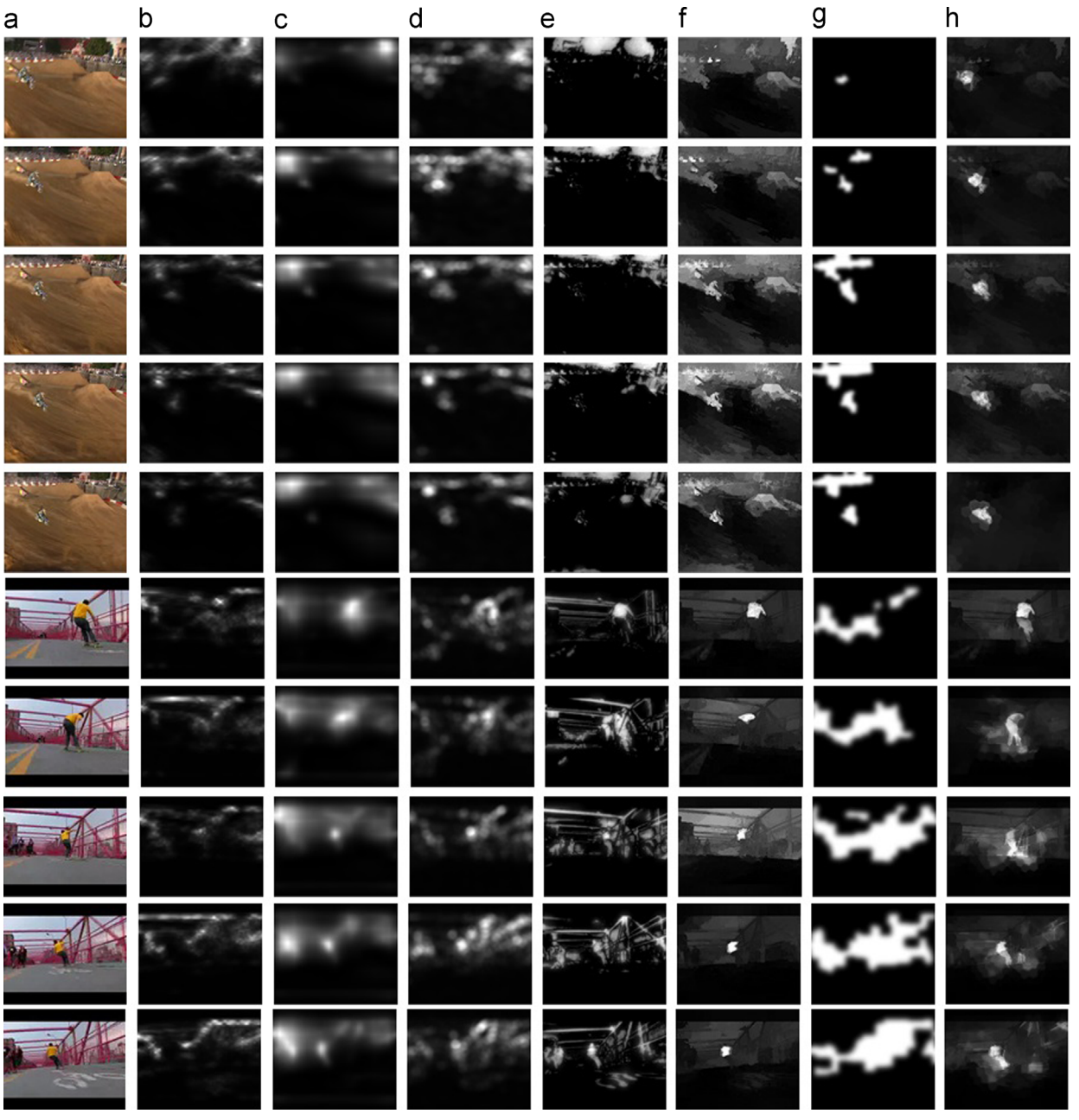

Fig. 9. Spatiotemporal saliency detection on two challenging videos: *MotorRolling* (top) and *SkateBoarding* (bottom). From left to right: (a) video frames (shown with an interval of 5 and 10 frames, respectively, for the top and bottom example); spatiotemporal saliency maps generated using (b) SR, (c) CE, (d) QFT, (e) MB, (f) SP, (g) DCMR and (h) SLT, respectively.

object motion is difficult to identify, and further results in a lot of noisy trajectories with short durations, which decreases the reliability of trajectory-level temporal saliency measurement in our SLT model. For such challenging videos shown in Fig. 10, all the other six saliency models also cannot handle severe camera shaking and cluttered motion of both object and background. Nonetheless, due to the effective use of trajectory in temporal saliency measurement, our SLT model generates more coherent spatiotemporal saliency maps compared to other saliency models.

3.5. Running time

We compared the computational cost with all the other six saliency models, and Table 1 reports for each model the average running time per frame for videos with a resolution of 352×288. Our SLT model was implemented using MATLAB, and all experiments were performed on a PC with Intel Core i7 2600 3.4 GHz CPU and 4 GB RAM. The average running time per frame taken by our SLT model is 11.814 s, and Fig. 11 further shows the average processing time taken by each component of our SLT model. It is obvious that SLT

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

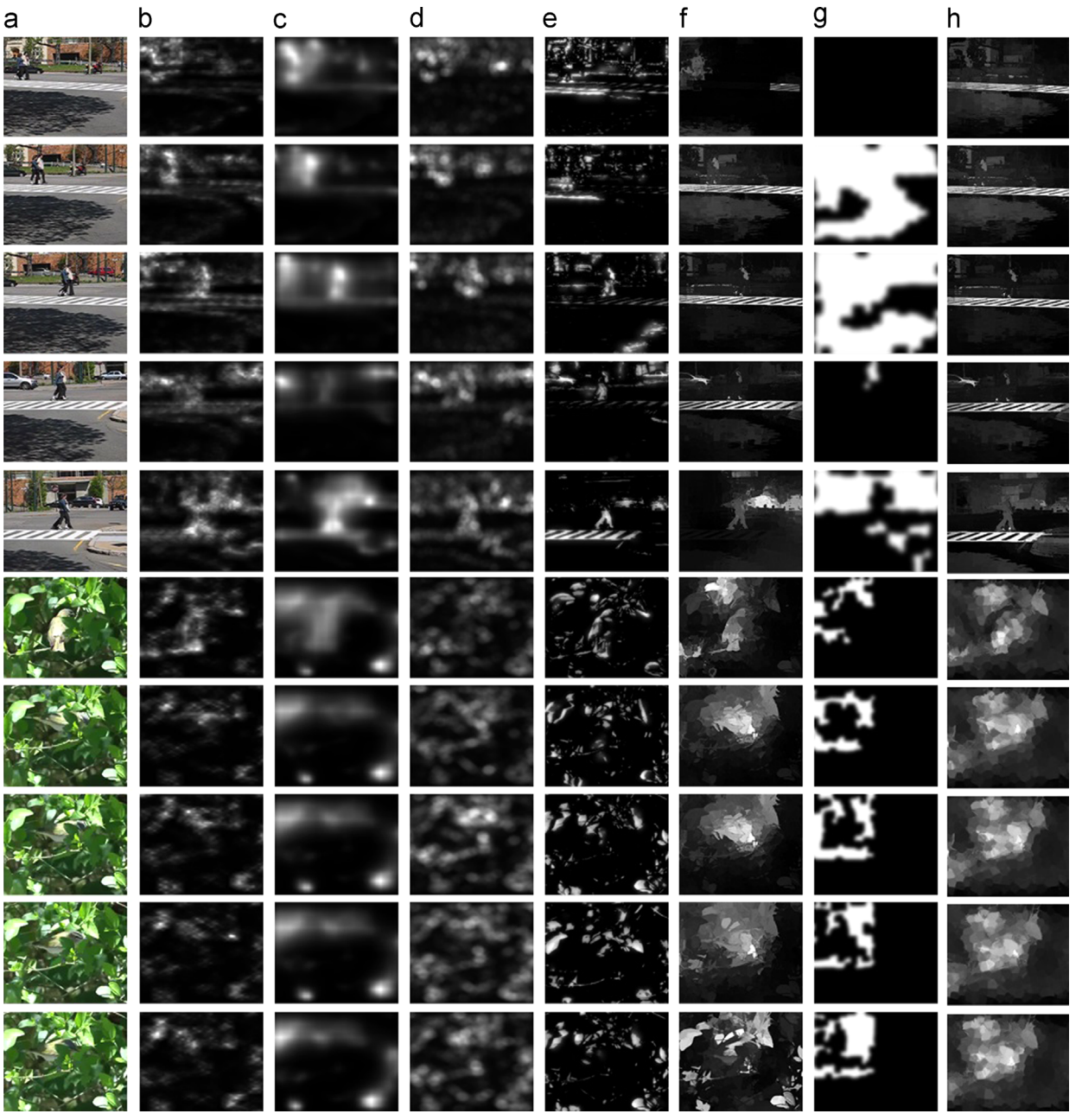

Fig. 10. Spatiotemporal saliency detection on two challenging videos: *Couple* (top) and *Red-eyed Vireo* (bottom). From left to right: (a) video frames (shown with an interval of 20 and 10 frames, respectively, for the top and bottom example); spatiotemporal saliency maps generated using (b) SR, (c) CE, (d) QFT, (e) MB, (f) SP, (g) DCMR and (h) SLT, respectively.

and SP have the higher running time than the other five models, since both models use the time-consuming optical flow estimation method to obtain motion vector field. In our SLT model, the optical flow estimation process occupies 86.1% of the total processing time. In order to improve the computational efficiency of our SLT model, the optical flow estimation can be bypassed by using the block-level motion vectors directly extracted from the video bitstream, with a compromise of saliency detection performance. Furthermore, the computation efficiency of our SLT model can be accelerated with an optimized C/C++ implementation and a parallel GPU implementation.

4. Conclusion

This paper presents an effective spatiotemporal saliency model based on trajectories of temporally consistent superpixels, aiming to improve saliency detection performance especially for videos with complicated motions. Two trajectory descriptors, i.e., accumulated motion histogram and

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014

Table 1
Comparison of average processing time per frame taken by different spatiotemporal saliency models.

Model	SR	CE	QFT	MB	SP	DCMR	SLT
Time (s)	0.427	4.377	0.117	0.174	11.850	0.045	11.814
Code	Matlab	Matlab	Matlab	C++	Matlab	C++	Matlab

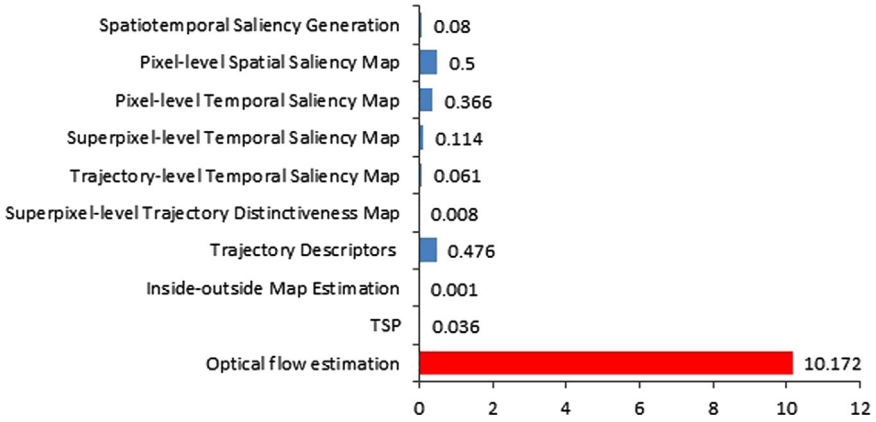

Fig. 11. Average processing time per frame taken by each component of our SLT model.

trajectory velocity entropy, are introduced to characterize the short-term and long-term temporal features of superpixel-level trajectories. Besides, a novel pipeline starting from the estimation of superpixel-level trajectory distinctiveness to the measurement of trajectory-level, superpixel-level and pixel-level temporal saliency is exploited to obtain temporally coherent saliency maps with high quality. Finally, a quality-guided fusion method reasonably integrates temporal saliency maps with spatial saliency maps to generate pixel-level spatiotemporal saliency maps, which achieve consistently better saliency detection performance than the state-of-the-art spatiotemporal saliency models. In our future work, we will investigate an effective extension of saliency aggregation methods [61,62], which have been reasonably exploited to fuse saliency maps generated by different saliency models, for spatiotemporal saliency generation, and will explore the use of spatiotemporal saliency model in applications such as salient object detection and segmentation from videos, video retargeting and video stabilization.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61471230, 61171144 and 61105001, a Marie Curie International Incoming Fellowship within the 7th European Community Framework Programme under Grant No. 911202, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

[1] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 353–367.

[2] A. Borji, L. Itti, Exploiting local and global patch rarities for saliency detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 478–485.

[3] R. Shi, Z. Liu, H. Du, X. Zhang, L. Shen, Region diversity maximization for salient object detection, IEEE Signal Process. Lett. 19 (4) (2012) 215–218.

[4] Z. Liu, W. Zou, O. Le Meur, Saliency tree: A novel saliency detection framework, IEEE Trans. Image Process. 23 (5) (2014) 1937–1952.

[5] Z. Liu, R. Shi, L. Shen, Y. Xue, K.N. Ngan, Z. Zhang, Unsupervised salient object segmentation based on kernel density estimation and two phase graph cut, IEEE Trans. Multimedia 14 (4) (2012) 1275–1289.

[6] 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. 37 (3) (2015) 569–582.

[7] A. Shamir, S. Avidan, Seam carving for media retargeting, Commun. ACM 52 (1) (2009) 77–85.

[8] H. Du, Z. Liu, J. Jiang, L. Shen, Stretchability-aware block scaling for image retargeting, J. Visual Commun. Image Represent. 24 (4) (2013) 499–508.

[9] J. Sun, H. Ling, Scale and object aware image retargeting for thumbnail browsing, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1511–1518.

[10] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (1) (2010) 185–198.

[11] L. Shen, Z. Liu, Z. Zhang, A novel H.264 rate control algorithm with consideration of visual attention, Multimedia Tools Appl. 63 (3) (2013) 709–727.

[12] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokric, V. Crnojević, D. Kukolj, Salient motion features for video quality assessment, IEEE Trans. Image Process. 20 (4) (2011) 948–958.

[13] O. Le Meur, Z. Liu, Saccadic Model of Eye Movements for Free-viewing Condition, Vision Research (2015) doi.: 10.1016/j.visres.2014.12.026.

[14] T. Chen, M.M. Cheng, P. Tan, A. Shami, S. Hu, Sketch2Photo: internet image montage, ACM Trans. Graphics 28 (5) (2009) 124.

[15] M. Mancas, O. Le Meur, Memorability of natural scenes: The role of attention, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2013, pp. 196–200.

[16] J. Han, C. Chen, L. Shao, X. Hu, J. Han, T. Liu, Learning computational models of video memorability from fMRI brain imaging, IEEE Trans. Cybern. (2014), http://dx.doi.org/10.1109/TCYB.2014.2358647.

[17] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and

high-level feature learning, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3325–3337.
[18] J. Han, P. Zhou, D. Zhang, G. Cheng, L. Guo, Z. Liu, S. Bu, J. Wu, Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding, ISPRS J. Photogramm. Remote Sens. 89 (2014) 37–48.
[19] A. Borji, D.N. Sihite, L. Itti, Salient object detection: A benchmark, in: Proceedings of European Conference on Computer Vision (ECCV), 2012, pp. 414–429.
[20] R.A. Abrams, S.E. Christ, Motion onset captures attention, Psychol. Sci. 14 (5) (2003) 427–432.
[21] D. Rudoy, D.B. Goldman, E. Shechtman, L. Zelnik-Manor, Learning video saliency from human gaze using candidate selection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1147–1154.
[22] C. Liu, P. Yuen, G. Qiu, Object motion detection using information theoretic spatio-temporal saliency, Pattern Recognit. 42 (11) (2009) 2897–2906.
[23] Y. Li, Y. Zhou, J. Yan, Z. Niu, J. Yang, Visual saliency based on conditional entropy, in: Proceedings of Asian Conference on Computer Vision (ACCV), 2009, pp. 246–257.
[24] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, in: Advances in Neural Information Processing Systems (NIPS), 2009, pp. 681–688.
[25] Y. Li, Y. Zhou, L. Xu, X. Yang, J. Yang, Incremental sparse saliency detection, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2009, pp. 3093–3096.
[26] V. Gopalakrishnan, D. Rajan, Y. Hu, A linear dynamical system framework for salient motion detection, IEEE Trans. Circuits Syst. Video Technol. 22 (5) (2012) 683–692.
[27] K. Muthuswamy, D. Rajan, Salient motion detection through state controllability, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 1465–1468.
[28] X. Hou, L. Zhang, Saliency detection: A spectral residual approach, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.
[29] X. Cui, Q. Liu, D.N. Metaxas, Temporal spectral residual: Fast motion saliency detection, in: Proceedings of the ACM International Conference on Multimedia, 2009, pp. 617–620.
[30] J. Han, S. He, X. Qian, D. Wang, L. Guo, T. Liu, An object-oriented visual saliency detection framework based on sparse coding representations, IEEE Trans. Circuits Syst. Video Technol. 23 (12) (2013) 2009–2021.
[31] Z. Ren, S. Gao, L. Chia, D. Rajan, Regularized feature reconstruction for spatiotemporal saliency detection, IEEE Trans. Image Process. 22 (8) (2013) 3120–3132.
[32] Y. Xue, X. Guo, X. Cao, Motion saliency detection using low rank and sparse decomposition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 1485–1488.
[33] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, F. Wu, Background prior based salient object detection via deep reconstruction residual, IEEE Trans. Circuits Syst. Video Technol. (2014), http://dx.doi.org/10.1109/TCSVT.2014.2381471.
[34] P. Siva, C. Russell, T. Xiang, L. Agapito, Looking beyond the image: Unsupervised learning for object saliency and detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3238–3245.
[35] C. Huang, Y. Chang, Z. Yang, Y. Lin, Video saliency map detection by dominant camera motion removal, IEEE Trans. Circuits Syst. Video Technol. 24 (8) (2014) 1336–1349.
[36] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[37] H. Seo, P. Milanfar, Static and space time visual saliency detection by self-resemblance, J. Vision 9 (12) (2009) 15.
[38] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.
[39] K. Muthuswamy, D. Rajan, Salient motion detection in compressed domain, IEEE Signal Process. Lett. 20 (10) (2013) 996–999.
[40] Y. Fang, W. Lin, Z. Chen, C. Tsai, C. Lin, A video saliency detection model in compressed domain, IEEE Trans. Circuits Syst. Video Technol. 24 (1) (2014) 27–38.
[41] D. Mahapatra, S.O. Gilani, M.K. Saini, Coherency based spatio-temporal saliency detection for video object segmentation, IEEE J. Sel. Top. Sign. Process. 8 (3) (2014) 454–462.
[42] W. Kim, C. Kim, Spatiotemporal saliency detection using textural contrast and its applications, IEEE Trans. Circuits Syst. Video Technol. 24 (4) (2014) 646–659.
[43] Z. Liu, Y. Xue, L. Shen, Z. Zhang, Nonparametric saliency detection using kernel density estimation, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2010, pp. 253–256.
[44] Z. Ren, Y. Hu, L.-T. Chia, D. Rajan, Improved saliency detection based on superpixel clustering and saliency propagation, in: Proceedings of the ACM International Conference on Multimedia, 2010, pp. 1099–1102.
[45] Z. Liu, Y. Xue, H. Yan, Z. Zhang, Efficient saliency detection based on Gaussian models, IET Image Proc. 5 (2) (2011) 122–131.
[46] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, S.M. Hu, Global contrast based salient region detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 409–416.
[47] F. Perazzi1, P. Krähenbül, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 733–740.
[48] Z. Liu, O. Le Meur, S. Luo, Superpixel-based saliency detection, in: Proceedings of IEEE International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS), 2013, pp. 1–4.
[49] Z. Liu, O. Le Meur, S. Luo, L. Shen, Saliency detection using regional histograms, Opt. Lett. 38 (5) (2013) 700–702.
[50] H. Li, L. Xu, G. Liu, Two-layer average-to-peak ratio based saliency detection, Signal Process. Image Commun. 28 (1) (2013) 55–68.
[51] Z. Liu, X. Zhang, S. Luo, O. Le Meur, Superpixel-based spatiotemporal saliency detection, IEEE Trans. Circuits Syst. Video Technol. 24 (9) (2014) 1522–1540.
[52] H. Wang, A. Klaser, C. Schmid, C. Liu, Action recognition by dense trajectories, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3169–3176.
[53] K. Fragkiadaki, G. Zhang, J. Shi, Video segmentation by tracing discontinuities in a trajectory embedding, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1846–1853.
[54] T. Brox, J. Malik, Large displacement optical flow: Descriptor matching in variational motion estimation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 500–513.
[55] J. Chang, D. Wei, J.W. Fisher III, A video representation using temporal superpixels, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2051–2058.
[56] A. Papazoglou, V. Ferrari, Fast object segmentation in unconstrained video, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1777–1784.
[57] N. Otsu, A threshold selection method from gray level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[58] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, J. Yamato, Saliency based video segmentation with graph cuts and sequentially updated priors, in: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2009, pp. 638–641.
[59] D. Tsai, M. Flagg, J.M. Rehg, Motion coherent tracking with multi-label MRF optimization, in: Proceedings of British Machine Vision Conference (BMVC), 2010, pp. 1–11.
[60] Y. Lee, J. Kim, K. Grauman, Key segments for video object segmentation, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1995–2002.
[61] L. Mai, Y. Niu, F. Liu, Saliency aggregation: A data-driven approach, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1131–1138.
[62] O. Le Meur, Z. Liu, Saliency aggregation: Does unity make strength? in: Proceedings of Asian Conference on Computer Vision (ACCV), 2014, pp. 1–15.

Please cite this article as: J. Li, et al., Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Processing-Image Communication (2015), http://dx.doi.org/10.1016/j.image.2015.04.014