Journal of Visual Communication and Image Representation 51 (2018) 131–143

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Video saliency detection via bagging-based prediction and spatiotemporal propagation☆

Xiaofei Zhou[a,b], Zhi Liu[a,b,*], Kai Li[a,b], Guangling Sun[b]

[a] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[b] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Keywords:
Spatiotemporal saliency
Unconstrained video
Bagging
Prediction
Propagation

ABSTRACT

The task of spatiotemporal saliency detection is to distinguish the salient objects from background across all the frames in the video. Although many spatiotemporal models have been designed from various aspects, it is still a very challenging task for handing the unconstrained videos with complicated motions and complex scenes. Therefore, in this paper we propose a novel spatiotemporal saliency model to estimate salient objects in unconstrained videos. Specifically, a bagging-based saliency prediction model, *i.e.* an ensembling regressor, which is the combination of random forest regressors learned from undersampled training sets, is first used to perform saliency prediction for each current frame. Then, both forward and backward propagation within a local temporal window are deployed on each current frame to make a complement to the predicted saliency map and yield the temporal saliency map, in which the backward propagation is constructed based on the temporary saliency estimation of the following frames. Finally, by building the appearance and motion based graphs in a parallel way, spatial propagation is employed over the temporal saliency map to generate the final spatiotemporal saliency map. Through experiments on two challenging datasets, the proposed model consistently outperforms the state-of-the-art models for popping out salient objects in unconstrained videos.

1. Introduction

The human visual system (HVS) can effortlessly capture visually salient objects in the complicated static and dynamic scenes using the inherent visual attention mechanism. The traditional saliency model was based on biologically plausible architecture [1] and feature integration theory [2], and was exploited to predict human fixations [3,4]. Afterwards, saliency models have been extended to estimate salient objects and a number of models based on different theories have been proposed in the past decades. Meanwhile, saliency models have benefited a wide range of applications such as salient object detection and segmentation [5–10], content-aware image/video retargeting [11–13], content-based image/video compression [14–16], image/video quality assessment [17,18,77], and visual scanpath prediction [19,20].

In recent years, saliency model for images is a booming topic, and a lot of efforts have been made and some recent benchmarks have been reported [21,22]. Meanwhile, compared to images, the temporal information in videos is a crucial cue for spatiotemporal saliency model. Recently, the research on spatiotemporal saliency models for videos also received increasing attention. Certainly, many prior efforts have been made from various aspects such as the center-surround scheme [23–29], information theory [30–32], control theory [33,34], frequency domain analysis [35,36], machine learning [5,37–40], sparse representation [41–44], information fusion [45–56], and regional saliency computation [58–62]. Although the aforementioned progress can achieve the decent effect to a certain degree, their performances will degrade when handling a variety of unconstrained videos with complicated motion and complex scenes such as nonlinear deformation, fast motion, dynamic background, and occlusion. Specifically, the spatiotemporal saliency maps generated using these models are insufficient to highlight salient objects uniformly and suppress background effectively.

With the aim to improve the performance of saliency detection in unconstrained videos, this paper proposes a novel spatiotemporal saliency model to effectively separate salient objects from background. Our model proceeds on a per-frame basis, operates within a local temporal window centered on each current frame, and is guided by the outputs of previous frames towards the salient objects in the following frames. Besides, in contrast to previous models, the salient object mask

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China.
E-mail addresses: liuzhisjtu@163.com, liuzhi@staff.shu.edu.cn (Z. Liu).

https://doi.org/10.1016/j.jvcir.2018.01.014
Received 7 September 2017; Received in revised form 24 December 2017; Accepted 20 January 2018

of the first frame is given as a prior in our model, which serves as a saliency prior and indicates the potential salient regions. The advantages of our model lie in the following three aspects:

Firstly, inspired by visual tracking [63,64], we extend online learning to saliency detection and learn prediction models using the random forest regressor based on previous frames. However, the class imbalance problem often occurs on the training data. It has a negative effect on the performance of prediction model, since the model tends to be overwhelmed by the majority class and ignores the minority class. Some efforts try to solve the class imbalance problem based on an ensemble based method [72] or a cost-sensitive learning strategy [73]. To mitigate the aforementioned issues, we propose a bagging-based prediction model, which ensembles the regressors built on multiple undersampled training sets, to perform saliency prediction. The predicted saliency map can indicate most part of salient objects on each current frame.

Secondly, motivated by previous works [58–62], a novel bidirectional temporal propagation method, operated within a local temporal window, is constructed to enhance temporal consistency and complement to saliency prediction. Specifically, the forward propagation, which is constructed on the obtained final spatiotemporal saliency maps of previous frames, is first performed for each current frame, and then the backward propagation is executed in a particular way, which is constructed on the temporary saliency estimation of the following frames. The generated temporal saliency map discovers more salient object regions compared to the predicted saliency map.

Thirdly, reviewing previous works such as graph based image saliency models [65] and two-phase spatial propagation in [60], it can be seen that the graph for saliency detection is usually constructed using appearance information (color, texture, etc.) only. However, the motion information between frames is also an important cue for video saliency detection. Driven by this point, we propose a novel spatial propagation method, which brings the two complementary sources of information together in a unified manner. Concretely, two graphs are first constructed using color and motion features. The propagation is performed over the temporal saliency map in a parallel way on the two graphs. The output generally shows substantially stronger results with the better highlighted salient objects and suppressed background regions.

Overall, the main contributions of this paper are threefold:

(1) Driven by the online learning and the class imbalance problem, we propose a bagging-based saliency prediction model, *i.e.* an ensembling regressor, which is the integration of random forest regressors learned from multiple undersampled training sets.
(2) To enhance temporal consistency and complement to saliency prediction, we propose a novel bidirectional temporal propagation method. In particular, the backward propagation to each current frame is constructed based on the temporary saliency estimation of the following frames.
(3) To fully exploit the complementary effect between appearance and motion information in a unified manner, we propose a novel spatial propagation method, which is performed via two graphs based on appearance and motion, respectively, in a parallel way.

The rest of this paper is organized as follows. The related work is reviewed in Section 2, and the proposed model is described in Section 3. Experimental results and analyses are presented in Section 4, and conclusions are given in Section 5.

2. Related work

The research on saliency detection for still images has continued for decades and amounts of effective models have been proposed as mentioned in [21,22]. For example, the incorporation of low-level and high-level prior learning is employed by [78] to compute the visual saliency. In [79], the manifold ranking-based matrix factorization model is proposed to incorporate the features extracted from each superpixel. In [82], a saliency integration approach via the use of similar images is proposed to elevate the saliency detection performance. Besides, the deep-learning based models [75,80] and multiple-instance learning based models [81] are proposed to improve the saliency detection performance. Generally speaking, the aforementioned efforts major in image saliency detection, thus they are inappropriate to perform video saliency detection. According to the issue of this paper, in this section, we will mainly review the state-of-the-art spatiotemporal saliency models designed from various aspects in the recent years.

As a pioneering work, Itti et al. [3] proposed a well-known center-surround scheme, which exploits luminance, color and orientation across different scales to generate the saliency map. Subsequently, in [23], a surprise model was further designed, in which a set of features including color, luminance, orientation, flicker and motion energy are exploited. Along this way, there are also some other models which estimate the difference of each patch/volume and its spatiotemporal surroundings. Based on the discriminant center-surround hypothesis [24,25], the Kullback-Leibler divergence on dynamic texture feature is exploited in [26]. In [27], local regression kernels based self-resemblance is used to measure saliency. In [18], a multiscale background model represented by Gaussian pyramid is used to detect objects undergoing salient motion. In [28], the earth mover's distance (EMD) is used for computing the center-surround difference in the spatiotemporal receptive field. The contrast of luminance and directional coherence is adopted to measure spatiotemporal saliency in [29].

Besides the classical center-surround scheme based models, lots of spatiotemporal saliency models, which are based on various mechanisms including information theory, control theory, frequency domain analysis, machine learning, sparse representation and the fusion scheme of spatial and temporal saliency, are also proposed in recent years.

The knowledge of information theory can be used for saliency measurement. For instance, self-information in [30] is used to represent the object saliency. The minimum conditional entropy in [31] follows a local approach to estimate saliency. The incremental coding length in [32] is exploited to measure the perspective entropy gain of each feature and achieve attention selectivity. On the basis of control theory, some effective works first model the video sequence as a linear dynamic system, and then use the observability of output [33] and the controllability of states [34] to measure salient motion, respectively.

The frequency domain analysis method is also exploited for video saliency detection. Inspired by spectral residual for image saliency detection [35], the temporal spectral residual on video slices along X-T and Y-T planes is exploited to perform motion saliency detection in [36]. Except for spectral residual, a spatiotemporal saliency model based on phase spectrum of quaternion Fourier transforms is proposed in [14].

With the increasing attention on machine learning, some efforts have also been attempted for video saliency detection. In [37], stimulus-driven and task-related factors are exploited to model video saliency under the framework of probabilistic multi-task learning. A standard soft margin support vector machine with Gaussian kernels is adopted to predict interesting locations in [38]. In [39], both low-level and high-level features are used to predict visual attention from videos by using support vector regression. In [40], the one-class support vector machine is used to remove the consistent trajectories in motion, yielding salient regions in videos. Besides, the conditional random field (CRF) is exploited to integrate multiscale contrast, center-surround histogram and spatial distribution of color and motion vector field in [5].

The sparse representation method is also adopted in spatiotemporal saliency models. In [41], a sparse feature selection model, which incorporates temporal consistence and temporal difference, is used to generate saliency maps for videos. Besides, the regularized feature reconstruction [42] and the sparse low-rank decomposition [43,44] are employed in video saliency measurement.

Fusion scheme is widely used in some saliency models. Some models utilize simple fusion schemes such as dynamic fusion technique [45], mean, maximum and multiplication [47], and weighted linear summation [50,51]. Some other models adopt slightly complicated fusion schemes. In [46], intra-map competition and inter-map competition are combined for fusion. In [48], the color contrast based and motion based saliency maps are first combined via linear summation, and then multiplied by the location prior to generate the final saliency map. In [49], temporal and spatial saliency maps are combined based on a center-surround scheme. In [76], the model fuses the color saliency based on global motion clues in a batch-wise fashion. In the compressed domain, a simple multiplication method [52] and a parameterized normalization, sum and product method [53] are used to fuse spatial and temporal saliency maps. Besides, spatial and temporal saliency maps can also be fused by using conditional random field [54], weighted combination method [55] and adaptive entropy-based uncertainty weighting [56].

In the recent years, more and more efforts have been made for video saliency detection. For example, in [57], the spatiotemporal saliency is measured by finding the steady-state distribution of a random walker with restart. Besides, the segmented regions/superpixels are also employed in some spatiotemporal saliency models. In [58], motion distinctiveness, global contrast and spatial sparsity are first used to measure temporal saliency and spatial saliency, respectively, and then the spatiotemporal saliency map is obtained via an adaptive fusion scheme. In [59], trajectory descriptors and inside-outside maps are exploited to generate the trajectory-level temporal saliency, and a quality-guided fusion scheme is then exploited to integrate temporal and spatial saliency maps, yielding the final spatiotemporal saliency map. In [60], the motion saliency is first measured in an iterative way based on a superpixel-level graph, and then bidirectional temporal propagation and local-global spatial propagation are performed successively to obtain the final spatiotemporal saliency map. In [61], intra-frame boundary information and inter-frame motion information are first combined to obtain the gradient flow field, which serves as the input of contrast saliency measurement to obtain the coarse spatiotemporal saliency map, and then an energy optimization method is introduced to enhance the spatiotemporal consistency. In [62], based on spatial edges and temporal motion boundaries, initial spatiotemporal saliency maps are first generated using geodesic distance on an intra-frame graph, and then the final spatiotemporal saliency maps are obtained via the geodesic distances to background regions on an inter-frame graph in the subsequent frames. More recently, in [76], the color-based saliency is fused with global motion cues in a batch-wise manner.

All the aforementioned efforts can achieve visually promising results in some cases, but their performances will degrade when coping with a variety of unconstrained videos with complicated motion and complex scenes. In order to elevate the saliency detection performance in videos, we design a novel spatiotemporal saliency model using bagging strategy and spatiotemporal propagation.

3. Proposed spatiotemporal saliency model

The main flowchart of our proposed model is illustrated in Fig. 1, for each current frame $\mathbf{F}_t$, a local temporal window $\mathrm{WT}_t = \{\mathbf{F}_{t-2},\mathbf{F}_{t-1},\mathbf{F}_t,\mathbf{F}_{t+1},\mathbf{F}_{t+2}\}$ is centered on $\mathbf{F}_t$, in which $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$ are the previous two frames with spatiotemporal saliency maps $\{\mathbf{STS}_{t-2},\mathbf{STS}_{t-1}\}$ and $\{\mathbf{F}_{t+1},\mathbf{F}_{t+2}\}$ are the following two frames. The pipeline of saliency computation for each current frame $\mathbf{F}_t$ consists of three components including saliency prediction, temporal propagation and spatial propagation, which will be detailed from Section 3.2–3.4 in turn. First of all, feature descriptors, such as location, color, texture and motion features, are extracted on the superpixel segmentation results of video frames in Section 3.1. The prediction models $\{\mathrm{M}_{t-2},\mathrm{M}_{t-1}\}$ are then learned from previous frames $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$ and utilized to perform saliency prediction for $\mathbf{F}_t$ in Section 3.2. The temporal propagation in Section 3.3 is performed to propagate saliency maps of $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$ and $\{\mathbf{F}_{t+1},\mathbf{F}_{t+2}\}$ to $\mathbf{F}_t$. The final spatiotemporal saliency map of $\mathbf{F}_t$ is obtained via the spatial propagation in Section 3.4. To summarize, the overall process of generating saliency map for each current frame $\mathbf{F}_t$ is described in Section 3.5.

3.1. Feature extraction

As demonstrated in the recent works [58–62], features extracted from superpixels are effective and efficient for spatiotemporal saliency detection. For each current frame $\mathbf{F}_t$, the simple linear iterative clustering (SLIC) algorithm [67] is employed to generate a set of perceptually homogenous superpixels $\{sp_t^i\}_{i=1}^{n_t}$, where n_t is the number of the generated superpixels and approximately equals to 400 in our implementation. Note that the notation sp_t without superscript denotes all superpixels in $\mathbf{F}_t$.

Many different kinds of features are adopted in our model. We first calculate the horizontal and vertical locations of superpixels as the location feature, which indicates the spatial distribution of salient object and background. Second, the color feature is extracted since this is one of the most important cues in human visual system and certain colors tend to draw more attention than the others [68]. Third, we adopt the local binary pattern (LBP) [69], which is widely used in many vision tasks, as the texture feature in our model. The above two features are used to describe the distribution of colors and textures, respectively, in each superpixel. Besides, the crucial hint for video sequence, motion information, which is complementary to the location, color and texture features, is also incorporated in our model. For each current frame $\mathbf{F}_t$, its pixel-level forward motion vector field $\mathbf{MVF}_{t,t+1}$ is calculated using the large displacement optical flow (LDOF) method [70]. The motion feature of each superpixel is extracted based on the amplitudes and orientations of pixels in $\mathbf{MVF}_{t,t+1}$. The aforementioned features are summarized in Table 1. From the perspective of efficiency and effectiveness, for each superpixel, a 20-dimension feature vector $\mathbf{x} = [x_1,x_2,..,x_{20}] \in \mathrm{R}^{20}$ is generated by concatenating all features, and is used for saliency prediction.

3.2. Saliency prediction

With the features extracted for superpixels, we propose a bagging-based prediction model, in which each base prediction model is learned on an individual training set using random forest regressor. Next, we will show how to obtain the prediction model and to perform saliency prediction.

For each current frame $\mathbf{F}_t$, its previous two frames $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$ with spatiotemporal saliency maps $\{\mathbf{STS}_{t-2},\mathbf{STS}_{t-1}\}$ are exploited to learn the prediction models. Here we take $\mathbf{F}_{t-2}$ as an instance. A binary mask $\mathbf{BSTS}_{t-2}$ is first obtained by applying the Otsu's method [71] on $\mathbf{STS}_{t-2}$. To select good training samples, a confidence score is defined as follows:

$$a_{t-2}^i = \frac{|sp_{t-2}^i \cap \mathbf{BSTS}_{t-2}|}{|sp_{t-2}^i|}, \tag{1}$$

where a_{t-2}^i denotes the confidence score measuring the percentage of the pixels in the superpixel sp_{t-2}^i assigned to the salient object, and $|\cdot|$ denotes the number of pixels in the superpixel or the overlapped region. Then the corresponding saliency score A_{t-2}^i is defined as follows:

$$A_{t-2}^i = \begin{cases} 1 & a_{t-2}^i \geqslant q_h \\ 0 & a_{t-2}^i = q_l \end{cases}. \tag{2}$$

If a_{t-2}^i is not less than q_h, sp_{t-2}^i is regarded as a positive sample and the saliency score A_{t-2}^i is set to 1. If a_{t-2}^i equals to q_l, sp_{t-2}^i is treated as a negative sample and A_{t-2}^i is set to 0. Here, in order to obtain confidence superpixels, we set the high threshold q_h and the low threshold q_l as 0.8 and 0, respectively. In this way, for $\mathbf{F}_{t-2}$, we obtain a total of Q_{t-2}

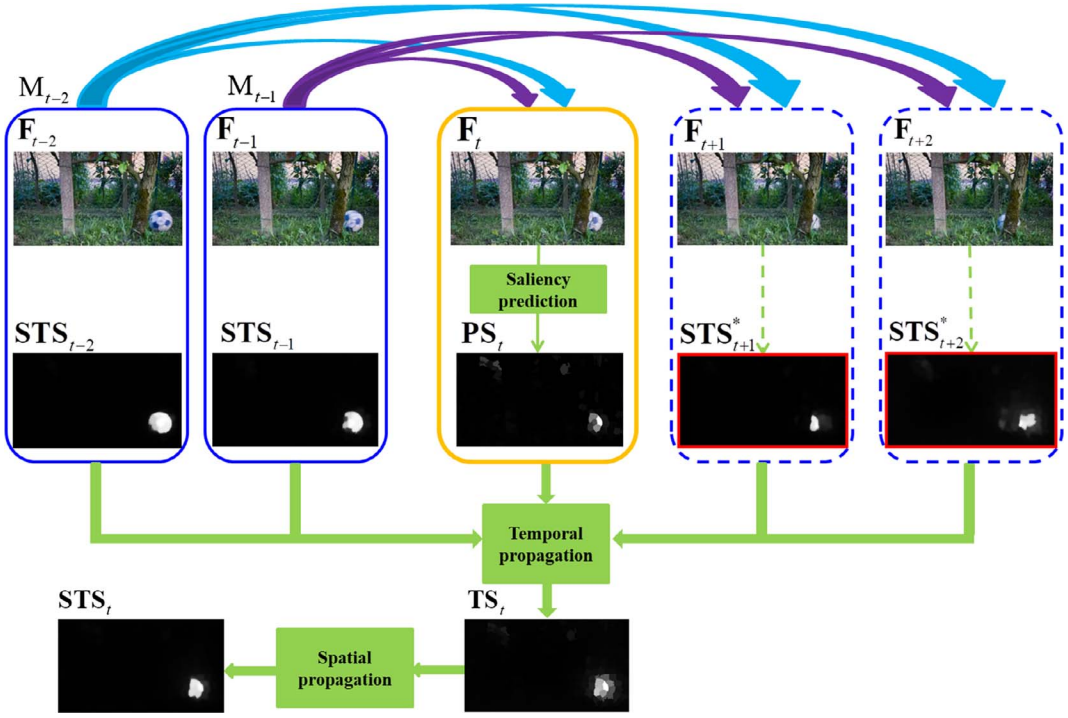

Fig. 1. Illustration of the proposed saliency model. For each current frame $\mathbf{F}_t$, the saliency prediction is first performed to obtain the predicted saliency map $\mathbf{PS}_t$ located in the box with yellow solid line. Then, the bidirectional temporal propagation is deployed to obtain the temporal saliency map $\mathbf{TS}_t$. Lastly, the spatial propagation is performed to obtain the final spatiotemporal saliency map $\mathbf{STS}_t$. Note that, grayscale images within boxes with blue solid line and boxes with blue dash line denote the spatiotemporal saliency maps of previous frames $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$ and the temporary saliency maps of the following frames $\{\mathbf{F}_{t+1},\mathbf{F}_{t+2}\}$, respectively. For a clear illustration, the temporary saliency maps are marked with red solid line, and are exploited in the bidirectional temporal propagation.

Table 1
Features extracted for each superpixel.

Feature descriptions	Notation	Dim
Location features		
The average of normalized x coordinates	x_1	1
The average of normalized y coordinates	x_2	1
Color features		
The average of RGB values	$x_3 \sim x_5$	3
The variances of RGB values	$x_6 \sim x_8$	3
The average of CIELab values	$x_9 \sim x_{11}$	3
The variances of CIELab values	$x_{12} \sim x_{14}$	3
Texture features		
The average of LBP values	x_{15}	1
The variances of LBP values	x_{16}	1
Motion features		
The average of motion amplitude values	x_{17}	1
The variances of motion amplitudes values	x_{18}	1
The average of motion orientation values	x_{19}	1
The variances of motion orientation values	x_{20}	1

($Q_{t-2} \leqslant n_{t-2}$) confident superpixels $R_{t-2} = \{sp^1_{t-2}, sp^2_{t-2}, ..., sp^{Q_{t-2}}_{t-2}\}$ with the corresponding saliency scores $A_{t-2} = \{A^1_{t-2}, A^2_{t-2}, ..., A^{Q_{t-2}}_{t-2}\}$ and features $\mathbf{X}^*_{t-2} = \{\mathbf{x}^1_{t-2}, \mathbf{x}^2_{t-2}, ..., \mathbf{x}^{Q_{t-2}}_{t-2}\}$.

Next, a random forest regressor, which maps the feature vector of each superpixel to a saliency score, is exploited to obtain the prediction model on the training data, $\mathbf{X}^*_{t-2} = \{\mathbf{x}^1_{t-2}, \mathbf{x}^2_{t-2}, ..., \mathbf{x}^{Q_{t-2}}_{t-2}\}$ and $A_{t-2} = \{A^1_{t-2}, A^2_{t-2}, ..., A^{Q_{t-2}}_{t-2}\}$. However, the number of positive samples (for salient objects) is often much less than negative samples (for background). For example, as shown in Fig. 1, the proportion of salient object is much smaller than background. Certainly, it may be vice versa for other videos with a larger background. This phenomenon, *i.e.* class imbalance, directly results in performance degradation for the prediction model. To address such a problem, we design a novel training scheme with the introduction of bagging strategy. Here we denote the superpixels belonging to salient object and background as minority class P and majority class N, respectively. The majority class is downsampled for m times with a certain proportion q, resulting in m subsets $\{\mathrm{N}^1, \mathrm{N}^2, ..., \mathrm{N}^m\}$ of majority class. As shown in Fig. 2, combining with the minority class P, we obtain a total of m undersampled training sets, $\{\{\mathrm{P},\mathrm{N}^1\},\{\mathrm{P},\mathrm{N}^2\},...,\{\mathrm{P},\mathrm{N}^m\}\}$. Then we train a total of m random forest regressors, $\{\mathrm{RF}^1, \mathrm{RF}^2, ..., \mathrm{RF}^m\}$, on these training sets. Therefore, the prediction model M, *i.e.* an ensembling regressor, is the combination of these random forest regressors. Specifically, the prediction model for $\mathbf{F}_{t-2}$ is denoted as M_{t-2} and the corresponding set of random forest regressors is denoted as $\{\mathrm{RF}^1_{t-2}, \mathrm{RF}^2_{t-2}, ..., \mathrm{RF}^m_{t-2}\}$. Here, from the perspective of efficiency and effectiveness, the time of undersampling m and the certain propagation q are empirically set as 15 and 2/3, respectively.

Finally, based on the learned prediction models, $\{\mathrm{M}_{t-2}, \mathrm{M}_{t-1}\}$, the saliency prediction for each current frame $\mathbf{F}_t$ is performed as follows:

$$\mathbf{PS}_t = \frac{1}{2m}\left[\sum_{r1=1}^{m} \mathrm{RF}^{r1}_{t-1}(\mathbf{X}_t) + \sum_{r2=1}^{m} \mathrm{RF}^{r2}_{t-2}(\mathbf{X}_t)\right], \tag{3}$$

where $\mathbf{PS}_t$ denotes the predicted saliency map, which is marked with a yellow box in Fig. 1. $\mathbf{X}_t = \{\mathbf{x}^1_t, \mathbf{x}^2_t, ..., \mathbf{x}^{n_t}_t\}$ indicates that the features of all superpixels in $\mathbf{F}_t$ are used as the testing data of the learned random forest regressors.

3.3. Temporal propagation

A reasonable assumption of saliency in videos is that the correlated regions in the adjacent frames should have similar saliency scores, and thus the temporal consistency between saliency maps should be enhanced. For this purpose, a novel temporal propagation within a local temporal window, $\mathrm{WT}_t = \{\mathbf{F}_{t-2}, \mathbf{F}_{t-1}, \mathbf{F}_t, \mathbf{F}_{t+1}, \mathbf{F}_{t+2}\}$, is proposed. A detailed description of the proposed temporal propagation method consists of the following three steps.

(a) Forward temporal propagation. Based on the previous frames $\{\mathbf{F}_{t-2}, \mathbf{F}_{t-1}\}$ with the available spatiotemporal saliency maps $\{\mathbf{STS}_{t-2}, \mathbf{STS}_{t-1}\}$, the forward temporal propagation is performed for

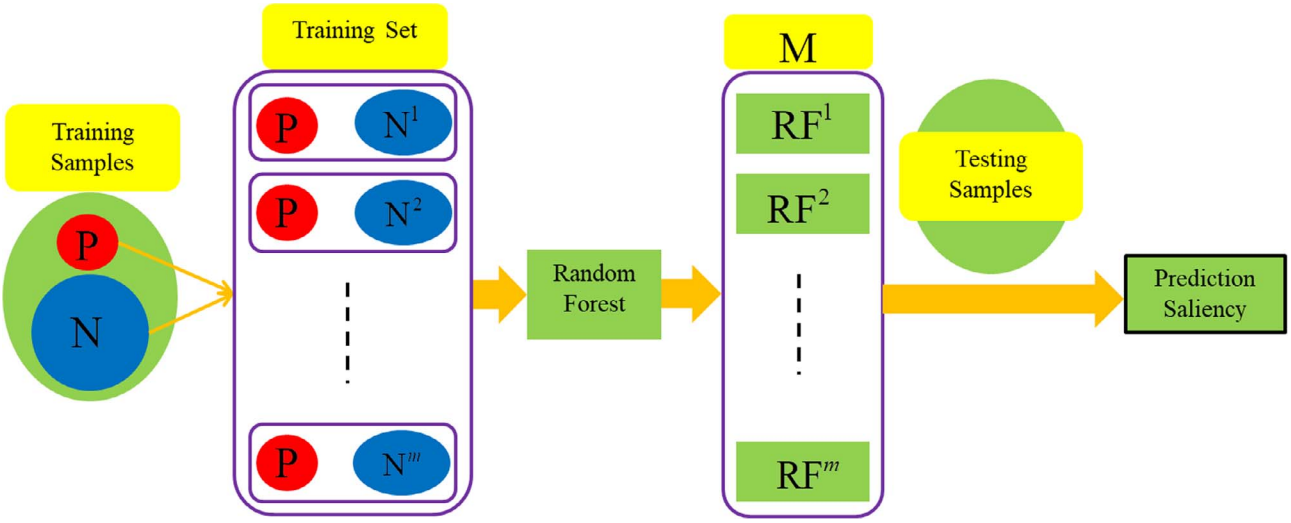

Fig. 2. Illustration of bagging-based saliency prediction.

each current frame $\mathbf{F}_t$. Here we take $\mathbf{F}_{t-1}$ as the previous frame for an instance. For each superpixel sp_t^i in $\mathbf{F}_t$, the most matched superpixel in $\mathbf{F}_{t-1}$ is defined as follows:

$$sp_{t-1}^{j*} = \text{argmin}_{sp_{t-1}^j \in \mathbf{F}_{t-1}} [O_{t,t-1}(i,j) \times D_{t,t-1}^S(i,j)], \tag{4}$$

where $O_{t,t-1}(i,j)$ and $D_{t,t-1}^S(i,j)$ are the overlap degree and spatial distance, between sp_{t-1}^j and the projected region $Pr_{t-1}(sp_t^i)$ of sp_t^i in $\mathbf{F}_{t-1}$, respectively. $Pr_{t-1}(sp_t^i)$ is obtained by mapping each pixel in sp_t^i into $\mathbf{F}_{t-1}$ using the motion vector field $\mathbf{MVF}_{t,t-1}$. The overlap degree and spatial distance are defined as follows:

$$O_{t,t-1}(i,j) = \frac{|Pr_{t-1}(sp_t^i) \cap sp_{t-1}^j|}{|sp_{t-1}^j|}, \tag{5}$$

$$D_{t,t-1}^S(i,j) = \|\mu_{t-1}^j - (\mu_t^i + \mathbf{mvf}_{t,t-1}^i)\|, \tag{6}$$

where μ_{t-1}^j and μ_t^i denote the spatial center position of sp_{t-1}^j and sp_t^i, respectively, and $\mathbf{mvf}_{t,t-1}^i$ is the displacement of sp_t^i, *i.e.* the mean of pixel-level motion vectors in sp_t^i.

The most matched superpixel sp_{t-1}^{j*} and its adjacent superpixels constitute the matching set $\mathbb{N}_{t,t-1}^i$, and then the similarity between the superpixel sp_t^i and each superpixel sp_{t-1}^j in $\mathbb{N}_{t,t-1}^i$ is calculated by using color feature and motion feature, respectively,

$$\mathrm{w}_{t,t-1}^C(i,j) = \frac{1}{Z_1^C}\left(e^{-\frac{D_{t,t-1}^S(i,j)}{\sigma^S}} \Big/ Z^S\right) \bullet \left(e^{-\frac{\|\mathbf{C}_t^i - \mathbf{C}_{t-1}^j\|}{\sigma^C}} \Big/ Z_2^C\right), \tag{7}$$

$$\mathrm{w}_{t,t-1}^M(i,j) = \frac{1}{Z_1^M}\left(e^{-\frac{D_{t,t-1}^S(i,j)}{\sigma^S}} \Big/ Z^S\right) \bullet \left(e^{-\frac{\|\mathbf{M}_t^i - \mathbf{M}_{t-1}^j\|}{\sigma^M}} / Z_2^M\right), \tag{8}$$

where $\mathrm{w}_{t,t-1}^C(i,j)$ and $\mathrm{w}_{t,t-1}^M(i,j)$ are two similarities between sp_t^i and sp_{t-1}^j. $\mathbf{C}_t^i$ and $\mathbf{M}_t^i$ denote the color feature and the motion feature, respectively, of sp_t^i. Here $\mathbf{C}_t^i$ is the average of CIELab values of pixels in sp_t^i, *i.e.* $\mathbf{C}_t^i = [x_9, x_{10}, x_{11}]$, and $\mathbf{M}_t^i$ is the average of motion amplitude and orientation values of pixels in sp_t^i, *i.e.* $\mathbf{M}_t^i = [x_{17}, x_{19}]$. Besides, σ^S and Z^S are set to the mean and sum of all spatial distances between sp_t^i and those superpixels in its matching set $\mathbb{N}_{t,t-1}^i$. Analogously, $\{\sigma^C, Z_2^C\}$ and $\{\sigma^M, Z_2^M\}$ are computed by using the distance based on color and motion feature, respectively. The two normalization factors, Z_1^C and Z_1^M, are the summation of all similarity values calculated using Eqs. (7) and (8), respectively.

The forward temporal propagation is performed for the current frame $\mathbf{F}_t$ on the basis of superpixel. For each superpixel sp_t^i in $\mathbf{F}_t$, the forward propagated saliency is defined as follows:

$$\mathbf{FS}_t^i = \sum_{j \in \mathbb{N}_{t,t-2}^i} [\mathrm{w}_{t,t-2}^C(i,j) + \mathrm{w}_{t,t-2}^M(i,j)] \bullet \mathbf{STS}_{t-2}^j + \sum_{j \in \mathbb{N}_{t,t-1}^i} [\mathrm{w}_{t,t-1}^C(i,j) + \mathrm{w}_{t,t-1}^M(i,j)] \bullet \mathbf{STS}_{t-1}^j. \tag{9}$$

(b) Backward temporal propagation. Based on the previous frames $\{\mathbf{F}_{t-2}, \mathbf{F}_{t-1}\}$ with the available saliency maps $\{\mathbf{STS}_{t-2}, \mathbf{STS}_{t-1}\}$ and the learned prediction models $\{\mathrm{M}_{t-2}, \mathrm{M}_{t-1}\}$, the saliency prediction, forward temporal propagation, and spatial propagation, which will be described in Section 3.4, are performed for the following frames $\{\mathbf{F}_{t+1}, \mathbf{F}_{t+2}\}$, yielding the temporary saliency maps. Here we take $\mathbf{F}_{t+1}$ as an instance. The saliency prediction is first performed for $\mathbf{F}_{t+1}$ using Eq. (3) to obtain the predicted saliency map $\mathbf{PS}_{t+1}$ as follows:

$$\mathbf{PS}_{t+1} = \frac{1}{2m}\left[\sum_{r1=1}^{m} \mathrm{RF}_{t-1}^{r1}(\mathbf{X}_{t+1}) + \sum_{r2=1}^{m} \mathrm{RF}_{t-2}^{r2}(\mathbf{X}_{t+1})\right], \tag{10}$$

where $\mathbf{X}_{t+1}$ denotes the features of all superpixels in frame $\mathbf{F}_{t+1}$. Similar to Eq. (9), for each superpixel sp_{t+1}^i in $\mathbf{F}_{t+1}$, the forward temporal propagation is defined as follows:

$$\mathbf{FS}_{t+1}^i = \sum_{j \in \mathbb{N}_{t+1,t-2}^i} [\mathrm{w}_{t+1,t-2}^C(i,j) + \mathrm{w}_{t+1,t-2}^M(i,j)] \bullet \mathbf{STS}_{t-2}^j + \sum_{j \in \mathbb{N}_{t+1,t-1}^i} [\mathrm{w}_{t+1,t-1}^C(i,j) + \mathrm{w}_{t+1,t-1}^M(i,j)] \bullet \mathbf{STS}_{t-1}^j. \tag{11}$$

For frame $\mathbf{F}_{t+1}$, $\mathbf{PS}_{t+1}$ and $\mathbf{FS}_{t+1}$ are combined and fed into the spatial propagation to generate the temporary saliency map $\mathbf{STS}_{t+1}^*$ as follows:

$$\begin{aligned} &\mathbf{FS}_{t+1}^* = \mathbf{PS}_{t+1} + \mathbf{FS}_{t+1} \\ &\mathbf{FS}_{t+1}^* \Rightarrow \mathbf{STS}_{t+1}^*, \end{aligned} \tag{12}$$

where the arrow denotes the process of spatial propagation.

Based on the temporary saliency maps $\{\mathbf{STS}_{t+1}^*, \mathbf{STS}_{t+2}^*\}$ of $\{\mathbf{F}_{t+1}, \mathbf{F}_{t+2}\}$, the backward temporal propagation is performed for each current frame $\mathbf{F}_t$ similar to the forward temporal propagation. For each superpixel sp_t^i in $\mathbf{F}_t$, the backward propagated saliency is defined as follows:

$$\mathbf{BS}_t^i = \sum_{j \in \mathbb{N}_{t,t+1}^i} [\mathrm{w}_{t,t+1}^C(i,j) + \mathrm{w}_{t,t+1}^M(i,j)] \bullet \mathbf{STS}_{t+1}^{*j} + \sum_{j \in \mathbb{N}_{t,t+2}^i} [\mathrm{w}_{t,t+2}^C(i,j) + \mathrm{w}_{t,t+2}^M(i,j)] \bullet \mathbf{STS}_{t+2}^{*j}. \tag{13}$$

(c) Temporal saliency map generation. For each current frame $\mathbf{F}_t$, the predicted saliency map $\mathbf{PS}_t$, the forward propagated saliency map $\mathbf{FS}_t$ and the backward propagated saliency map $\mathbf{BS}_t$ are integrated into the temporal saliency map by

$$\mathbf{TS}_t = \mathbf{PS}_t + \mathbf{FS}_t + \mathbf{BS}_t. \tag{14}$$

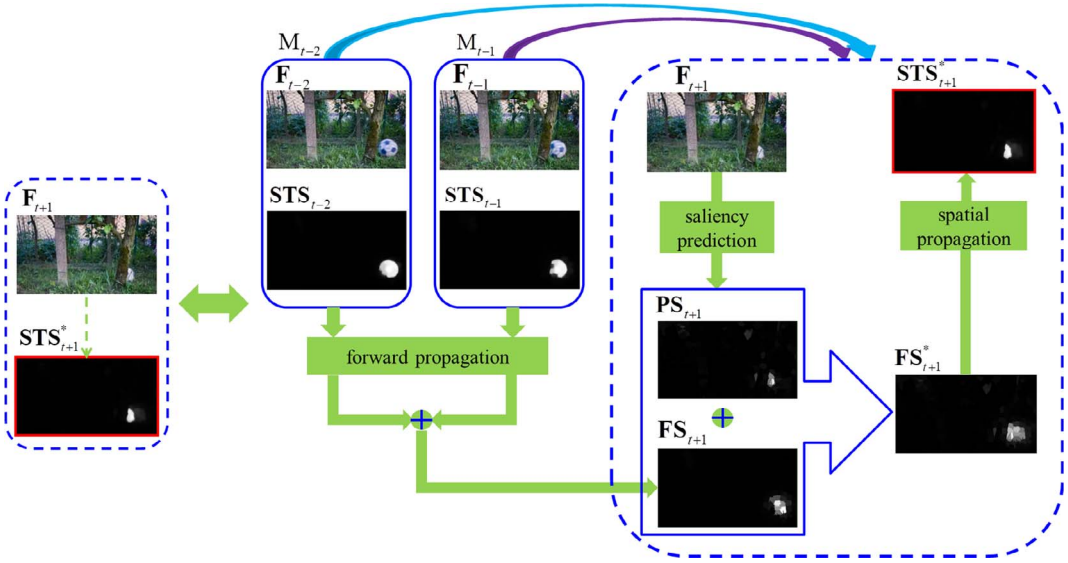

Fig. 3. Illustration of temporary saliency map estimation for the frame $\mathbf{F}_{t+1}$. Forward propagation is based on the previous frames $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$. The temporary saliency map is marked by the box with red solid line.

It can be seen from Fig. 3 that more salient object regions are highlighted in the temporal saliency map $\mathbf{TS}_t$ when compared to the predicted saliency map $\mathbf{PS}_t$. This demonstrates the effectiveness of bidirectional temporal propagation, which takes full advantage of temporal coherence within the local temporal window. Note that the temporal propagation exploits the estimated motion vector fields, but unfortunately, motion vectors are not accurate enough, particularly for complicated scenes. Furthermore, the error will be accumulated frame by frame when the temporal propagation is performed along only one direction. To eliminate the accumulated error as far as possible, the bidirectional propagation within a local temporal window is effectively utilized in our temporal propagation method, in which the backward propagation is based on the generated temporary saliency maps. The combination of forward and backward propagation weakens the accumulated errors to a certain extent, and thus can improve the quality of temporal saliency maps effectively.

3.4. Spatial propagation

To further facilitate the saliency inference in videos, we propose a spatial propagation method, which incorporates appearance and motion information in a unified manner. The flowchart of the proposed spatial propagation method is shown in Fig. 4.

For each current frame $\mathbf{F}_t$, the temporal saliency map $\mathbf{TS}_t$ is thresholded using the Otsu's method, yielding a binary salient object mask $\mathbf{BTS}_t$, as shown in Fig. 4(b). Meanwhile, using the graph-based manifold ranking (GMR) method [65], two similar graphs are constructed based on appearance and motion, denoted as Appearance-GMR and Motion-GMR, respectively. Specifically, for an undirected weighted graph $G = \{V,\varepsilon\}$, V denotes the set of all nodes in G, and ε denotes the set of all edges in G. Each node representing a superpixel is not only connected to the nodes representing its neighboring superpixels, but also connected to the nodes who share the common boundaries with its neighboring nodes. Besides, the nodes representing the superpixels on the four borders of video frame are also connected. The edge between each pair of connected nodes, which represent a pair of superpixels, sp_t^i and sp_t^j in $\mathbf{F}_t$, is assigned with the weight $\mathrm{w}_t^C(i,j)$ and $\mathrm{w}_t^M(i,j)$, which measure the color similarity for the graph Appearance-GMR and the motion similarity for the graph Motion-GMR, respectively,

$$\mathrm{w}_t^C(i,j) = e^{-\frac{\|\mathbf{C}_t^i-\mathbf{C}_t^j\|}{\mu^C}}, \tag{15}$$

$$\mathrm{w}_t^M(i,j) = e^{-\frac{\|\mathbf{M}_t^i-\mathbf{M}_t^j\|}{\mu^M}}, \tag{16}$$

where μ^C and μ^M are used to control the strength of weight between a pair of nodes, which are all set as 0.1 empirically. According to the foreground query step in GMR, the appearance-based foreground query

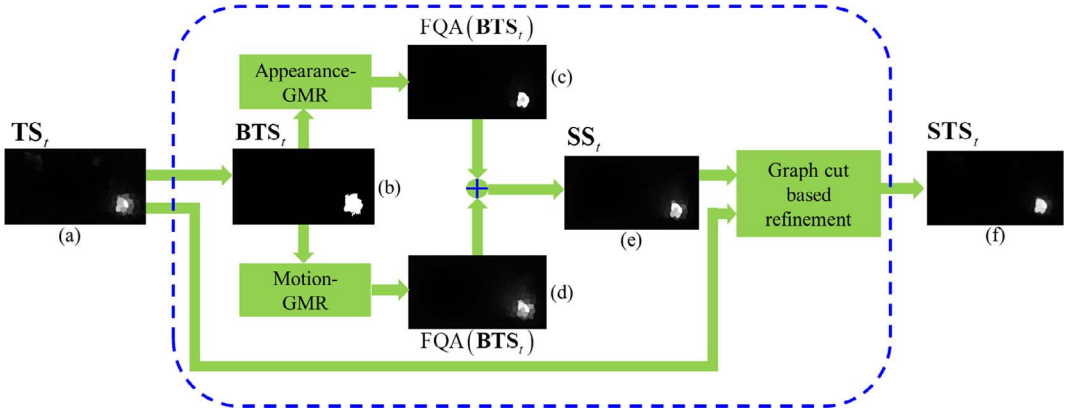

Fig. 4. Illustration of spatial propagation. (a) Temporal saliency map $\mathbf{TS}_t$; (b) binary map $\mathbf{BTS}_t$; output of foreground query based on (c) appearance-GMR and (d) motion-GMR, respectively; (e) initial result of spatial propagation, $\mathbf{SS}_t$; (f) final spatiotemporal saliency map $\mathbf{STS}_t$.

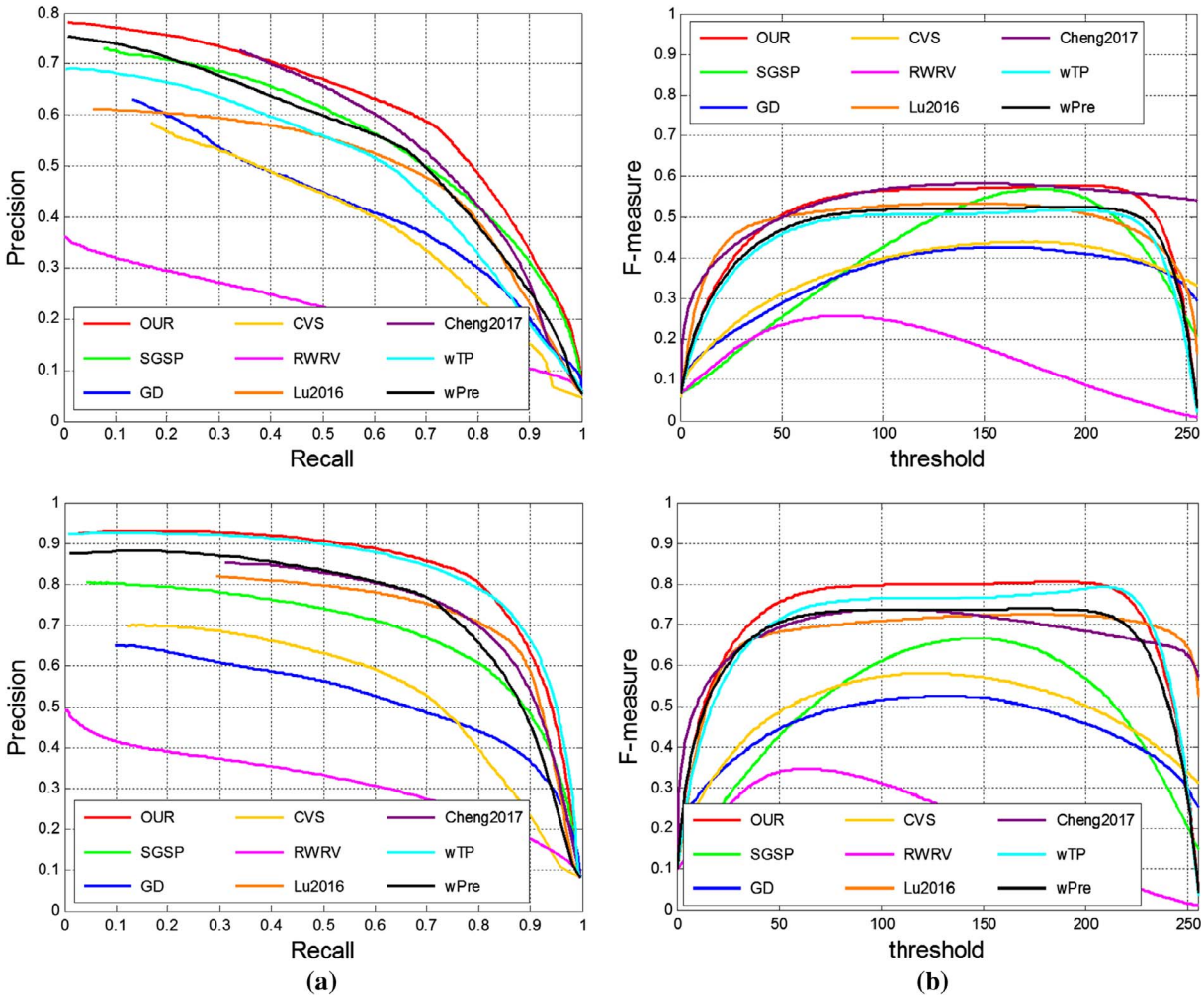

Fig. 5. (Better viewed in color) quantitative evaluation results of different saliency models. (a) PR curves, (b) F-measure curves. The first row shows the results on UVSD dataset and the second row shows the results on DAVIS dataset.

(FQA) and motion-based foreground query (FQM) are performed on the binary salient object mask $\mathbf{BTS}_t$. The outputs of FQA and FQM, as shown in Fig. 4(c) and (d), are combined to generate the initial result of spatial propagation, $\mathbf{SS}_t$, as shown in Fig. 4(e), which is defined as follows:

$$\mathbf{SS}_t = \mathrm{FQA}(\mathbf{BTS}_t) + \mathrm{FQM}(\mathbf{BTS}_t). \tag{17}$$

Finally, by feeding $\mathbf{SS}_t$ and the output of temporal propagation $\mathbf{TS}_t$ into a graph cut (GC) based refinement module [74], a salient object mask $\mathbf{FM}_t$ is generated and used to integrate with $\mathbf{SS}_t$ and $\mathbf{TS}_t$. Specifically, the final spatiotemporal saliency map $\mathbf{STS}_t$ is defined as follows:

$$\mathbf{STS}_t = \frac{1}{3}(\mathbf{FM}_t + \mathbf{TS}_t + \mathbf{SS}_t). \tag{18}$$

As shown in Fig. 4, compared to $\mathbf{TS}_t$, the initial result of spatial propagation, $\mathbf{SS}_t$, better highlights salient object and suppresses background regions via the two graphs based propagation. Furthermore, the final result of spatial propagation, *i.e.* $\mathbf{STS}_t$, suppresses background regions and highlights salient object more uniformly with the well-defined boundaries.

3.5. Overall process

To summarize, the overall process for each current frame $\mathbf{F}_t$ consists of three stages including saliency prediction, temporal propagation and spatial propagation, as shown in Fig. 1. Specifically, the prediction models $\{\mathrm{M}_{t-2},\mathrm{M}_{t-1}\}$, which are learned from previous frames $\{\mathbf{F}_{t-2},\mathbf{F}_{t-1}\}$, are employed to perform saliency prediction for $\mathbf{F}_t$, yielding the predicted saliency map $\mathbf{PS}_t$ (Section 3.2). Then, the bidirectional temporal propagation is deployed on $\mathbf{F}_t$ to generate the temporal saliency map $\mathbf{TS}_t$ (Section 3.3). Finally, spatial propagation is performed over the appearance and motion based graphs, yielding the final spatiotemporal saliency map $\mathbf{STS}_t$ (Section 3.4). Afterwards, based on the saliency map $\mathbf{STS}_t$ of $\mathbf{F}_t$, we can obtain the corresponding prediction model M_t using the method described in Section 3.2. Then, combining with the prediction model M_{t-1} of the previous frame $\mathbf{F}_{t-1}$, we produce a new prediction model pair, $\{\mathrm{M}_{t-1},\mathrm{M}_t\}$, which can be employed to perform saliency prediction for the following frames. In this way, our model works on a per-frame basis, which is guided by the saliency maps of previous frames.

Here, it should be pointed out that our model is initialized with the given salient object mask of the first frame, which is a prior introduced into our model. However, such a prior is not sufficient for spatiotemporal saliency detection in videos. The key insight is how to utilize such a prior reasonably and effectively. In the following experimental results in Section 4, we will demonstrate the reasonability and effectiveness of our model. Besides, it should be noted that the our model processes the frames from the second frame to the penultimate frame in each video, and the generation of saliency maps for the second frame and the penultimate frame are calculated within a local temporal window, which contains only three frames, *i.e.* the previous frame, the current frame and the following frame.

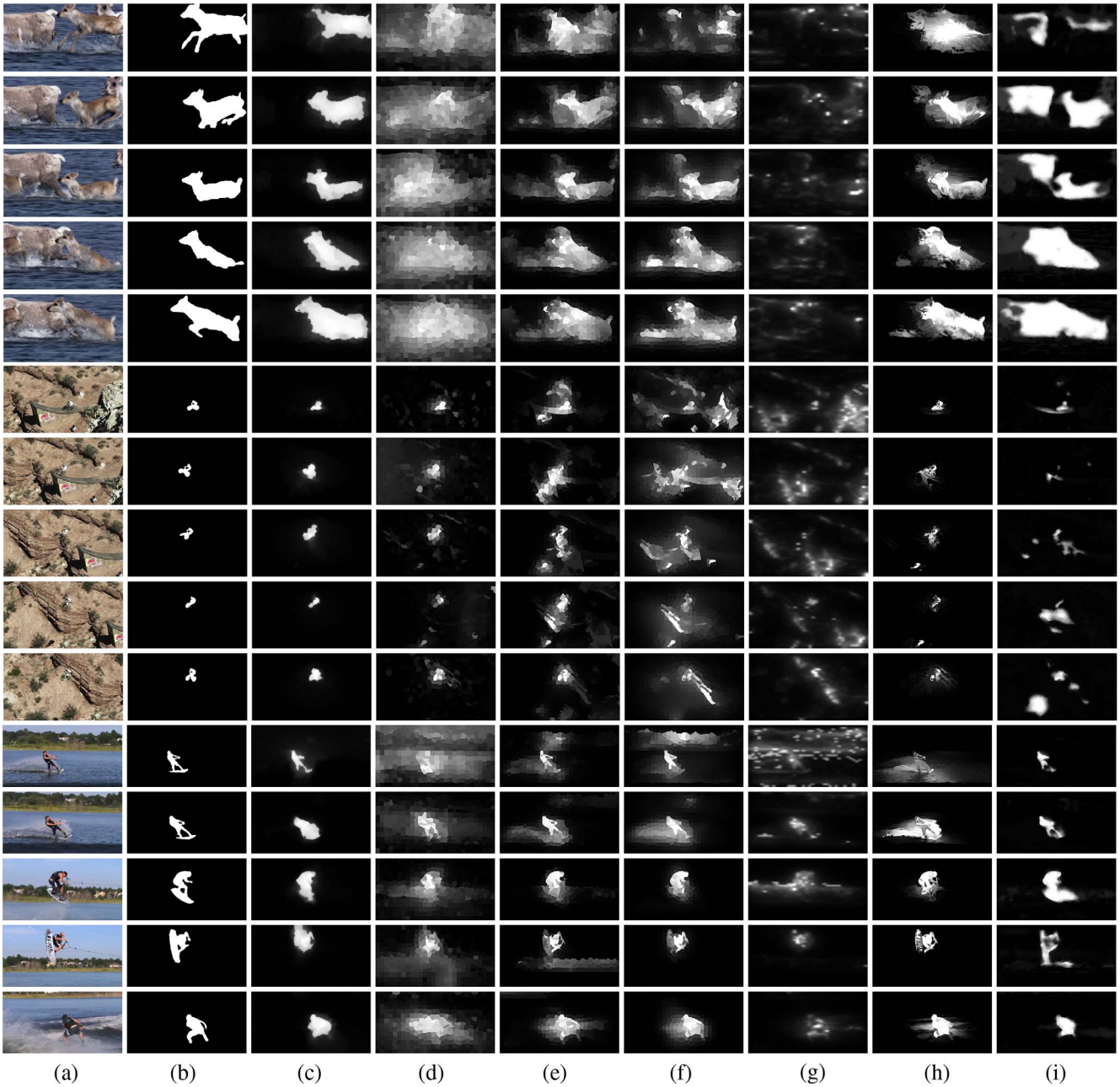

Fig. 6. Examples of spatiotemporal saliency maps for some videos in UVSD (shown with an interval of 3, 30 and 25 frames, respectively, from top to bottom). (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) OUR, (d) SGSP, (e) GD, (f) CVS, (g) RWRV, (h) Cheng2017, (i) Lu2016, respectively.

4. Experimental results

In this section, we perform comprehensive experiments for the proposed model on two public video datasets. First, the video datasets and experimental settings are introduced in Section 4.1. Then the quantitative performance comparison with state-of-the-art spatiotemporal saliency models is shown in Section 4.2. The qualitative evaluation of different models with detailed analysis is presented in Section 4.3. Next, some failure cases are analyzed in Section 4.4. Finally, the computation issue of our model is discussed in Section 4.5.

4.1. Datasets and experimental settings

We performed extensive experiments on two video datasets with manually annotated binary ground truths for salient objects. The first dataset UVSD [60] contains a total of 18 challenging videos with complicated motions. The second dataset DAVIS [66] is one of the latest dataset for video object segmentation, which contains 50 high-quality videos with different motions of human, animal and vehicle in challenging circumstances.

We compared our model with five state-of-the-art spatiotemporal saliency models including GD [62], CVS [61], RWRV [57], SGSP [60], and Cheng2017 [76]. Besides, the state-of-the-art deep learning based image saliency model, Lu2016 [75], is also adopted in performance comparison. Meanwhile, to evaluate the effect of saliency prediction and temporal propagation in our model, we evaluated the two variants of our model including “wTP” (our model without temporal propagation) and “wPre” (our model without saliency prediction). For the six state-of-the-art saliency models, the source codes with default parameter settings or executables provided by the authors were used for all videos in the two datasets. Besides, for a fair comparison, all saliency maps generated using different models are normalized into the same

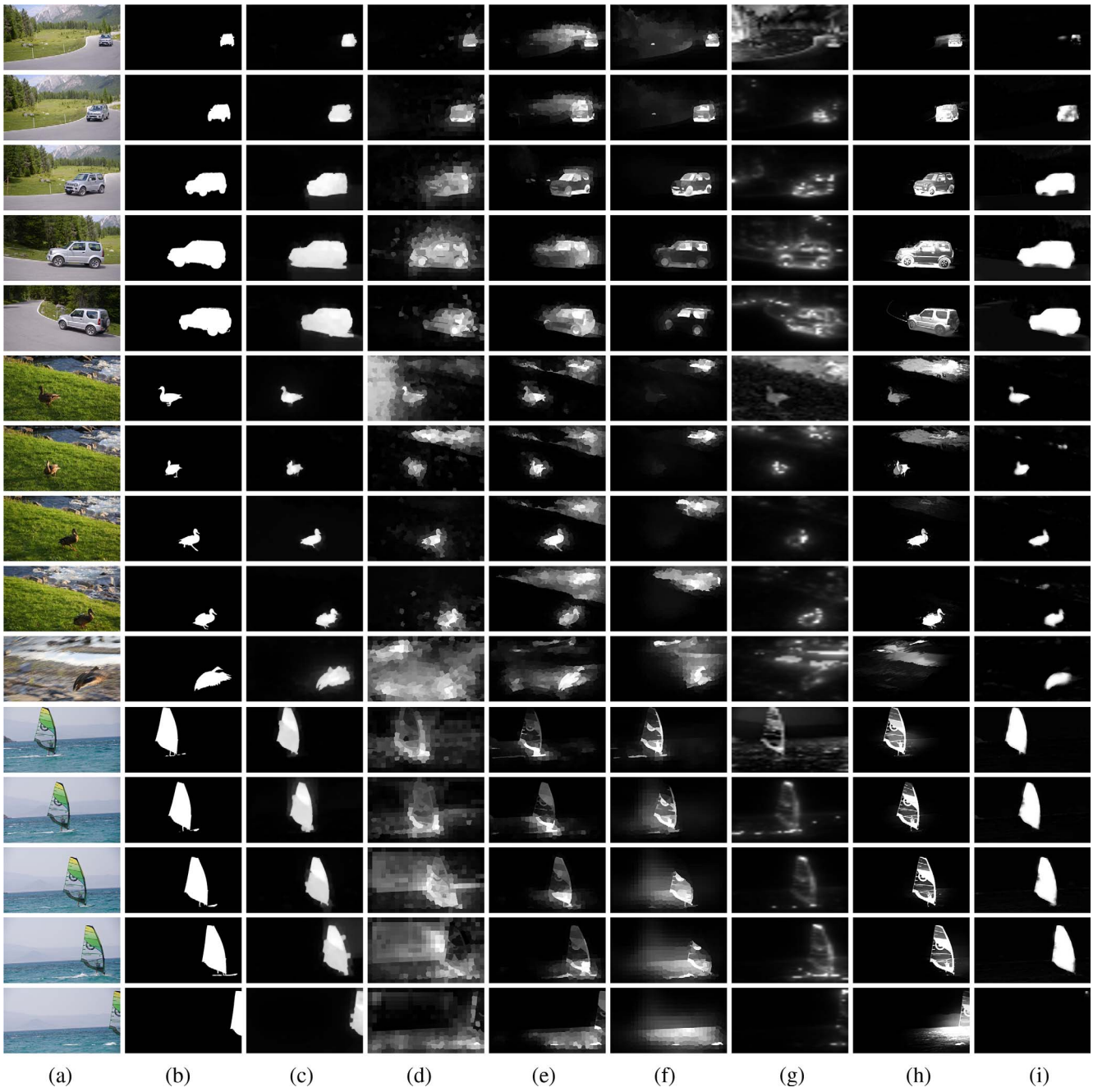

Fig. 7. Examples of spatiotemporal saliency maps for some videos in DAVIS (shown with an interval of 20, 12 and 12 frames, respectively, from top to bottom). (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) OUR, (d) SGSP, (e) GD, (f) CVS, (g) RWRV, (h) Cheng2017, (i) Lu2016, respectively.

range of [0, 255] with the full resolution of original videos.

4.2. Quantitative evaluation

In order to objectively evaluate saliency detection performance of different models, we adopt three commonly used metrics, including precision-recall (PR) curve and F-measure curve. Specifically, F-measure is defined as the weighted harmonic mean of precision and recall for a comprehensive evaluation, with the following form:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \tag{19}$$

where β^2 is set to 0.3 indicating more importance of precision than recall. For plotting the curve, the saliency maps are binarized with thresholds ranging from 0 to 255, and then 256 pairs of precision-recall combination and F-measure against thresholds are generated to plot the curves.

The PR curves of all the other models and our model on the two video datasets are plotted in Fig. 5(a) for comparison. It can be seen that our model consistently outperforms all the other models on both datasets with a great margin. In terms of F-measure curves, as shown in Fig. 5(b), our model still achieves the best performance with a wider range of higher F-measure values compared to all the other models, except that Cheng2017 achieves a comparable performance on UVSD dataset. The PR curves and F-measure curves shown in Fig. 5 demonstrate the validity and superiority of our model, which is able to distinguish the actual salient objects in unconstrained video.

As shown in Fig. 5, the performances of the two variants of our model, *i.e.* wTP and wPre, are not consistent on both datasets. For UVSD dataset, wPre consistently outweighs wTP on all the two metrics. For

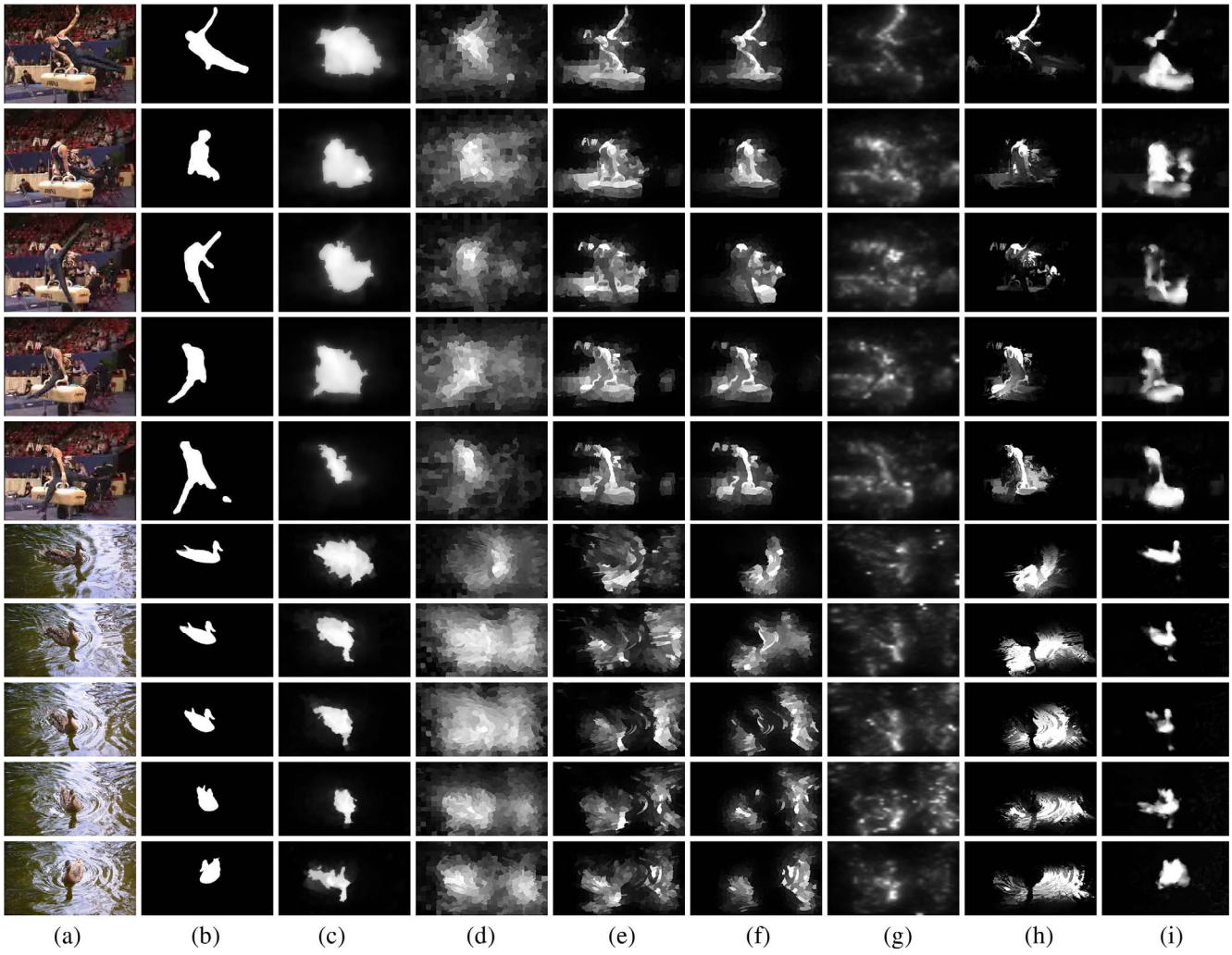

Fig. 8. Failure examples of spatiotemporal saliency maps for some challenging videos (shown with an interval of 50 and 15 frames, respectively, from top to bottom). (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) OUR, (d) SGSP, (e) GD, (f) CVS, (g) RWRV, (h) Cheng2017, (i) Lu2016, respectively.

DAVIS dataset, wTP performs better than wPre in terms of PR and F-measure curves. Although wPre and wTP performs better when compared to some top-performing models, we can find that our model consistently outperforms its two variants, wTP and wPre, on both datasets. This demonstrates that the integration of saliency prediction and temporal propagation is reasonable and sufficiently improves the saliency detection performance.

4.3. Qualitative evaluation

Figs. 6 and 7 provide the qualitative evaluation for our model and the six state-of-the-art models on UVSD and DAVIS, respectively. All these videos exhibit various challenges such as motion blur with deformation, camera-shake and shape complexity (top example in Fig. 6); small object with cluttered background and fast motion (middle example in Fig. 6); heterogeneous object with fast motion and non-rigid deformation (bottom example in Fig. 6); scale-variation with cluttered scene (top example in Fig. 7); appearance change, dynamic background, non-rigid deformation and motion blur (middle example in Fig. 7); fast motion with dynamic background, interacting objects and camera-shake (bottom example in Fig. 7). Obviously, spatiotemporal saliency detection in such videos is a challenging task.

Compared with the other models, it can be seen that our model can better highlight the complete salient objects and suppress background regions more effectively. RWRV only highlights regions around salient object boundaries or falsely highlights some background regions, because RWRV operates on the basis of patch/volume, which in essence cannot highlight salient object completely. SGSP, GD, CVS, Lu2016 and Cheng2017 are built on the basis of superpixels and the results of them are usually better than RWRV, but they also fail on such videos. In contrast, our model can better handle these challenging videos, because the saliency prediction and the temporal propagation are complementary and effective. Specifically, when optical flow estimation is not accurate enough due to complicated scenes, the temporal propagation fails but saliency prediction may work effectively thanks to the adoption of different features. When the saliency prediction model is invalid, the temporal propagation may work well based on forward and backward propagation. However, the gradient flow field of CVS is calculated based on edges of appearance and motion, it is insufficient for the complicated scenes. The results mainly highlight parts of salient object, as shown in Figs. 6(f) and 7(f). The spatiotemporal edge probability map in GD is constructed based on optical flow and object boundaries, which suffers from falsely highlighting some background regions in videos with fast motion and dynamic/cluttered background. As shown in Figs. 6(e) and 7(e), the results of GD mainly highlight parts of salient object, and falsely highlight parts of background. As for SGSP, motion histogram based motion saliency is also incapable of highlighting salient object for such complicated scenes. The results of SGSP also falsely highlight parts of background regions as shown in Figs. 6(d) and 7(d). As for the results of Cheng2017, shown in Figs. 6(h) and 7(h),

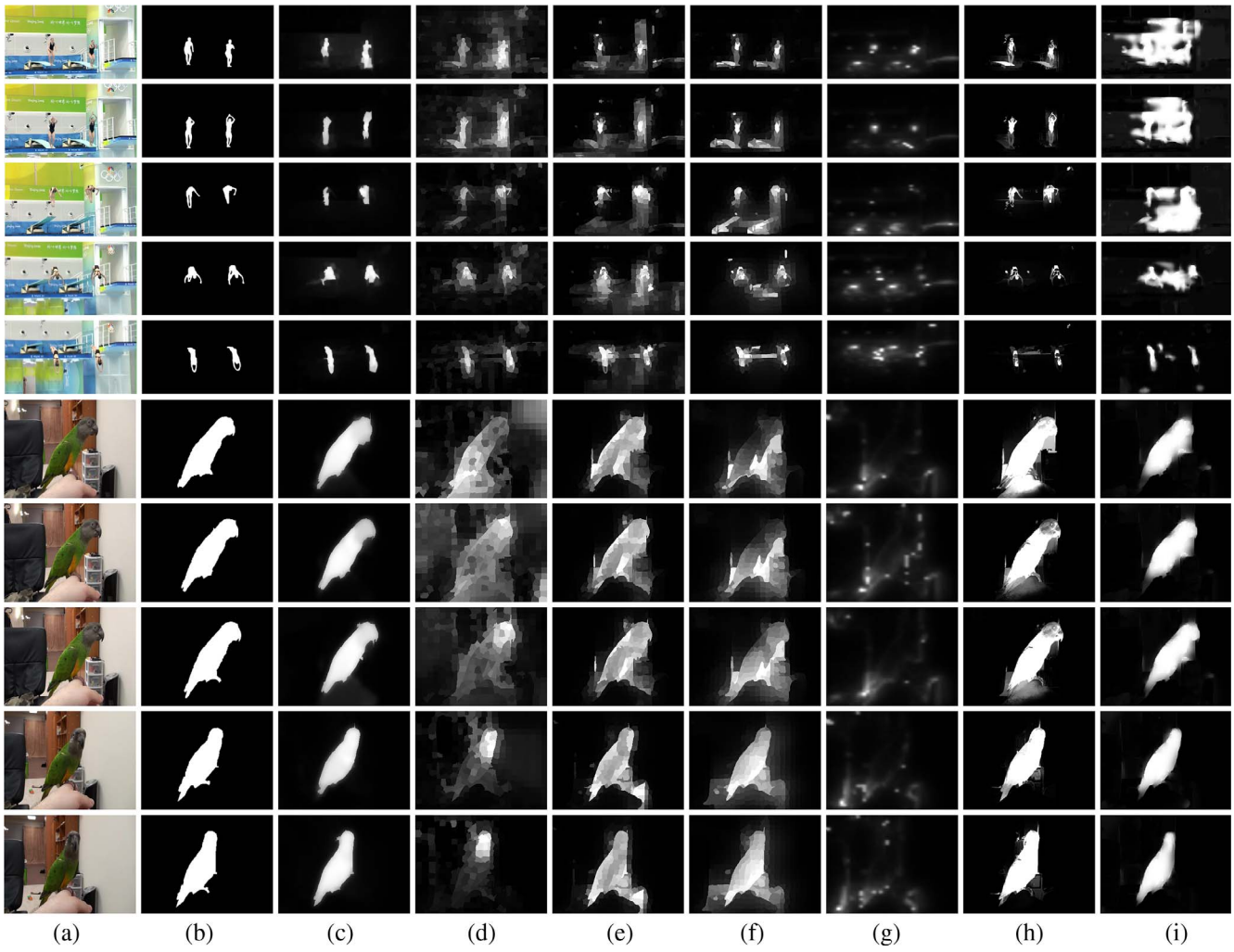

Fig. 9. Examples of spatiotemporal saliency maps for indoor videos. (a) Video frames, (b) binary ground truths, and spatiotemporal saliency maps generated using (c) OUR, (d) SGSP, (e) GD, (f) CVS, (g) RWRV, (h) Cheng2017, (i) Lu2016, respectively.

Table 2
Average processing time per frame and percentage taken by each component of our model.

Component	Optical flow estimation (LDOF)	Feature extraction	Learn/update prediction model	Test prediction model	Temporal propagation	Spatial propagation	Total time
Time (s)	17.040	2.088	47.324	0.654	2.264	5.202	74.572
Percentage (%)	22.9	2.8	63.4	0.9	3.1	6.9	100

it can be seen that the salient objects are not highlighted uniformly and only parts of salient objects are discovered. The reason behind such phenomena is that the model works in a batch-wise fashion and the motion gradient map is introduced into the raw color saliency map. Due to the lack of inter-frame information in Lu2016, its results, shown in Figs. 6(i) and 7(i), are often unable to highlight some parts of objects especially in case of non-rigid deformation. On the contrary, the spatial propagation in our model further enhances the quality of saliency map via the graphs constructed by using appearance and motion features. The results of our model, shown in Figs. 6(c) and 7(c), are the most similar to the ground truths.

Furthermore, the results on videos with indoor scene, *i.e.* diving and bird, are shown in Fig. 9. Concretely, in the top example, which suffers from fast motion and non-rigid deformation, the results generated by our model shown in Fig. 9(c) can pop out the athletes obviously compared to other models. In contrast, the results of SGSP, GD and CVS falsely highlight the background regions around the athletes, and the results of RWRV and Lu2016 are completely failed in highlighting the salient objects. As for the results of Cheng2017, they cannot highlight the athletes uniformly. In the bottom example, the best three results are generated by our model, Cheng2017 and Lu2016. Among them, the results of our model are the closest to the ground truth, and highlight the bird more uniformly and suppress the background more effectively. Thus, we can conclude that our model is suitable for the indoor scene and can achieve the best performance compared to the state-of-the-art models. This also demonstrates the effectiveness and robustness of our model.

4.4. Failure cases and analysis

As aforementioned, our model achieves the best performance compared to the state-of-the-art saliency models on both quantitative and qualitative evaluations. However, it is difficult for our model to deal with some challenging videos such as the examples shown in

Fig. 8. In the top example, a gymnast with black trousers is performing on a pommel horse. It is obvious that this example is with cluttered background, deformation and heterogeneous object. Since the complicated non-rigid deformation and the distinct colors of the gymnast's body, our model fails to highlight the salient object with well-defined boundaries. In the bottom example, a mallard is swimming in the water. This example suffers from edge ambiguity caused by water ripple and plume of mallard. Due to the reflection of mallard's head and neck in the water, our results falsely highlight such reflection regions as salient. As shown in Fig. 8(d)–(i), all the other models also cannot handle such challenging videos. Nonetheless, our model can better suppress background regions and uniformly highlight most parts of salient objects in such challenging videos.

4.5. Computation cost

Besides the performance comparison in both quantitative and qualitative evaluations of different models, we also report the computation cost of our model. Our model is implemented on a PC with Intel Core i7-4790 K 4 GHz CPU and 16 GB RAM. Table 2 shows the average processing time per frame with a resolution of 320 × 240. The average processing time per frame taken by the Matlab implementation of our model is 74.572 s. It can be observed that the two most time-consuming components are the optical flow estimation (22.9% of the total processing time), which needs to compute the optical flow for six pairs of frames, and the learn/update process of prediction model (63.4% of the total processing time). Therefore, in order to alleviate the time complexity of our model and make it more practical for applications, one of the effective solutions is the GPU-accelerated based LDOF and learn/update of prediction model. Besides, an optimized C/C++ implementation can also be adopted to improve the efficiency of our model.

5. Conclusion

This paper proposes a novel spatiotemporal saliency model for video saliency detection, which consists of three steps including bagging-based saliency prediction, bidirectional temporal propagation and spatial propagation. Firstly, the bagging-based saliency prediction, *i.e.* an ensembling regressor, identifies salient objects in an unconstrained video, yielding the predicted saliency map. Secondly, to ensure the temporal consistency, the bidirectional temporal propagation is deployed to generate the temporal saliency map, in which the backward propagation is constructed based on the generation of temporary saliency maps of following frames. Finally, the spatial propagation is conducted on the graphs constructed by using appearance and motion features in a parallel way, to generate the final spatiotemporal saliency map, which further promotes the saliency detection performance. Extensive experiments are performed on two public unconstrained video datasets, and the proposed model consistently outperforms the existing state-of-the-art spatiotemporal saliency models on both datasets, thereby improves the saliency detection performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 61471230 and No. 61601278, and by Shanghai Municipal Natural Science Foundation under Grant No. 16ZR1411100.

References

[1] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (1985) 219–227.
[2] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychol. 12 (1980) 97–136.
[3] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259.
[4] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 802–817.
[5] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 353–367.
[6] R. Shi, Z. Liu, H. Du, X. Zhang, L. Shen, Region diversity maximization for salient object detection, IEEE Signal Process. Lett. 19 (2012) 215–218.
[7] X. Zhou, Z. Liu, G. Sun, L. Ye, X. Wang, Improving saliency detection via multiple kernel boosting and adaptive fusion, IEEE Signal Process. Lett. 23 (2016) 517–521.
[8] Z. Liu, W. Zou, O. Le Meur, Saliency tree: A novel saliency detection framework, IEEE Trans. Image Process. 23 (2014) 1937–1952.
[9] 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 (2012) 1275–1289.
[10] 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 (2015) 569–582.
[11] A. Shamir, S. Avidan, Seam carving for media retargeting, Comm. ACM 52 (2009) 77–85.
[12] Z. Yuan, T. Lu, Y. Huang, D. Wu, H. Yu, Addressing visual consistency in video retargeting: A refined homogeneous approach, IEEE Trans. Circuits Syst. Video Technol. 22 (2012) 890–903.
[13] H. Du, Z. Liu, J. Jiang, L. Shen, Stretchability-aware block scaling for image retargeting, J. Vis. Commun. Image Represent. 24 (2013) 499–508.
[14] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (2010) 185–198.
[15] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image Vision Comput. 29 (2011) 1–14.
[16] L. Shen, Z. Liu, Z. Zhang, A novel H.264 rate control algorithm with consideration of visual attention, Multimedia Tools Appl. 63 (2013) 709–727.
[17] H. Liu, I. Heynderickx, Visual attention in objective image quality assessment: Based on eye-tracking data, IEEE Trans. Circuits Syst. Video Technol. 21 (2011) 971–982.
[18] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokric, V. Crnojević, D. Kukolj, Salient motion features for video quality assessment, IEEE Trans. Image Process. 20 (2011) 948–958.
[19] O. Le Meur, Z. Liu, Saccadic model of eye movements for free-viewing condition, Vision Res. 116 (2015) 152–164.
[20] O. Le Meur, A. Coutrot, Introducing context-dependent and spatially-variant viewing biases in saccadic models, Vision Res. 121 (2016) 72–84.
[21] A. Borji, M.M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Process. 24 (2015) 5706–5722.
[22] A. Borji, D.N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study, IEEE Trans. Image Process. 22 (2013) 55–69.
[23] L. Itti, P. Baldi, A principled approach to detecting surprising events in video, Proc. IEEE CVPR (2005) 631–637.
[24] D. Gao, V. Mahadevan, N. Vasconcelos, On the plausibility of the discriminant center-surround hypothesis for visual saliency, J. Vision 8 (2008).
[25] D. Gao, V. Mahadevan, N. Vasconcelos, The discriminant center-surround hypothesis for bottom-up saliency, Proc. NIPS (2008) 497–504.
[26] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 171–177.
[27] H.J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vision 9 (2009).
[28] Y. Lin, Y. Tang, B. Fang, Z. Shang, Y. Huang, S. Wang, A visual-attention model using earth mover's distance based saliency measurement and nonlinear feature combination, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 314–328.
[29] W. Kim, C. Kim, Spatiotemporal saliency detection using textural contrast and its applications, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 646–659.
[30] C. Liu, P.C. Yuen, G. Qiu, Object motion detection using information theoretic spatio-temporal saliency, Pattern Recognit. 42 (2009) 2897–2906.
[31] Y. Li, Y. Zhou, J. Yan, Z. Niu, J. Yang, Visual saliency based on conditional entropy, Proc. ACCV (2009) 246–257.
[32] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, Proc. NIPS (2008) 681–688.
[33] V. Gopalakrishnan, D. Rajan, Y. Hu, A linear dynamical system framework for salient motion detection, IEEE Trans. Circuits Syst. Video Technol. 22 (2012) 683–692.
[34] K. Muthuswamy, D. Rajan, Salient motion detection through state controllability, Proc. IEEE ICASSP (2012) 1465–1468.
[35] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, Proc. IEEE CVPR (2007) 1–8.
[36] X. Cui, Q. Liu, D.N. Metaxas, Temporal spectral residual: fast motion saliency detection, Proc. ACM MM (2009) 617–620.
[37] J. Li, Y. Tian, T. Huang, W. Gao, Probabilistic multi-task learning for visual saliency estimation in video, Int. J. Comput. Vis. 90 (2010) 150–165.
[38] E. Vig, M. Dorr, T. Martinetz, E. Barth, Intrinsic dimensionality predicts the saliency of natural dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 1080–1091.
[39] W.F. Lee, T.H. Huang, S.L. Yeh, H.H. Chen, Learning-based prediction of visual attention for video signals, IEEE Trans. Image Process. 20 (2011) 3028–3038.
[40] C. Huang, Y. Chang, Z. Yang, Y. Lin, Video saliency map detection by dominant camera motion removal, IEEE Trans. Circuits Syst. Video Technol. 24 (2014)

1336–1349.
[41] Y. Luo, Q. Tian, Spatio-temporal enhanced sparse feature selection for video saliency estimation, Proc. IEEE CVPR Workshops (2012) 33–38.
[42] Z. Ren, S. Gao, L. Chia, D. Rajan, Regularized feature reconstruction for spatio-temporal saliency detection, IEEE Trans. Image Process. 22 (2013) 3120–3132.
[43] Z. Ren, L.-T. Chia, D. Rajan, Video saliency detection with robust temporal alignment and local-global spatial contrast, Proc. ACM ICMR (2012) article 47.
[44] Y. Xue, X. Guo, X. Cao, Motion saliency detection using low-rank and sparse decomposition, Proc. IEEE ICASSP (2012) 1485–1488.
[45] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatio-temporal cues, in: ACM Intl. Conf. on Multimedia, 2006, pp. 815–824.
[46] O. Le Meur, P. Le Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Res. 47 (2007) 2483–2498.
[47] S. Marat, T.H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos, Int. J. Comput. Vis. 82 (2009) 231–243.
[48] G. Abdollahian, C.M. Taskiran, Z. Pizlo, E.J. Delp, Camera motion-based analysis of user generated video, IEEE Trans. Multimedia 12 (2010) 28–41.
[49] Y. Tong, F.A. Cheikh, F.F.E. Guraya, H. Konik, A. Trémeau, A spatiotemporal saliency model for video surveillance, Cognit. Comput. 3 (2011) 241–263.
[50] W. Kim, C. Jung, C. Kim, Spatiotemporal saliency detection and its applications in static and dynamic scenes, IEEE Trans. Circuits Syst. Video Technol. 21 (2011) 446–456.
[51] Q. Li, S. Chen, B. Zhang, Predictive video saliency detection, Commun. Comput. Inf. Sci. 321 (2012) 178–185.
[52] K. Muthuswamy, D. Rajan, Salient motion detection in compressed domain, IEEE Signal Process. Lett. 20 (2013) 996–999.
[53] 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 (2014) 27–38.
[54] W.T. Li, H.S. Chang, K.C. Lien, Exploring visual and motion saliency for automatic video object extraction, IEEE Trans. Image Process. 22 (2013) 2600–2610.
[55] D. Mahapatra, S.O. Gilani, M.K. Saini, Coherency based spatio-temporal saliency detection for video object segmentation, IEEE J. Sel. Top. Signal Process. 8 (2014) 454–462.
[56] Y. Fang, Z. Wang, W. Lin, Z. Fang, Video saliency incorporating spatiotemporal cues and uncertainty weighting, IEEE Trans. Image Process. 23 (2014) 3910–3921.
[57] H. Kim, Y. Kim, J.Y. Sim, C.S. Kim, Spatiotemporal saliency detection for video sequences based on random walk with restart, IEEE Trans. Image Process. 24 (2015) 2552–2564.
[58] Z. Liu, X. Zhang, S. Luo, O. Le Meur, Superpixel-based spatiotemporal saliency detection, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 1522–1540.
[59] J. Li, Z. Liu, X. Zhang, O. Le Meur, L. Shen, Spatiotemporal saliency detection based on superpixel-level trajectory, Signal Process.: Image Commun. 38 (2015) 100–114.
[60] Z. Liu, J. Li, L. Ye, G. Sun, L. Shen, Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation, IEEE Trans. Circuits Syst. Video Technol. (2016), http://dx.doi.org/10.1109/TCSVT.2016.2595324.
[61] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Trans. Image Process. 24 (2015) 4185–4196.
[62] W. Wang, J. Shen, F. Porikli, Saliency-aware geodesic video object segmentation, Proc. IEEE CVPR (2015) 3395–3402.
[63] W. Zhong, H. Lu, M.H. Yang, Robust object tracking via sparse collaborative appearance model, IEEE Trans. Image Process. 23 (2014) 2356–2368.
[64] Z. Xiao, H. Lu, D. Wang, L2-RLS-based object tracking, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 1301–1309.
[65] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, Proc. IEEE CVPR (2013) 3166–3173.
[66] F. Perazzi, J. Pont-Tuset, B. Mc Williams, L.V. Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, Proc. IEEE CVPR (2016) 724–732.
[67] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2274–2282.
[68] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, Proc. IEEE CVPR (2012) 853–860.
[69] M. Heikkilä, M. Pietikäinen, C. Schmid, Description of interest regions with local binary patterns, Pattern Recogn. 42 (2009) 425–436.
[70] T. Brox, J. Malik, Large displacement optical flow: Descriptor matching in variational motion estimation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 500–513.
[71] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1979) 62–66.
[72] X.Y. Liu, J. Wu, Z.H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. 39 (2009) 539–550.
[73] P. Cao, B. Li, D. Zhao, A novel cost sensitive neural network ensemble for multiclass imbalance data learning, Proc. IEEE IJCNN (2013) 1–8.
[74] H. Lu, X. Zhang, J. Qi, N. Tong, X. Ruan, M.H. Yang, Co-bootstrapping saliency, IEEE Trans. Image Process. 26 (2017) 414–425.
[75] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, Proc ECCV (2016) 825–841.
[76] C. Chen, S. Li, Y. Wang, H. Qin, A. Hao, Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion, IEEE Trans. Image Process. 26 (2017) 3156–3170.
[77] M.L. Mele, D. Millar, C.E. Rijnders, Using spatio-temporal saliency to predict subjective video quality: a new high-speed objective assessment metric, International Conference on Human-Computer Interaction (2017) 353–368.
[78] M. Song, C. Chen, S. Wang, Y. Yang, Low-level and high-level prior learning for visual saliency estimation, Inf. Sci. 281 (2014) 573–585.
[79] D. Tao, J. Cheng, M. Song, X. Lin, Manifold ranking-based matrix factorization for saliency detection, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 1122–1134.
[80] N. Liu, J. Han, DHSNet: deep hierarchical saliency network for salient object detection, Proc. IEEE CVPR (2016) 678–686.
[81] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 865–878.
[82] J. Ren, Z. Liu, X. Zhou, G. Sun, C. Bai, Saliency integration driven by similar images, J. Vis. Commun. Image Represent. 50 (2018) 227–236.