Pattern Recognition 72 (2017) 285–299

Contents lists available at ScienceDirect

Pattern Recognition

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

Fusing disparate object signatures for salient object detection in video

Zhigang Tu[a,d], Zuwei Guo[d], Wei Xie[b,*], Mengjia Yan[a], Remco C. Veltkamp[c], Baoxin Li[d], Junsong Yuan[a,*]

[a] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore*
[b] *School of Computer, Central China Normal University, LuoyuRoad 152, Wuhan, China*
[c] *Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands*
[d] *School of Computing, Informatics, Decision System Engineering, Arizona State University, Tempe, AZ 85287, USA*

ARTICLE INFO

Article history:
Received 25 March 2017
Revised 30 June 2017
Accepted 26 July 2017
Available online 28 July 2017

Keywords:
Spatiotemporal saliency computation
Salient video object detection
Object signatures
Fusion

ABSTRACT

We present a novel spatiotemporal saliency model for object detection in videos. In contrast to previous methods focusing on exploiting or incorporating different saliency cues, the proposed method aims to use object signatures which can be identified by any kinds of object segmentation methods. We integrate two distinctive saliency maps, which are respectively computed from object proposals of an appearance-dominated method and a motion-dominated algorithm, to obtain a refined spatiotemporal saliency maps. This enables the method to achieve good robustness and precision in identifying salient objects in videos under various challenging conditions. First, an improved appearance-based and a modified motion-based segmentation approaches are separately utilized to extract two kinds of candidate foreground objects. Second, with these captured object signatures, we design a new approach to filter the extracted noisy object pixels and label foreground superpixels in each object signature channel. Third, we introduce a foreground connectivity saliency measure to compute two types of saliency maps, from which an adaptive fusion strategy is exploited to obtain the final spatiotemporal saliency maps for salient object detection in a video. Both quantitative and qualitative experiments on several challenging video benchmarks demonstrate that the proposed method outperforms existing state-of-the-art approaches.

1. Introduction

The human vision system is able to select visual information of interest and ignore the rest in its visual field. This mechanism is remarkably helpful for humans to focus quickly on objects of importance in a complex scene rapidly. Visual saliency studies have gained much attention in the passed few decades owing to its wide range of applications such as image segmentation [1], image retargeting [2], image cropping [3], video compression [4], video object segmentation [5], and video object detection [6,7]. The studies originate from the task of detecting regions of interest where an observer may fixate [8], In this work, we address the issue of salient video object detection. Detecting salient objects in videos acquired under uncontrolled imaging conditions remains to be a challenging task.

Salient object detection, in both still image and video, aims to identify foreground objects from the background. It is based on the assumption that objects are usually distinctive in color, texture, motion, pattern, etc., compared to the background [9,10]. The output in one frame is a saliency map, where each value represents the probability of its corresponding pixel belongs to the salient object. Those pixels with high probability are identified as potential objects.

Detecting salient object in video is a difficult problem due to the challenges like how to integrate the motion cues with the spatial cues, how to deal with the problem if one or some adjacent frames are static, and how to tackle the case when the motion features are not available. Furthermore, in reality, an acquired video typically is under the influence of additional complicating factors like intensity variation and shadowing due to illumination changes, fast object and large background motion, cluttered background, etc. Till now, there are a few methods have been presented to consider the afore-mentioned issues for spatiotemporal salient object detection in a video [5,6].

Except for the saliency-based technique, some other methods, which aim to efficiently detect and localize video objects in diverse content, have been investigated [11,12]. For example, the

[*] Corresponding authors at: School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore. School of Computer, Central China Normal University, Wuhan, China.

E-mail addresses: TUZG@ntu.edu.sg, tuzhigang1986@gmail.com (Z. Tu), zguo32@asu.edu (Z. Guo), XW@mail.ccnu.edu.cn (W. Xie), YANM0006@e.ntu.edu.sg (M. Yan), R.C.Veltkamp@uu.nl (R.C. Veltkamp), Baoxin.Li@asu.edu (B. Li), jsyuan@ntu.edu.sg (J. Yuan).

http://dx.doi.org/10.1016/j.patcog.2017.07.028
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

motion-dominated fast object segmentation (FOS) scheme [11] and the appearance-dominated block-sparse robust principal component analysis (B-RPCA) technique [12] are two outstanding methods, although they also have their respective limitations when facing practical imaging challenges.

Both the saliency-based and other kinds of techniques are difficult to tackle the complex realistic challenges jointly. Towards addressing various difficulties in a unified model, and taking best advantages of the saliency technique and other kinds of object detection approaches, we present a method for detecting salient objects in videos by developing a saliency estimation framework that fuses saliency maps from complementary appearance signatures and motion signatures that are respectively extracted by improved versions of state-of-the-art object detection algorithms in the literature. Briefly, exploiting complementary object detection methods which are able to extract object signatures efficiently and accurately under complicated conditions is one critical task (see Section 3). Appropriately fusing these signatures-derived saliency maps for final salient object detection is another significant work (see Section 4). The highlights of the proposed method are summarized below.

Firstly, to detect potential foreground object signatures robustly in complex scenes, we consider two famous video object segmentation methods: the FOS [11] and the B-RPCA [12], although the overall fusion idea works for more object segmentation methods, due to three reasons. (1) They are complementary, where the FOS method is motion-dominated and the B-RPCA method is appearance-dominated. (2) Both the FOS method and the B-RPCA method are able to handle multiple complex conditions. (3) Their segmentation results in videos are state-of-the-art. The FOS method provides a reasonable baseline for segmenting foreground objects from the background in videos. It is relatively reliable under challenging conditions like the fast moving background, nonrigid deformations, objects with arbitrary appearance and motion types. However, the motion boundaries, which heavily depend on the estimated optical flow [13–15], usually do not correspond to the entire object boundary, because the estimated optical flow is typically very noisy, e.g, the flow is inherently inaccurate at occlusion boundaries [16]. To handle this situation, we utilize a learning method [16] to compute the motion boundaries. Furthermore, we incorporate one additional feature – the motion direction difference to further improve the learning performance, as it is helpful when the object is moving in a modest speed [11]. The appearance-dominated B-RPCA approach are able to address various realistic challenges, e.g., background motions, illumination changes, camouflage, etc, in a unified framework. But it is of low efficiency. Besides, this scheme removes all the correct foreground pixels if any frame is static, causing the B-RPCA to fail in detecting any object in this case. To this end, we improve its motion saliency estimation (MSE) in three aspects (to be elaborated later) to handle these problems.

Secondly, we compute different saliency maps from the appearance and motion signatures derived from the previous step. After obtaining those types of foreground object signatures, we design a novel method to refine them by removing noisy foreground object pixels and then label the foreground superpixels accurately in every object signature channel. We employ the detected foreground superpixels as object priors for a foreground connectivity measure [10] so as to improve saliency detection in the respective channels. In each channel, by using a principled saliency optimization technique, which integrates foreground weights and background measure [17], a much more accurate and smooth saliency map can be estimated.

Finally, based on the estimated appearance-dominated and motion-dominated saliency maps, we detect two category of refined foreground objects by utilizing adaptive thresholding [18]. In each map, higher weights will be added to the places where objects are extracted in both saliency maps, while lower weights are added to those identified as backgrounds in both. Then, we fuse these two complementary maps using a novel object and saliency guided fusion strategy to obtain a spatiotemporal consistent video saliency maps, which is able to capture the final salient objects in each frame of the given video precisely.

Compared to the existing methods, the main contributions of this paper are:

- Unlike existing methods that focus on exploiting low-level or high-level features for refining saliency detection, in the first time, we propose to use object signatures from complementary video-based object segmentation methods to supply cues for salient object detection in videos under various complex scenes.
- An efficient foreground superpixels labeling method, which depends on the identified object signatures, is exploited to find more accurate object labels to compute foreground weights for saliency maps computation.
- A learning-based method that combines various appearance and motion cues is introduced to predict motion boundaries, which are stable and helpful for foreground video objects detection in complex scenes.
- A novel fusion method, which depends on our object signature-derived saliency maps and the refined object signatures, is presented to integrate saliency maps appropriately from different channels to form a higher-quality spatiotemporal saliency maps for final salient object detection.

The remainder of the paper is organized as follows. We review related work in Section 2. Sections 3 and4 describe the framework of the proposed method. We report the experimental results with both qualitative and quantitative evaluation by comparing to the state-of-the-art algorithms on four widely used benchmark datasets in Section 5. Section 6 concludes the paper.

2. Related work

Salient object detection in still image. Here the goal is to extract the most visual-attention-catching objects in a static scene [9,10,17,19]. Based on their mechanisms to represent the saliency, these techniques can be classified into two main categories: top-down method and bottom-up method. The top-down methods [19,20] are goal directed, and some high-level priors are applied to guide the detection. The bottom-up methods [5,9,10,21] are independent of the high-level knowledge of the image content, and focus on using low-level visual cues, such as texture, location and local contrast to detect salient objects. Contrast priors are widely applied in such approaches. However, the contrast-prior-based approaches often fail to detect the objects in their entirety and are only good at identifying high-contrast edges. Some methods exploited boundary priors [22] to enhance saliency estimation. But the boundary prior is typically fragile and prone to failing [17]. In our approach, based on the availability of estimated foreground objects (and the background), we introduce a foreground connectivity measure that employs both a contrast prior and a boundary prior to improve the performance of saliency estimation.

Salient object detection in videos. In this condition, motion plays a dominant role on grabbing human visual attention in most videos. Accordingly, a motion is usually exploited as a temporal feature and it provides the strong indication for the salient objects. In earlier years, salient motion detection approaches tried to extract moving objects as salient foreground regions [23]. More recent approaches attempt to combine motion information with spatial information. Some approaches first compute temporal and spatial saliency map respectively, and then merge them with certain

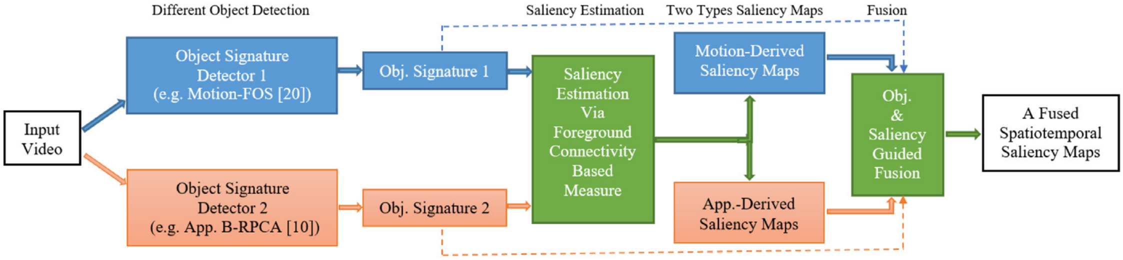
Different Object Detection
Saliency Estimation
Two Types Saliency Maps
Fusion
Input Video
Object Signature Detector 1 (e.g. Motion-FOS [20])
Object Signature Detector 2 (e.g. App. B-RPCA [10])
Obj. Signature 1
Obj. Signature 2
Saliency Estimation Via Foreground Connectivity Based Measure
Motion-Derived Saliency Maps
App.-Derived Saliency Maps
Obj. & Saliency Guided Fusion
A Fused Spatiotemporal Saliency Maps

Fig. 1. The framework of the proposed method.

rules to produce a spatiotemporal saliency map [24]. In such a simplistic fusion strategy, the final spatiotemporal map may be easily contaminated by noise in either the spatial or the temporal map. Some approaches simply extend the image saliency algorithms to deal with video saliency by adding motion cues [7]. However, the performance of these algorithms is not very good, as they do not model well temporal saliency exhibited between video frames.

3. Object signatures detection

Most of the current methods for detecting salient visual objects rely on various appearance or motion features and they may perform poorly if the video contains multiple challenging conditions [6]. The key idea of our approach is to utilize some object signatures that can be extracted by any reasonably good object segmentation methods, hence allowing multiple complementary channels of saliency maps to be estimated. To the best of our knowledge, this is the first to use complementary object signatures from different video-based object segmentation algorithms to guide saliency computation in videos. Fig. 1 outlines the general framework schematically.

Two object signatures, which are captured by two video object segmentation methods, are regarded in this work: we select the motion-dominated FOS method and the appearance-dominated B-RPCA technique as the baselines. We further improve these baselines (elaborated below) to enhance their usefulness for our purpose.

3.1. Object signature estimation with IFOS

Using the method of [11], objects can be efficiently and automatically segmented in a video. This method includes two main phases: initial foreground estimation and foreground-background labelling refinement. The first phase is the key contribution, where pixels inside an object are roughly identified according to motion boundaries. In other words, the performance of the motion boundary detection determines the success of the FOS method. We follow largely the FOS approach and detail our implementation and improvements below.

Optical flow is the apparent motion of brightness patterns in a video [13,25], supplies the motion information of each pixel [14]. We compute optical flow for each successive pair of frames based on the method of [15], which is able to handle large displacements and has a fast implementation. We then estimate the motion boundaries based on the calculated optical flow. Two simple features, i.e., the magnitude of the gradient of the optical flow (Eq. (1)) and the difference in direction (Eq. (2)), are applied in [11] to estimate the motion boundaries.

$$B_i^m = 1 - \exp(-\lambda^m \|\nabla w_i\|) \quad (1)$$

where i denotes a pixel position in the image domain. $\mathbf{w} = (u, v)$ is the flow field, while u and v are the flow component in x- and y-direction. $B_i^m \in [0, 1]$ is the strength of the motion boundary at pixel i, λ^m is a scaling parameter used to control the steepness of the exponential function.

$$B_i^\theta = 1 - \exp(-\lambda^\theta \max_{j \in N}(\delta\theta_{i,j}^2)) \quad (2)$$

$B_i^\theta \in [0, 1]$, $\delta\theta_{i,j}$ represents the difference in direction between $\mathbf{w}_i$ and its neighbors $\mathbf{w}_j$. Then, these two measures are combined to construct a more reliable measure:

$$B_i = \begin{cases} B_i^m, & \text{if} \quad B_i^m > T \\ B_i^m \cdot B_i^\theta, & \text{if} \quad B_i^m \leq T \end{cases} \quad (3)$$

where T is a threshold. Finally, an empirical value $T = 0.5$ is applied to B_i in [11] to get a binary motion boundary.

Learning-based motion boundary detection (LMBD). To improve the performance of FOS, we incorporate the LMBD method [16] to estimate the motion boundaries in the first phase. Since the structured random forest (SRF) [26] leverages several cues – appearance, motion, and confidence in motion at the patch level, the LMBD is more robust. Specifically, the SRF predicts boundaries at the patch level, making the LMBD is robust to failures in the optical flow. Besides, SRF learns the correlation between local features and motion boundaries, and thus textures in the background and boundaries can be distinguished. Since the feature representation at the patch-level includes several cues, the LMBD is able to detect edges of the object in static frames.

Improving LMBD. The success of LMBD depends on the selection and design of features. We further improve its performance by using a new feature: the motion difference in direction (see Eq. (2)). We use this feature as a new channel to construct a more effective feature representation. The motion direction difference is helpful for detecting boundaries of moving object, as the foreground object and the background typically go in different directions. The comparisons between our improved LMBD and the original LMBD are displayed in Figs. 2 and3. As shown in Fig. 2, the motion boundaries of the girl can be completely captured by our method, while many parts are missed in LMBD. The ROC curves in Fig. 3 also demonstrate our improved LMBD due to the motion difference in direction is effective. Fig. 4 (b) (top) shows the object segmentation results of our IFOS.

3.2. Object signature estimation with IB-RPCA

Gao et al. [12] proposed a B-RPCA method, which imposes few specific assumptions to the background and only supposes that its appearance variation is highly constrained. The background can be extracted according to a low-rank matrix. Mathematically, they consider the observed video frames as a matrix $\boldsymbol{M}$, which is a sum of a low-rank matrix $\boldsymbol{L}$ that denotes the background, and a sparse outlier matrix $\boldsymbol{S}$ that consists of the moving objects. In this way, even the videos contain diverse challenges, the foreground moving objects can be accurately detected by solving the decomposition according to the RPCA technique [27]. The B-RPCA method has three main steps: (1) First-pass RPCA; (2) Motion Saliency Estimation (MSE); (3) Second-pass RPCA. We modify the second step

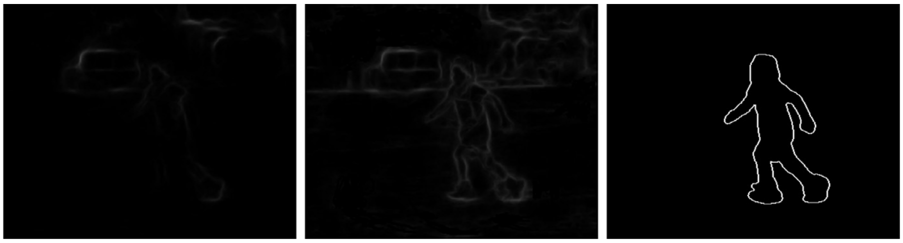

Fig. 2. Results of LMBD [16] (left) and our method (middle) of sequence *girl* on the SegTrack v2 dataset.

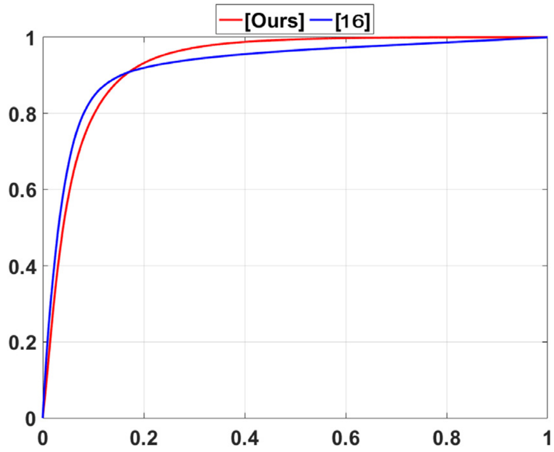

Fig. 3. Comparison of ROC curves on the SegTrack v2 dataset.

to improve the performance of the B-RPCA method in respects of efficiency and accuracy as detailed below.

Drawbacks: The MSE operation is effective to filter off or suppress the non-stationary background motions and identify the foreground object that keeps moving spatiotemporally constant. However, if the object incidentally stops or moves approximate to zero, or if some video frames are static, all the foreground objects will be removed by the second step. Besides, the MSE is low efficiency, especially when the background contains some non-stationary motions, the computational time will increase exponentially.

Improving MSE. To handle these two issues caused by MSE, we improve the MSE step in the following three aspects similar to Tu et al. [28]:

(1) We decrease the constraint of the trajectory length to 5 frames to reduce the time complexity.

(2) To suppress the wrongly identified non-stationary motions, we add a velocity angle constraint to the motion direction consistency measure. Not only the negative or positive direction of the flow components u and v along the trajectory is concerned, but also the variation in direction is considered. The velocity angle measure is expressed as:

$$\triangle\theta = \arctan(u_{t+1}/v_{t+1}) - \arctan(u_t/v_t) \in [-\pi/4, \pi/4] \quad (4)$$

where $(u_t,\ v_t) \neq 0$ denotes the optical flow of the current time t. Same as the operation of motion direction consistency, this measure is conducted at places where the velocity is no-zero along the trajectory (refer to the MSE conduction for detail in [12]).

(3) We explicitly consider the condition that objects stop or slowly move. In this case, we do not perform the second step and only execute the first step. To further remove the wrongly detected foreground objects, we discard the small size of motion coherent blocks. In general, salient foreground objects has a non-trivial size. If the size of one block smaller than τ (We set $\tau = 10 \times 10$ experimentally in this paper), it can be regarded as the background. Fig. 4 (b) (bottom) displays the object detection results of our IB-RPCA method.

4. Saliency estimation

After obtaining two types of foreground object signatures, we now apply the foreground connectivity method [10] to each channel to compute two saliency maps. This is motivated by the good performance of this method, which considered different saliency cues including contrast prior, boundary prior, and foreground prior. Its most important foreground prior can be effectively computed according to our detected foreground candidates. Besides using the presumably better foreground prior, we further improve the method of [10] in the following two aspects.

4.1. Labeling foreground superpixels

In one previous work, Srivatsa and Babu [10] used objectness proposals which are identified according to BING [29] to compute an objectness map. Then, a rough estimate of foreground is obtained by thresholding the objectness map, and superpixels that are part of the foreground are accordingly captured. Since the extracted foreground in [10] is rough, the captured foreground superpixels are not precise. Our detected foreground objects are much more accurate, and thus in each frame, to identify whether a su-

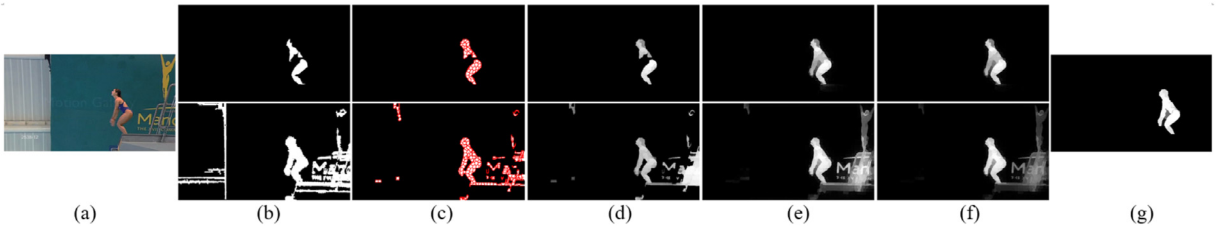

Fig. 4. An illustration of our saliency estimation on the UCF-Sports dataset [31]. (a) One frame of the input video. (b) Detected foreground object signatures by methods IFOS (top) and IB-RPCA (bottom). (c) Identified foreground superpixels. (d) Computed foreground weights. (e) Saliency estimation based on detected object signatures and computed foreground weights. (f) Saliency maps refinement. (g) A spatiotemporal saliency map by the proposed fusion strategy.

perpixel belongs to the foreground, we present a measure as:

$$FP_n \leftarrow oP_n > \eta \cdot nP_n \tag{5}$$

where $P_n \in P$ denotes a superpixel. nP_n represents the number of pixels belong to a superpixel region R. oP_n denotes the number of overlapped pixels between the superpixel region R of P_n and our detected foreground objects. η is a ratio parameter $\in$ [0, 1]. $FP_n \in FP$ represents a superpixel which is identified as belong to the foreground. Generally, if more than half region of a superpixel locates on our detected foregrounds, we will label this superpixel as belonging to the foreground, i.e., FP_n. In the paper, we set $\eta = 0.55$. The SLIC algorithm [30] is utilized to abstract each video frame into superpixels.

4.2. Foreground connectivity

In this step, a robust saliency measure called foreground connectivity is employed to assign saliency values based on superpixel connectivity to the identified foreground. An undirected weighted graph is constructed by using superpixels as nodes. All adjacent superpixels (P_n, P_m) in the image are connected and the weight of $d(P_n, P_m)$ is set as the Euclidean distance between their mean CIE-Lab values. The geodesic distance between any two superpixels $d_{geo}(P_n, P_m)$ is calculated as the accumulated edge weights along their shortest distance on the graph [17]:

$$d_{geo}(P_n, P_m) = \min_{P_1=P_n, P_2, \ldots, P_l=P_m} \sum_{i=1}^{l-1} d(P_i, P_{i+1}) \tag{6}$$

where l is the number of superpixels along one path of two superpixels (P_n, P_m). The foreground connectivity of a superpixel P_n is defined:

$$F_{GC}(P_n) = \frac{\sum_{k=1}^{N} d_{geo}(P_n, P_k) \cdot \delta(P_k)}{\sum_{k=1}^{N} d_{geo}(P_n, P_k) \cdot (1 - \delta(P_k))} \tag{7}$$

where $\delta(\cdot)$ is 1 if a superpixel is identified as foreground superpixel by Eq. (5), and N is the total number of superpixels. Same as [10], we also take the reciprocal of F_{GC} and apply it as the foreground weights w^{fg}:

$$w^{fg}(P_n) = 1/F_{GC} \tag{8}$$

Eq. (8) computes foreground weights for all superpixels. In general, the foreground weight of a superpixel should be assigned to zero if it is not detected as foreground:

$$w^{fg}(P_n) = 0, \forall P_n \notin FP \tag{9}$$

We find that a superpixel who is not detected as the foreground according to Eq. (5) but with high w^{fg} value, is also located on the foreground. This is due to the detected foreground objects by IFOS or IB-RPCA are not perfect. Some foreground objects or some parts of an object are not extracted. We reset a superpixel to the foreground if it satisfies:

$$F_2(P_n) \leftarrow w^{fg}(P_n) > median(w^{fg}(FP)) \tag{10}$$

where $F_2(P_n) \in F_2P$ denotes a detected foreground superpixel. *median* is an operation, which computes the median value of the foreground weights of the captured foreground superpixels according to Eq. (9).

We find that if a superpixel has a very low w^{fg} value, even if it is identified as a foreground superpixel based on Eq. (5), it should be set to zero. This is because some noises in the background are wrongly detected as foreground objects in IFOS and IB-RPCA. We detect the background superpixels roughly based on w^{fg}, which is expressed as:

$$RB(P_n) \leftarrow w^{fg}(P_n) < median(w^{fg}(P)) \tag{11}$$

where $RB(P_n) \in RBP$ denotes a superpixel roughly identified as belonging to the background by a small threshold.

We detect the incorrectly captured foreground superpixels in Eq. (5) as following:

$$B(P_n) \leftarrow (P_n \in FP) \wedge (P_n \in RBP) \tag{12}$$

where $B(P_n) \in BP$ denotes an identified background superpixel. Then, we update foreground superpixels:

$$FP' \leftarrow (FP \vee F_2P) - BP \tag{13}$$

Fig. 4 (c) shows the identified foreground superpixels according to our method. Instead of the general method Eq. (9), the new valid foreground weights are calculated as (see Fig. 4 (d)):

$$w^{fg}(P_n) = 0, \forall P_n \notin FP' \tag{14}$$

4.3. Saliency optimization

At last, we adopt the saliency optimization framework of [10], which integrates our foreground weights with background measure of [17], to estimate the accurate and smooth final saliency maps. Accordingly, a motion-dominated saliency maps from IFOS (we call it IFOS saliency maps S_M) and an appearance-dominated saliency maps from IB-RPCA (we call it IB-RPCA saliency maps S_A) are obtained (see Fig. 4 (e)).

4.4. Saliency maps fusion

The estimated two types of saliency maps complement each other. However, simply combining different saliency maps, e.g., taking the product [9] or average [24], does not necessarily produce a better map, unless we can incorporate them in a proper way. In this section, a fusion approach is proposed which is implemented in three steps:

4.4.1. Segmenting the object regions

We use adaptive thresholding to roughly extract two types of foreground object regions in each frame in S_M and S_A, respectively:

$$\begin{aligned} FG_m &\leftarrow S_M > graythresh(S_M) \\ FG_a &\leftarrow S_A > graythresh(S_A) \end{aligned} \tag{15}$$

where *graythresh* is the Matlab built-in function according to [18].

4.4.2. Saliency maps refinement

After obtained the foreground regions of IFOS and IB-RPCA, we look for the overlapped regions between them and utilize the intersection-over-union (IOU) score to refine the IFOS saliency maps and the IB-RPCA saliency maps. Based on FG_m and FG_a, we can extract box regions in each of them, i.e., $mB = \{mB_1, \ldots, mB_M\}$ and $aB = \{aB_1, \ldots, aB_N\}$. Then, we re-weight either the saliency map of IFOS or IB-RPCA in two steps.

For the IFOS saliency map in one frame, for one box region of IFOS $mB_m \in mB$, if its IOU score–between mB_m and aB_n is higher than a given threshold T_B, i.e., IOU(mB_m, aB_n) > T_B, we combine mB_m with aB_n by selecting the larger box region between them:

$$mB_m = \begin{cases} mB_m & \text{if} \quad size(mB_m) >= size(aB_n) \\ aB_n & \text{if} \quad size(mB_m) < size(aB_n) \end{cases} \tag{16}$$

If its IOU score less than or equal to T_B, we do not combine mB_m with aB_n. We set $T_B = 0.75$ experimentally in this work.

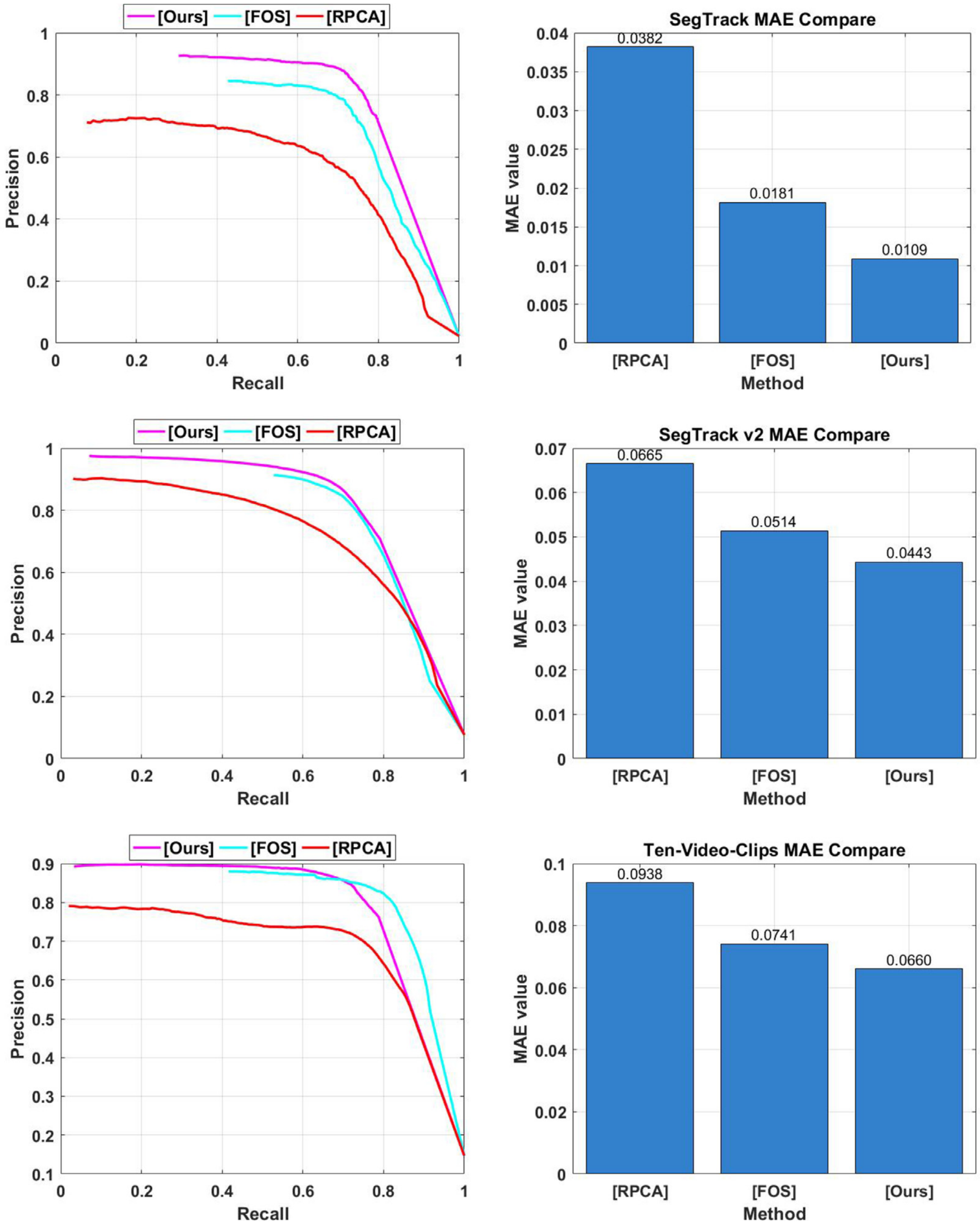

Fig. 5. Comparison of PR curves (left column) and MAE (right column) on datasets SegTrack (top row), SegTrack v2 (middle row) and Ten-Video-Clips (bottom row) of our method with the original techniques FOS [11] and B-RPCA [12].

For the IB-RPCA saliency map in one frame, for one box region of IB-RPCA $aB_n \in aB$, if its IOU score is higher than T_B, we select the larger region between aB_n and mB_m in the same way as Eq. (16). If its IOU score is in the range of (0, T_B], we select the smaller box region between aB_n and mB_m:

$$aB_n = \begin{cases} aB_n & \text{if} \quad size(aB_n) <= size(mB_m) \\ mB_m & \text{if} \quad size(aB_n) > size(mB_m) \end{cases} \tag{17}$$

We increase the saliency values in the refined box regions of saliency maps S_M and S_A as follows:

$$\begin{aligned} S_M(i) &= 2 \cdot S_M(i) \\ S_A(i) &= 2 \cdot S_A(i) \end{aligned} \tag{18}$$

where pixel $i \in mB_m$ on the up and $i \in aB_n$ in the bottom in Eq. (18).

The two refined saliency maps are normalized linearly to the range between 0 and 1. In this way, the object regions that are detected in both IFOS and IB-RPCA are enhanced (see Fig. 4 (f)).

4.4.3. Fusing complementary saliency maps

We threshold the two refined saliency maps adaptively as above again to find the foreground regions of IFOS maps (i.e., FG_m) and IB-RPCA maps (i.e., FG_a). Firstly, for the coherent box regions in each frame, if anyone of its IOU score is higher than T_B, we select the corresponding larger region. The complete foreground region can be labeled as the combination of all the refined separate regions:

$$LB = B_{1L} \cup \ldots B_{kL} \cup \ldots B_{KL}, \; (K \leq \min(M, N)) \tag{19}$$

$B_{kL} = max(aB_n, mB_m)$. Secondly, we find other foreground pixels that are overlapped in other regions where the IOU scores are

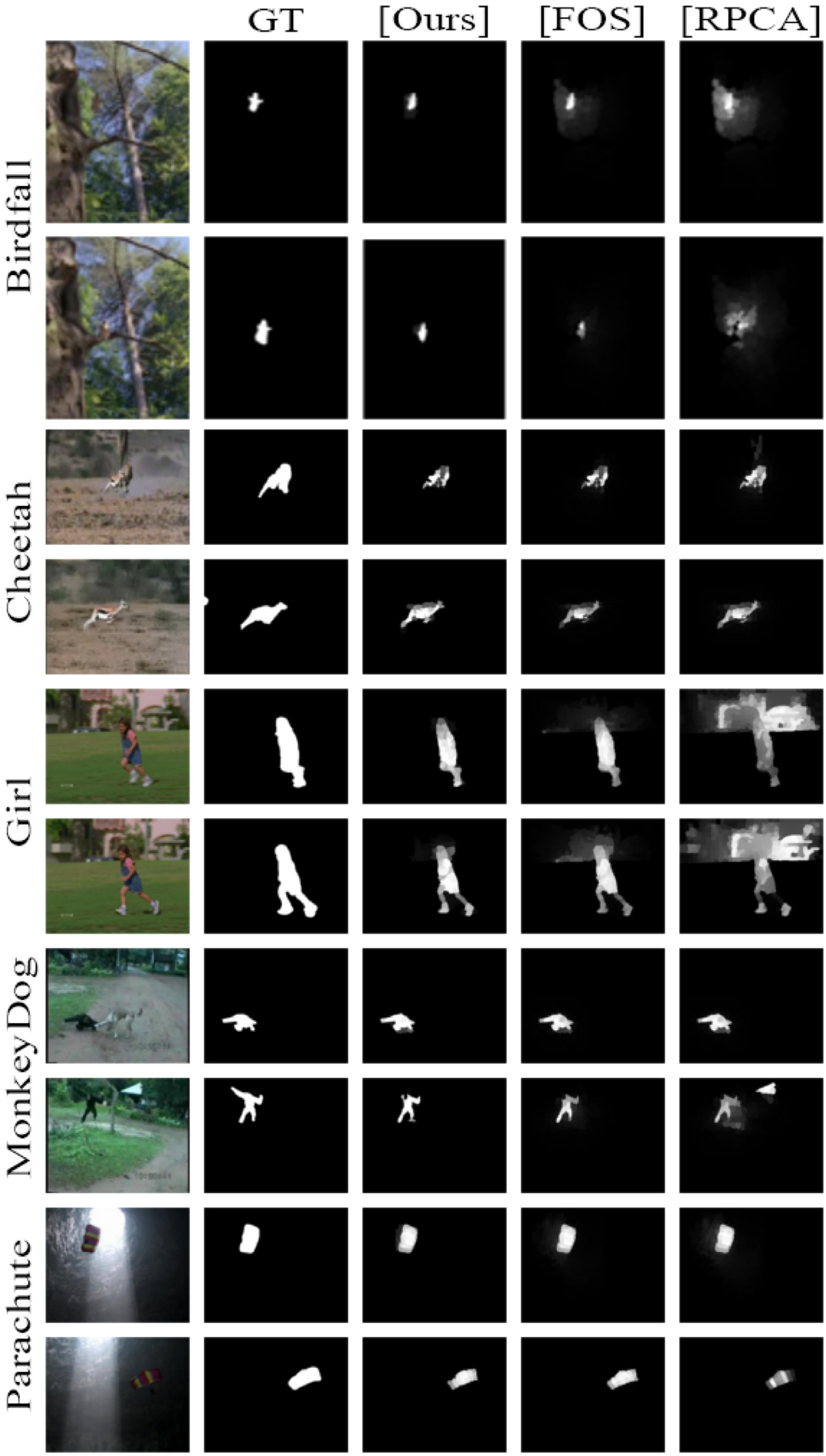

Fig. 6. Video saliency maps comparison between our method and the FOS, B-RPCA approaches on the SegTrack dataset.

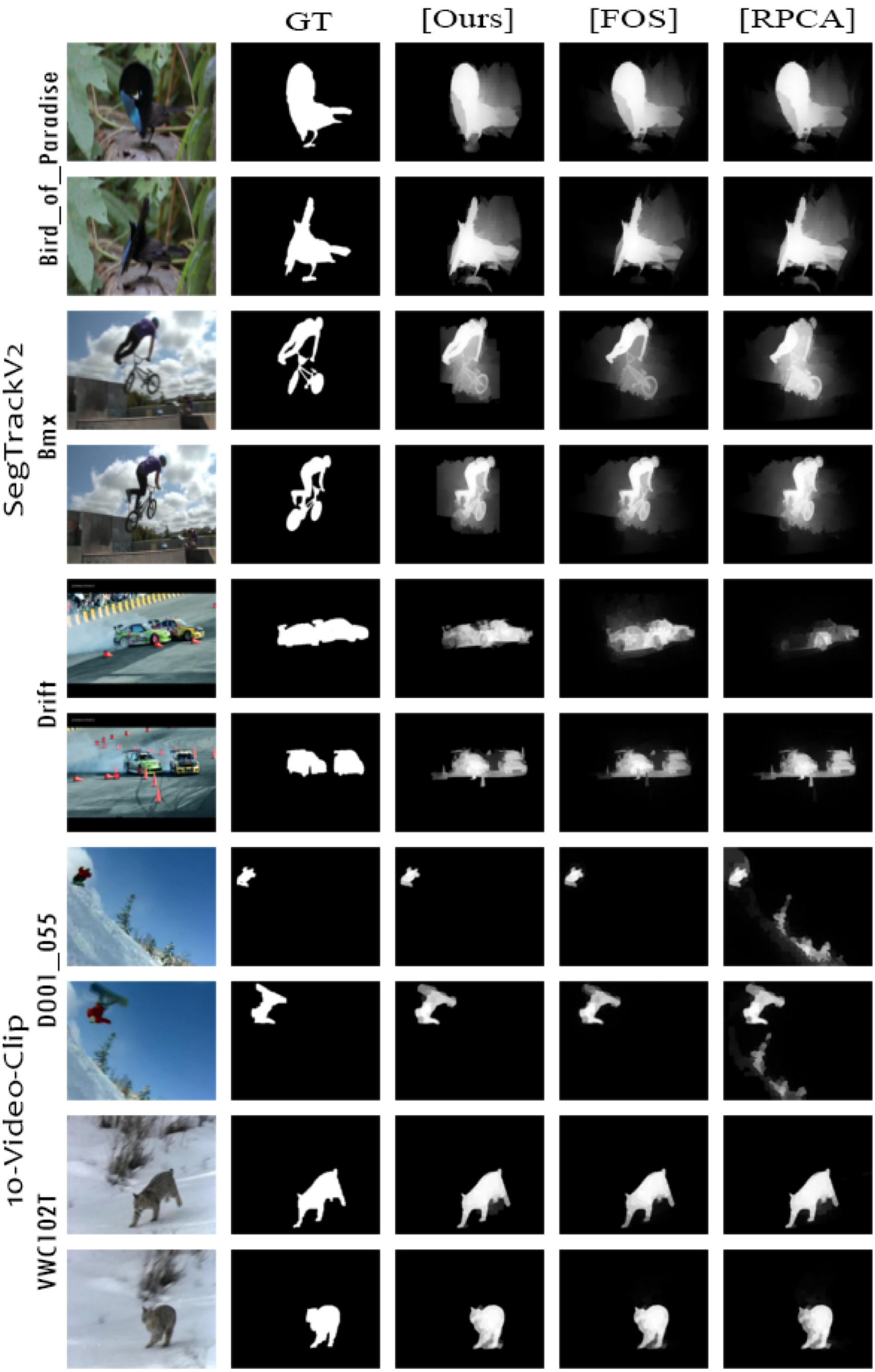

Fig. 7. Video saliency maps comparison between our method and the FOS, B-RPCA approaches on the SegTrack v2 and Ten-Video-Clips datasets.

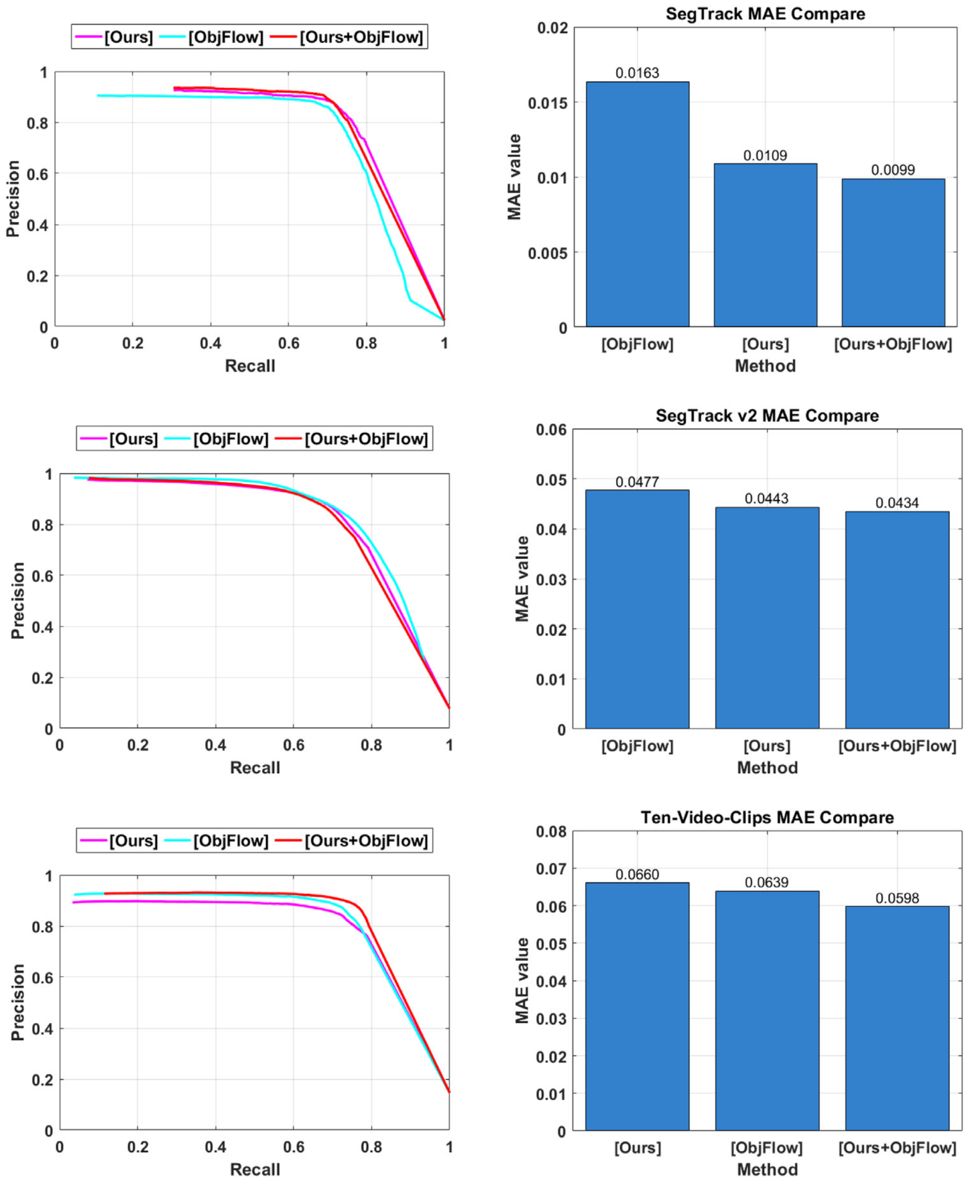

Fig. 8. Comparison of PR curves (left column) and MAE (right column) on datasets SegTrack (top row), SegTrack v2 (middle row) and Ten-Video-Clips (bottom row) of our method with the state-of-the-art deep learning method ObjFlow [36].

smaller than T_B:

$$LP \leftarrow (FG_m \cdot FG_a) > 0 \tag{20}$$

The final foreground pixels are labelled as:

$$FG \leftarrow LB \cup LP \tag{21}$$

We fuse the motion-dominated saliency map S_M and the appearance-dominated saliency map S_A frame by frame to obtain a spatiotemporal saliency maps of a video (see Fig. 4 (g)):

$$STSacy(i) = \begin{cases} S_M(i) & \text{if} \quad i \in FG\&S_M(i) > \Gamma \\ \max(S_M(i), S_A(i)) & \text{if} \quad i \in FG\&S_M(i) \le \Gamma \\ S_M(i) \cdot S_A(i) & \text{if} \quad i \notin FG \end{cases} \tag{22}$$

We set $\Gamma = 0.8$ to select the good quality saliency map $S_M(i)$ same as [6]. High quality motion saliency features are more reliable than appearance saliency features in a video since it is more robust to cluttered background.

5. Experiments

To evaluate the performance of the proposed method in video saliency detection, we now report comparisons from three aspects in term of commonly-used evaluation metrics on standard publicly video datasets.

Datasets: Four datasets are utilized for experimenting: SegTrack [32], SegTrack v2 [33], Ten-Video-Clips [34], and UCF-Sport [31]. The SegTrack dataset was initially produced to evaluate tracking methods, and then it is widely used for video segmentation as well as video saliency detection. The videos contain diverse challenges, such as similar color between objects and background, non-rigid deformations, and fast camera motion. Li et al. [33] introduced more sequences to construct SegTrack v2. The Ten-Video-Clips dataset includes 10 short video clips of 5–10 s each, and one video clip focuses on one primary object in the natural scene. The UCF-Sports contains 150 realistic videos of sports broadcasts that are captured in dynamic and cluttered environments. There

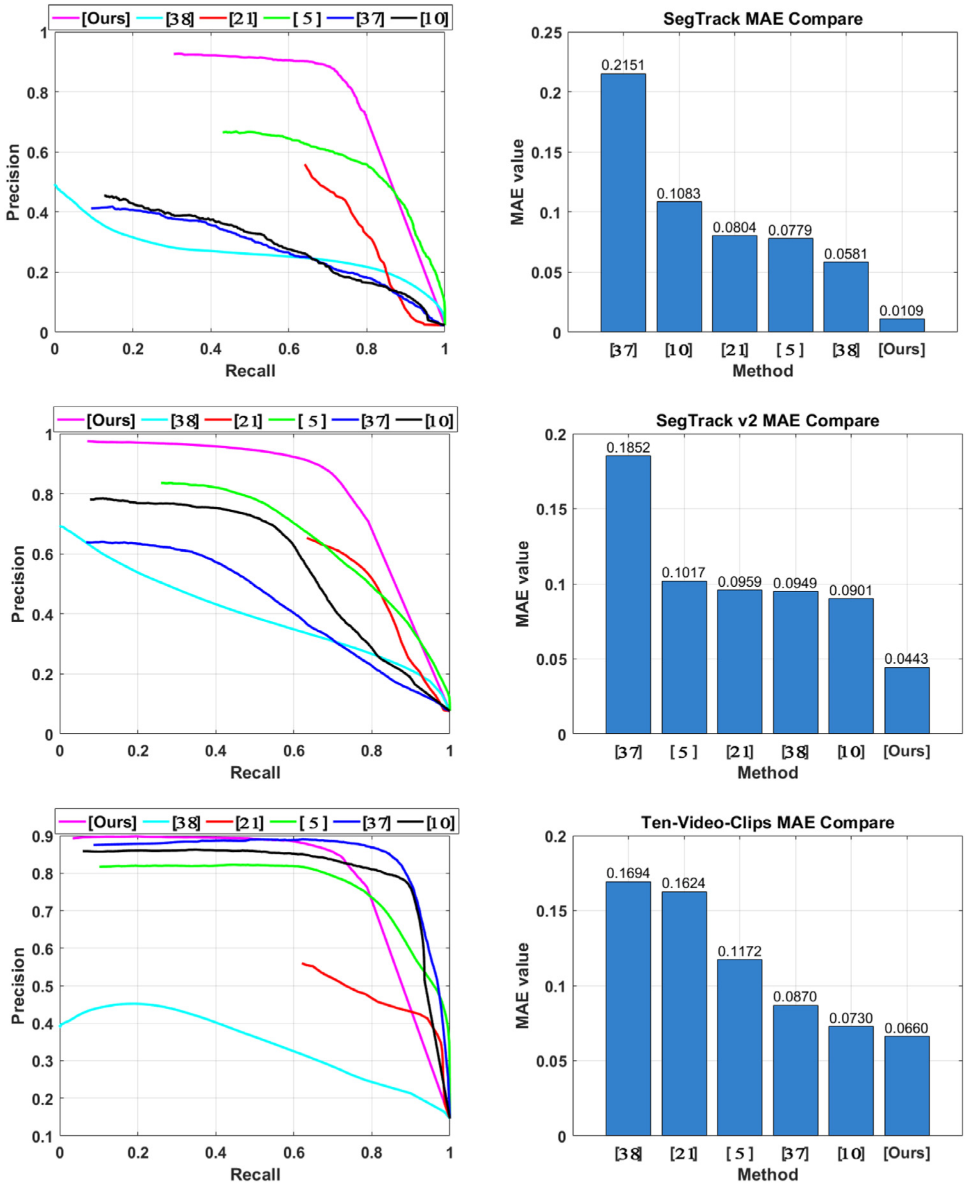

Fig. 9. Comparison of PR curves (left column) and MAE (right column) on datasets SegTrack (top row), SegTrack v2 (middle row) and Ten-Video-Clips (bottom row).

are 10 action classes, and each video corresponds to one action performed by one or several people, which are the salient objects.

Evaluation metrics: For performance testing, we first use the precision-recall curves (PR curves). A curve is produced by normalizing the saliency map in the range of [0, 255], producing binary masks with a threshold varies from 0 to 255, and comparing the quality of different binary masks against the ground truth (GT). The curves are then averaged on each video sequence. Precision defines as the ratio of salient pixels assigned correctly, while recall denotes the ratio of salient pixel detected. PR curves concern only the case where the object saliency is higher than the background saliency, and pixels incorrectly assigned as salient degrades the performance. Therefore, we further introduce the mean absolute error (MAE) measure [35]. The MAE computes the average per-pixel difference between the saliency map and the GT, which is normalized to [0, 1]. It gives a better estimation of how close a saliency map is to the GT. To evaluate the performance of our method in video object segmentation, we utilize the average per-frame pixel error rate [32] for evaluation. Which represents the number of mislabelled pixels according to the ground truth segmentation.

5.1. Evaluation of fusing different saliency maps from disparate object signatures

To test the effectiveness of our idea – integrating disparate saliency maps which are derived from complementary object signatures can get a refined spatiotemporal saliency maps, we compare the results of our method with the motion-dominated FOS and the appearance-dominated B-RPCA which we used in this work. In addition, we introduce another state-of-the-art deep learning based video object detection method–ObjFlow [36] to evaluate that incorporating other saliency maps computed from other object signatures, whether the proposed fusion strategy is available and the performance of salient object detection can be further improved?

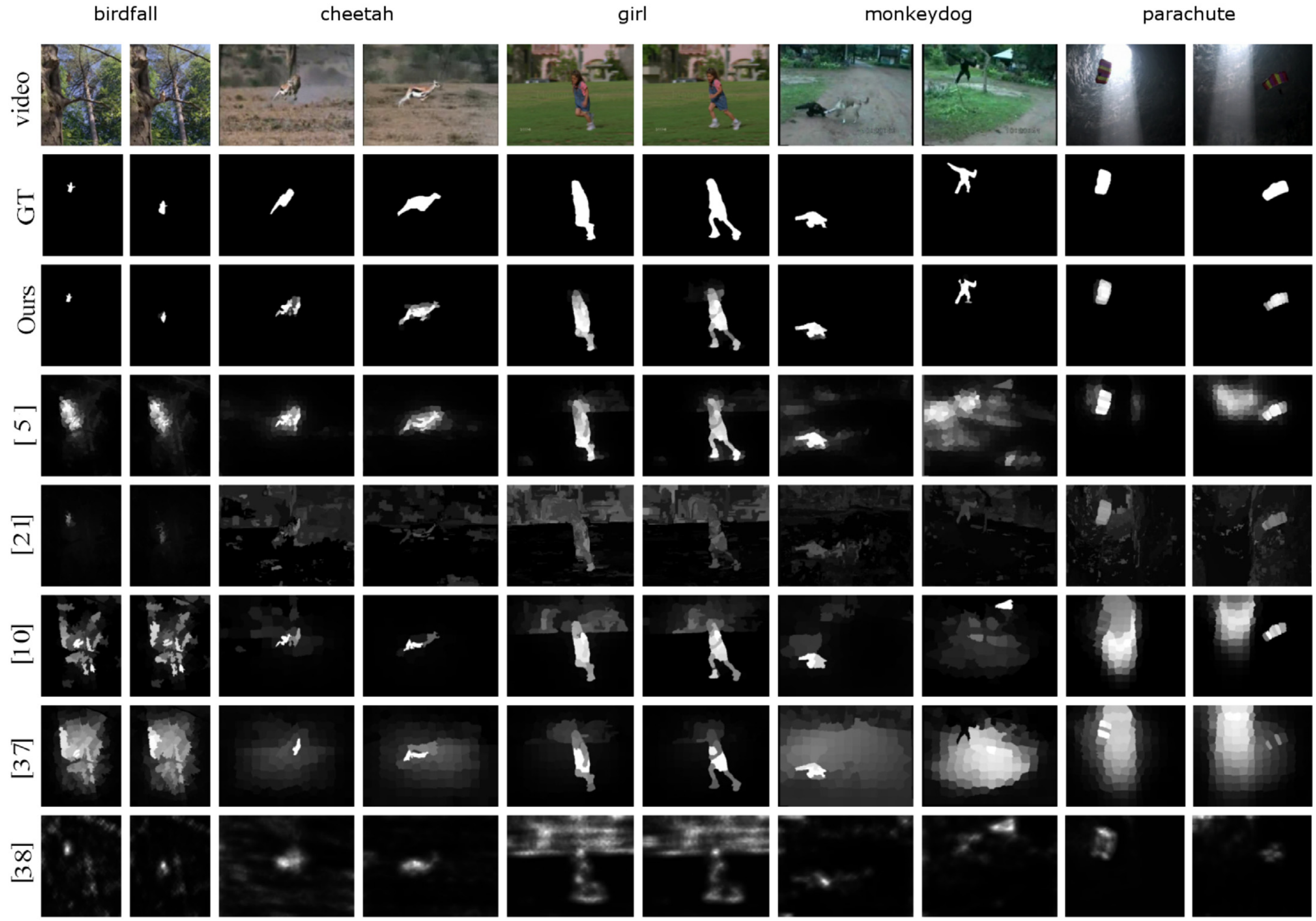

Fig. 10. Video saliency maps comparison between our method and the state-of-the-art works on the SegTrack dataset.

5.1.1. Evaluation of fusing saliency maps from motion-dominated and appearance-dominated object signatures

In this experiment, the estimated saliency results of FOS [11], B-RPCA [12], our modified IFOS and IB-RPCA, as well as the proposed fusion method are compared quantitatively and visually.

Quantitative evaluation: As shown in Fig. 5, both the PR cures and the MAE values demonstrate that combining disparate saliency maps properly is valid, and our fusion scheme is effective. Besides, from Fig. 5, we can find that our exploited IFOS and IB-RPCA outperform the original FOS and B-RPCA. Different object detection methods can tackle different challenges, no approach can handle all the problems in the object detection domain till now. (1) in the SegTrack dataset, our saliency based fusion results are much better than the classical FOS and B-RPCA methods as well as the modified IFOS and IB-RPCA approaches. For example, the MAE value of us is significantly boosted (0.0109 (our) vs. 0.0382 (B-RPCA), 0.0356(IB-RPCA), 0.0181 (FOS) and 0.0170 (IFOS)), where about 78% more accurate than B-RPCA, 69% more accurate than IB-RPCA, 40% more accurate than FOS, and 36% more accurate than IFOS. (2) in the SegTrack v2 and Ten-Video-Clips datasets, our method outperforms the four approaches not as high as in the SegTrack dataset, this may because these two datasets contain more challenges, the performance of B-RPCA is not good at some cases which badly affects the fused results. Therefore, finding more effective object segmentation methods and fusion strategies are the future tasks.

Visual evaluation: Fig. 6 shows the video saliency maps of FOS, B-RPCA and our methods on the SegTrack dataset. Our methods performs best among them. The shapes of the foreground objects of us are more approximate to the ground truth than other methods. This can be also demonstrated in Fig. 7, where the results are computed on the SegTrack v2 and Ten-Video-Clips datasets. In particular, for the birdfall video in Fig. 6, the small bird can be accurately identified in our method. In other two methods, however, the bird is heavily violated by the background, it is hard to be distinguished. For the B-RPCA method (the fourth column), the bird is totally confused with the background. For the bird_of_paradise, bmx, and drift videos in Fig. 7, the background noise is higherly removed in our method when compared to FOS and B-RPCA. FOS can locate the foreground objects in video DO01_055 and VWC102T well in the Ten-Video-Clips dataset, but for some parts of the object, like the slider, it is poorly discovered. In contrast, due to the appropriately fusion, our method is able to highlight these parts better.

5.1.2. Evaluation of fusing saliency maps computed from other complementary object signatures

In this experiment, we integrate the object signatures which are estimated by the last deep learning-related video segmentation technique into our framework. The object flow (ObjFlow) method [36] is introduced for testing. As shown in Fig. 8, when fusing the video saliency maps of ObjFlow with our video saliency maps (Ours+ObjFlow), better results are obtained. For example, in the SegTrack dataset, the ObjFlow method performs worse than our method (Ours), when combing the result of ObjFlow with our result, the performance of the fused Ours+ObjFlow can also

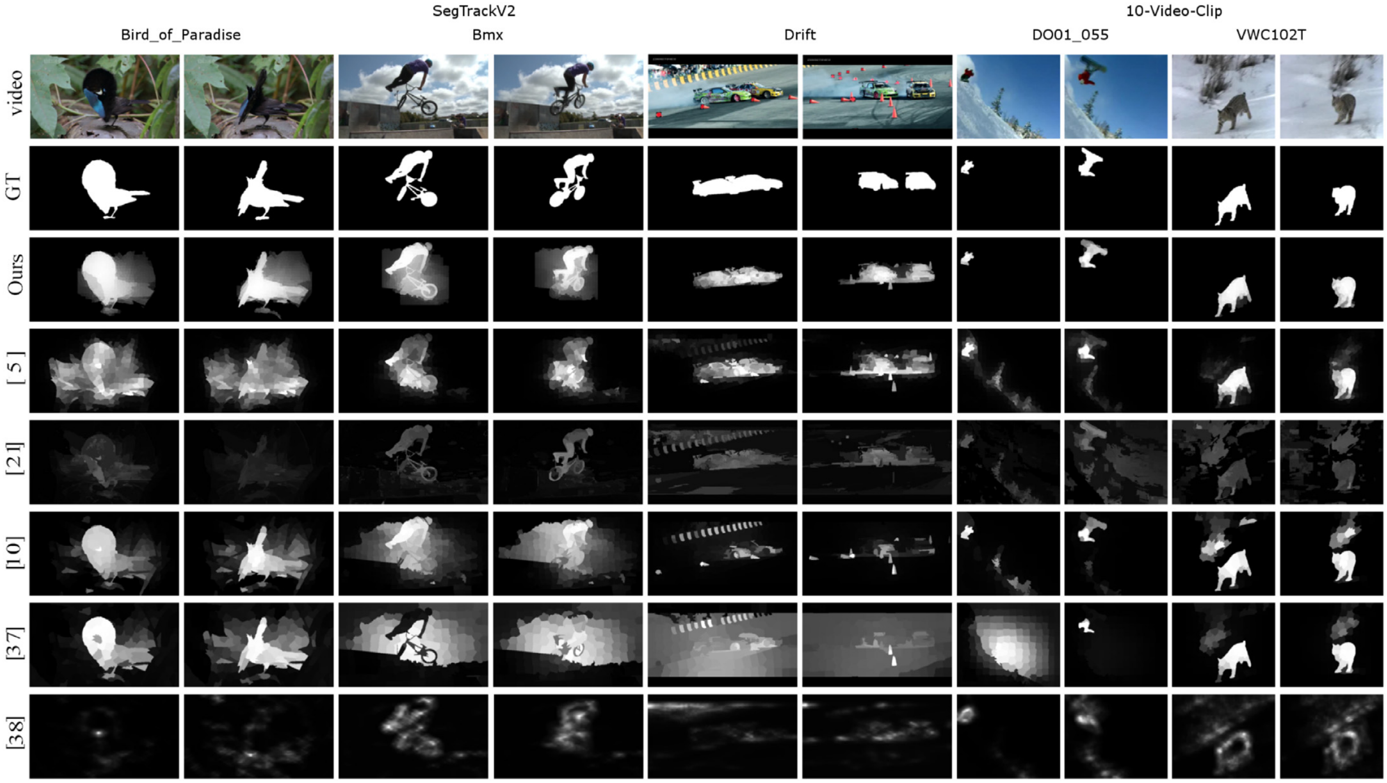

Fig. 11. Video saliency maps comparison between our method and the state-of-the-art works on the SegTrack v2 and Ten-Video-Clips datasets.

been improved. The accuracy enhancement of MAE of the fused Ours+ObjFlow method is 9.17% (vs. Ours) and 39.26% (vs. ObjFlow). In the SegTrack v2, the accuracy of MAE is boosted 2.03% (vs. Ours) and 9.01% (vs. ObjFlow). In the Ten-Video-Clips dataset, the ObjFlow method performs better than our method, the accuracy gain of MAE of the fused is Ours+ObjFlow method 9.39% (vs. Ours) and 6.42% (vs. ObjFlow). Due to the ObjFlow method is inefficient, we do not integrate it into our object signatures based saliency fusion model.

5.2. Comparison with other methods

We compare our spatiotemporal saliency results with five state-of-the-art image and video saliency methods [5,10,21,37,38]. The first two methods [10,37] focuses on image saliency detection. Specially, [10] is closely related to our approach as we apply its foreground connectivity measure to assign weights to foreground superpixels. The other methods aim at video saliency detection. All the saliency maps are computed by directly running the source code supplied by the authors. The source code of our work is available online at.[1]

Quantitative evaluation: Fig. 9 gives quantitative comparisons between our method and 5 competitive algorithms on 3 well-known datasets. It demonstrates that the proposed method outperforms the others. The PR curves show our method is able to highlight the complete salient objects more effectively and preserve salient object boundaries more precisely in most of the cases. The precision rates of ours are the highest on both SegTrack and SegTrack v2 datasets. Specially, in SegTrack v2, our result reaches to above 0.95. However, on the Ten-Video-Clips dataset, the PR curve of us is approximate to [10] and worse than [37]. This is due to the IB-RPCA method is able to detect the foreground objects completely, but it also falsely captures some background noises (see Fig. 4 (b)). Therefore, some background noises are wrongly treated as foreground proposals. On the other hand, the MAE values of our approach are the lowest in all the three tests, which indicates our saliency maps are closer to the GT. In SegTrack, comparing to the best result [38] of other methods, the MAE of ours is more than 80% more accurate.

Visual evaluation: Fig. 10 shows the estimated saliency maps on SegTrack dataset. The brighter pixels represent higher saliency probabilities. It can be observed that the proposed method can not only detect the foreground salient objects with well-preserved boundaries but also suppress the background regions much better than others. The image saliency method [10], to which our method is most similar, performs poorly for these videos. Especially, in birdfall (fast object motion) and monkeydog (camera motion), both [10] and [37] cannot capture the foreground object correctly. One main reason is that they lack the motion cue. The spatiotemporal model of [5] performs best among others, but its performance degrades when encountering complex conditions, e.g., the birdfall (small object with fast motion), parachute (illumination changes). In contrast, the IFOS and the IB-RPCA approaches which we used to detect object candidates can deal with these challenges, making our method much more robust in extracting salient objects in diverse video contents spatiotemporally.

Fig. 11 shows the results on SegTrack v2 and Ten-Video-Clips datasets, and Fig. 12 shows the results on the UCF-Sports dataset. Again, compared with other methods, our spatiotemporal model can highlight the salient objects more accurately with complex visual scenes. The shape of the foreground objects in our method is better defined. Moreover, the background is better separated in our results while most other methods incorrectly detect part of the background as being salient.

[1] https://github.com/ZhigangTU/Spatiotemporal-Salient-Object-Detection-in-Video

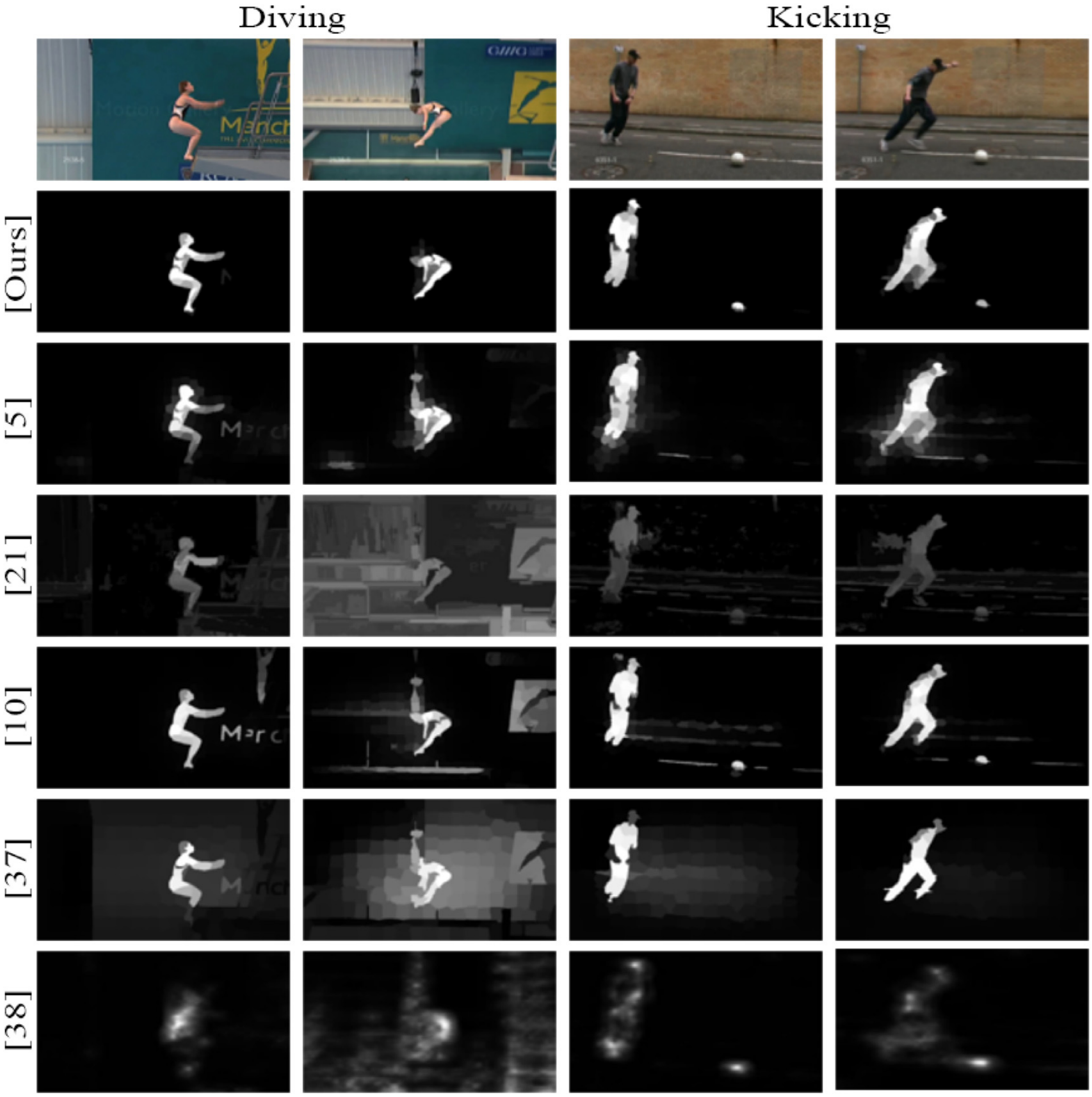

Fig. 12. Video saliency maps comparison between our method and the state-of-the-art works on the UCF-Sports dataset.

6. Conclusions

In this paper, we proposed a spatiotemporal method for salient object segmentation in videos. Based on the identified foreground object candidates that are computed with some complementary object detection algorithms, we introduce a foreground connectivity method to estimate saliency maps of each channel of the extracted objects. Then, employing our fusion strategy, a more robust and accurate spatiotemporal saliency maps can be obtained. This method brings a new research perspective to video saliency detection: instead of exploiting/improving various features directly designed for saliency detection, it is able to use object signatures from suitable object segmentation algorithms and build a saliency model on top of them. Future work includes exploiting other suitable complementary object signatures and proposing more effective fusion techniques.

Acknowledgments

This work is supported in part by Singapore Ministry of Education Academic Research Fund Tier 2 MOE2015-T2-2-114, and in part by grants from Office of Naval Research, US. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ONR. It was also supported by the National Natural Science Foundation of China (61501198), Natural Science Foundation of Hubei Province (2014CFB461), and Wuhan Youth Science and Technology Chenguang program (2014072704011248).

References

[1] M. Donoser, M. Urschler, M. Hirzer, H. Bischof, Saliency driven total variation segmentation, in: Proceedings of the International Conference on Computer Vision, 2009, pp. 817–824.
[2] Y. Fang, Z. Chen, W. Lin, C.W. Lin, Saliency detection in the compressed domain for adaptive image retargeting, IEEE Trans. Image Process. 21 (9) (2012) 3888–3901.
[3] L. Marchesotti, C. Cifarelli, G. Csurka, A framework for visual saliency detection with applications to image thumbnailing, in: Proceedings of the International Conference on Computer Vision, 2009, pp. 2232–2239.
[4] 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.
[5] W. Wang, J. Shen, F. Porikli, Saliency-aware geodesic video object segmentation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2015, pp. 3395–3402.
[6] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, J. Brandt, Discovering primary objects in videos by saliency fusion and iterative appearance estimation, IEEE Trans. Cir. Syst. Video Technol. 26 (6) (2016) 1070–1083.

[7] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.

[8] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (4) (1985) 219–227.

[9] R. Margolin, A. Tal, L.Z. Manor, What makes a patch distinct? in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2013, pp. 1139–1146.

[10] R. Srivatsa, R. Babu, Salient object detection via objectness measure, in: Proceedings of the International Conference Image Process., 2015.

[11] A. Papazoglou, V. Ferrari, Fast object segmentation in unconstrained video, in: Proceedings of the International Conference on Computer Vision, 2013, pp. 1777–1784.

[12] Z. Gao, L.F. Cheong, Y.X. Wang, Block-sparse RPCA for salient motion detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (10) (2014) 1975–1987.

[13] Z. Tu, N. Aa, C.V. Gemeren, R.C. Veltkamp, A combined post-filtering method to improve accuracy of variational optical flow estimation, Pattern Recognit. 47 (5) (2014) 1926–1940.

[14] Z. Tu, R. Poppe, R.C. Veltkamp, Weighted local intensity fusion method for variational optical flow estimation, Pattern Recognit. 50 (2016) 223–232.

[15] 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.

[16] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid, Learning to detect motion boundaries, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2015, pp. 2578–2586.

[17] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2014, pp. 2814–2821.

[18] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Sys., Man, and Cyb. 9 (1) (1979) 62–66.

[19] A. Borji, Boosting bottom-up and top-down visual features for saliency estimation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2012, pp. 438–445.

[20] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Proceedings of the International Conference on Computer Vision, 2009, pp. 2106–2113.

[21] F. Zhou, S.B. Kang, M.F. Cohen, Time-mapping using space-time saliency, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2014, pp. 3358–3365.

[22] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 29–42.

[23] L.E. Wixson, Detecting salient motion by accumulating directionally consistent flow, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 774–780.

[24] W. Kim, C. Jung, C. Kim, Spatiotemporal saliency detection and its applications in static and dynamic scenes, IEEE Trans. Cir. Syst. Video Technol. 21 (4) (2011) 446–456.

[25] Z. Tu, W. Xie, C. Gemeren, R.C. Veltkamp, Variational method for joint optical flow estimation and edge-aware image restoration, Pattern Recognit. 65 (2017) 11–25.

[26] P. Dollar, C.L. Zitnick, Structured forests for fast edge detection, in: Proceedings of the International Conference on Computer Vision, 2013, pp. 1841–1848.

[27] E. Candes, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 1–37.

[28] Z. Tu, J. Cao, Y. Li, B. Li, MSR-CNN: applying motion salient region based descriptors for action recognition, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2016, pp. 3524–3529.

[29] M. Cheng, Z. Zhang, W. Lin, P. Torr, BING: binarized normed gradients for objectness estimation at 300fps, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2014, pp. 3286–3293.

[30] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.

[31] M.D. Rodriguez, J. Ahmed, M. Shah, Action MACH a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[32] D. Tsai, M. Flagg, A. Nakazawa, J.M. Rehg, Motion coherent tracking using multi-label MRF optimization, Int. J. Comput. Vis. 100 (2) (2012) 190–202.

[33] F. Li, T. Kim, A. Humayun, D. Tsai, J. Rehg, Video segmentation by tracking many figure-ground segments, in: Proceedings of the International Conference on Computer Vision, 2013, pp. 2192–2199.

[34] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, J. Yamato, Saliency-based video segmentation with graph cuts and sequentially updated priors, in: Proceedings of the International Congress on Mathematical Education, 2009, pp. 638–641.

[35] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.

[36] Y. Tsai, M. Yang, M.J. Black, Video segmentation via object flow, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2016.

[37] C. Yang, L. Zhang, H. Lu, X. Ruan, M.H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.

[38] H.J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis. 9 (12) (2009) 1–27.

Zhigang Tu started his M.Phil. Ph.D. in image processing at the School of Electronic Information, Wuhan University, China, 2008. He received a Ph.D. degree in Communication and Information System from Wuhan University, 2013. In 2015, he received a Ph.D. degree in Computer Science from Utrecht University, Netherlands. He is currently a postdoctoral researcher at school of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include motion estimation, object segmentation, object tracking, action recognition, and human–computer interaction.

Wei Xie is an associate professor at Computer School of Central China Normal University, China. His research interests include motion estimation, superresolution reconstruction, image fusion and image enhancement. He received his B.E. degree in electronic information engineering and Ph.D. degree in communication and information system from Wuhan University, China, in 2004 and 2010, respectively. Then, from 2010 to 2013, he served as an assistant professor at Computer School of Wuhan University, China.

Remco C. Veltkamp is full professor of Multimedia at Utrecht University, Netherlands. His research interests are the analysis, recognition and retrieval of, and interaction with, music, images, and 3D objects and scenes, in particular the algorithmic and experimentation aspects. He has written over 150 refereed papers in reviewed journals and conferences, and supervised 15 Ph.D. theses. He was director of the national project GATE – Game Research for Training and Entertainment.

Baoxin Li received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2000. He is currently a full professor and chair of computer science and engineering with Arizona State University, US. From 2000 to 2004, he was a senior researcher with SHARP Laboratories of America, Camas, WA, where he was the technical Lead in developing SHARP's HiIMPACT Sports technologies. From 2003 to 2004, he was also an adjunct professor with the Portland State University, Portland, OR. He holds nine issued US patents. His current research interests include computer vision and pattern recognition, image/video processing, multimedia, medical image processing, and statistical methods in visual computing. He won the SHARP Laboratories' President Award twice, in 2001 and 2004. He also received the SHARP Laboratories' Inventor of the Year Award in 2002. He received the National Science Foundation's CAREER Award from 2008 to 2009. He is a senior member of the IEEE.

Junsong Yuan received the B.Eng. degree from the Special Class for the Gifted Young, Huazhong University of Science and Technology, the M.Eng. degree from the National University of Singapore, and the Ph.D. degree from Northwestern University. He is currently an Associate Professor and the Program Director of video analytics with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore. His research interests include computer vision, video analytics, gesture and action analysis. He serves as the Program Co-Chair of the IEEE Conference on Visual Communications and Image Processing (2015), the Organizing Co-Chair of the ACCV 2014 and the ICME (2014 and 2015), the Area-Chair of the CVPR 2017. He serves as a Guest Editor of the IJCV and an Associate Editor of IEEE T-IP, IEEE T-CSVT and The Visual Computer journal (TVC). He received Nanyang Assistant Professorship from Nanyang Technological University, Outstanding EECS Ph.D. thesis award from Northwestern University, and National Outstanding Student from Ministry of Education, PR China. He is a senior member of IEEE.