Multimed Tools Appl (2018) 77:10685–10699
DOI 10.1007/s11042-017-4822-7

Video-based salient object detection via spatio-temporal difference and coherence

Lei Huang[1] · **Bin Luo**[1]

Received: 27 February 2017 / Revised: 30 April 2017 / Accepted: 4 May 2017 /
Published online: 24 May 2017

Abstract Salient object detection aims to extract the attractive objects in images and videos. It can support various robotics tasks and multimedia applications, such as object detection, action recognition and scene analysis. However, efficient detection of salient objects in videos still faces many challenges as compared to that in still images. In this paper, we propose a novel video-based salient object detection method by exploring spatio-temporal characteristics of video content, i.e., *spatial-temporal difference* and *spatial-temporal coherence*. First, we initialize the saliency map for each keyframe by deriving spatial-temporal difference from color cue and motion cue. Next, we generate the saliency maps of other frames by propagating the saliency intra and inter frames with the constraint of spatio-temporal coherence. Finally, the saliency maps of both keyframes and non-keyframes are refined in the saliency propagation. In this way, we can detect salient objects in videos efficiently by exploring their spatio-temporal characteristics. We evaluate the proposed method on two public datasets, named *SegTrackV2* and *UVSD*. The experimental results show that our method outperforms the state-of-the-art methods when taking account of both effectiveness and efficiency.

Keywords Salient object detection · Spatio-temporal difference · Spatio-temporal coherence · Saliency propagation

✉ Bin Luo
luobin@nju.edu.cn

Lei Huang
leihuang@nju.edu.cn

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

1 Introduction

Salient object detection aims to extract the attractive objects in images and videos [5]. It can support many robotics tasks, such as object detection [12, 38], action recognition [11, 26], and scene analysis [10, 50]. It can also serve as a fundamental of various multimedia applications, including image/video classification [4, 29, 48], summarization [28, 34], quality assessment [27, 47], retrieval [8, 16, 41, 52], content-aware editing [32, 33, 39], and social media analysis [3, 42]. As compared to the prosperity of image-based salient object detection [5], efficiently detecting salient objects in videos still faces many challenges.

A primary idea for video-based salient object detection is to apply the existing image-based methods independently on each video frame and generate saliency maps frame by frame. However, such a solution suffers from several problems. First, image-based salient object detection methods only consider spatial difference among different image regions, such as color contrast, but ignore their temporal difference, which is important in video-based salient object detection by providing object motion cue. Second, serious incoherence exists among the generated saliency maps when processing each frame independently; this is because the appearances of objects and background may be quite diverse on different frames. Finally, obvious content redundancy exists in videos because successive video frames need to contain sufficiently similar content to provide smooth viewing experience. It leads to unnecessary computational cost if ignoring video content redundancy simply. Figure 1 shows an example of the difference between image-based and video-based salient object detection, in which Fig. 1b is the result generated by applying a typical image salient object method independently on each video frame and Fig. 1c is the result of our method. We can see that video-based salient object detection is superior to image-based salient object detection in emphasizing salient objects and suppressing background.

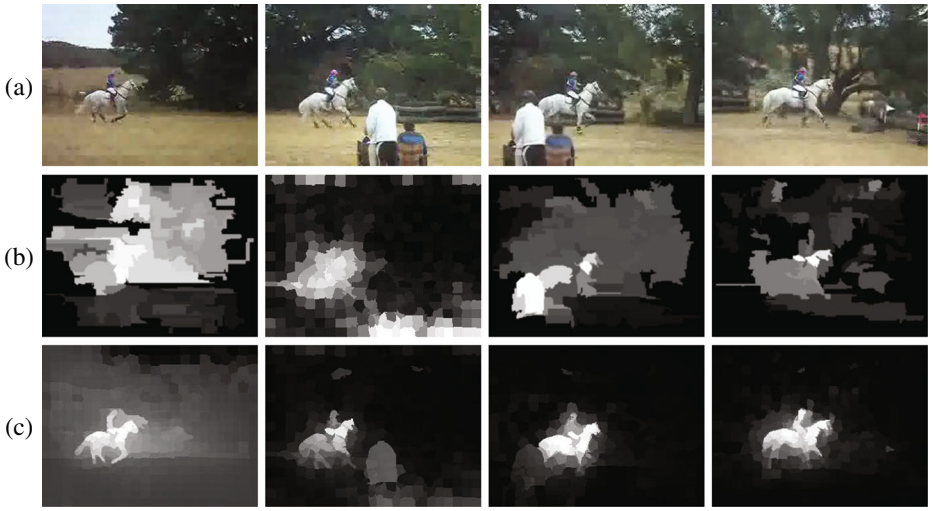

Fig. 1 An example of visual comparison between image-based and video-based salient object detection. **a** Video frames. **b** Result of image-based salient object detection by applying [7] independently on each video frame. **c** Result of video-based salient object detection by using our method

To tackle these problems, we propose a novel video-based salient object detection method, which explores the potential of both *spatio-temporal difference* and *spatio-temporal coherence* of video content. Figure 2 shows an overview of our proposed method. Based on the super-pixel representation of video frames, we first calculate the saliency values of super-pixels on the keyframes based on spatial difference, i.e., color contrast among adjacent super-pixels in the same keyframe, and temporal difference, i.e., object motion extracted from adjacent frames. Second, we construct the relationships between the super-pixels in the same frame or adjacent frames based on their color similarity and motion vector. Next, we propagate saliency values from the keyframes to non-keyframes and improve saliency coherence on all the frames with the constraint of *spatio-temporal coherence*, i.e., making the similar super-pixels in the same frame and adjacent frames have coherent saliency values. Finally, we globally normalize all the saliency maps and obtain the salient object detection result of the given video. Some preliminary results of our method were presented in [17]. In this paper, we improve the saliency propagation mechanism in our proposed method, which obtains better salient object detection performance. Moreover, we supplement the performance analysis of the key parameters in our method and provide more comprehensive evaluation on two public datasets.

Our major contributions mainly include: First, we present a novel video-based salient detection method, which can detect both static and moving salient objects effectively based on spatial-temporal difference. Second, we make use of video content redundancy to improve the efficiency of our method by combining salient object detection on keyframes and saliency propagation among frames.

2 Related work

2.1 Saliency cues

Most exiting salient object detection methods focus on dealing with various images and videos without the constraints of specific applications. Hence, they use general saliency cues, including color, depth, location and motion, rather than specific cues, such as vehicle

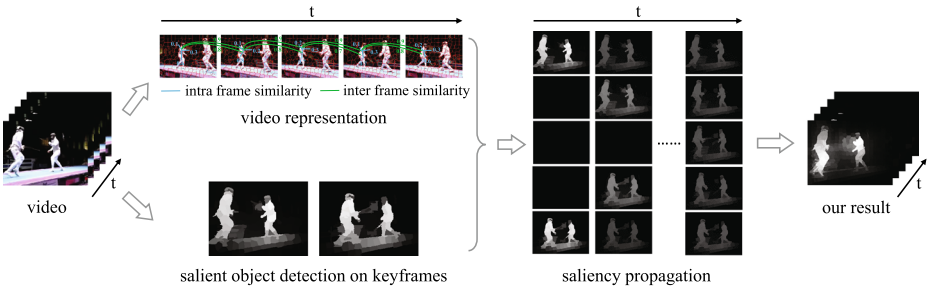

Fig. 2 An overview of our method. A given video is first represented with super-pixels and their relationships. Next, the saliency maps on keyframes are generated based on spatio-temporal difference. After this, the saliency values are propagated from the keyframes to non-keyframes with the constraint of spatio-temporal coherence, and the saliency coherence within each salient object and background on all the frames is improved during saliency propagation. Finally, the salient object detection result of the given video is generated after global normalization

detection in surveillance video analysis. Color cue, especially color contrast, is widely used in both image-based and video-based salient object detection. Color contrast can be measured globally or locally based on different features, such as average color and color histogram. Specifically, Achanta et al. [1] calculated the saliency value of each pixel according to the difference between its color and the average color of the whole image. Cheng et al. [7] decomposed a given image into regions and measured the saliency value of each image region with global color contrast weighted spatial distance.

Depth is an effective cue for salient object detection on RGB-D images and videos [9]. Depth contrast can distinguish salient objects to their surrounding background even they have similar colors [35], which performs well as the only feature in salient object detection [20]. Moreover, depth prior, i.e., assigning higher saliency values to near regions, is also effective in detecting salient object by combining with other features [21, 36].

Content location is utilized as a supplement in saliency detection, especially on natural images [40]. Both center-bias [7] and boundary-bias [51] can improve the performance salient object detection effectively. It is because salient objects are easily placed near to image center in photography and the objects near to image center attract more attention.

Motion is an important cue in salient object detection on videos, because moving objects easily attract viewer's attention [49]. To remove the influence of camera motion, motion contrast is used to represent object motion based on block-level motion vector fields [13, 18] or optical flow [30, 44].

Recently, the features extracted by deep neural network show their outstanding performance in many applications [25, 45]. It is also brought in salient object detection. For example, Li et al. [22] proposed an end-to-end deep contrast network, which consists of a pixel-level fully convolutional stream and a segment-wise spatial pooling stream, to general pixel-level saliency maps. Hou et al. [15] presented a top-down method by bringing short connections into skip-layer structures within the architecture of Holisitcally-Nested Edge Detector, which achieved accurate saliency detection results based on the reasonable combination of both low-level and high-level features.

2.2 Saliency coherence

A prominent problem in salient object detection is difficult to keep coherent saliency within each salient object. It hampers successive applications of salient object detection, such as image/video editing.

Graph-based models, such as random walk [19] and manifold ranking [46], are used to address the problem by propagating saliency values. For example, Qin et al. [37] utilized cellular automata for saliency map optimization, which shows robust to input saliency maps generated by different methods. Liu et al. [30] used both forward and backward saliency propagation based on inter-frame similarity matrices to improve temporal coherence.

Another solution of saliency coherence improvement is to treat saliency maps as the input of segmentation algorithms and generate binary saliency maps. Such a technique is named *saliency cuts*. Saliency cuts methods may utilize only saliency maps or saliency maps together with original images as their input. For instance, Achanta et al. calculated a threshold from saliency value distribution of the input saliency map and binarized the saliency map according to the threshold [1]. Li et al. produced segmentation seeds using adaptive triple thresholding, and fed the seeds to GrabCut algorithm [24].

3 Our method

3.1 Video representation

To a given video, we first over-segment each frame F^t into super-pixels using simple linear iterative clustering (SLIC) algorithm [2] in order to reduce computational cost while retaining the intrinsic structure of video content. Based on the over-segmentation, we can represent F^t with a set of super-pixels $F^t = \left\{p_1^t, p_2^t, \ldots, p_{N_t}^t\right\}$, where p_i^t is a super-pixel on F^t and N_t is the number of super-pixels on F^t. Inspired by [7], we extract a 12^3-bin color histogram $\boldsymbol{h}_i^t$ on $L^*a^*b^*$ color space from each super-pixel p_i^t, and retain the top-85 dominant bins in each histogram to improve efficiency.

Based on the super-pixel representation of video content, We construct the relationships between super-pixels according to their similarities. We first define the relationship between two super-pixels on the same frame. Suppose that p_i^t and p_j^t are on F^t, we define their relationship based on their color similarity weighted by spatial distance:

$$r_{i,j}^t = \left(1 - ||\boldsymbol{h}_i^t - \boldsymbol{h}_j^t||_2\right) \cdot \exp\left(-{d_{i,j}^t}^2\right), \tag{1}$$

where $\boldsymbol{h}_i^t$ and $\boldsymbol{h}_j^t$ are the color histograms of p_i^t and p_j^t, respectively; $||.||_2$ denotes Euclidean distance; $d_{i,j}^t$ is the spatial distance between the centers of p_i^t and p_j^t, which is divided by the length of image diagonal for normalization. To avoid the cumulative effect of small $r_{i,j}^t$ values, we filter the small $r_{i,j}^t$ values with a threshold τ, i.e., $r_{i,j}^t$ is set to 0 if $r_{i,j}^t$ is smaller than τ. In our experiments, τ is set to 0.3.

We also define the relationship between two super-pixels on adjacent frames. Suppose that p_i^t and p_j^{t+1} are on the adjacent frames F^t and F^{t+1} respectively, we seek a matching super-pixel $p_{i'}^{t+1}$ for p_i^t on F^{t+1} using object tracking strategy:

$$p_{i'}^{t+1} = \underset{p_k^{t+1} \in \Omega_i^{t+1}}{\arg\min} ||\boldsymbol{h}_i^t - \boldsymbol{h}_k^{t+1}||_2, \tag{2}$$

where $\boldsymbol{h}_i^t$ and $\boldsymbol{h}_k^{t+1}$ are the color histograms of p_i^t and p_k^{t+1}, respectively; Ω_i^{t+1} is a surrounding region on F^{t+1} with the same center coordinate to that of p_i^t; p_k^{t+1} is a super-pixel whose center is located within Ω_i^{t+1}. In our experiment, the size of Ω_i^{t+1} is set to 64×64 pixels. Assisted by $p_{i'}^{t+1}$, we define the relationship between p_i^t and p_j^{t+1} as follows:

$$r_{i,j}^{t,t+1} = \left(1 - ||\boldsymbol{h}_i^t - \boldsymbol{h}_j^{t+1}||_2\right) \cdot \exp\left(-{d_{i',j}^{t+1}}^2\right), \tag{3}$$

where $\boldsymbol{h}_i^t$ and $\boldsymbol{h}_j^{t+1}$ are the color histograms of p_i^t and p_j^{t+1}, respectively; $d_{i',j}^{t+1}$ is the normalized distance between the centers of $p_{i'}^{t+1}$ and p_j^{t+1}. Similarly, we filter the small $r_{i,j}^{t,t+1}$ values if they are smaller than the threshold τ, which equals 0.3 in our experiments.

3.2 Salient object detection on keyframes

Referring to [30], the dominant time cost in video-based salient object detection is caused by optical flow estimation, which plays a significant role in saliency calculation in videos. To achieve high efficiency, we explore the content redundancy of video frames by calculating

saliency values of super-pixels directly on several keyframes and generating saliency maps for other video frames with saliency propagation. In this way, we can elide optical flow estimation on a major proportion of video frames and improve the efficiency of our method. Though there have been amounts of methods for video keyframe extraction, we simply use uniform sampling because of efficiency requirement, i.e., sampling a keyframe every k video frames. In our experiments, k is set to 4.

On each keyframe, we detect a saliency map based on spatio-temporal difference. In saliency calculation based on spatial difference, we calculate the saliency value of each super-pixel p_i^t using color contrast with boundary connectivity [14] as follows:

$$c_i^t = \sum_{k=1}^{N_t} r_{i,k}^t \cdot \left(1 - \exp\left(-B_k^2/2\right)\right), \tag{4}$$

where c_i^t is the saliency value of p_i^t calculated based on color contrast; B_k is the boundary connectivity strength of p_k^t, which denotes the length ratio of p_j^t's edge on image boundary to its whole edge; N_t is the number of super-pixels on frame F^t.

In saliency calculation based on temporal difference, we estimate optical flow using large displacement optical flow algorithm [6] to obtain pixel-level motion vector, and further calculate the saliency value of each super-pixel p_i^t as follows:

$$o_i^t = 1 - \left|\left|\boldsymbol{m}_i^t - \boldsymbol{m}_g^t\right|\right|_2, \tag{5}$$

where o_i^t is the saliency value of p_i^t calculated based on object motion; $\boldsymbol{m}_i^t$ is the normalized motion vector of p_i^t with eight uniform intervals from $[-\pi, \pi]$; $\boldsymbol{m}_g^t$ represents the global motion on F^t, which is calculated as follows:

$$\boldsymbol{m}_g^t = \frac{1}{N_t} \sum_{k=1}^{N_t} \boldsymbol{m}_k^t \cdot \left(1 - \exp\left(-B_k^2/2\right)\right), \tag{6}$$

where B_k and N_t are same to that in (4).

We linearly combine the saliency values of each super-pixel calculated based on color contrast and object motion:

$$s_i^t = \alpha \cdot c_i^t + (1 - \alpha) \cdot o_i^t, \tag{7}$$

where α is a parameter to emphasize the effect of object motion, which is set to 0.3 in our experiments.

3.3 Saliency propagation

Once the saliency maps are generated on all the keyframes, we propagate saliency values from keyframes to their adjacent frames and further to other frames. To avoid unconstrained increasing of saliency values in propagation, we keep the sum of the saliency values propagated from each super-pixel constant; otherwise, all super-pixels will have high saliency values after sufficient times of saliency propagation. Hence, we define the propagation weight between two super-pixels p_i^t and p_j^* on adjacent frames as follows:

$$\omega_{i,j}^{t,*} = \frac{r_{i,j}^{t,*}}{1 + \sum_{m=1}^{N_{t-1}} r_{i,m}^{t,t-1} + \sum_{n=1}^{N_{t+1}} r_{i,n}^{t,t+1}}, \tag{8}$$

where $*$ can be set to $t-1$ or $t+1$; $r_{i,m}^{t,t-1}$ and $r_{i,n}^{t,t+1}$ are the relationship scores between p_i^t and a super-pixel on F^{t-1} and F^{t+1}, respectively; the "1" in the denominator of $\omega_{i,j}^{t,*}$ denotes the ratio of the retained saliency value for p_i^t in propagation.

We also propagate saliency values among the super-pixels on the same frame to improve saliency coherence within each salient object. The propagation weight between two super-pixels p_i^t and p_j^t is defined as follows:

$$\omega_{i,j}^t = \frac{r_{i,j}^t}{\sum_{k=1}^{N_t} r_{i,k}^t}, \tag{9}$$

where the ratio of the retained saliency for p_i^t in propagation is set to 1, i.e., $r_{i,i}^t$ is equal to 1.

According to (8) and (9), we calculate the saliency value of each super-pixel p_i^t after once saliency propagation among the super-pixels on the adjacent frames and the same frame as follows:

$$s_i^t = \mathbf{w}_i^t \mathbf{s}^t, \tag{10}$$

where $\mathbf{w}_i^t = [\omega_{1,i}^{t-1,t}, \ldots, \omega_{N_{k-1},i}^{t-1,t}, \omega_{i,1}^t, \ldots, \omega_{i,i}^t, \ldots, \omega_{i,N_k}^t, \omega_{1,i}^{t+1,t}, \ldots, \omega_{N_{k+1},i}^{t+1,t}]$ and $\mathbf{s}^t = \left[s_1^{t-1}, \ldots, s_{N_{k-1}}^{t-1}, s_1^t, \ldots, s_i^t, \ldots, s_{N_k}^t, s_1^{t+1}, \ldots, s_{N_{k+1}}^{t+1}\right]^T$. Specifically, $\omega_{i,i}^t$ is equal to $\frac{1}{1+\sum_{m=1}^{N_{t-1}} r_{i,m}^{t,t-1}+\sum_{n=1}^{N_{t+1}} r_{i,n}^{t,t+1}} + \frac{1}{\sum_{k=1}^{N_t} r_{i,k}^t}$.

We iteratively propagate saliency values among super-pixels on all the video frames till reaching the predefined iteration number or stable saliency values of all the super-pixels. Because the initial saliency values of all the super-pixels on non-keyframes are set to zero and the sum of saliency values is constrained to be constant in propagation, the saliency values of the super-pixels within salient objects are not high enough after propagation. Meanwhile, some super-pixels in background on keyframes may be assigned high saliency values mistakenly in salient object detection on keyframes. It will leave residual saliency values to super-pixels in background after propagation. In order to enhance the saliency difference between salient objects and background, we globally normalize all the saliency maps and obtain the final salient object detection result for the given video.

4 Experiments

4.1 Datasets and experimental settings

We validated the performance of our method on two public datasets, named *SegTrackV2* [23] and *UVSD* [30]. They contain 14 and 18 videos with manually labeled ground truth of salient objects on pixel level, respectively. Specifically, the videos in SegTrackV2 include various motion activities and scenes, while the videos in UVSD contain complicated motion and complex scenes. The diversity of these datasets increases the difficulty in salient object detection. We compared the proposed method with the state-of-the-art methods of video-based salient object detection: DCMR [18], GD [44], SGSP [30], SP [31] and SR [43].

All the experiments were conducted on a computer with Intel i5 2.8GHz CPU and 8GB memory. To all the other methods engaged in comparison, we used their default settings suggested by the authors, and normalized their generated saliency maps for a fair comparison.

4.2 Performance analysis

There are several key parameters influencing the performance of our proposed method, including the threshold for relationship filtering in video representation, the interval in keyframe selection and the combination weight of color contrast based saliency and object motion based saliency. We analyze the influence of these parameters as follows.

Threshold for relationship filtering The filtering of small relationship values in (1) and (3) aims to avoid propagating the saliency of a super-pixel to its unrelated super-pixels, which hampers the effect of saliency map initialization. However, too strict relationship reduces the effect and robustness of saliency propagation, which may lead to the failure in generating high quality saliency maps, especially for the non-keyframes. We validate the performance of our method using different filtering thresholds from 0.1 to 0.6 with the step of 0.1. Figure 3 shows the validation results. It shows that the low filtering thresholds, such as 0.1 and 0.2, may cause slight decline of performance and the high filtering threshold, such as 0.6, will prevent saliency propagation among super-pixels. Hence, we choose 0.3 as the default filtering threshold in our experiments.

Interval in keyframe selection It is a trade-off between effectiveness and efficiency to determine a suitable interval in keyframe selection. Large interval leads to a small number of keyframes and low time cost of salient object detection on these keyframes, but brings in the risk of generating low quality saliency maps on non-keyframes. In contrast, small interval

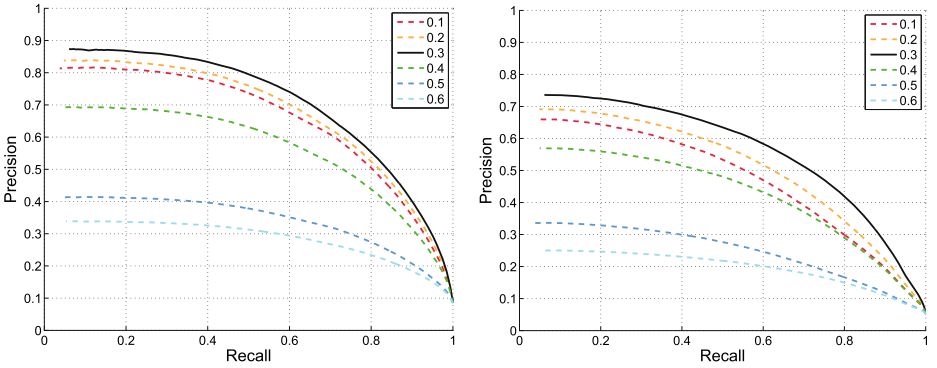

Fig. 3 Performance of our method using different filtering thresholds. **a** Results on SegTrackV2. **b** Results on UVSD

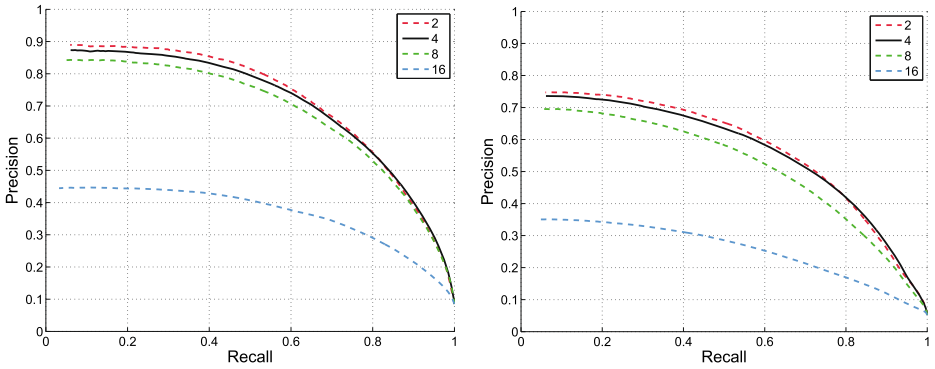

Fig. 4 Performance of our method using different keyframe selection interval. **a** Results on SegTrackV2. **b** Results on UVSD

increases the number of keyframes, whose saliency maps usually have high quality based on direct salient object detection, but decreases the efficiency. We validate the performance of the proposed method with different keyframe selection intervals, namely 2, 4, 8, 16. Figure 4 shows the validation results and Table 1 shows the corresponding time cost. It shows that small selection interval (such as 2) may slightly improve performance but cause obvious increase in time cost, and large selection interval (such as 16) may cause serious performance degradation. We choose 4 as the default keyframe selection interval to make a trade-off between effectiveness and efficiency.

Combination weight of color contrast based saliency and object motion based saliency We combine the saliency values based on color contrast and object motion in (7) to generate saliency maps on keyframes. Object motion usually plays a more important role than color contrast in video-based salient object detection, because moving objects easily attract viewer's attention. We validate the performance of our method using different combination weights from 0.1 to 0.5 with the step of 0.1. Figure 5 shows the validation results. It shows that smaller combination weight leads to better performance. However, both these two datasets, SegTrackV2 and UVSD, have some bias on spatio-temporal characteristics of video content, i.e., they focus on emphasizing the objects with obvious and complex motion. In order to handle different types of videos, such as nearly static videos, we choose 0.3 as the default combination weight in our experiments.

Table 1 Running time per frame of our method using different keyframe selection interval

Interval (frames)	2	4	8	16
SegTrackV2 (s)	5.95	3.52	2.69	2.32
UVSD (s)	6.14	3.36	2.60	1.65

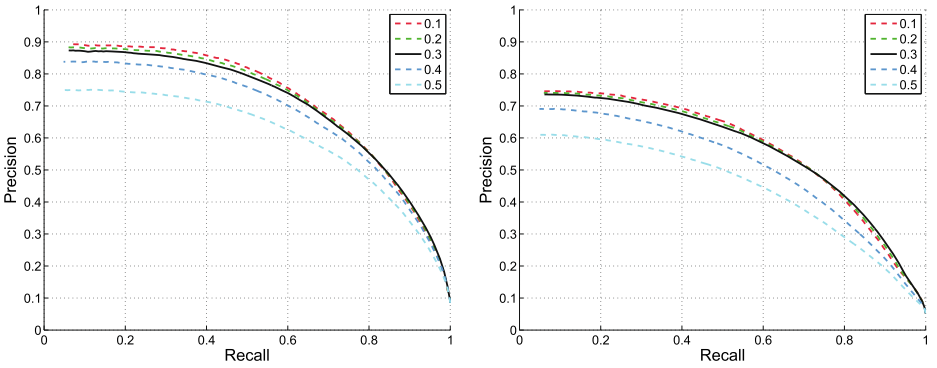

Fig. 5 Performance of our method using different combination weights. **a** Results on SegTrackV2. **b** Results on UVSD

4.3 Experimental results

We compare our method with five state-of-the-art video-based salient object detection methods. The quantitative results and the qualitative results are shown in Figs. 6 and 7, respectively. From Fig. 6, we can see that our method has similar performance to SGSP, and outperforms other methods. An important reason that our method can obtain good performance is the exploration of both spatial and temporal characteristics of video content in salient object detection on keyframes and bidirectional saliency propagation. In comparison, the methods without using optical flow, such as DCMR and SR, have obviously worse performance than other methods because of the lack of relatively accurate temporal characteristics. Meanwhile, the effective saliency propagation strategy helps our method to

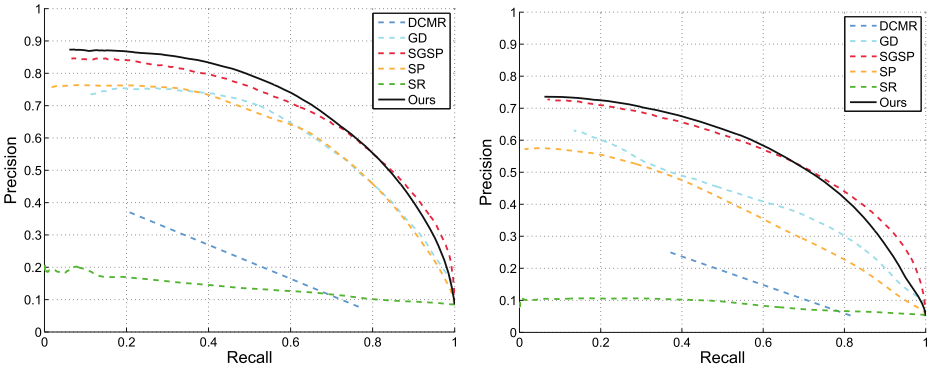

Fig. 6 Comparison of different methods with PR curves. **a** Results on SegTrackV2. **b** Results on UVSD

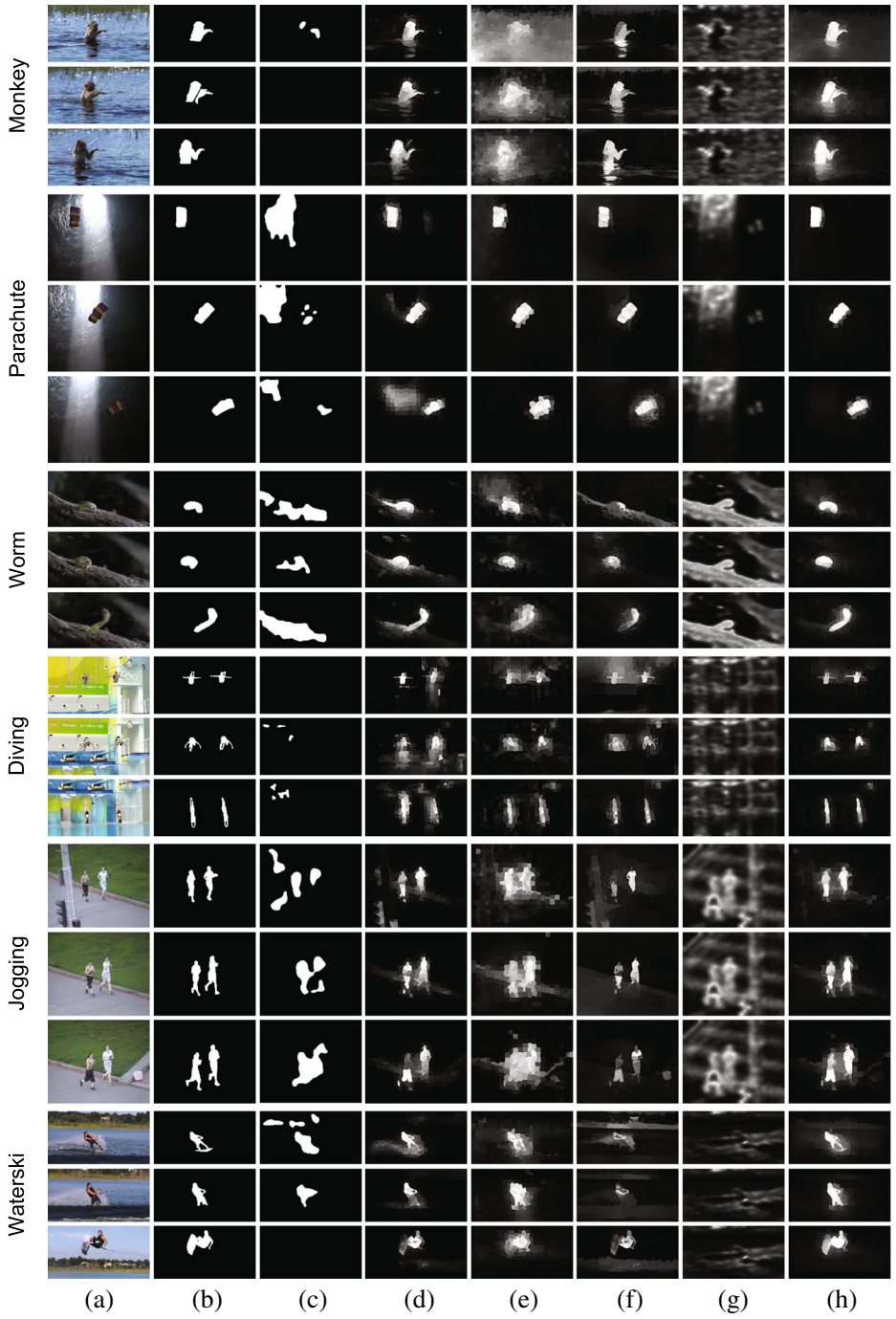

Fig. 7 Qualitative comparison of salient object detection results using different methods. **a** Video frames. **b** Ground truth. **c**–**g** Results of DCMR [18], GD [44], SGSP [30], SP [31] and SR [43]. **h** Our results

Table 2 Comparison with running time per frame of different methods

Method	Language	SegTrackV2 (s)	UVSD (s)
DCMR	C++	0.04	0.03
GD	Matlab	8.31	10.02
SGSP	Matlab	10.42	9.98
SP	Matlab	10.83	10.24
SR	Matlab	0.13	0.12
Ours	Matlab	3.52	3.36

obtain better performance than other methods using similar spatio-temporal characteristics of video content, such as GD and SP.

In Fig. 7, the top three videos are from SegTrackV2, and the bottom three ones are from UVSD. We can see that our method can effectively detect nearly complete salient objects under complicated object motion and complex scenes. In comparison, other methods fail in emphasizing salient objects, such as the results of SR on the videos Diving and Waterski, and suppressing residual saliency in background, such as the results of SGSP on Monkey and Jogging.

Table 2 shows the running time per frame of different methods. We can see that the methods without using optical flow, such as DCMR and SR, require less time cost than other methods, but they have obviously worse performance (as shown in Fig. 6). As compared to the methods with similar effectiveness, such as SGSP and GD, our method is more efficient because it only needs to detect salient objects on keyframes. Therefore, our method outperforms the existing video-based salient object detection methods when taking account of both effectiveness and efficiency.

4.4 Discussion

We also find some limitations of our method in the experiments. Figure 8 shows an example of our failure results. In the example, our method fails in generate high quality saliency maps on non-keyframes because too complicated object motion prevents effect saliency propagation. In this situation, our method needs to use smaller keyframe selection interval for performance improvement.

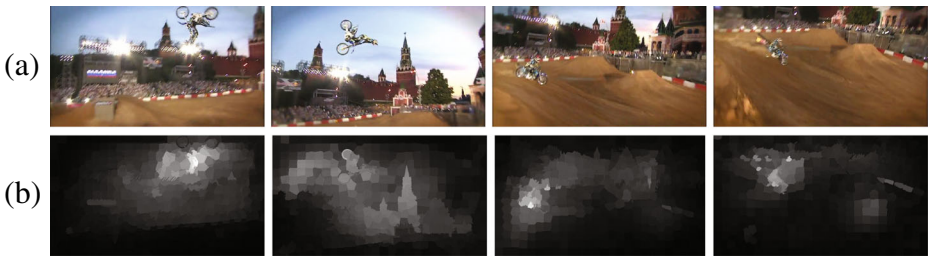

Fig. 8 A failure example of our method. **a** Video frames. **b** Our result consisted of low quality saliency maps

5 Conclusion

In this paper, we presented a video-based salient object detection method by fully exploring the potential of spatio-temporal difference and coherence in video content. Specifically, we detected salient objects on keyframes based on the combination of color contrast and object motion, and further propagate saliency values intra and inter frames. Finally, we generated the saliency maps with high saliency coherence for all the frames in a given video. The experimental results showed that our proposed method can achieve better salient object detection results with higher efficiency as compared to the state-of-the-art methods.

In future, we will focus on exploring more spatio-temporal characteristics of video content for salient object detection and improve the efficiency of our method by adaptive keyframe selection.

Acknowledgements The authors would like to thank the anonymous reviews for their helpful suggestion. This work is supported by National Science Foundation of China (61202320) and Research Project of Excellent State Key Laboratory (61223003).

References

1. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE conference on computer vision and pattern recognition, pp 1597–1604
2. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
3. Bao BK, Min W, Lu K, Xu C (2013) Social event detection with robust high-order co-clustering. In: ACM conference on international conference on multimedia retrieval. ACM, pp 135–142
4. Bao BK, Zhu G, Shen J, Yan S (2013) Robust image analysis with sparse representation on quantized visual features. IEEE Trans Image Process 22(3):860–871
5. Borji A, Cheng MM, Jiang H, Li J (2015) Salient object detection: a benchmark. IEEE Trans Image Process 24(12):5706–5722
6. Brox T, Malik J (2011) Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans Pattern Anal Mach Intell 33(3):500–513
7. Cheng MM, Mitra NJ, Huang X, Torr PH, Hu SM (2015) Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell 37(3):569–582
8. Cheng Z, Li X, Shen J, Hauptmann AG (2016) Which information sources are more effective and reliable in video search. In: International ACM SIGIR conference on research and development in information retrieval. ACM, pp 1069–1072
9. Desingh K, K MK, Rajan D, Jawahar C (2013) Depth really matters: improving visual salient region detection with depth. In: British machine vision conference
10. Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35(8):1915–1929
11. Gao Z, Zhang H, Xu G, Xue YB (2015) Multi-perspective and multi-modality joint representation and recognition model for 3d action recognition. Neurocomputing 151:554–564
12. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE conference on computer vision and pattern recognition, pp 580–587
13. Guo C, Zhang L (2010) A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans Image Process 19(1):185–198
14. Guo J, Ren T, Bei J (2016) Salient object detection for rgb-d image via saliency evolution. In: IEEE international conference on multimedia and expo. IEEE
15. Hou Q, Cheng MM, Hu XW, Borji A, Tu Z, Torr P (2016) Deeply supervised salient object detection with short connections. arXiv:1611.04849
16. Hu X, Wang G, Wu H, Lu H (2014) Rotation-invariant texture retrieval based on complementary features. In: International symposium on computer, consumer and control. IEEE, pp 311–314

17. Huang L, Luo B (2016) Salient object detection via video spatio-temporal difference and coherence. In: International conference on computational intelligence and security. IEEE, pp 218–222
18. Huang CR, Chang YJ, Yang ZX, Lin YY (2014) Video saliency map detection by dominant camera motion removal. IEEE Trans Circ Syst Video Technol 24(8):1336–1349
19. Jiang B, Zhang L, Lu H, Yang C, Yang MH (2013) Saliency detection via absorbing markov chain. In: IEEE international conference on computer vision, pp 1665–1672
20. Ju R, Liu Y, Ren T, Ge L, Wu G (2015) Depth-aware salient object detection using anisotropic center-surround difference. Signal Process Image Commun 38:115–126
21. Lang C, Nguyen TV, Katti H, Yadati K, Kankanhalli M, Yan S (2012) Depth matters: influence of depth cues on visual saliency. In: European conference on computer vision, pp 101–115
22. Li G, Yu Y (2016) Deep contrast learning for salient object detection. In: IEEE conference on computer vision and pattern recognition, pp 478–487
23. Li F, Kim T, Humayun A, Tsai D, Rehg JM (2013) Video segmentation by tracking many figure-ground segments. In: IEEE international conference on computer vision, pp 2192–2199
24. Li S, Ju R, Ren T, Wu G (2015) Saliency cuts based on adaptive triple thresholding. In: IEEE international conference on image processing. IEEE, pp 4609–4613
25. Li Y, Lu H, Li J, Li X, Li Y, Serikawa S (2016) Underwater image de-scattering and classification by deep neural network. Comput Electric Eng 54:68–77
26. Liu AA, Su YT, Nie WZ, Kankanhalli M (2017) Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans Pattern Anal Mach Intell 39(1):102–114
27. Liu H, Heynderickx I (2011) Visual attention in objective image quality assessment: based on eye-tracking data. IEEE Trans Circ Syst Vid Technol 21(7):971–982
28. Liu Y, Zhou F, Liu W, De la Torre F, Liu Y (2010) Unsupervised summarization of rushes videos. In: ACM international conference on multimedia. ACM, pp 751–754
29. Liu Y, Liu Y, Chan KC (2011) Tensor-based locally maximum margin classifier for image and video classification. Comput Vis Image Underst 115(3):300–309
30. Liu Z, Li J, Ye L, Sun G, Shen L (2015) Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. In: IEEE transactions on circuits and systems for video technology
31. Liu Z, Zhang X, Luo S, Le Meur O (2014) Superpixel-based spatiotemporal saliency detection. IEEE Trans Circ Syst Vid Technol 24(9):1522–1540
32. Lu H, Li Y, Nakashima S, Serikawa S (2016) Single image dehazing through improved atmospheric light estimation. Multimed Tools Appl 75(24):17081–17096
33. Lu H, Serikawa S (2014) Underwater scene enhancement using weighted guided median filter. In: IEEE international conference on multimedia and expo. IEEE, pp 1–6
34. Nie L, Hong R, Zhang L, Xia Y, Tao D, Sebe N (2016) Perceptual attributes optimization for multivideo summarization. IEEE Trans Cybern 46(12):2991–3003
35. Niu Y, Geng Y, Li X, Liu F (2012) Leveraging stereopsis for saliency analysis. In: IEEE conference on computer vision and pattern recognition, pp 454–461
36. Peng H, Li B, Xiong W, Hu W, Ji R (2014) RGBD salient object detection: a benchmark and algorithms. In: European conference on computer vision, pp 92–109
37. Qin Y, Lu H, Xu Y, Wang H (2015) Saliency detection via cellular automata. In: IEEE conference on computer vision and pattern recognition, pp 110–119
38. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
39. Ren T, Liu Y, Wu G (2009) Image retargeting based on global energy optimization. In: IEEE international conference on multimedia and expo, pp 406–409
40. Ren T, Liu Y, Ju R, Wu G (2016) How important is location information in saliency detection of natural images. Multimed Tools Appl 75(5):2543–2564
41. Sang J, Xu C (2011) Browse by chunks: topic mining and organizing on web-scale social media. ACM Trans Multimed Comput Commun Appl 7(1):30
42. Sang J, Xu C (2012) Right buddy makes the difference: an early exploration of social relation analysis in multimedia applications. In: ACM international conference on multimedia. ACM, pp 19–28
43. Seo HJ, Milanfar P (2009) Static and space-time visual saliency detection by self-resemblance. J Vis 9(12):15
44. Wang W, Shen J, Porikli F (2015) Saliency-aware geodesic video object segmentation. In: IEEE conference on computer vision and pattern recognition, pp 3395–3402
45. Xu Z, Yang Y, Hauptmann AG (2015) A discriminative cnn video representation for event detection. In: IEEE conference on computer vision and pattern recognition, pp 1798–1807

46. Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. In: IEEE conference on computer vision and pattern recognition, pp 3166–3173
47. Zhang L, Hong R, Nie L, Hong C (2016) A biologically inspired automatic system for media quality assessment. IEEE Trans Autom Sci Eng 13(2):894–902
48. Zhong SH, Liu Y, Liu Y (2011) Bilinear deep learning for image classification. In: ACM international conference on multimedia. ACM, pp 343–352
49. Zhong SH, Liu Y, Ren F, Zhang J, Ren T (2013) Video saliency detection via dynamic consistent spatio-temporal attention modelling. In: AAAI conference on artificial intelligence
50. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. In: Advances in neural information processing systems, pp 487–495
51. Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In: IEEE conference on computer vision and pattern recognition, pp 2814–2821
52. Zhu L, Shen J, Xie L (2017) Unsupervised visual hashing with semantic assistant for content-based image retrieval. IEEE Trans Knowl Data Eng 29(2):472–486

Lei Huang is a Ph.D. student of Software Institute, Nanjing University. Her current research interests include multimedia content analysis and social media.

Bin Luo is a Professor with Software Institute, Nanjing University. He has over 50 refereed research papers. Dr. Luo is a CCF Distinguished Member. His current research interests include intelligent information system and software engineering.