

Video Saliency Detection Using Spatiotemporal Cues

Yu CHEN[†], *Nonmember*, Jing XIAO^{†a)}, *Member*, Liuyi HU[†], Dan CHEN[†], Zhongyuan WANG[†],
and Dengshi LI^{††}, *Nonmembers*

SUMMARY Saliency detection for videos has been paid great attention and extensively studied in recent years. However, various visual scene with complicated motions leads to noticeable background noise and non-uniformly highlighting the foreground objects. In this paper, we proposed a video saliency detection model using spatio-temporal cues. In spatial domain, the location of foreground region is utilized as spatial cue to constrain the accumulation of contrast for background regions. In temporal domain, the spatial distribution of motion-similar regions is adopted as temporal cue to further suppress the background noise. Moreover, a backward matching based temporal prediction method is developed to adjust the temporal saliency according to its corresponding prediction from the previous frame, thus enforcing the consistency along time axis. The performance evaluation on several popular benchmark data sets validates that our approach outperforms existing state-of-the-arts.

key words: saliency detection, spatio-temporal cues, foreground prior, spatial distribution, background noise suppression

1. Introduction

Saliency detection for images and videos has been used in wide series of applications, such as object recognition, video surveillance and mobile robotics [1]. It aims to automatically select the visual important regions or objects by exploiting the inherent visual attention mechanism. In the past decades, a number of saliency models based on different schemes and theories have been proposed, which can be roughly classified into two categories: a rapid, stimulus-driven, bottom-up pattern and a slower, goal-driven, top-down pattern. The top-down pattern is driven by task and relatively less explored due to the complexity and variety of daily tasks and behaviors [2]. Considering the efficiency and the generality, we focus on bottom-up, features driven saliency detection for videos.

Perceptual research shows that contrast (e.g., color contrast, luminance contrast, motion contrast) plays a significant role in low-level visual saliency [3] and is widely applied not only to image saliency detection but also to video saliency detection. Most of the existing methods utilized contrast as the most important feature to compute saliency based on the fundamental assumption that an image unit, i.e.

a pixel or patch, is salient if it is highly distinctive compared with its surroundings. For spatial saliency detection [4]–[7], the center-surround principle [8] is generally utilized to introduce spatial relations between image units into saliency models. Kim et al. [9] measured spatial saliency using the center-surround difference on ordinal signatures of edge and color orientation histograms. Differently, Goferman et al. [10] took the contrast from both local and global perspectives into consideration and built a content-aware saliency detection model. Cheng et al. [11] proposed a global contrast based algorithm, which employed the spatial distance between the current region and other regions as a factor to weigh the color histogram contrast. For temporal saliency detection [12]–[15], motion information is considered as the primarily influential factor. Gao et al. [16] extended the image saliency model in [17] by adding the motion channel to predict human eye fixations in dynamic scenes. Rahtu et al. [18] proposed a statistical framework and formulate temporal saliency measure with motion histogram contrast. Zhou et al. [19] employed retiming and temporal filtering to enhance the rendering of salient motion. Although the abovementioned methods have achieved remarkable performance, they still introduce background noise to some extent. From the spatial aspect, calculating contrast of every image unit within the same scope of neighborhood tends to accumulate overmuch color significance for background image units, thus resulting in overemphasized background regions. From the temporal aspect, using motion contrast alone as saliency measure is not sufficient enough to distinguish the foreground objects and background regions when the motion of them are highly similar. Hence, lots of background fragments are produced. In addition, these methods some issues in maintaining spatiotemporal consistency of videos, which probably generate discontinuous salient objects in saliency maps. Some examples are shown in Fig. 1.

To address the above problems, we proposed a novel saliency detection framework for videos using spatio-temporal cues. On the one hand, we exploit the detection results of the previous frames as foreground prior and propose a contrast calculation method to reduce the background noise in spatial domain. We treat the foreground and background regions differently when calculating contrast. Color significance is accumulated in a large extent for foreground superpixels while constrained within a small neighborhood scope for background superpixels. In this way, the contrast accumulation of the background is restricted. On the

Manuscript received December 15, 2017.

Manuscript revised March 9, 2018.

Manuscript publicized June 20, 2018.

[†]The authors are with the School of Computer, Wuhan University, Wuhan, 430072 China.

^{††}The author is with the School of Mathematics and Computer, Jiangnan University, Wuhan, 430072 China.

a) E-mail: jing@whu.edu.cn

DOI: 10.1587/transinf.2017PCP0011

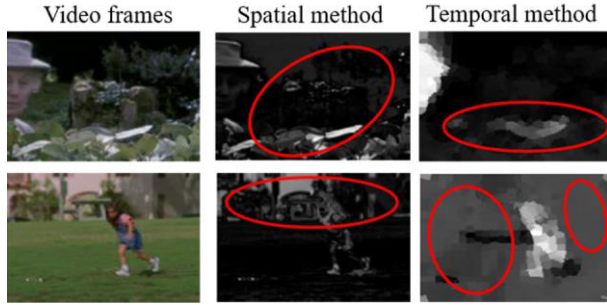


Fig. 1 Example results of existing saliency detection method. Background fragments are marked with red circles.

other hand, motion similarity is introduced in the temporal saliency calculation process to further suppress the background noise. Based on the observation that motions of the superpixels in the foreground regions should be more consistent than that of the superpixels in the background regions, we compute the variance of the spatial distribution of superpixels having similar motion contrast to distinguish the salient objects and the fragments in the background regions. Moreover, we develop a prediction method to cope with the spatiotemporal inconsistency.

Compared with our previous work [22], two improvements are introduced: 1) The motion histogram of background regions is utilized rather than that of the whole frame when calculating the motion histogram contrast, which can reduce the contribution from foreground regions to background regions and suppress the background fragments. 2) A backward matching based temporal prediction method is developed to adjust the temporal saliency according to its corresponding prediction from the previous frame. Foreground prior is adopted to constrain the matching area from the foreground regions in the current frame to the previous frame, thus decreasing the probability of mismatch and improving the accuracy of the temporal prediction.

The remainder of this paper is organized as follows: we introduce our model in Sect. 2. Experimental results and analysis are presented in Sect. 3, and the conclusion is given in Sect. 4.

2. Proposed Method

2.1 Overall Framework

We propose a novel saliency detection method making full use of foreground prior. We use superpixel as the image unit throughout this paper. The proposed spatiotemporal saliency model is shown in Fig. 2. Each module is introduced in detail in the following sections.

2.2 Spatial Saliency Detection

2.2.1 Spatial Feature Extraction

We choose color significance as feature to measure the initial spatial saliency. Considering that significant color of

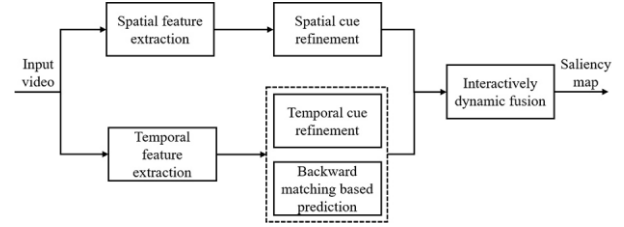


Fig. 2 Overall saliency detection framework.

low appearance frequency usually attracts more visual attention, color significance is simply measured by accumulation of color differences weighted by the corresponding appearance frequency.

Directly processing with the original videos requires a large amount of computation due to the variety of colors. Hence, minimum variance quantization proposed in [8] is used to reduce the number of colors of the frame without losing visual fidelity. Color cube is firstly separated into several bins of different sizes based on color distribution in the frame. Then the average color in each bin is used to create a new reduced colormap. We discard infrequently occurring colors and replace them with the most similar remained colors in colormap. Each superpixel can be represented by the corresponding principal color for its high color consistency. So we obtain superpixel significance in accordance with the significance of principal color, which is computed as

$$CS(sp_t^i) = \sum_{j=1}^{np_t} f_j \cdot \|(L_i, a_i, b_i) - (L_j, a_j, b_j)\|_2. \quad (1)$$

The above function indicates the color significance with respect to principal color c_i in $L^*a^*b^*$ color space of sp_t^i . sp_t^i denotes the i th superpixel in the t th frame I_t . f_j represents the appearance probability of c_j in the whole frame and np_t is the number of principal colors of superpixels in I_t .

2.2.2 Spatial Cue Refinement

According to the assumption [20] that a region highly different from its surrounding background is salient, we argue that superpixels should be treated differently instead of equally in previous approaches. To be specific, the contrast calculation of superpixels belonging to foreground regions can be done in the area of whole frame while the contrast calculation of superpixels belonging to background regions should be constrained within a small range of neighborhood to avoid emphasizing the background resulting from the overmuch color significance accumulation. Detailed explanation can be found in Fig. 3. Hence, we first extract the location of foreground as spatial cue and then apply it to the contrast calculation. Considering the computational cost, we utilize the salient regions from the previous detection results to act as the foreground regions in the current frame. In order to restrict the contribution to the background superpixels, an exponential term is employed, which can be

written as

$$S_S(sp_t^i) = \sum_{j=1}^{n_t} CS(sp_t^i) \cdot \exp \left[-\frac{D(sp_t^i, sp_t^j)}{D_k(sp_t^i, sp_t^j)} \right], \quad (2)$$

$$D_k(sp_t^i, sp_t^j) = \begin{cases} \min_{j=1,2,\dots,n_t} (D(sp_t^i, sp_t^j)) & \text{if } sp_t^i \in B_t \\ \text{sqrt}(\text{area}(F_t)) & \text{if } sp_t^i \in F_t \end{cases}. \quad (3)$$

In Eq. (2), $S_S(sp_t^i)$ is the spatial saliency of sp_t^i . n_t presents the number of superpixels in the t th frame. $D(sp_t^i, sp_t^j)$ describes the Euclidean distance between the centers of sp_t^i and sp_t^j . From Eq. (2), it is obvious that the exponential term controls the saliency contribution from other superpixels to sp_t^i . The distant superpixels contribute less than the close ones. Equation (3) defines the key distance D_k . B_t and F_t present the background and foreground regions respectively. When calculating the spatial saliency of background superpixels, D_k reaches its minimum value and so is to the exponential weight. Hence, the contribution from other superpixels is constrained. For foreground superpixels, we set D_k to a relatively large value, which is the square root of the area of F_t in this paper, to enhance the contribution from other superpixels, thus achieving background suppression while emphasizing foreground.

2.3 Temporal Saliency Detection

2.3.1 Motion Feature Extraction

We choose motion histogram contrast as feature to measure the initial temporal saliency. The motion vector field MVF_t of I_t is produced by optical flow method [21]. According to [22], we extract the accumulated amplitude histograms $MH_t^0(k)/MH_t^i(k)$ and mean orientation histograms $OH_t^0(k)/OH_t^i(k)$ of motion at the frame-level/superpixel-level as global/local motion features respectively. Orientations of motion vectors are uniformly quantized into $b = 8$ intervals, which are indexed by k . As we have obtained the location of foreground regions in advance, the accumulated amplitude histogram $BMH_t^0(k)$ and the mean orientation histogram $BOH_t^0(k)$ of background are utilized in the place of frame-level histograms to reduce the influence from the foreground regions. The motion histogram contrast is expressed as



Fig. 3 Spatial cue illustration. White foreground and black background based on superpixel level. The dark purple superpixel accumulates sufficient differences from other ones, which is demonstrated by the orange solid lines. The green superpixels have a small scope of surrounding background. Orange dotted lines denote differences from other superpixels cannot contribute to the certain superpixel.

$$MHC(sp_t^i) = \sum_{j=1}^b MH_t^i(j) \sum_{k=1}^b DOH \cdot BMH_t^0(k), \quad (4)$$

$$DOH = \|OH_t^i(j) - BOH_t^0(k)\|_2. \quad (5)$$

The initial temporal saliency tends to have noticeable background noise. The main reason is that motion histogram contrast only takes the distinctiveness between local feature and global feature into consideration while the distinctiveness between local features, which can provide more information, is ignored.

2.3.2 Temporal Cue Refinement

According to the spatial element distribution assumption [7], the distinct elements are more likely to belong to the same object when they have similar motion characteristics. Inspired by this assumption, we further propose that superpixels grouped into a certain region are more likely to belong to the foreground while background superpixels tend to uniformly scatter in the whole scene. Figure 4 gives an intuitively explanation. Based on the above observations, spatial distribution of the similar motion vectors is naturally introduced to refine the temporal saliency map. Since the grouped foreground superpixels have low variance while the scattered background superpixels have high variance, we choose the variance of spatial distribution to measure the probability of being foreground or foreground superpixel, which can be formulated as

$$VM(sp_t^i) = \sum_{j=1}^{n_t} D(csp_t^i, sp_t^j) \cdot \exp \left[-\frac{(MHC(sp_t^i) - MHC(sp_t^j))^2}{2\sigma_D^2} \right], \quad (6)$$

$$csp_t^i = \sum_{j=1}^{n_t} c_t^j \cdot \exp \left[-\frac{(MHC(sp_t^i) - MHC(sp_t^j))^2}{2\sigma_D^2} \right]. \quad (7)$$

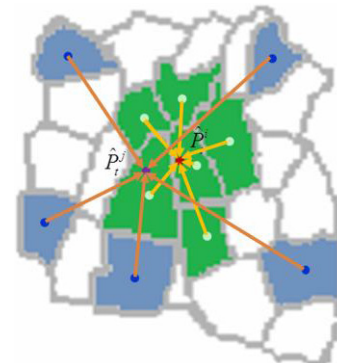


Fig. 4 Temporal cue illustration. Superpixels that have similar motion contrast are marked with the same color. The red point and the purple point are the center of green and blue superpixels respectively. The blue superpixels are far away from their center hence are more likely to be the background superpixels while the green ones are close to their center hence are more likely to be the foreground superpixels.

In Eq. (6), VM is the variance of the spatial distribution. csp_t^j is the weighted average center of sp_t^j . In Eq. (7), c_t^j is the center of sp_t^j . σ_D is a parameter that controls the motion sensitivity of the distribution. $MHC(sp_t^j)$ denotes the motion histogram contrast of the superpixel sp_t^j with respect to other superpixels which reflects the degree of similarity between them. The exponential term allows motion similar superpixels to contribute to the variance calculation while constrains the contribution from other irrelevant superpixels.

It is widely acknowledged that most of the salient regions are foreground. Hence, a negative exponential function is utilized to incorporate the variance of spatial distribution into temporal saliency calculation, which is expressed as

$$S_T^{inter}(sp_t^j) = MHC(sp_t^j) \cdot \exp(-VM(sp_t^j)). \quad (8)$$

In Eq. (8), $S_T^{inter}(sp_t^j)$ denotes the intermediate temporal saliency of sp_t^j calculated from the current frame.

2.3.3 Backward Matching Based Prediction

To further enhance the spatiotemporal consistency of temporal saliency, we match the foreground regions between adjacent frames and utilize the temporal saliency from the corresponding superpixels in the previous frame to adjust the saliency of the foreground superpixels in the current frame.

We adopt the motion vector field from the current frame to the previous frame and map superpixels in the current frame to superpixels in the previous frame to guarantee each superpixel in the current frame has at least one match. We also utilize Euclidean distance preferring the nearest matching candidate to ensure the one to one correspondence for every superpixel in the current frame. Since we only care about the spatiotemporal consistency of foreground objects, the matching and prediction process is just applied to foreground regions in the current frame. Moreover, we set the constrain that the superpixels in the current frame can only map to the foreground region in the previous frame. Our backward matching method is shown in Fig.5 and the index of superpixel in the previous frame which is matched with the current superpixel sp_t^j can be calculated by

$$ind_{t-1} = \arg \min_j \|c_{t-1}^j - c_t^{i*}\|_2, \quad c_{t-1}^j, c_t^{i*} \in F_{t-1}, \quad (9)$$

where c_t^{i*} is the center of the superpixel in the previous frame where sp_t^j is mapped. c_{t-1}^j presents the center of sp_{t-1}^j belonging to the foreground region F_{t-1} in the previous frame.

After matching the foreground superpixels in the adjacent frames, we predict the temporal saliency of superpixels in the current frame by straightforwardly copying the temporal saliency of the corresponding superpixels in the previous frame. The final temporal saliency map S_T is generated by fusing the predicted saliency and the intermediate temporal saliency. The formulation is given as follows

$$S_T(sp_t^j) = \lambda_T \cdot S_T^p(sp_{t-1}^j) + (1 - \lambda_T) \cdot S_T^{inter}(sp_t^j), \quad (10)$$

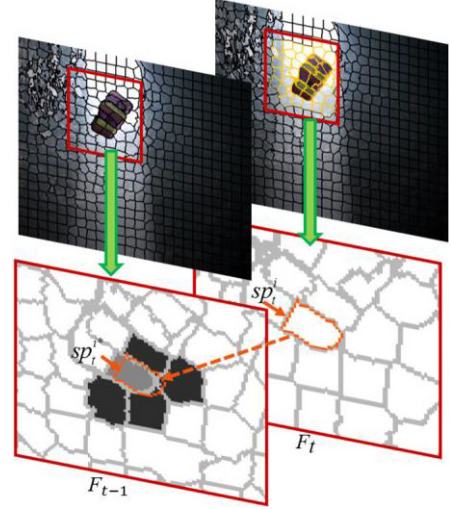


Fig. 5 Backward matching illustration. The red rectangles are foreground regions extract according to foreground prior. The orange superpixels present the superpixel to be predicted in the current frame and its corresponding location in the previous frame. The grey superpixels are matching candidates. The final match is the light grey one according to the criterion.

where $S_T^p(sp_{t-1}^j)$ is the prediction of sp_t^j , λ_T is a constant employed to balance the influence between the prediction and motion features.

2.4 Interactively Dynamic Fusion

Since motions extracted by optical flow often suffer a lot from complex environment variations, such as moving background, non-rigid deformation and occlusion, it is necessary to collaborate temporal and spatial saliency maps to achieve better performance of video saliency detection. If high mutual consistency exists between temporal and spatial saliency maps, it indicates that the interaction confidence between saliency maps from different domains is high. In order to measure interactive consistency between temporal and spatial saliency maps, the commonness of these two maps is measured as

$$C(sp_t^j) = S_S(sp_t^j) \cdot S_T(sp_t^j). \quad (11)$$

$C(sp_t^j)$ is normalized to $[0, 1]$, we utilize two interactive consistency indicators, namely consistency of temporal domain to spatial domain SIT and consistency of spatial domain to temporal domain SIS , to measure the spatiotemporal consistency, which are computed as

$$SIT = \sum_{i=1}^{n_t} C(sp_t^i) / \sum_{i=1}^{n_t} S_T(sp_t^i), \quad (12)$$

$$SIS = \sum_{i=1}^{n_s} C(sp_s^i) / \sum_{i=1}^{n_s} S_S(sp_s^i). \quad (13)$$

Spatial and temporal saliency are dynamically fused using the above two parameters to generate the final spatiotemporal saliency map, which is written as

$$S(sp_t^j) = \frac{SIT \cdot S_T(sp_t^j) + SIS \cdot S_S(sp_t^j)}{SIT + SIS}. \quad (14)$$

3. Experimental Results

The proposed method makes utilization of both spatial and temporal information among video frames. In this section, we demonstrate the advantages of our method compared with other methods by experiment.

3.1 Data Sets and Experimental Settings

We evaluate different approaches on two well-known data sets, SegTrack v2 and Freiburg Berkeley Motion Segmentation (FBMS). SegTrack v2 data set has eight video sequences of single object and the number of frames in each sequence varies from 21 to 279. FBMS is a data set of 59 video sequences, in which contain diverse kinds of objects including animals, vehicles and humans. Both of them pose various challenges such as large complicated background, significant shape deformation, and large camera motion on video saliency detection. Also, they provide pixel-level binary ground truths of salient objects for objectively evaluating the performance.

In our experiments, five state-of-the-art saliency detection models are chosen as the comparison methods, namely region based contrast (RC) [11], saliency filters (SF) [7], energy minimization based saliency (SEG) [18], Quaternion Fourier Transform based saliency (QS) [12] and cluster

based co-saliency (CS) [13].

There are several internationally recognized standards for evaluating saliency detection performance [1], including precision versus recall curves (PR curves), receiver operating characteristic curve (ROC curves), area under ROC curve (AUC), F-Measure and mean absolute error (MAE) and so on. We adopt the commonly used PR curves, AUC, F-Measure and MAE in the experiments.

3.2 Performance Evaluation and Analysis

We first present the qualitative results of different saliency models on SegTrack v2 data set. Figure 6 gives the visual comparison. As can be seen, our method highlights the salient objects and effectively suppresses background noise. Due to the lack of information about the sizes and the locations of foreground objects, RC and SF tend to highlight the background region in complicated visual scenes. Especially in the test sequence birdfall, the sky is highlighted while the falling bird is hardly detected because its color features are very much similar with that of the background. QS overemphasizes small and local features such as edges. SEG utilizes local features to compute saliency. Hence, to the sequences girl and parachute where the salient objects are highly distinctive from the background in local regions, it performs well. However, it introduces severe background noise in the meantime. CS calculates the global consistency of the cluster based features between adjacent frames. If there are cluster features that have high temporal

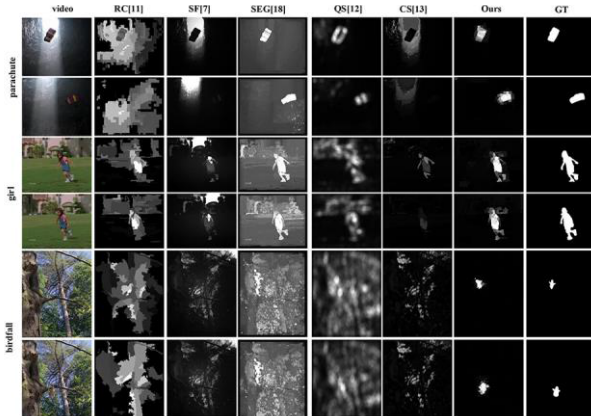


Fig. 6 Qualitative results on SegTrack v2 data set.



Fig. 8 Qualitative results on FBMS data set.

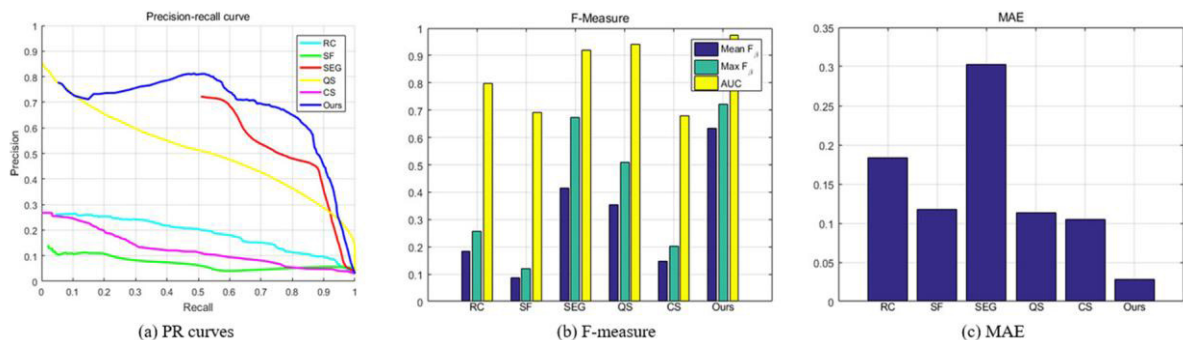


Fig. 7 Quantitative results on SegTrack v2 data set.

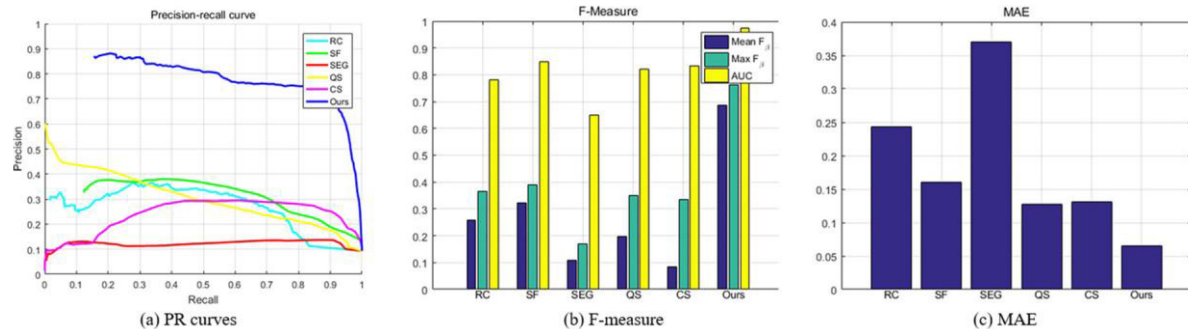


Fig. 9 Quantitative results on FBMS data set.

consistency and similarity with salient objects in the background, then CS prefers to highlight those background regions. Therefore, CS produces rather good results in the test sequence girl but constantly highlights the non-salient objects in other sequences. The proposed method efficiently utilizes the foreground prior and multiple features extracted in both spatial and temporal domains to enforce the spatiotemporal consistency and suppress the background noise, thus achieving better results.

The quantitative results on SegTrack v2 data set are shown in Fig. 7. We can see that our method outperforms the others on all the four evaluating measures. RC and SF provide relatively low precision and F-Measure, due to the complex motion of objects and insufficient temporal cues. Also, the frequently highlighting non-salient objects and background regions results in low precision and F-Measure for CS and SEG respectively. Compared with other methods, the proposed model has low MAE on account of the effective elimination of background noise.

Then, the performance comparisons on FBMS data set are provided as well. Since the proposed approach is designed for detecting single object, sequences containing single object in FBMS data set are chosen to conduct the experiments. Figure 8 shows the qualitative results. The FBMS data set also dramatically varies in both background and motion. Hence, the problems exhibiting in the SegTrack v2 data set still exist. RC tends to highlight non-salient regions. The salient objects change with time in SF results. SEG still has difficulties in suppressing background while CS constantly highlight the bottle in the background and fail to maintain the consistency of the cat. Our results are closest to the ground truths.

Figure 9 presents the quantitative results on FBMS data set. Our approach still outperforms the others.

The proposed method proves to have great ability in suppressing background noise. Both subjective and objective results on SegTrack v2 and FBMS data sets demonstrate the advantages of our proposed method over other methods. However, the proposed method heavily relies on the foreground locating results. False detection may lead to overmuch background contrast accumulation in spatial domain and mismatch of foreground superpixels in temporal domain, thus resulting in background noise and temporal inconsistency. Hence, we measure the distance between the

centers of the foreground region and the salient region. If the distance is above a certain threshold, then we reset the foreground region to the whole frame.

4. Conclusion

In this paper, we proposed a novel video saliency detection method to produce high-quality saliency maps. On the one hand, we exploit the detection results of the previous frames as spatial cues and propose a foreground prior based contrast calculation method to reduce the background noise in spatial domain. On the other hand, motion similarity is introduced as temporal cues in the temporal saliency calculation process to further suppress the background noise. Moreover, we develop a backward matching based superpixel prediction method to cope with the spatiotemporal inconsistency. Experimental results on two public datasets demonstrate that the proposed model outperforms the state-of-the-art spatiotemporal saliency models.

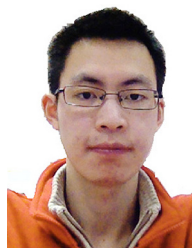
Acknowledgements

This work was supported by the National Nature Science Foundation of China (No. 61502348), the Hubei Province Technological Innovation Major Project (No. 2016AAA015), the EU FP7 QUICK project under Grant Agreement No. PIRSES-GA-2013-612652 and the Fundamental Research Funds for the Central Universities (413000048).

References

- [1] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: a benchmark," *IEEE Trans. Image Process.*, vol.24, no.12, pp.5706–5722, Dec. 2015.
- [2] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: a survey," *Cornell University Library*, <https://arxiv.org/abs/1411.5878>, accessed Nov. 18, 2014.
- [3] W. Einhäuser and P. König, "Does luminance-contrast contribute to a saliency map for overt visual attention," *Eur. J. Neurosci.*, vol.17, no.5, pp.1089–1097, March 2003.
- [4] O.L. Meur, P.L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.28, no.5, pp.802–817, May 2006.
- [5] D.A. Klein and S. Frntrop, "Center-surround divergence of feature statistics for salient object detection," *Proc. IEEE ICCV*,

- Barcelona, Spain, pp.2214–2219, Nov. 2011.
- [6] Y. Xie, H. Lu, and M.-H. Yang, “Bayesian saliency via low and mid level cues,” *IEEE Trans. Image Process.*, vol.22, no.5, pp.1689–1698, May 2013.
 - [7] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” *IEEE Conf. CVPR*, Providence, USA, pp.733–740, June 2012.
 - [8] J. Lou, M. Ren, H. Wang, and D. Hu, “Regional Principal Color Based Saliency Detection,” *PLoS ONE*, vol.9, no.11, p.e112475, 2014.
 - [9] W. Kim, C. Jung, and C. Kim, “Spatiotemporal saliency detection and its applications in static and dynamic scenes,” *IEEE Trans. Circuits Syst. Video Technol.*, vol.21, no.4, pp.446–456, April 2011.
 - [10] S. Goferman, L.Z. Manor, and A. Tal, “Context-aware saliency detection,” *Proc. IEEE Conf. CVPR*, San Francisco, USA, pp.1915–1926, June 2010.
 - [11] M. Cheng, G. Zhang, N.J. Mitra, X. Huang, and S. Hu, “Global contrast based salient region detection,” *Proc. IEEE Conf. CVPR*, Colorado Springs, USA, pp.569–582, 2015.
 - [12] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform,” *Proc. IEEE Conf. CVPR*, Anchorage, USA, pp.1–8, June 2008.
 - [13] H. Fu, X. Cao, and Z. Tu, “Cluster-based co-saliency detection,” *IEEE Trans. Image Process.*, vol.22, no.10, pp.3766–3778, Oct. 2013.
 - [14] V. Mahadevan and N. Vasconcelos, “Spatiotemporal saliency in dynamic scenes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.32, no.1, pp.171–177, Jan. 2010.
 - [15] W. Wang, J. Shen, X. Li, and F. Porikli, “Robust video object cosegmentation,” *IEEE Trans. Image Process.*, vol.24, no.10, pp.3137–3148, Oct. 2015.
 - [16] D. Gao, V. Mahadevan, and N. Vasconcelos, “The discriminant center-surround hypothesis for bottom-up saliency,” *Proc. Adv. NIPS*, Vancouver, Canada, pp.497–504, Jan. 2007.
 - [17] D. Gao and N. Vasconcelos, “Bottom-up saliency is a discriminant process,” *Proc. IEEE 11th ICCV*, Rio de Janeiro, Brazil, pp.1–6, Oct. 2007.
 - [18] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, “Segmenting salient objects from images and videos,” *Proc. 11th ECCV*, Heraklion, Greece, vol.6315, pp.366–379, Sept. 2010.
 - [19] F. Zhou, S.B. Kang, and M.F. Cohen, “Time-mapping using space-time saliency,” *Proc. IEEE Conf. CVPR*, Columbus, USA, pp.3358–3365, June 2014.
 - [20] Z. Liu, X. Zhang, S. Luo, and O.L. Meur, “Superpixel-based spatiotemporal saliency detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol.24, no.9, pp.1522–1540, Sept. 2014.
 - [21] T. Brox and J. Malik, “Large displacement optical flow: Descriptor matching in variational motion estimation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.33, no.3, pp.500–513, March 2010.
 - [22] L. Hu, Z. Wang, M. Ye, J. Xiao, and R. Hu, “Spatiotemporal saliency based on location prior model,” *IEEE IJCNN*, Vancouver, Canada, pp.2526–2533, July 2016.



Yu Chen received the B.S. degree from the School of Electronic Information, Wuhan University, China, in 2013. He is currently pursuing the Ph.D. degree in the School of Computer, Wuhan University, China. His current research interests include video coding and image processing.



Jing Xiao received the Ph.D. degree from Faculty of Geo-Information Science and Earth Observation, Twente University, Enschede, The Netherlands in 2013. She is currently an associate professor at School of Computer, Wuhan University, Wuhan, China. Her research interest is video coding and processing.



Liuyi Hu received the B.S. degree from the School of Electronic Information, Wuhan University, China, in 2014. He is currently pursuing the M.S. degree in the School of Computer, Wuhan University, China. Her current research interests include video coding and image processing.



Dan Chen is currently a professor at School of Computer Science, Wuhan University, Wuhan, China. He was a HEFCE Research Fellow with the University of Birmingham, U.K. His research interests include data science and engineering, high performance computing, and modelling and simulation of complex systems.



Zhongyuan Wang received the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 2008. Dr. Wang is now a professor with School of Computer, Wuhan University, Wuhan, China. His research interests include video compression, image processing, and multimedia communications etc.



Dengshi Li received the M.S. degree from Central China Normal University. She currently works at Jiangnan University and serves as an associate professor. She is also a Ph.D. candidate of the National Engineering Research Center for Multimedia Software, Wuhan University. Her research interests are sound field reproduction, audio and video coding and signal processing.