Video saliency detection by spatio-temporal sampling and sparse matrix decomposition

Yunfeng Pan, Qiuping Jiang, Zhutuan Li, Feng Shao*
Faculty of Information Science and Engineering
Ningbo University
Ningbo 315211
CHINA
shaofeng@nbu.edu.cn

Abstract: - In this paper, we present a video saliency detection method by spatio-temporal sampling and sparse matrix decomposition. In the method, we sample the input video sequence into three planes: X-T slice plane, Y-T slice plane, and X-Y slice plane. Then, motion saliency map is extracted from the X-T and Y-T slices, and static saliency map is extracted from the X-Y slices by low-rank matrix decomposition. Finally, these maps are retransformed into the X-Y image domain and combined with central bias prior to obtain the video saliency maps. Extensive results on ASCMN dataset demonstrate that the proposed video saliency model can achieve higher subjective and objective performances.

Key-Words: - saliency detection, spatio-temporal sampling, sparse matrix decomposition, motion saliency, static saliency

1 Introduction

Visual attention plays an important role in various visual applications by allocating the limited computational resources to those perceptually significant regions. Through detecting saliency map, we can assign relatively high saliency values to those perceptually significant regions while low saliency values to the other regions. The existing saliency methods can be classified into two categories: eye-fixation prediction and salient object detection. Currently, salient object detection models are widely applied in various computer vision and image processing tasks, such as object tracking [1], image/video segmentation [2], image/video quality assessment [3], content-aware compression [4], and image/video retargeting [5].

The state-of-the-art saliency detection models can be categorized into two groups: top-down and bottom-up models. Numerous researches have been conducted on this aspect and many bottom-up saliency detection models were developed [6-14]. Compared with image saliency detection, video saliency detection is much more challenging in the detection and utilization of temporal and motion information. More video saliency detection methods try to combine spatial and temporal cues [15-25]. As an effective way for motion saliency detection, the solutions for separating foreground objects from backgrounds have been proposed [26].

In this paper, we proposed a novel video saliency detection method by spatio-temporal sampling and sparse matrix decomposition. In the method, the input video sequence is first sampled into three planes: X-T slice plane, Y-T slice plane, and X-Y slice plane. Then, motion saliency map is extracted from the X-T and Y-T slices, and static saliency map is extracted from the X-Y slices. Finally, these maps are retransformed into the X-Y image domain and combined with central bias prior to obtain the final video saliency maps. Subjective and objective results on ASCMN dataset demonstrate that the proposed saliency model can achieve a consistently higher saliency detection performance than the state-of-the-art saliency models.

The rest of this paper is organized as follows. Section 2 reviews the relevant works in saliency detection. Section 3 describes the proposed video salient detection model. Experimental results are presented and analyzed in Section 4, and the conclusions are drawn in Section 5

2 Related work

In the past decades, numerous saliency detection models have been proposed to highlight the salient objects based on various low-level and high-level properties. The most representative work is the biologically plausible saliency detection model proposed by Itti *et al*., in which various simple low-level feature (e.g., color, intensity and orientation) are integrated using center-surround mechanism [6]. Inspired by Itti's model, various bottom-up saliency detection models were devised. Hou *et al*. [7]

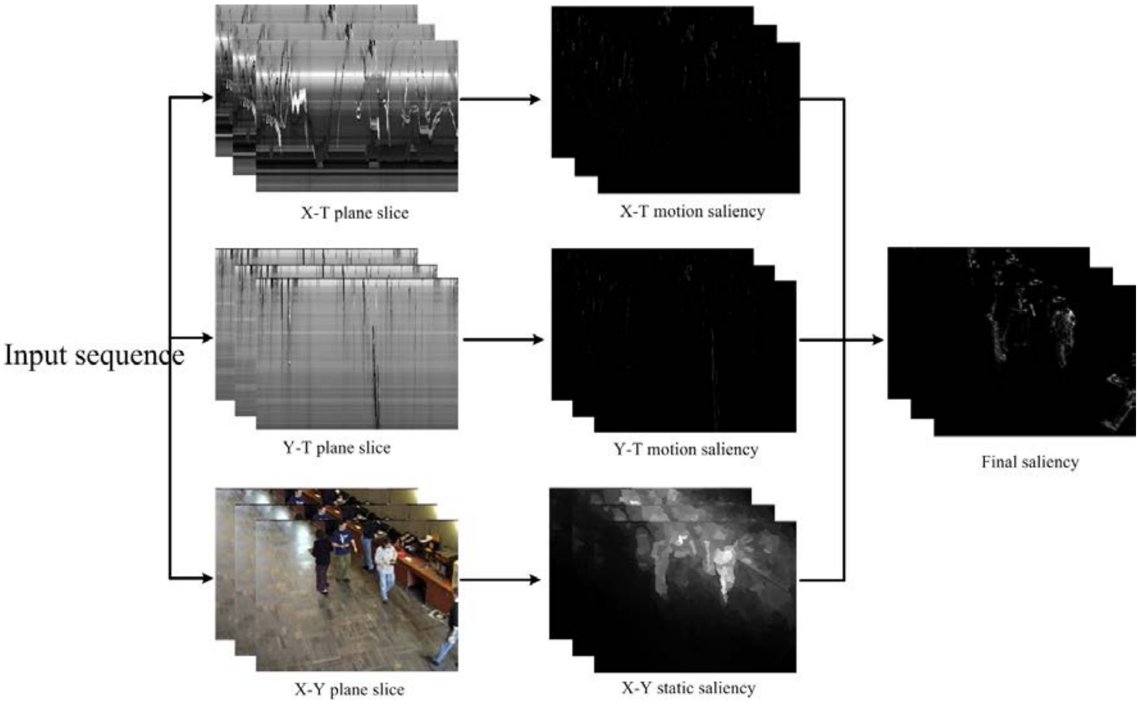


Fig.1. Framework of the proposed video saliency detection model.

proposed a Fourier spectrum residual analysis to detect salient regions based on the log magnitude spectrum representation of image. Cheng *et al*. [8] segment the input image into sub-regions, and calculate the saliency for each region by comparing global region contrast to all other regions in the image. Harel *et al*. [9] proposed a Graph-based algorithm by redefining the edge weights based on the feature dissimilarity and the spatial proximity between the two connected nodes. Achanta *et al*. [10] computed the conspicuity likelihood of each pixel based on its color contrast over the whole image. Erdem *et al*. [11] constructed covariance matrix for saliency estimation by incorporating low-level features. Margolin *et al*. [12] integrated both pattern and color distinctiveness for salient object detection. Many other models can be found in [13-14].

Recently, various video saliency detection models have been proposed, which is important to consider the inherent temporal information in addition to the spatial features. Some methods [15,16] perform saliency detection between two consecutive frames and take difference between frames as motion feature. Optical flow is the most widely used motion detection methods [17,18]. In some compressed domain video saliency detection models [19,20], the motion feature is extracted from motion vectors at the decoder. Other methods utilize the spatio-temporal information across frames without explicit motion extraction. Cui *et al*. [21] directly applied the spectrum residual analysis to the X-T plane and Y-T plane. Video saliency is also related to background subtraction. Gaussian mixture models can also be used to detect saliency [22]. Mahadevan *et al*. [23] took spatio-temporal video patches as dynamic textures and produced a saliency map by classifying non-salient points as background. Many other models can be found in [24-25].

To our knowledge, although a number of reports [15-25] has been appeared in video saliency detection, how to extract the temporal information and how to combine spatial and temporal cues effectively still remains not fully investigated. In this paper, we try to tackle the issue by spatio-temporal sampling and sparse matrix decomposition.

3 Proposed video saliency detection method

Fig.1 presents the framework of the proposed video saliency detection model. For an input video sequence, it is first sampled into three planes: X-T slice plane, Y-T slice plane, and X-Y slice plane. The produced slices contained different visual cues for saliency detection, e.g., the X-T and Y-T slices produce temporal saliency maps, and X-Y slice produce spatial saliency maps. After calculating the saliency maps in the different planes, these maps are

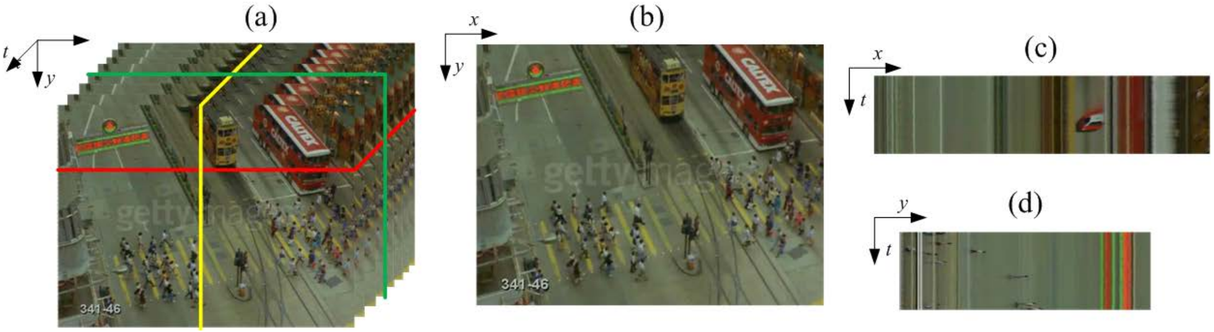


Fig.2. Examples of 3D volume and the resultant slices: (a) 3D volume of the input frames; (b) X-Y slice; (c) X-T slice; (d) Y-T slice.

retransformed into the X-Y image domain and combined with central bias prior. The important features of the proposed video saliency detection model are that: low-rank matrix decomposition is used to recover the background and motion information from the slices.

3.1 Motion saliency extraction

The motion feature is extracted by sampling spatio-temporal slices (X-T and Y-T slices) from a pixel volume (X-Y-T), which is obtained by stacking T input frames. After this step, three stacks of 2D plane images are available: the X-Y stack with T slices (to be analyzed in the next subsection), the X-T stack with Y slices, and the Y-T stack with X slices. The sampled temporal slices (X-T and Y-T) contain enough motion behavior of objects, such as moving in the horizontal or vertical direction. As an example, the 3D volume and the resultant slices are shown in Fig.2. It is obvious that the moving objects appeared in the X-T and Y-T slices are intuitive.

In order to extract motion cues from the X-T and Y-T slices, the most direct means is to separate the moving objects from the slices (as done in region-based visual attention [27]). Observed from Figs.2(c) and (d), static background produces straight lines in the X-T and Y-T slices, while the moving object produces a motion signature. Since our aim is to separate the moving objects from these backgrounds, low-rank matrix decomposition is an effective way to recover the backgrounds, and thus separate the moving objects from backgrounds, which assumes that the background pixels generally demonstrate similar appearance.

We give a short overview of the low-rank matrix decomposition. For a feature matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, it can be decomposed into two parts, a low-rank matrix **D** and a sparse one **E**

$$\mathbf{A} = \mathbf{D} + \mathbf{E} \tag{1}$$

Applying this model to saliency detection, the background is naturally represented by low-rank matrix **D**, and moving the objects might be captured by the sparse matrix **E**.

To recover the matrix **D** and **E**, the above problem can be formulated by [28]

$$(\mathbf{D}^*, \mathbf{E}^*) = \min \;\; rank(\mathbf{D}) + \lambda \|\mathbf{E}\|_0 \quad s.t.\ \mathbf{A}=\mathbf{D}+\mathbf{E} \tag{2}$$

where λ is a coefficient to balance **D** and **E**, and $\|\cdot\|_0$ indicates l_0-norm. However, such problem is intractable as the matrix rank and l_0-norm are not convex. Thus, the problem is resolved through another way

$$(\mathbf{D}^*, \mathbf{E}^*) = \min \;\; \|\mathbf{D}\|_* + \lambda \|\mathbf{E}\|_1 \quad s.t.\ \mathbf{A}=\mathbf{D}+\mathbf{E} \tag{3}$$

where $\|\cdot\|_*$ is the nuclear norm of matrix **D** (the sum of singular values of **D**), and $\|\cdot\|_1$ indicates l_1-norm. The optimal matrices $\mathbf{D}^*$ and $\mathbf{E}^*$ can be obtained by solving the above formulation via alternative iterations. More detail about the low-rank sparse matrix decomposition can be found in [28].

Based on the above formulation, each X-T and Y-T slice **S** is decomposed respectively as

$$(\mathbf{B}^*, \mathbf{M}^*) = \min \;\; \|\mathbf{B}\|_* + \lambda \|\mathbf{M}\|_1 \quad s.t.\ \mathbf{S}=\mathbf{B}+\mathbf{M} \tag{4}$$

The above process is repeated for each slice. Then, the Y motion matrices of X-T slices and X motion matrices of Y-T slices are grouped into 3D volumes (X-Y-T), respectively, i.e., $\mathbf{V}_{X-T}$ and $\mathbf{V}_{Y-T}$, and then integrated into a 3D volume as $norm(\mathbf{V}_{X-T} \cdot * \mathbf{V}_{Y-T})$, where '$\cdot *$' is the element-wise product operator, and $norm(\cdot)$ denotes normalization processing. The final temporal saliency maps are obtained by sampling the 3D volume in X-Y image domain. Figs.3(a) and (b) show the motion saliency detection results on the temporal slices in Figs.2(c)

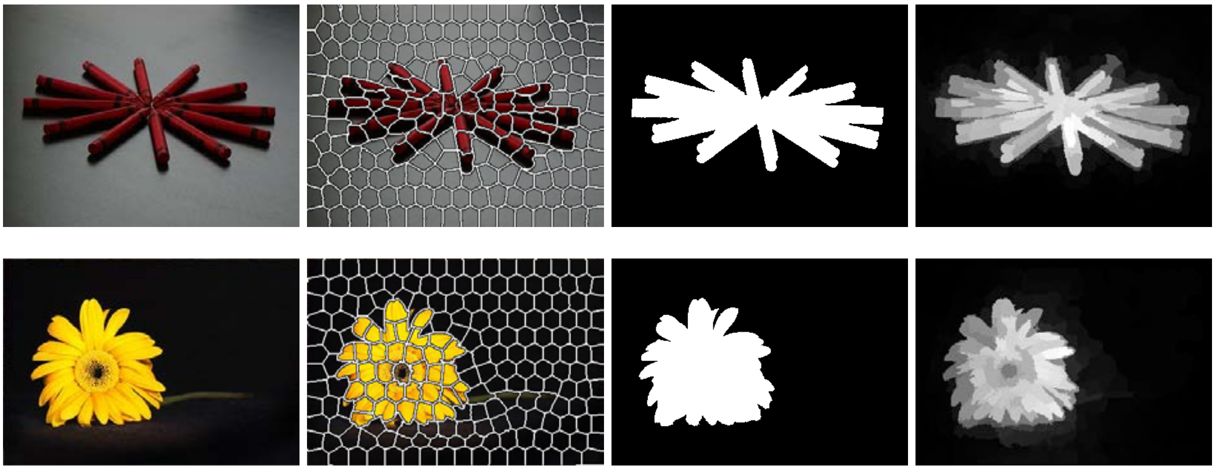
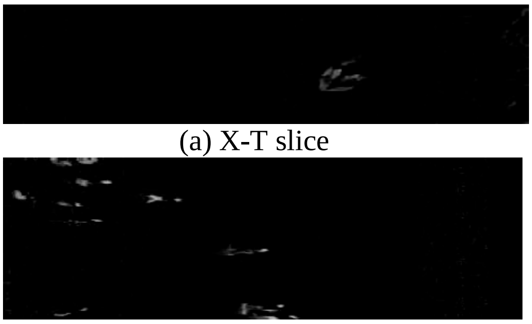


Fig.4. Examples of saliency detection: (a) Original image; (b) superpixel segmentation result, (c) ground truth mask, and (d) the final saliency map.

and (d), respectively (To favor display, the contrast of the maps is adjusted.). As shown in the figure, the proposed method can precisely capture the motion.



(a) X-T slice

(b) Y-T slice

Fig.3. Example of motion saliency detection on a temporal slice.

3.2 Static saliency extraction

The most important operation in image saliency detection is to integrate different features or priors to detect saliency. For the X-Y slices, some common used low-level features (e.g., color, texture and spatial features) are extracted. For simplicity, the color features are represented by 3 components in CIELab color space, the texture features are represented by Gabor filter responses with 3 scales and 12 orientations, and the spatial features are represented by horizontal and vertical locations. Then, all these features are formed a 41 dimensional feature vector for each pixel, which captures color, texture and spatial information that are most common low-level visual features.

Then, we select these extracted features to over-segment the X-Y plane slice so that the non-salient background regions can also contain multiple segments (to increase background prior). We advocate superpixels as basic units in saliency detection. Given an input image, we employ SLIC segmentation algorithm [29] to extract superpixels, which generates superpixels by clustering pixels based on their feature similarity. In the experiment, the number of superpixels is set to 200. The image accordingly is decomposed into N superpixels $\{p_i\}_{i=1,\ldots,N}$, where N is the number of superpixels. For each superpixel p_i, let $\mathbf{f}_i \in \mathbb{R}^{d\times 1}$ be the mean feature vector of the superpixel, where d is the dimension of feature description. The X-Y slice is represented by a matrix $\mathbf{F} = \left[\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N\right] \in \mathbb{R}^{d\times N}$.

Similar as in Eq.(4), we decompose the feature matrix $\mathbf{F}$ into a low-rank matrix and a sparse one

$$(\mathbf{U}^*, \mathbf{E}^*) = \min \ \|\mathbf{U}\|_* + \lambda\|\mathbf{E}\|_1 \qquad s.t.\ \mathbf{F}{=}\mathbf{U}{+}\mathbf{E} \tag{5}$$

With the optimal sparse matrix $\mathbf{E}^*$, the saliency value of the superpixel is given by the l_1-norm.

$$s_i = \left\|\mathbf{e}_i^*\right\|_1 \tag{6}$$

where $\mathbf{e}_i^*$ is the feature vector of the superpixel in the sparse matrix $\mathbf{E}^*$. The saliency value s_i represents the probability of the superpixel belonging to an object, i.e., larger value for higher probability, and vice versa. Finally, the saliency map of the X-Y slice is generated by assigning the saliency value for all pixels in the superpixel. See Fig.4 as example, superpixel segmentation can generate regular-sized regions with better boundary contour, and the proposed method can produce higher saliency along the edges of the object.

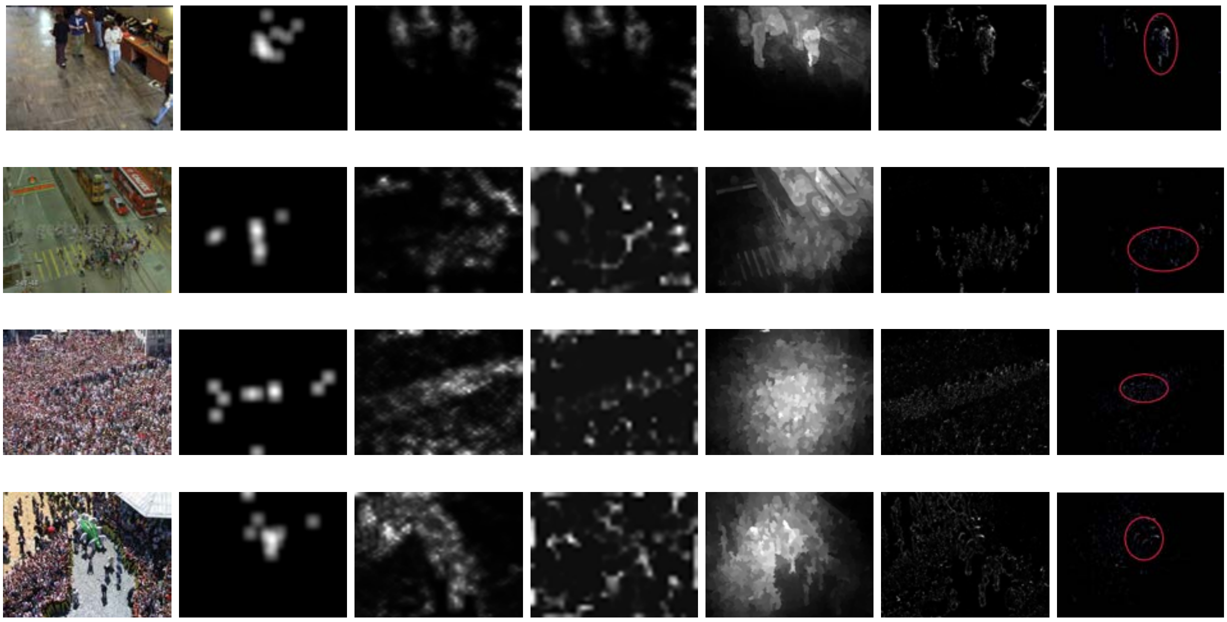


Fig.5. Comparison of saliency maps from different models.

3.3 Combining with central bias

By integrating the above motion and static saliency maps (to reflect temporal and spatial visual cues), the final saliency for each pixel (*x*,*y*) is defined as

$$Sal(x,y) = S_{motion}(x,y) \cdot S_{static}(x,y) \qquad (7)$$

Based on the results of the literatures and subjective experiments, the viewers tend to focus on the central fixation location than other locations. That is, pixels located near to the center may provide more information than the other pixels, thus becoming more salient. In this paper, central bias is modeled by 2D Gaussian with the strong central fixation distribution on the center and then spreads to the neighbors

$$CB(x,y) = \exp\left\{-\left(\frac{(x-x_c)^2}{2\sigma_x^2} + \frac{(y-y_c)^2}{2\sigma_y^2}\right)\right\} \qquad (8)$$

where (x_c, y_c) is the center of the image, σ_x^2 and σ_y^2 are the variance along the two directions respectively. In the experiment, σ_x^2 is set to $0.5W_{im}$ and σ_y^2 is set to $0.5H_{im}$.

4 Experimental results and analyses

In order to evaluate the performance of the proposed video saliency detection model, we conduct experiments on commonly used ASCMN dataset [30]: it contains 24 videos (5 classes: Abnormal, Surveillance, Crowd, Moving, Noise), together with eye tracking data from 13 viewers. For similarity measures, we use the Normalized Scanpath Saliency (NSS), and the area under the Receiver Operating Characteristics Curve (AUC-ROC). These measures have been designed to use fixations. We compare our method with two methods: Self-Resemblance model (SR) [31], and Bayesian Surprise model (BS) [32]. For the existing methods, we use the source codes or executable codes provided by the authors.

Subjective comparison of the proposed method with the state-of-the-art methods are shown in Fig.5. The original 100-th images with ground-truth fixation data are shown in Figs.5(a)-(b), while the results of the two state-of-the-art methods are presented in Figs.5(c)-(d). The spatial and temporal saliency maps obtained by the proposed method are given in Figs.5(e)-(f), and the final combination is given in Fig.5(g). From the results of SR and BS in Figs.4(c) and (d), we find that for those uncertain motion regions, these two methods cannot perform well (deviate from the correct fixations), while the proposed method can correctly locate the fixations in video (marked by red circles in the figures).

We also accomplish an objective comparison by measuring the effectiveness of the extracted saliency maps from different methods with the ground-truth fixation data as criterion. Table.1 shows the AUC-ROC and NSS values for different saliency detection models. From Table.1, we can see that the AUC-ROC and NSS values of the proposed model are larger than those of SR [31] and BS [32], and thus the performance of the proposed model is the best among these compared models. Since the proposed model considers both spatial and temporal cues for saliency calculation, it achieves better performance than others.

Table 1 Comparison results for different video saliency detection models.

Models	SR [31]	BS [32]	Proposed
AUC-ROC	0.7154	0.6897	0.7425
NSS	0.4240	0.5551	0.4309

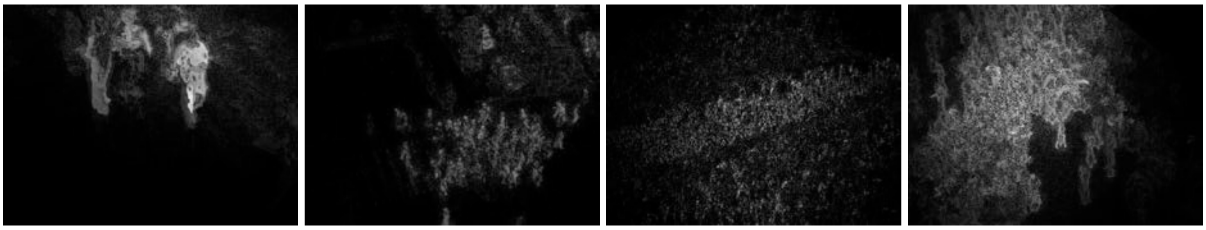


Fig.6. The saliency maps combining with additive fusion.

We also investigate the performance of another fusion strategy (i.e., additive combination) for video saliency detection. Fig.6 shows the saliency maps combining with additive fusion strategy for the same test images in Fig.5. The weight between temporal and spatial saliency is obtained by optimal training. We can see that the additive combination strategy cannot get good performance. Since the distributions of temporal and spatial saliency maps are somewhat different, the additive combination cannot suppress each other.

However, the proposed method has the following limitations: 1) the salient motion regions are not very prominent in some video sequences, because the assumption of background in low-rank matrix decomposition may be not appropriate; 2) the computation complexity of the proposed method is somewhat high, because videos should be resampled first and retransformed. These limitations should be properly considered in the future work.

5 Conclusion

In this paper, we presented a novel video saliency detection method by spatio-temporal sampling and sparse matrix decomposition. The main features of the proposed method are that: (1) we sample the input video sequence into three planes: X-T slice plane, Y-T slice plane, and X-Y slice plane; (2) we use sparse matrix decomposition to recover the background and motion features from the slices; (3) motion and static saliency maps are multiplicatively combined with central bias prior. Experimental results show the effectiveness of the proposed method.

Although the proposed scheme exhibit good performance in saliency detection, some aspects still deserve further research and improvement: 1) the connectivity of foreground and background should be further considered; 2) more low-level and high-level feature cues should be incorporated; 3) computational complexity should be an important indictor of the proposed method.

Acknowledgement:
This work was supported by the Natural Science Foundation of China (grant 61271021), and the Natural Science Foundation of Ningbo (grant 2012A610039).

References:

[1] N. J. Butko, L. Y. Zhang, G. W. Cottrell, and J. R. Movellan, “Visual saliency model for robot cameras,” in Proc. of IEEE International Conference on Robotics and Automation, pp. 2398-2403, 2008.

[2] Q. Zhang, K. N. Ngan, “Multi-view video based multiple objects segmentation using graph cut and spatiotemporal projections,” Journal of Visual Communication and Image Representation, vol. 21, no. 5, pp. 453-461, 2010.

[3] Z. Wang, Q. Li, “Information content weighting for perceptual image quality assessment,” IEEE Transactions on Image Processing, vol. 20, no. 5, pp. 1185-1198, May 2011.

[4] Z. Li, S. Qin, and L. Itti, “Visual attention guided bit allocation in video compression,” Image and Vision Computing, vol. 29, no. 1, pp. 1-14, Jan 2011.

[5] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, “Retargeting images and video for preserving information saliency,” IEEE Computer Graphics and Applications, vol. 27, no. 5, pp. 80-88, 2007.

[6] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene

analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, 1998.

[7] X. Hou, L. Zhang, "Saliency detection: a spectral residual approach," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[8] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, S. Hu, "Global contrast based salient region detection," in Proc. of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 409-416, 2011.

[9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in Proc. of Advances in Neural Information Processing Systems, pp. 545-552, 2006.

[10] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency tuned salient region detection," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1597-1604, 2009.

[11] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariance," Journal of Vision, vol.13, no. 4, article 11, 2013.

[12] R. Margolin, A. Tal, and L. Manor, "What makes a patch distinct?" in Proc. of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1139-1146, 2013.

[13] W. Zhang, Q. Xiong, and S. Chen, "Data-driven saliency region detection based on undirected graph ranking," WSEAS Transactions on Computers, vol. 13, pp. 310-319, 2014.

[14] X. Lv, D. Zou, L. Zhang, and S. Jia, "Feature coding for image classification based on saliency detection and fuzzy reasoning and its application in elevator videos," WSEAS Transactions on Computers, vol. 13, pp. 266-276, 2014.

[15] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in Proc. of ACM Multimedia, pp. 815-824, 2006.

[16] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in Proc. of European Conference on Computer Vision, pp. 366-379, 2010.

[17] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1293-1300, 2010.

[18] S. Mathe, C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in Proc. of European Conference on Computer Vision, pp. 842-856, 2012.

[19] Y. Fang, W. Lin, Z. Chen, C. Tsai, C. Lin, "A video saliency detection model in compressed domain," IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 1, pp. 27-38, 2014.

[20] K. Muthuswamy, and D. Rajan, "Salient motion detection in compressed domain," IEEE Signal Processing Letters, vol. 20, no. 10, pp. 996-999, 2013.

[21] X. Cui, Q. Liu, and D. Metaxas, "Temporal spectral residual: fast motion saliency detection," in Proc. of ACM Multimedia, pp. 617-620, 2009.

[22] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in Proc. of International Conference on Pattern Recognition, pp. 28-31, 2004.

[23] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 171-177, 2010.

[24] Y. Luo, J. Yuan, "Salient object detection in videos by optimal spatio-temporal path discovery," in Proc. of ACM Multimedia, pp. 509-512, 2013.

[25] Y. Xue, X. Guo, X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1485-1488, 2012.

[26] W.-T. Li, H.-S. Chang, K.-C. Lien, H.-T. Chang, Y. F. Wang, "Exploring visual and motion saliency for automatic video object extraction," IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2600-2610, 2013.

[27] J. Tünnermann, B. Mertsching, "Region-based artificial visual attention in space and time," Cognitive Computation, vol. 6, no. 1, pp 125-143, 2014.

[28] E. J. Candès, X. Li, Y. Ma, J. Wright, "Robust principal component analysis," Journal of the ACM, vol. 58, no. 3, Article No. 11, 2011.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274-2282, 2012.

[30] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit, "Dynamic saliency models and human

attention: a comparative study on videos," in Proc. of Asian Conference on Computer Vision, pp. 586-598, 2013.

[31] L. Itti, P. Baldi, "A principled approach to detecting surprising events in video," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 631-637, 2005.

[32] H. J. Seo, P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," Journal of Vision, vol. 9, no. 12, pp. 1-17, 2009.