MOTION SALIENCY DETECTION USING LOW-RANK AND SPARSE DECOMPOSITION

*Yawen Xue§,†, *Xiaojie Guo§ and Xiaochun Cao§*

{yxue, xguo, xcao}@tju.edu.cn
§ School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
†School of Computer Software, Tianjin University, Tianjin 300072, China

ABSTRACT

Motion saliency detection has an important impact on further video processing tasks, such as video segmentation, object recognition and adaptive compression. Different to image saliency, in videos, moving regions (objects) catch human beings' attention much easier than static ones. Based on this observation, we propose a novel method of motion saliency detection, which makes use of the low-rank and sparse decomposition on video slices along X-T and Y-T planes to achieve the goal, *i.e.* separating foreground moving objects from backgrounds. In addition, we adopt the spatial information to preserve the completeness of the detected motion objects. In virtue of adaptive threshold selection and efficient noise elimination, the proposed approach is suitable for different video scenes, and robust to low resolution and noisy cases. The experiments demonstrate the performance of our method compared with the state-of-the-art.

***Index Terms*—** Motion Saliency Detection, Low-rank and Sparse Decomposition, Video Analysis

1. INTRODUCTION

Visual attention analysis provides an intuitive methodology to semantic content understanding and important information capture in both images and videos. Most primates, including humans, can divert their mind subconsciously to the salient objects in images or to the motion objects in videos. Such a remarkable ability leads to that they can sample the most "interesting" features and interpret complex scenes while spending limited processing. In other words, visual saliency makes a distinguishing region stand out and thus catch special attention quickly, which provides an alternative solution to many tasks, such as video segmentation [1], adaptive content delivery [2] and adaptive compression [3].

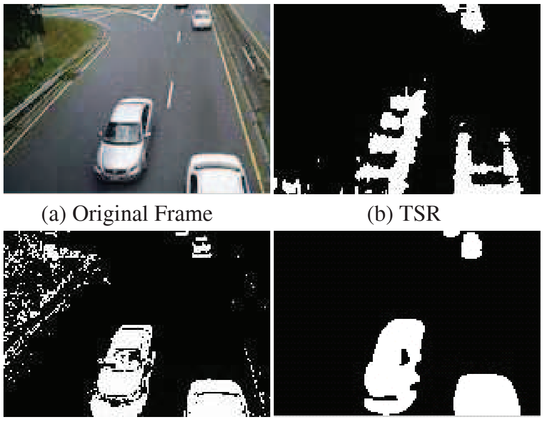

(a) Original Frame (b) TSR
(c) GMM (d) Our Method

Fig. 1. Example of result comparison between the proposed method and the state-of-the-art. (a) Original frame, (b)-(d) are the results obtained by Temporal Spectrum Residual [4], GMM [5] and our method, respectively.

In last decade, saliency detection has attracted much attention. For static images, the widely used model for capturing salient regions is introduced by Itti *et al.* [6], which breaks down such a complex problem of saliency detection into several blocks. The core of their designed model is feature selection (*e.g.* color, texture, local or global contrast). Following this model, a number of techniques have beem developed in the literature. For instance, Ma and Zhang [2] propose an approach to attention area detection in images based on local contrast analysis and fuzzy growing. Achanta *et al.* adopt low-level features of luminance and color to generate saliency maps [7]. Graph-Based Visual Saliency [8] computes activation maps on certain feature channels and then normalizes and combines them to form the final saliency map. Hou and Zhang [9] propose a Fourier spectrum residual analysis method to compute the regions that attract humans' attention in images. In [10], the authors adopt a two-stage method to accomplish the task, which is actually an extension of [9]. More recently, Cheng *et al.* [11] take into account several features simultaneously including regional contrast, global contrast and spatial coherence.

To detect motion saliency in videos, however, most of the

⋆ THESE AUTHORS CONTRIBUTED EQUALLY TO THIS WORK. THIS WORK WAS SUPPORTED BY NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA (NO. 60905019), PROGRAM FOR NEW CENTURY EXCELLENT TALENTS IN UNIVERSITY, TIANJIN KEY TECHNOLOGIES R&D PROGRAM (NO. 11ZCKFGX00800) TSINGHUA-TENCENT JOINT LABORATORY FOR INTERNET INNOVATION TECHNOLOGY, AND SKL OF CG&CAD.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

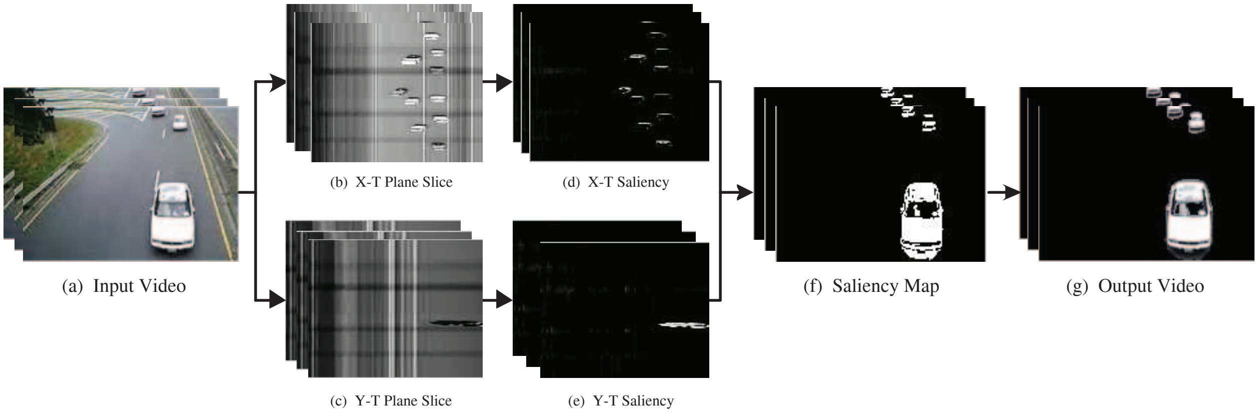


Fig. 2. Illustration of the main stages of the proposed method

techniques for images (mentioned above) are not available. Since, different with image saliency detection, moving regions (objects) alternatively catch more human beings' attention than static ones, even though which have large contrast to their neighbors in static images. That is to say, the focal point changes from the regions with large contrast to their neighbors for images to those with motion discrimination for videos. Therefore, the contrast based methods are hardly applied to videos directly. An exception exists in [4], which extends the spectrum residual analysis [9] in images to videos. Actually, the goal of moving object separation from background is the same as that of motion saliency detection. A few solutions for separating foreground moving objects from backgrounds have been proposed, such as Gaussian Mixture Model [5]. In this work, we introduce a novel method to detect motion saliency by using low-rank and sparse decomposition. Prior to detailing the stages of our proposed method, we first post an example of the performance comparison between our method and the state-of-the-art in Fig. 1. As can be seen, the result obtained by our method is significantly better than the others. (More experiments and results can be found in Sec. 4.)

2. LOW-RANK AND SPARSE DECOMPOSITION

Suppose we have a data matrix $\mathbf{A} \in \mathbb{R}^{n*m}$, and know that it can be decomposed as $\mathbf{A} = \mathbf{D} + \mathbf{E}$, where $\mathbf{D}$ has low rank and $\mathbf{E}$ is sparse. Both $\mathbf{D}$ and $\mathbf{E}$ are of arbitrary magnitude.

Recall that the Classical Principle Component Analysis (CPCA) seeks the rank-k estimate of $\mathbf{D}$ to approximate $\mathbf{A}$, or to say, reduce the dimensionality of the observations in $\mathbf{A}$ by optimally solving:

$$\min ||\mathbf{A} - \mathbf{D}||, \;\; s.t. \;\; rank(\mathbf{D}) \leq k, \qquad (1)$$

where $|| * ||$ stands for the 2-norm, *i.e.* the largest singular value of it. Note that CPCA is under the assumption that the observations in $\mathbf{A}$ are polluted by noise slightly, *i.e.* the absolute values of the elements in $\mathbf{E}$ are small. Otherwise, the solution of CPCA is far away from the optimal $\mathbf{D}$. However, in real world problems, data pollution is ubiquitous and arbitrary. As a consequence, CPCA may somewhat lose its power to deal with many real world problems.

To overcome the drawback of CPCA, Robust Principal Component Analysis (RPCA) is proposed by Candès *et al.* [12]. The goal of RPCA is to optimize the problem:

$$\min rank(\mathbf{D}) + ||\mathbf{E}||_0, \;\; s.t. \;\; \mathbf{A} = \mathbf{D} + \mathbf{E}, \qquad (2)$$

where $|| * ||_0$ denotes the ℓ_0-norm. But such a problem is intractable in polynomial-time. Instead, one can solve its convex relaxation as follows:

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

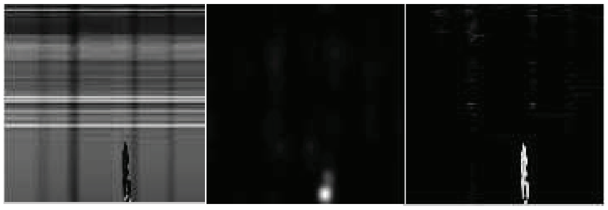
where λ is the coefficient controlling the weight of the sparse matrix $\mathbf{E}$, and $|| * ||_*$ and $|| * ||_1$ represent the nuclear norm and the ℓ_1-norm of the matrix, respectively. This formulation performs well in practice, which recovers the true low-rank solution even when up to a third of the observations are grossly corrupted. More detail about the proof of the low-rank and sparse decomposition using RPCA can be found in [12].

3. OUR METHOD

In this section, we first formulate the problem this work intends to solve, and then detail the stages of our proposed method (Fig. 2).

Problem Formulation. Different to static image saliency detection, the motion regions in a video intensively attract humans' attention instead of the regions with large contrast in every single image. And due to the correlation between frames, the motion regions in the video[1] can be identified from the background by low-rank and sparse decomposition. Note that foreground motion objects, such as cars and pedestrians, usually occupy only a fraction of the image pixels and hence can be treated as sparse errors. In this work, we stack

[1] Assume the background of the video is stationary or with small flutter.



(a) Y-T Plane Slice (b) TSR (c) Our Method

Fig. 3. Visual comparison of the middle results for motion saliency detection on a temporal slice. (a) shows a Y-T plane slice. (b) and (c) are the results of detecting motion saliency on (a) using TSR [4] and our method, respectively.

the temporal slices along X-T and Y-T as the matrices $\mathbf{S}$. Naturally, the low-rank component $\mathbf{B}$ corresponds to the background and the sparse component $\mathbf{M}$ captures the motion objects in the foreground. Figure 3 (a) shows a temporal slice to confirm our observation. As shown in Fig. 3 (b) and (c), our method significantly outperforms TSR in terms of capturing the motion. Moreover, we take adaptive threshold selection and refinement to reduce the effect of noise and missing pixels (because of the ignore of spatial consideration).

Decomposition. Based on the problem formulation, each X-T and Y-T slice $\mathbf{S}$ is decomposed, similar with Eq. 3, as:

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

Then, the motion matrices, *i.e.* $abs(\mathbf{M})$, obtained from the X-T (Y-T) slices are integrated together as S_{cubeXT} (S_{cubeYT}) along X-Y-T. Then, using $norm(S_{cubeXT} .* S_{cubeYT})$ to form the initial saliency map S_{cube}, where $.*$ is the element-wise product operator, and $norm(*)$ represents normalization processing. The size of T in our experiments equals the size of the video, it also can be defined as the size of a sub-video.

Refinement. To reduce the effect of missing pixels on the motion objects and refine the results, we further take into account the spatial information. Intuitively, the pixels belonging to the same motion object are always locally coherent. This indicates that a pixel $p_{i,j,k}$ is very likely to be missing salient pixel when its neighbors in frame k are motion salient. Inspired by this observation, we use a Gaussian function to recall the missing pixels as follows (we omit subscript k for short):

$$S_{cube}(i,j) = \sum_{||p_{x,y} - p_{i,j}||_2 < \tau} S_{cube}(x,y) * f(||p_{x,y} - p_{i,j}||_2), \tag{5}$$

where τ is the radius of the support region centered on $p_{i,j}$, $|| * ||_2$ is the ℓ_2-norm, and f is a Gaussian function: $f(d) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{d^2}{2\sigma^2}$.

Adaptive Threshold Selection. The procedure of optimizing Eq. 4 may bring some noise, which means some salient pixels with small absolute values should belong to the background. To handle this problem, we employ an adaptive threshold selection step to eliminate the noise. Similar to [4], we assume the distribution of the values of salient pixels in S_{cube} satisfies the Gaussian distribution (μ, σ). Therefore, we adaptively adopt $T_{global} = \mu + \sigma$ as the global threshold to eliminate the noise in S_{cube}:

$$S_{cube}(x,y,t) = \begin{cases} 1, \ \ if S_{cube}(x,y,t) \geq T_{global}, \\ 0, \qquad\qquad\qquad\qquad otherwise. \end{cases} \tag{6}$$

Actually, the relative small regions are rejected in our implementation based on the observation that the small motion regions are likely to be false positive, and even human eyes hardly capture the tiny motion objects. Therefore we employ a threshold of the region size $T_{size} = (h * w)/1500$ to reject the regions too small, where h and w are the height and the width of video frame, respectively.

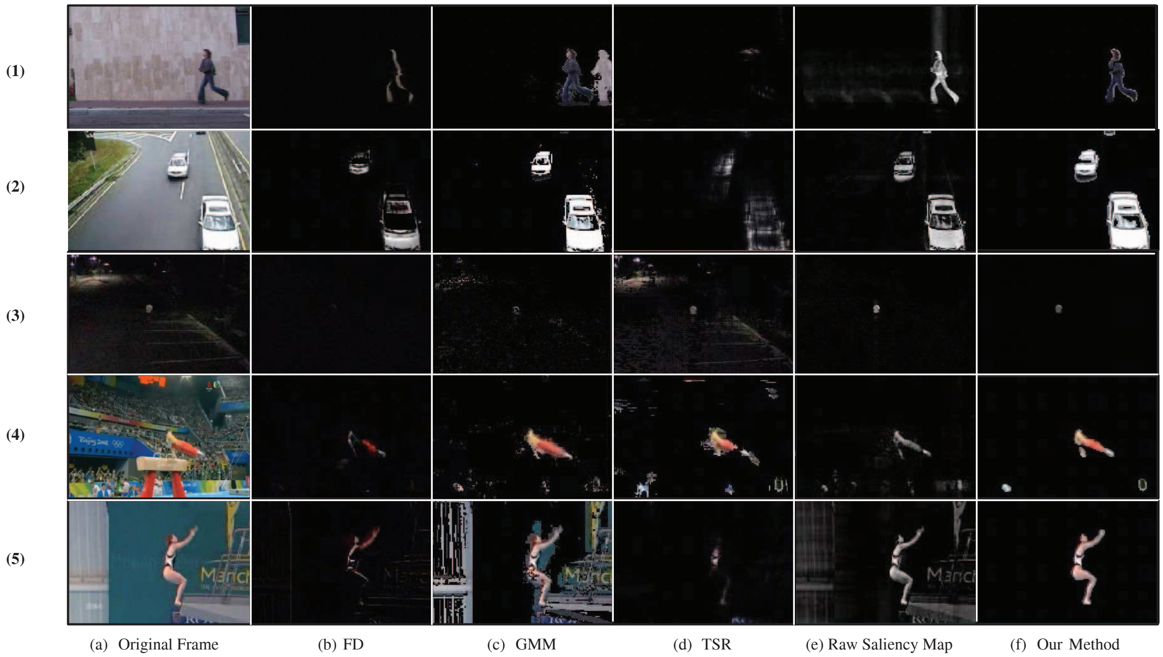
4. EXPERIMENTS

To reveal the performance of our method, we compare our proposed method with the state-of-the-art, including Frame Difference (FD), Gaussian Mixture Model (GMM) [5] and Temporal Spectrum Residual (TSR) [4]. Five types of videos are carried on: (1) single motion object, (2) multiple motion objects, (3) low quality, (4) cluttered background and (5) moving camera with repeating background[2].

The implementation involves three input parameters including λ in decomposition, and τ and σ in refinement. In our experiments, we use $\lambda = 0.3$, $\tau = 5$ and $\sigma = 1$ uniformly.

Figure 4 gives the experimental results. As can be seen, the simple method like frame difference is not able to obtain reasonable results and sensitive to noise as shown in Fig. 4 (b). With respect to GMM, for the first video sequence, although the major motion object (human) is discovered, there exists ghost, *i.e.* because of GMM modeling the background through first several frames, the performance of GMM during the background modeling is unreliable. Once it is well modeled, GMM performs well as shown in the second row of Fig. 4 (c). The third video is of low quality, the result of which is significantly ruined by noise. In the fourth case, the recall of GMM is superior to the others, but the precision is inferior to our method. Due to the natural of GMM, it fails in the context of moving camera, as shown in the last row of Fig. 4 (c). From the results of TSR shown in Fig. 4 (d), we find that TSR does not perform well on these videos: uncertain motion regions for the second case; noise sensitivity for the third; lower recall and precision, in experiment 4, compared with GMM and our method; missing the major parts of the motion objects for the rest two. All above result from the inaccuracy and sensitivity of the spectrum technique. Our method performs remarkably well on all the five types of videos. Note that the reasons why our method works on the moving camera video mainly are: 1) the repeating background and 2) the

[2] To see more results, please visit: `http://cs.tju.edu.cn/orgs/vision/msd/results.htm`



(a) Original Frame (b) FD (c) GMM (d) TSR (e) Raw Saliency Map (f) Our Method

Fig. 4. Experiment results. (a) original frames (OF) from different video scenes and types. (b)-(d) are the results by using frame difference (FD), GMM [5] and TSR [4], respectively. (e) shows the raw saliency maps obtained by our method. The final results by our method are shown in (f). The performance difference analysis please see the text.

effort of our adaptive threshold selection and refinement. Figure 4 (e) shows the raw saliency maps without executing the refinement and the adaptive threshold selection. The final results of our method are displayed in Fig. 4 (f).

5. CONCLUSION

In this work, we proposed a novel motion saliency detection method based on low-rank and sparse decomposition, which provides many video processing tasks, such as video segmentation and adaptive content delivery, with a powerful video pre-processing technique. The proposed method is able to distinguish foreground motion objects from backgrounds without any background modeling procedure. Thank to spatial consideration, we further reduce the effect of incompleteness. In addition, by employing adaptive threshold selection and noise elimination, the method can automatically and robustly accomplish the task. The experiments carried on different video qualities and scenes demonstrated that our proposed method outperforms the state-of-the-art.

6. REFERENCES

[1] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *IEEE ICME*, 2009, pp. 638–641.

[2] Y. Ma and H. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *ACM MM*, 2003, pp. 374–381.

[3] C. Christopoulos, A. Skodras, A. Koike, and T. Ebrahimi, "The jpeg2000 still image coding system: An overview," *IEEE Trans. on Consumer Electronics*, vol. 46, no. 4, pp. 1103–1127, 2000.

[4] X. Cui, Q. Liu, and D Metaxas, "Temporal spectral residual: Fast motion saliency detection," in *ACM MM*, 2009, pp. 617–620.

[5] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *ICPR*, 2004, pp. 28–31.

[6] L. Itti, C. Koch, , and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[7] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE CVPR*, 2009, pp. 1597–1604.

[8] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2007, pp. 545–552.

[9] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE CVPR*, 2007, pp. 1–8.

[10] Z. Wang and B. Li, "A two-stage approach to saliency detection in images," in *IEEE ICASSP*, 2008, pp. 965–968.

[11] M. Cheng, N. Zhang, G.and Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *IEEE CVPR*, 2011, pp. 409–416.

[12] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, 2011.