2014 22nd International Conference on Pattern Recognition

Graph Construction for Salient Object Detection in Videos

Keren Fu[A,B] Irene Y.H. Gu[B] Yixiao Yun[B] Chen Gong[A] Jie Yang[A]
[A] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[B] Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden

***Abstract*—Recently many *graph-based* salient region/object detection methods have been developed. They are rather effective for still images. However, little attention has been paid to salient region detection in videos. This paper addresses salient region detection in videos. A unified approach towards graph construction for salient object detection in videos is proposed. The proposed method combines static appearance and motion cues to construct graph, enabling a direct extension of original graph-based salient region detection to video processing. To maintain coherence in both intra- and inter-frames, a spatial-temporal smoothing operation is proposed on a structured graph derived from consecutive frames. The effectiveness of the proposed method is tested and validated using seven videos from two video datasets.**

***Keywords*—Salient region detection, Graph construction, Video processing, Optical flows**

I. Introduction

When viewing images, human eyes tend to first focus on important and informative regions. Modeling such selective strategies of human beings recently has drawn much attention in computer vision field. Such models/algorithms are very useful for content-based applications, e.g. image cropping [1], thumbnailing [2], resizing and re-targeting [3], [4].

Recently, *graph-based* salient region/object detection methods have been developed by exploiting *graphs* to solve this problem. They first construct a graph, and then some graph-based techniques are applied to it to obtain image saliency. For example, Wei *et al* [5] propose to treat boundary parts of an image as the background. The patch saliency is defined as the shortest geodesic distance for a graph to image boundaries. As a salient object is often isolated from the background, the geodesic distance between image boundaries and object parts is relatively large, leading to an object being popped out. Yang *et al* [6] utilize similar boundary priors as [5] but propagate saliency via graph-based manifold ranking. Their method is shown to be superior against state-of-the-art methods (including [5]) for salient object detection. Gopalakrishnan *et al* [7] perform random walks on graphs to model the saliency in images. More specifically, the global pop-out and compactness properties of salient objects are modeled in random walks by the equilibrium access time performed on a constructed complete and k-regular graph. In [8], salient region detection is achieved by maximizing a submodular objective function, which maximizes the total similarities computed by finding a closed-form harmonic solution on the constructed graph for an input image. The above methods model the saliency based on the structured graph in each individual input image, and some graph-based techniques (e.g. *shortest path*, *random walk*, *manifold-based methods*) are employed to solve saliency detection problems.

Although these methods are effective, they are all designed for still images and tested on still image datasets. Little attention has been paid to salient region detection in videos where more visual cues can be exploited to improve the performance. Based on their aims, bottom-up saliency detection methods can be roughly divided into two categories [9]: (a) eye fixation prediction; (b) salient region detection. Despite videos are widely exploited by the methods in category (a) (e.g. [10] exploits video processing), little attention is paid to using videos for detecting salient regions, where visual cues in videos could be integrated for enhancing salient region detection. Our study is focused on studying methods in category (b), since they are known to better highlight the entire object than that in the category (a), and are also able to benefit content-based applications (e.g. [1], [2], [3], [4]).

Although identifying salient regions in videos is very useful for video content extraction and summarization [11], it remains under-explored. This paper addresses salient region detection problems in videos. A unified approach towards graph construction for salient object detection in videos is proposed. We show a unified way to extend graph-based methods to video processing by incorporating more cues, e.g. motion, during graph construction. After a graph is constructed by integrating more features, saliency computation of previous graph-based methods [5], [6], [8], [7] for still images can be directly extended to video processing. In addition, to maintain spatial-temporal coherence, we propose a spatial-temporal smoothing operation on a structured graph derived from consecutive frames.

Main contributions of this paper are four-folds:

1) We extend graph-based salient region detection methods to video processing that may be realized in a unified way by integrating more cues in graph construction.

2) We propose a new feature to construct the graph for salient region detection in videos, referred to as *mean histogram of optical flows* (MHOF).

3) Spatial-temporal coherence is maintained via saliency smoothing on a structured graph derived from consecutive frames.

4) The proposed method is tested on two video datasets and validated by comparing with ground truth.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.411

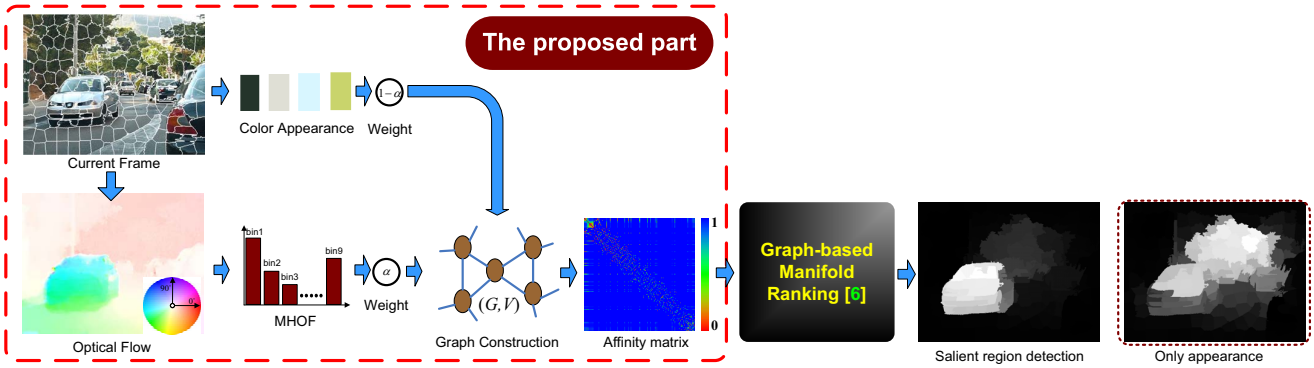

Fig. 1. The processing pipeline for each frame. The proposed part is highlighted in the red-dash rectangle. A frame is first segmented into Simple Linear Iterative Clustering (SLIC) superpixels [12]. Both color appearance and dynamic cue, i.e. mean histogram of optical flows (MHOF) are integrated for graph construction. The graph-based manifold ranking [6] is employed to subsequently process the graph. The result of the original method in [6] for still images (only appearance) is shown in the last for comparison.

II. Graph Construction for Salient Object Detection in Images and Videos

We first review the common way of graph construction for still images. We then extend the graph construction to videos and show results of directly employing an existing salient region detection method to the proposed graph construction.

A. Graph Construction for Still Images

In conventional cases of still images, a graph that represents local relationship is constructed by defining image patches [5], [7] / superpixels [5], [6], [8] as vertices and feature discrepancy [5] / affinity [8], [6], [7] as edges. Each vertex connects to its neighbors, which are either spatially adjacent to it or in its local neighborhood. Since superpixel representation for saliency detection is shown to be effective, e.g. used in [5], [8], [6] as a typical pre-processing to facilitate saliency computation. Motivated by this, we first segment a given image into Simple Linear Iterative Clustering (SLIC) superpixels [12] ($N \approx 200$ superpixels, where the i-th superpixel is denoted by $R_i, i = 1, 2, ..., N$). These superpixels are then deemed as the vertices of the graph.

Feature discrepancy can be modeled as the distance under a specific metric, e.g. the Euclidean distance between two feature vectors [5], [8], [6] or the χ^2 distance between two histograms. For our case, the Euclidean distance is chosen as the metric for measuring appearance differences between vertices (i.e. superpixels). Let c_i and c_j be mean color vectors of two *adjacent/neighbor* superpixel R_i and R_j, whose relationship is denoted by symbol "$R_i \asymp R_j$". The normalized feature discrepancy d_{ij} between R_i and R_j is defined as:

$$d_{ij} = \frac{||c_i - c_j||_2}{\max_{R_p \asymp R_q} ||c_p - c_q||_2} \tag{1}$$

where d_{ij} is normalized by dividing the global maximum.

In the cases where affinity needs to be defined as edges [8], [6], [7], the affinity w_{ij} between R_i and R_j could be derived from the difference defined in (1), leading to:

$$w_{ij} = \kappa(d_{ij}) \tag{2}$$

where $\kappa(\cdot)$ denotes a kernel function with respect to the distance d_{ij} between c_i and c_j. Noting that any function that fits *Mercer's condition* can be used as kernel, e.g. exponential functions $\exp(-\beta x)$ and Gaussian functions $\exp(-x^2/\sigma^2)$.

After a graph is constructed using the feature discrepancy (1) or affinity (2), a specific salient object detection method such as [8], [6], [5], [7] could be employed to assign saliency value to each superpixel. For details on how these methods cope with graph, readers are referred to the corresponding literatures.

B. Graph Construction for Videos

In video cases, although [8], [6], [5], [7] can be used to process each individual frame, additional cues can be incorporated to achieve more reliable results. We propose to leverage a novel feature: *mean histogram of optical flows* (MHOF), to compute discrepancy/affinity of vertices in a graph.

Since human perception tends to be attracted by foreground objects containing relative motion that is highly distinguishable from the background, motion cues can be salient features to describe moving objects [13], [14]. In each superpixel, we compute dense optical flows by using the method in [15], followed by extracting the *mean histogram of optical flows* (MHOF) (shown in the bottom left of Fig.1).

Assuming (u_j, v_j) is the forward optical flow at pixel I_j in a certain frame, the MHOF of a specific superpixel R_k is defined as:

$$h_k(i) = \sum_{I_j \in R_k} \sqrt{u_j^2 + v_j^2}\,\delta_i(u_j, v_j)/|R_k| \tag{3}$$

where $h_k(i)$ is the energy of the ith orientation bin of histogram h_k for R_k, and $\delta_i(u_j, v_j)$ is a binary function that equals 1 if the input (u_j, v_j) is quantized into the ith orientation, and 0 otherwise. In our case, we typically quantize orientation into nine bins (inspired by and similar to HOG [16]). The aim of the "mean" operation, i.e. dividing by $|R_k|$ ($|\cdot|$ denotes sum area), is to eliminate the effect of size discrepancy among different superpixels.

Since there is no normalization in (3), MHOF is not a rigid "histogram". The rationale is that, although optical flows in

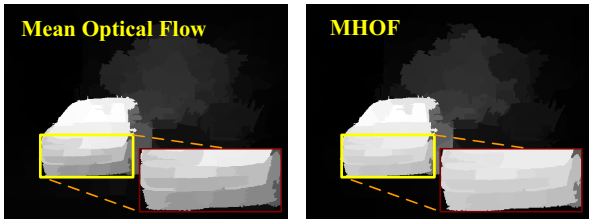

Fig. 2. Difference between using the mean optical flow (left) and the proposed MHOF (right). Comparison is conducted by changing the motion feature in Fig.1.

different superpixels may orient towards the same direction, their magnitude differences could still contribute to saliency (perceived as velocity distinction [14]).

The function of MHOF can be explained as capturing the average statistical motion information in each superpixel. Since the gradient of optical flow magnitude does not always coincide with the boundaries of superpixel segmentation, MHOF is better than a mean vector $(\sum_j u_j/|R_i|, \sum_j v_j/|R_i|)$ that could counterweigh optical flows of opposite direction. Comparing with [10] that uses optical flows to predict human fixation in videos, we reserve both orientation and magnitude information, while in [10] only the magnitude of optical flows is used. Fig.2 shows the results of MHOF and simple averaged optical flow. MHOF is observed better on measuring motion distinction.

Incorporating the idea of MHOF, the feature difference that integrates both the color and motion is defined as:

$$\hat{d}_{ij} = (1-\alpha)\frac{||c_i - c_j||_2}{\max_{R_p \asymp R_q} ||c_p - c_q||_2} + \alpha \frac{||h_i - h_j||_2}{\max_{R_p \asymp R_q} ||h_p - h_q||_2} \tag{4}$$

where $\alpha \in [0,1]$ is the relative importance of the motion component (set to 0.5 as default to render equal importance of the two components). Here, we have tried another distance metric for h_i i.e. the χ^2 distance, resulting in less better performance. Similar to (2), the affinity can be constructed as $\kappa(\hat{d}_{ij})$.

It is worth mentioning that α is adaptively selected for each frame. If the maximum magnitude of optical flows in a frame is lower than a predefined threshold T, i.e. the motion is negligible, α is set to a small number due to less reliability of optical flows. In our experiments, T is empirically set as 1% of the maximum length of image height and width.

Fig.1 shows the processing pipeline for each frame. We directly employ the manifold ranking algorithm in [6] on the proposed graph without any modification. Affinity $w_{ij} = \exp(-\beta \hat{d}_{ij})$ ($\beta = 10$) is set. One can observe from Fig.1 that the moving car is well highlighted. This is because the proposed graph construction integrates dynamic cue. In contrast, merely using static appearance cue (the original method in [6] for still images) detects the brightness presented in the background (Fig.1 last). Other graph-based salient region detection methods (e.g. [7], [5], [8]) could also be applied.

C. Spatial-Temporal Smoothing

In videos, two consecutive frames are usually similar, e.g. object position, shape, contour, appearance are expected

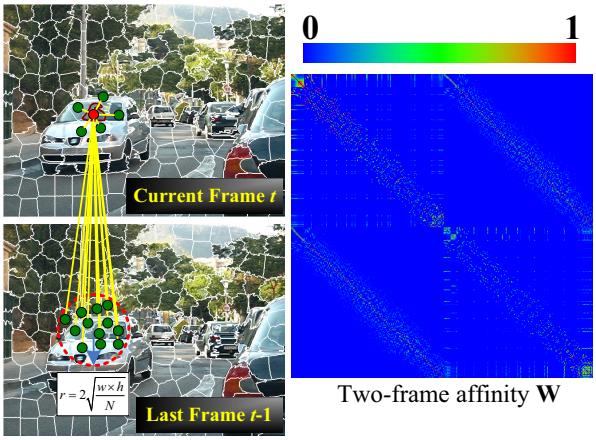

Fig. 3. Left: two-frame graph construction, where yellow lines indicate connection. Right: the visualized two-frame affinity matrix $\mathbf{W}$.

to vary smoothly without abrupt changes [18]. Hence, it is desirable that saliency maps of consecutive frames change smoothly. However, due to the noise introduced by pre-segmentation and optical flows, there could exist drastic difference between saliency maps of consecutive frames (see Fig.4 row 2 and row 3). This motivates us to explore a spatial-temporal smoothing operation.

To maintain the coherence between two consecutive frames, denoted as t and $t-1$, saliency energy of the previous frame $t-1$ is propagated to the current frame t. A graph connecting superpixels in two adjacent frames is constructed. As shown in Fig.3 (left), a superpixel R_k (denoted as a red dot) connects to both its adjacent superpixels in frame t and its spatial neighbors in frame $t-1$ (denoted as dark green dots). For inter-connections that cross two frames, we define a circle region whose center is at the spatial position of R_k. The region radius is set to two times the superpixel size (estimated by $\sqrt{(h \times w)/N}$ [12], h and w denote image height and width). The rationale is to cover position shift of a target between adjacent frames. In this case, object saliency in the previous frame $t-1$ would be correctly propagated to the current frame t.

Based on the above, the affinity matrix $\mathbf{W}$ of a two-frame graph is divided into two parts: intra-affinity $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$, and inter-affinity $\mathbf{W}_{12}$ and $\mathbf{W}_{21}$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}_{(N_1+N_2)\times(N_1+N_2)} \tag{5}$$

where $\mathbf{W}_{11}$, $\mathbf{W}_{12}$, $\mathbf{W}_{21}$, $\mathbf{W}_{22}$ are $N_1 \times N_1$, $N_1 \times N_2$, $N_2 \times N_1$, $N_2 \times N_2$ matrices, respectively (N_1 and N_2 are superpixel numbers of the current and the former frame). Noting $\mathbf{W}_{21} = \mathbf{W}_{12}^T$, $\mathbf{W}_{11} = \mathbf{W}_{11}^T$ and $\mathbf{W}_{22} = \mathbf{W}_{22}^T$.

For intra-affinity values, we reuse the results from (4) ($\kappa(\hat{d}_{ij})$) that are previously used for saliency computation. For inter-affinity values, they are computed according to (4) with only the color term, for it suffices to identify the same object in two adjacent frames whereas the same object in adjacent frames may not present the same motion. Using the color cue for smoothing also enhances the robustness against the

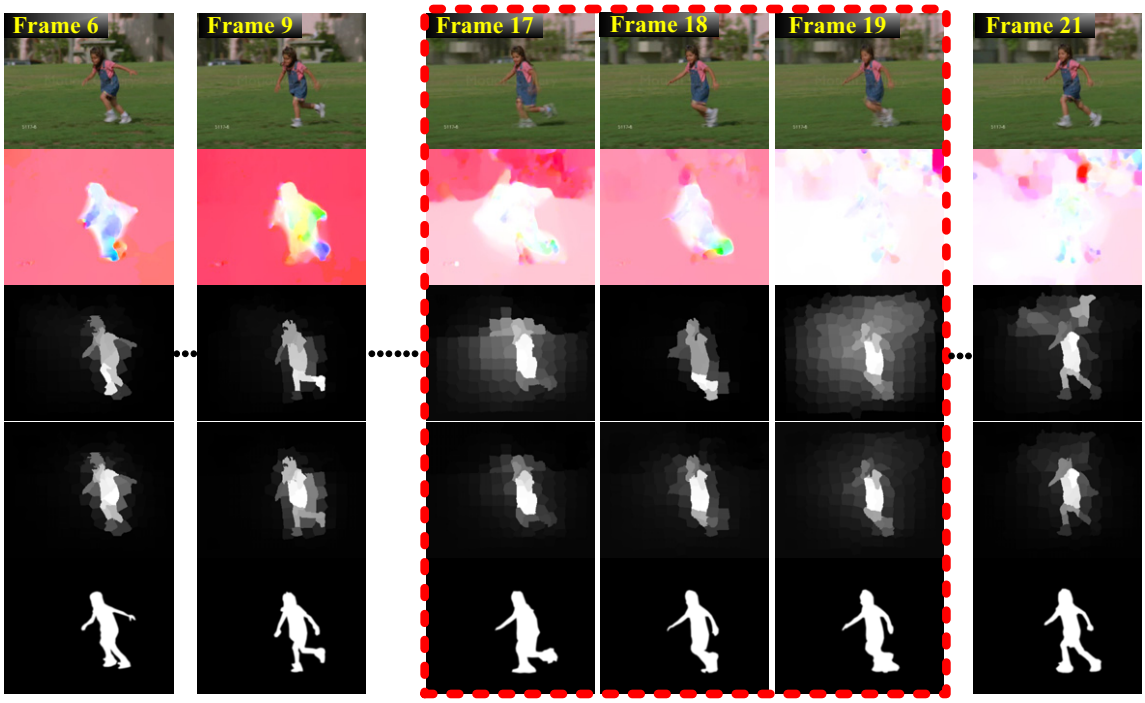

Fig. 4. The effectiveness of spatial-temporal smoothing by using consecutive frames (video "*girl*" in [17]). Rows from top to bottom: original frames, dense optical flow maps generated by [15], saliency detection results without smoothing, results with spatial-temporal smoothing, ground truth. For three consecutive frames 17-19 (highlighted in the red dot rectangle), due to the noise introduced by optical flows, results without smoothing (the 3rd row) vary drastically. Employing smoothing makes results more stable and coherent (the 4th row).

fluctuation of optical flows. Fig.3 (right) shows $\mathbf{W}$ from an image pair.

Let $\mathbf{m}^t$ and $\mathbf{m}^{t-1}$ be the saliency detection results (vectorized into column forms) of current frame t and previous frame $t-1$, respectively. Under the pair-wise smoothness constraint $\mathbf{x}^T\mathbf{L}\mathbf{x}$, where $\mathbf{L}=\mathbf{D}-\mathbf{W}$ and $\mathbf{D}$ is the degree matrix of $\mathbf{W}$, the following energy is minimized:

$$\begin{aligned}&\min_{\mathbf{f}^t,\mathbf{f}^{t-1}} \{(1-\mu)\sum_{k\in\{1:N_1\}}(f_k^t-m_k^t)^2\\&+\mu\sum_{j\in\{1:N_2\}}(f_j^{t-1}-m_j^{t-1})^2+\lambda\mathbf{x}^T\mathbf{L}\mathbf{x}\}\\&=\min_{\mathbf{x}}\{(\mathbf{x}-\mathbf{s})^T\mathbf{M}(\mathbf{x}-\mathbf{s})+\lambda\mathbf{x}^T\mathbf{L}\mathbf{x}\}\end{aligned}\tag{6}$$

where $\mu\in[0,1]$ specifies the relative importance of the previous frame, $\mathbf{x}$ is the concatenated vector of $\mathbf{f}^t$ and $\mathbf{f}^{t-1}$, and $\mathbf{f}^t$ and $\mathbf{f}^{t-1}$ denote the smoothed results for the current and previous frame, respectively. $\mathbf{s}$ is the concatenated vector of $\mathbf{m}^t$ and $\mathbf{m}^{t-1}$. $\mathbf{M}$ is a diagonal matrix with the first N_1 diagonal entries be $1-\mu$ and the subsequent N_2 diagonal entries be μ. λ is a tradeoff between the data term and the smoothing term. In our experiments, λ is set to be a large number (e.g. 1000) to emphasize the smoothing. By differentiating (6) and setting the result to 0, the closed-form solution is obtained as:

$$\mathbf{x}=(\mathbf{M}+\lambda\mathbf{L})^{-1}\mathbf{s}\tag{7}$$

Fig.4 shows the effectiveness of our spatial-temporal smoothing on the "*girl*" video from a public video dataset: SegTrack dataset [17]. $\mu=0.5$ is used to render an equal contribution of two consecutive frames. It is worth noting that although our saliency detection system is fully automatic and unsupervised, it still highlights the salient target in the video and resembles the binary ground truth (Fig.4).

III. Experiments and Comparisons

We have validated our graph construction for videos by using a manifold ranking-based method [6]. Other graph-based methods are possible to replace [6]. We compare the followings: results of [6] with only static appearance cue (referred to as "*appearance*"); results of [6] applied to the graph constructed by the proposed method (referred to as "*appearance+motion*"); results of [6] applied to the graph constructed by the proposed method and also processed with the proposed spatial-temporal saliency smoothing (referred to as "*appearance+motion+smoothing*").

Seven videos are used for comparisons. They are chosen from two video datasets: SegTrack dataset [17] ("*Birdfall*", "*Cheetah*", "*Girl*", "*Parachute*", "*Monkey and dog*)" and GaTech video segmentation dataset [19] ("*Skater*", "*Water skater*"). Note for the "*Skater*" and "*Water skater*" videos, we manually label the ground truth for moving targets since the original binary mask for the dataset is not available.

We use *Precision-Recall curves* [9] as quantitative evaluation criterion on different methods. Under each fixed threshold, precision P and its corresponding recall R are defined as:

$$P=|M\cap Gt|/|M|\ ;\ R=|M\cap Gt|/|Gt|\tag{8}$$

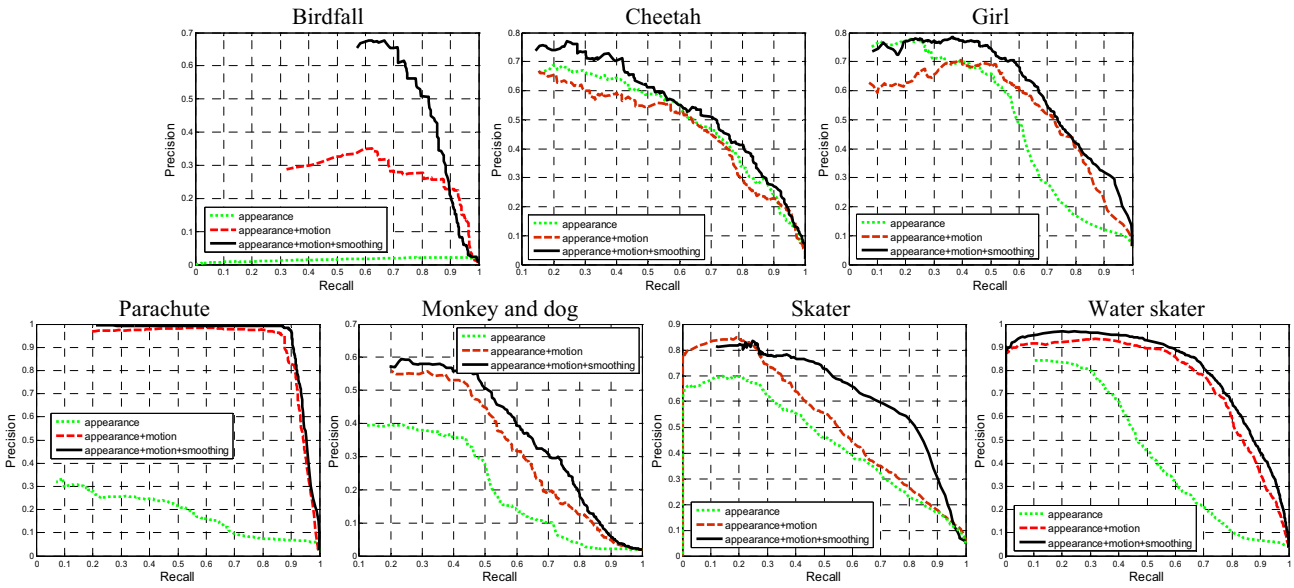

Fig. 5. Performance evaluation: Precision-Recall curves on seven videos from two datasets. In all plots, three curves and their related methods are: *appearance* (green dotted curve); *appearance* + *motion* (red dash line curve); and *appearance* + *motion* + *smoothing* (black solid line curve).

where M is a binary mask obtained by thresholding a saliency map using specific threshold. Gt is ground truth mask and $|\cdot|$ is the summed area of a mask. For each video, the precision and recall are averaged over all frames.

Fig.5 shows the comparisons of Precision-Recall curves. One can observe that for the proposed graph construction that integrates both appearance and motion cues, results are consistently better than the original method ("*appearance*"). Significant deficiency of the original method can be clearly observed in most videos including "*Birdfall*", "*Parachute*", "*Monkey and dog*", "*Skater*" and "*Water Skater*", where the effectiveness of our proposal is further validated. Our method ("*appearance+motion*") drastically outperforms "*appearance*" on "*Birdfall*" and "*Parachute*" videos since only color appearance cannot handle cluttered and textured background. In such cases, the appearance cue seems ambiguous and becomes difficult for distinguishing moving objects from background. Additionally, we have observed moderate improvement in "*Cheetah*" and "*Girl*" videos where the appearance cue provides certain support.

For the proposed spatial-temporal saliency smoothing ("*appearance+motion+smoothing*"), it shows useful for improving performance in a large margin. This can be observed in "*Birdfall*", "*Girl*" and "*Skater*" videos. This procedure generated both spatially and temporally smoothed and coherent saliency maps by taking advantage of structured graph from intra- and inter-frames.

Qualitative comparisons are shown by Fig.6, where gradual improvement is observed from left to right. In column 2, with only static appearance cue, the method cannot locate the moving foreground effectively in complex scenes ("*Birdfall*", "*Parachute*"), or it wrongly detects highly distinctive parts that belong to the background ("*Monkey and dog*", "*Skater*"). In column 3, incorporation with motion does help the detection of the moving foreground, but some results are less satisfactory since the performance of optical flow estimation is not always stable for consecutive frames ("*Birdfall*"). For column 4, better results are obtained after using our spatial-temporal smoothing, where the detection results are close to the binary ground truth.

IV. Conclusion

We have proposed a unified approach to construct graphs for salient region detection in videos. The proposed graph construction has integrated both static and motion cues by using a novel feature: *mean histogram of optical flows* (MHOF) that effectively captures the statistical motion information in each superpixel. The advantage of the proposed method in video processing is shown by applying the manifold ranking-based method [6] to constructed graphs on seven videos. The proposed spatial-temporal smoothing operation is shown to make saliency output more coherent, and to enhance the final performance.

In our experiments, we have observed some cases where the output of optical flow estimation is not stable (e.g. Fig.4). To obtain better performance, more robust optical flow estimation will be employed by our system in the future. Further, application of the proposed method to video summarization will also be investigated.

Acknowledgment

This research is partly supported by National Science Foundation, China (No: 61273258, 61105001), Ph.D. Programs Foundation of Ministry of Education of China (No. 20120073110018). Jie Yang is the corresponding author (email: jieyang@sjtu.edu.cn).

References

[1] F. Stentiford, "Attention based auto image cropping," in *Workshop on Computational Attention and Applications, ICVS*, 2007.

[2] L. Marchesotti et al, "A framework for visual saliency detection with applications to image thumbnailing," in *ICCV*, 2009.

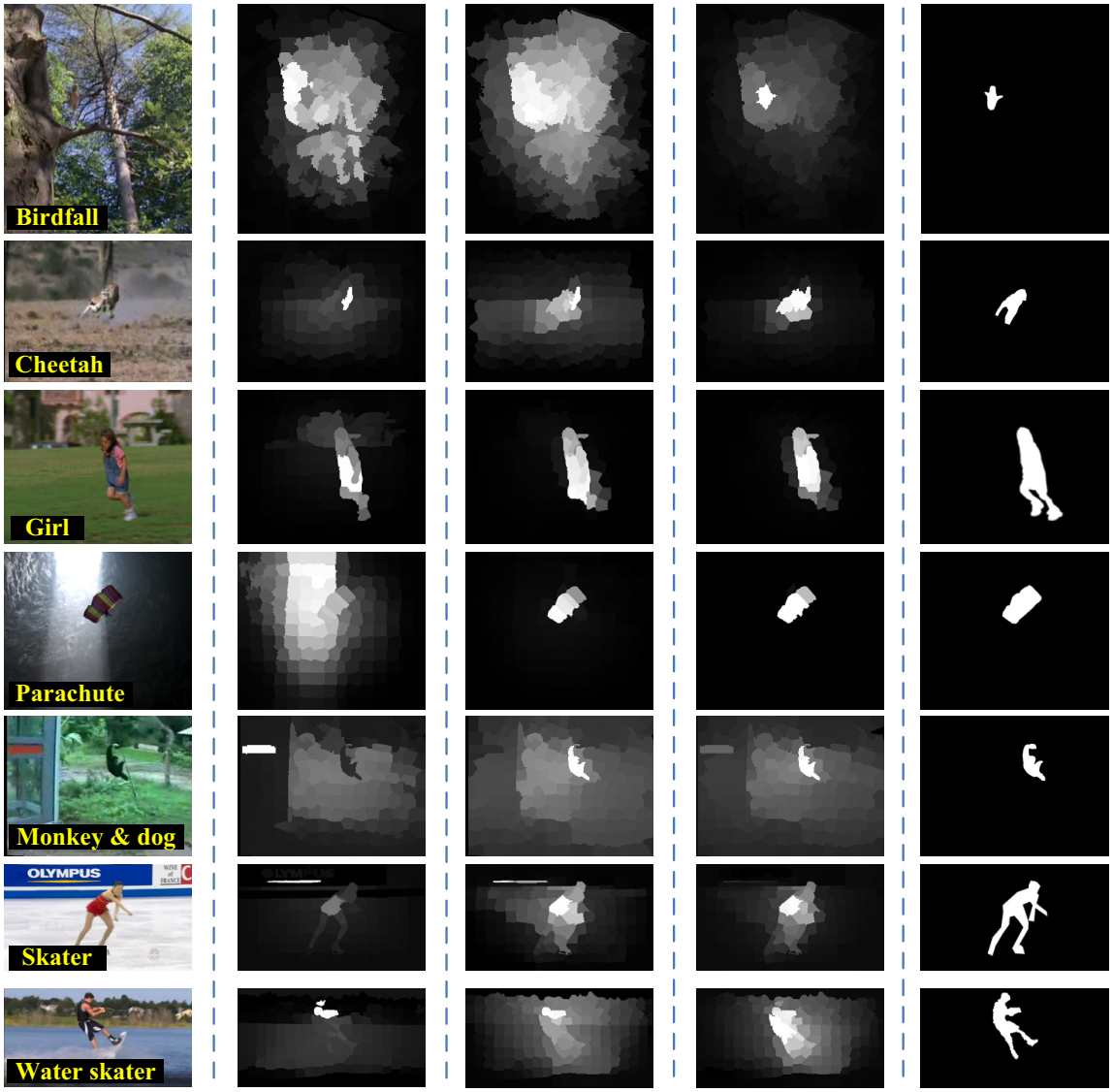

Fig. 6. Visual comparisons. Columns left to right: original frames, *appearance*, *appearance+motion*, *appearance+motion+smoothing*, ground truth.

[3] Y. Ding, X. Jing, and J. Yu, "Importance filtering for image retargeting," in *CVPR*, 2011.

[4] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *CVPR*, 2010.

[5] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *ECCV*, 2012.

[6] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013.

[7] V. Gopalakrishnan et al, "Random walks on graphs for salient object detection in images," *TIP*, vol. 19, no. 12, pp. 3232–3242, 2010.

[8] Z. Jiang and L. Davis, "Submodular salient region detection," in *CVPR*, 2013.

[9] A. Borji, D. Sihite, and L. Itti, "Salient object detection: A benchmark," in *ECCV*, 2012.

[10] D. Rudoy, D. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *CVPR*, 2013.

[11] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[13] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perceiving events and objects*, 1973.

[14] C. Healey and J. Enns, "Attention and visual memory in visualization and computer graphics," *IEEE Trans. Visualization and Computer Graphics*, vol. 18, no. 7, pp. 1170–1188, 2012.

[15] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," *PhD thesis, Massachusetts Institute of Technology*, 2009.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[17] D. Tsai, M. Flagg, and J. Rehg, "Motion coherent tracking with multi-label mrf optimization," in *BMVC*, 2010.

[18] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *CVPR*, 2013.

[19] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *CVPR*, 2010.