2011 Sixth International Conference on Image and Graphics

Video Reshuffling with Narratives toward Effective Video Browsing

Wei Fu, Jinqiao Wang, Xiaobin Zhu, Hanqing Lu *and* Songde Ma
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{wfu, jqwang, xbzhu, luhq}@nlpr.ia.ac.cn, masd@most.cn

***Abstract*—With the rapid increasing of video cameras, large amount of video data everyday brings the problem of video storage and browsing. In this paper, we propose a novel approach to video reshuffling with a group of static images to effectively summarize the video content. Each static image called narrative is generated to depict the behavior of a specific object or a special event. Firstly background subtraction and object tracking are employed to extract the segmentations of moving objects and corresponding trajectories. After that, we apply three sampling rules to optimized select representative object samples from the spatial-temporal object tube and stitch them to the background image by Poisson blending. Experimental results show the promise of the proposed approach.**

Keywords-video narratives; video editing; video summarization

I. INTRODUCTION

Nowadays, video data can be seen when and where and the quantity is still significantly increasing every day. Despite of the powerful ability to record mass data, large scale video data often confuses viewers due to too much irrelevant content contained. Taking the surveillance environment for example, large number of cameras record video data 24 hours a day. Although large information the recorded video contains, a little may be truly concerned by the viewers. Since browsing the whole video is a time-consuming process, a common solution to this problem is to summarize the origin video into a new still image or dynamic short video. Then we can get the desired information by quickly browsing the summary video.

What's more, we notice a fact that different viewers often have different requirement and they only need to pay attention to a specific object or event. If we can summarize each video segment as a static stroboscopic image respectively, viewers could flexibly obtain the information important for them and browse it at a glance. In this paper, we call such a still image a video narrative. Each video narrative maximizes the representation of one object's appearance and behavior (or event), while removes the redundant parts as much as possible in the video. Then all the narratives compose the compressed representation of the whole video.

In this paper, we achieve video reshuffling for optimized represent object behavior and event by a group of narratives. We combine background subtraction and object tracking to extract the moving objects and corresponding trajectories. After that spatio-temporal tubes are obtained. Then the representative object samples are optimized selected from these object spatio-temporal tubes to maximize object representation. Finally we map them to the background image to generate a synopsis image as video narrative.

The rest of the paper is organized as follows. In section 2 we give an overview about the related work. Then Section 3 describes our proposed method in detail. Experimental results are given in the Section 4. Section 5 concludes with applications, limitations and future work.

II. RELATED WORK

State-of-the-art research on video summarization could be classified to two categories. One is still-image based summarization and the other is summarization based on dynamic video.

For the still-image based summarization, earlier research usually focuses on selecting a set of salient images (key frames) to form a new image. Recently Goldman et al. [5] employed schematic storyboards to convey a significant time interval of a video content. In their approach, a storyboard was organized and annotated like a filmstrip but with more continuity and directionality. Another approach is presented by Mei et al. [11] as "Video Collage". A video sequence was compacted as an energy minimization problem to get a single image with seamlessly arranging ROIs (regions of interest) on a given canvas. Similar to their work, our method can also express a video clip using one or a set of static image(s). The difference is that our narratives mainly focus on how to depict the object behavior more explicitly.

As a typical work for the video based summarization, Pritch et al. [2][3][4] made a long video short by dynamic video synopsis. All moving objects are extracted and recorded as an object-based description of the video. Then the object tube is selected by energy minimization to generate a synopsis. However a synopsis video sometimes seems disordered and makes the viewers confused in a complicated scenario.

Most recently, Correa and Ma [6] have developed an interactive system for creating seamless summaries of video. A panoramic background is constructed and the foreground objects are extracted based spatio-temporal masks. After that

978-0-7695-4541-7/11 $26.00 © 2011 IEEE
DOI 10.1109/ICIG.2011.49

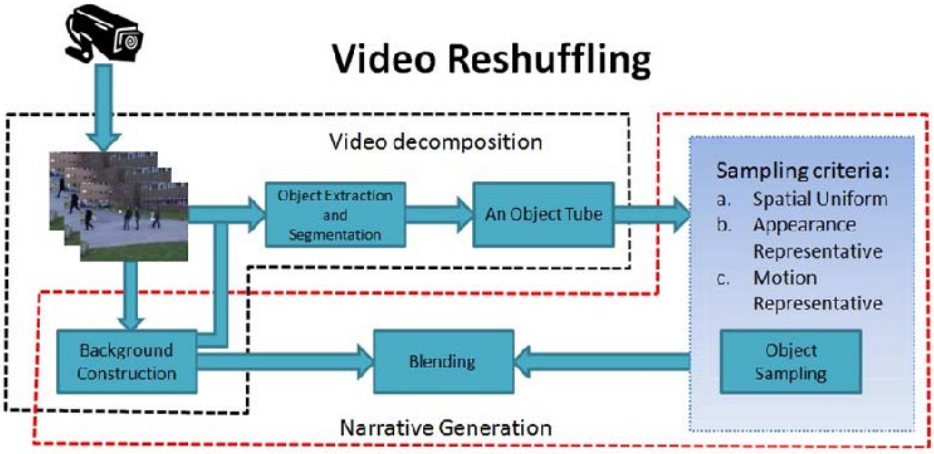

Figure 1 Proposed approach

they compose matted foreground objects on the background. Unlike our approach their work focuses on interactive authoring and animation facilities. Our work in contrast focuses on reshuffling video into a set of static narratives to optimized represent objects' behavior and event with a fully automatic approach. In addition, our approach can handle a more complicated video containing more than one moving objects by decomposing the video into several video segments based on objects.

III. VIDEO RESHUFFLING

To better represent the object appearance, behavior or event in one image, we propose to video reshuffled with narratives for effective large scale video browsing.

Since a whole video is reshuffled to be discomposed to a set of narratives, we can phrase the generation of each narrative as a process of selective synopsis. As a process of synopsis, the video reshuffling could be divided to two processes: video decomposition and narrative generation, as shown in Figure 1. The video decomposition process includes background construction, object extraction and segmentation, and object tracking. The narrative generation process includes object sampling and object blending.

A. Blackground Construction

In our approach, the construction of background is necessary for two aspects. On the one hand, the background image provide a prior for the next process of object extraction and segmentation. On the other hand, an background image conveying circumstance around objects is utilized in the blending process as a common target image.

A lot of approaches have been developed to improve the background models. In this paper, a Gaussian Mixture Model [13] is adopted to generate the background.

B. Object Extraction and Segmentation

After background modeling, the background cut method proposed in [1] is applied to segment all foreground objects in the current frame. Then a "fragtrack" method described in [10] follows to obtain the trajectory of the moving object in the subsequent sequence.

In order to get the segmentations of moving objects more accurately and smoothly, we employ the approach described in [3] and [4], considering the problem as an energy function minimization. We denote the set of all pixels in the frame by V and a label function by L_r. L_r is set to be 1 when the pixel r belongs to a foreground object and 0 when belonging to the background. The Gibbs energy is defined as follow:

$$E(L) = \sum_{r\in V} E_1(L_r) + \lambda \sum_{(r,s)\in\varepsilon} E_2(L_r, L_s), \quad (1)$$

where $E_1(L_r)$ is the color term, denoting the cost when the label of pixel r is L_r. And the second term $E_2(L_r, L_s)$ is the contrast term between adjacent pixels r and s. The symbol ε denotes the set of adjacent pixel pairs and λ is a weight which can be changed to balance the effects of the two terms.

Similar to [3] and [4], we define the color term as:

$$E_1(L_r = 1) = \begin{cases} 0 & d_r > k_1 \\ k_1 - d_r & \text{otherwise} \end{cases}$$
$$E_1(L_r = 0) = \begin{cases} \infty & d_r > k_2 \\ d_r - k_1 & k_2 > d_r > k_1 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where $d_r = \|I(r) - B(r)\|$ denoting the color differences between the image I and the current background B, and $k_i, (i=1,2)$ are two thresholds defined by users.

As for the second term, we borrow the definition from Sun's work [1] as follows:

$$E_2(L_r, L_s) = |L_r - L_s| e^{(-\beta \cdot d_{rs})}, \quad (3)$$

where $\beta = \left(2\left\langle \|I_r - I_s\|^2 \right\rangle\right)^{-1}$, and $\langle \bullet \rangle$ is the expectation operator. The parameter d_{rs} is defined as:

$$d_{rs}=\|I_r-I_s\|^2\cdot\frac{1}{1+\left(\frac{\|B_r-B_s\|}{K}\right)^2\exp\left(-\frac{z_{rs}^2}{\sigma}\right)}, \quad (4)$$

where $z_{rs}=\max\{\|I_r-B_r\|,\|I_s-B_s\|\}$ measures the dissimilarity between pixel pair (I_r,I_s) in image I and (I_r,I_s) in the corresponding background B. The parameters K and σ are set to be 25 and 10 respectively in our experiments. By defining like this, we can adaptively attenuate the contrasts in the background while preserving the contrasts across foreground/background boundaries and get a more accurate segmentation of foreground for each frame.

After that the min-cut method [8] is applied to minimize the energy and then we obtain the desirable labeling L.

Since the trajectory of the object is obtained, only the foreground segmentation corresponding to the trajectory is preserved in each frame. Then a spatio-temporal tube of the tracked object is generated, as shown in Figure 2.

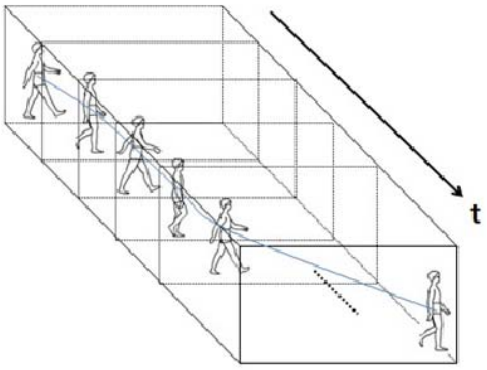

Figure 2 An illustration of a spatio-temporal object tube.

C. Optimized Narrative Generation

As described in [6], a narrative is defined as a composition of a set of objects and a common background. In this paper each narrative depicts a specific object behavior or event. Directly stitching the whole tube of an object to the background needs much computation and is not necessary. To achieve a compressive representation for the object appearance, behavior or event, we optimized sample a set of representative object duplications from the spatio-temporal object tube. And then we blend the selected representative objects to the background image obtained before.

1) Object sampling.

The simple sampling method is uniform temporal sampling for each moving object. However, this will lead to a non-uniform spatial distribution due to the change of motion speed or direction. In view of the change of object appearance information, motion information in space and time, we present three criteria for effective object sampling:

- The sample distribution along the object trajectory should be spatial uniform.
- The samples from the object tube should represent the change of appearance and behavior.
- The samples should represent the motion information such as speed, direction etc.

Let $\Omega=\{Q^1,Q^2,...,Q^i,...,Q^m\}$ denote all the extracted object tubes, where $Q^i=\{O_1^i,O_2^i,...,O_t^i...,O_n^i\}$ denoting the ith tube with n samples and $O_t^i=\{x_t^i,y_t^i,f_t\}$ stands for the object locating at (x_t^i,y_t^i) in the frame f_t. We construct three temporal histograms: (1) a uniform temporal histogram H_t, (2) a temporal appearance histogram H_a of the object tube, and (3) a temporal histogram H_v measuring the velocity of the moving object. By combining the three histograms we construct a new histogram for object sampling as follow:

$$\begin{aligned} &H=\lambda_1\cdot H_t+\lambda_2\cdot H_a+\lambda_3\cdot H_v \\ &s.t.\quad 0<\lambda_1,\lambda_2,\lambda_3<1 \\ &\lambda_1+\lambda_2+\lambda_3=1 \end{aligned}, \quad (5)$$

To construct the histogram H_a, we use $Dist(O_i^k,O_j^k)$ to measure the continuity of object appearance in the kth tube, which can be computed as follow:

$$\begin{aligned} Dist(O_i^k,O_j^k)=&\left\|His(O_i^k)-His(O_j^k)\right\|_2 \\ &+\beta\left|S(O_i^k)-S(O_j^k)\right| \end{aligned}, \quad (6)$$

where $His(O_i^k)$ indicts the color histogram of the object O_i^k, $S(O_i^k)$ denotes the area of the current foreground object segmentation and the parameter β is defined by users to balance the influence of two terms.

If $Dist(O_i^k,O_j^k)<T_1$, we assume that appearance of the object changes a little while $Dist(O_i^k,O_j^k)>T_2$ implies that occlusion occurs or the current object is merged by others. So H_a can be constructed as follow:

$$H_a(t)=\begin{cases} 0, & Dist(t)\le T_1 \\ \dfrac{Dist(t)}{\sum_{\tau}Dist(\tau)}, & T_1<Dist(t)<T_2 \\ 0, & \text{otherwise} \end{cases}, \quad (7)$$

where $Dist(t)=Dist(O_{t-1}^k,O_t^k)$ measures the change of the current object Q^k appearance.

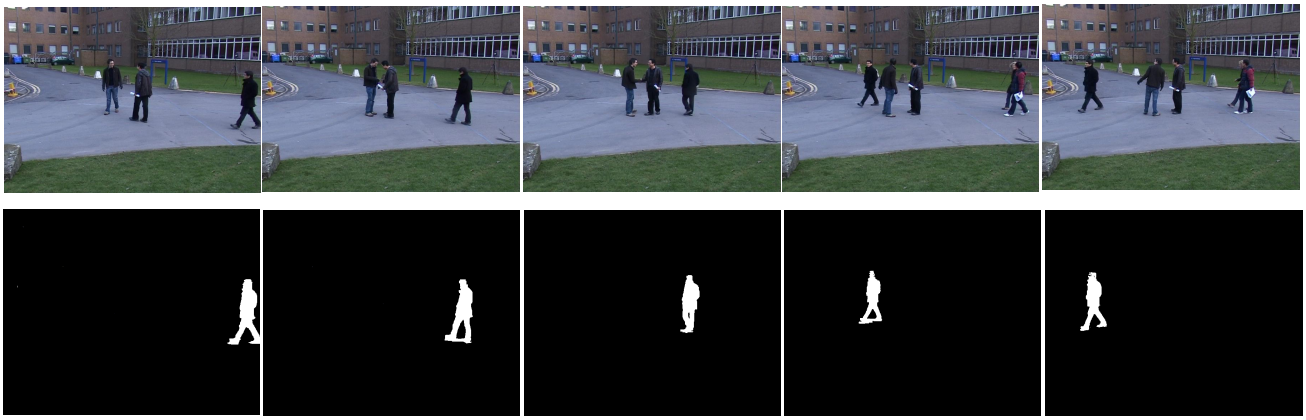

Figure 4 an illustrator of a sampled object tube

As for the third term H_v, we consider the object's velocity including speed and angular velocity. The sample rate should be large enough when the tube is with a large speed and small angular speed, and vice versa.

Let $\overrightarrow{v(t)}=\overrightarrow{P_{t-1}P_t}$ denote the instantaneous velocity of the object in the tth frame, where $P_t=(x_t,y_t)$. The speed of object is defined as $v(t)=\left\|(x_t,y_t)-(x_{t-1},y_{t-1})\right\|_2$, and the angular velocity is defined as $\theta(t)=\cos^{-1}\left(\frac{\overrightarrow{v(t)}\cdot\overrightarrow{v(t-1)}}{v(t)\cdot v(t-1)}\right)$. Then the histogram H_v is set as:

$$H_v(t)=\frac{1}{Z}\cdot\left(1+\cos\theta(t)+\alpha\cdot v(t)\right), \tag{8}$$

where $Z=\sum_{\tau}\left(1+\cos\theta(\tau)+\alpha\cdot v(\tau)\right)$ is a normalized coefficient and α is an experiential weighting parameter.

The representative objects are selected according to the final combined histogram H.

2) Object blending.

Once the representative objects are determined, we blend them with the extracted background image. The goal of object blending is to achieve a seamless fusion between the foreground and the background. In this paper, we adopt the "Poisson Blending" method described in [7].

Specifically, let g and b be the extracted object and background pixel respectively and Ω be the domain of blending. By solving the following Poisson equations, we get f denoting the values of the pixels inner Ω.

$$\min_{s}\iint_{\Omega}(\Delta f-\Delta b)^2, with\ f|_{\partial\Omega}=g|_{\partial\Omega}, \tag{9}$$

IV. EXPERIMENTS

To demonstrate the performance of the proposed approach, we test our approach on several surveillance video clips selected from the PETS 2009 database.

Figure 3 shows two examples of object extraction and segmentation.

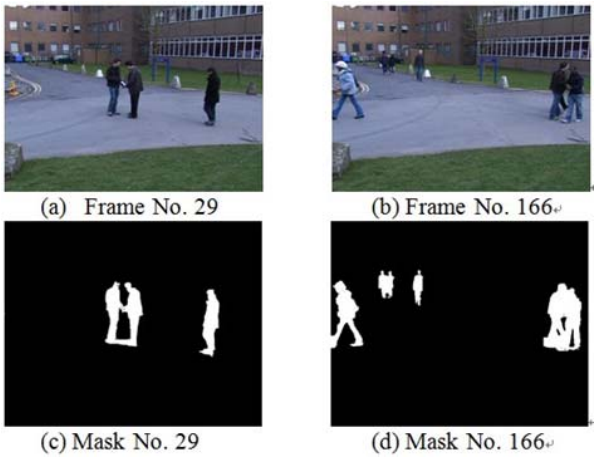

Figure 3 (a) and (b) are frames from the origin video. (c) and (d) are the corresponding foreground mask images.

Figure 4 shows an instance of a spatio-temporal object tube, extracted according to the sampling criteria. The top row shows several source frames in the database. With the trajectory obtained, only the one moving person mask corresponding to the trajectory is preserved in every frame. And then an optimized object tube is obtained after sampling process, as shown in the bottom row of Figure 4.

Figure 5 and Figure 6 give some examples of narratives generated from video reshuffling under two different scenes.

For Figure 6a the sparseness in the middle is due to occlusion and merger with the other two persons. The appearance of the person changes a lot when occlusion occurs, so we assume that the current object should be eliminated in the sampling process. Narrative Fig.6b shows a

woman walking from right to left. The distribution of the object samples is spatial uniform until her pose changes suddenly at last. In Figure 6c the pose of the man changes all the time and most of the typical poses are selected to be illustrated in one narrative on the premise of spatial uniform. Figure 6d shows the sampling results related to the velocity and direction. The rate of sampling is lower at the inflexion point of the trajectory.

In addition, we test our method on several real surveillance video clips. As shown in Fig.7, six narratives are generated from a video clip. In the source video, each object moves across the scene during 40 to 60 frames. Using a narrative, we represent these frames as one still image. We believe this method is suitable for video browsing and searching for a specific event.

V. CONCLUSION

In this paper, we have proposed a novel approach to generate object based narratives to summarize the source video. Three sampling rules are presented to maximized preserve the behavior or event information. This kind of video reshuffling has large applications in video indexing, fast browsing, and video summarization. There exist limitations in our work. Firstly focusing on a specific object unduly may discard the related behaviors or events which may have important influence on the desired object. Secondly, we discuss fixed scene only. Next we will consider the interaction information between objects, and also the dynamic scene based narratives.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 60833006 and 60905008), and 973 Program (Project No. 2010CB327905).

REFERENCES

[1] J. Sun, W. Zhang, X. Tang and H. Shum, "Background Cut", *Proc. Ninth European Conf. Computer Vision*, pp. 628-641, 2006.

[2] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a Long Video Short: Dynamic Video Synopsis", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 435-441, 2006.

[3] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam Synopsis: Peeking Around the World", *Proc. Int'l Conf. Computer Vision*, 2007.

[4] A. Rav-Acha, Y. Pritch, and S. Peleg, "Nonchronological Video Synopsis and Indexing", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971-1984, 2008.

[5] D.B. Goldman, B. Curless, D. Salesin, and S.M. Seitz, "Schematic storyboarding for video visualization and editing", *ACM Trans. Graphics*, vol. 35, no. 3, pp. 1971-1984, 2006.

[6] C.D. Correa, and K.-L Ma, "Dynamic Video Narratives", *ACM Trans. Graphics*, vol. 29, no. 3, 2010.

[7] M. Gangnet, P. Pérez, and A. Blake, "Poisson Image Editing", *ACM Trans. Graphics*, pp. 313-318, 2003.

[8] V. Kolmogorov, and R. Zabih, "What Energy Function can be Minimized via Graph Cuts", *Proc. European Conf. Computer Vision*, pp. 65-81, 2002.

[9] Y. Li, T. Zhang, and D. Tretter, "An Overview of Abstraction Techniques", Technical Report HPL-2001-191, HP Laboratory, 2001.

[10] A. Adam, E. Rivlin and I. Shimshoni, "Robust Fragments-based Tracking Using the Integral Histogram", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.798-805, 2006.

[11] T. Mei, B. Yang, S. Q. Yang and X. S. Hua, "Video Collage: Presenting a Video Sequence Using a Single Image", *The Visual Computer: International Journal of Computer Graphics,* vol.25, no.1, pp.39-51, 2009.

[12] B. T. Truong and S. Venkatesh, "Video Abstraction: A Systematic Review and Classification", *ACM TOMCCAP*, 2007.

[13] Z. Zivkovic, "Improved Gaussian Mixture Model for Background Subtraction", *Proc. ICPR*, 2004.

[14] A. Ellis, A. Shahrokni and J. M. Ferryman, "PETS2009 and Winter-PETS 2009 results: A Combine Evaluation", In Winter-PETS, 2009.

Figure 5 a set of narratives generated from a video clip

(a) (b) (c)

(e) (f) (g)

Figure 6 a set of narratives generated from a video clip

Figure 7 some narratives generated from real surveillance video clips