OBJECT-CENTERED NARRATIVES FOR VIDEO SURVEILLANCE

Wei Fu, Jinqiao Wang, Chaoyang Zhao, Hanqing Lu, Songde Ma

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{wfu,jqwang,luhq}@nlpr.ia.ac.cn, chaoyang1635@gmail.com, masd@most.cn

ABSTRACT

Effective video presentation and summarization techniques are critical for fast browsing of video content. In this paper, we propose a novel presentation approach to vividly depict the moving process of a specific object in a surveillance video, which aims at effectively summarizing video content by a static image named narrative. Firstly, the object of interest is extracted and segmented from the video to form a spatio-temporal object tube. Then three criteria are proposed to select the most representative objects from this tube. We formulate the object selecting process as an energy minimization problem, in which each energy term measures a corresponding criterion cost. We maximally preserve the changes of appearance and behavior while remove other redundant content as much as possible. Finally, the selected representative objects are stitched to the background image by Poisson editing. Experimental results show the promise of the proposed approach.

Index Terms— Video narratives, Video summarization

1. INTRODUCTION

With the proliferation of video data recorded by surveillance cameras, video presentation and summarization techniques are critical for effective video browsing. Especially under the scenario of surveillance, although a large amount of data is recorded 24 hours a day, little may be truly concerned by the viewers. A common solution to this problem is to abstract the origin video into a static image or dynamic short video, with which viewers can quickly capture the main idea of video. In terms of browsing and navigation, a good video abstract should enable the viewer to gain maximum information about the source video in the minimum time. Over the past years, various ideas and techniques have been proposed towards the effective abstraction of video contents [1].

State-of-the-art research broadly falls into two categories: static image based summarization [2, 3, 4] and dynamic video based summarization [5, 6]. For the static-image based summarization, earlier research usually focused on selecting a set of key frames to form a new image. Goldman *et al.* [2] 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.* [3] 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. As a typical work for the video based summarization, Pritch *et al.* [5] made a long video short by dynamic video synopsis. All moving objects were extracted and rearranged in a synopsis video. However, synopsis videos may seem disordered when too much information contained. Correa and Ma [6] developed an interactive system for creating seamless summaries of video. A panoramic background was constructed first and then matted foregrounds were composed on the background.

However, as discussed in our previous work [4], a fact is that different viewers may concern different aspects of video content, thus they only need to pay attention to some specific object of interest. In this condition, approaches mentioned above lack specificity to express "what is the interested". In order to enhance the user experience of browsing, we proposed a new video presentation technology to summarize only the part of interest in the video as a still stroboscopic image, by which viewers could obtain their desirable information more flexibly and explicitly at a glance. The main difference is that, instead of directly sampling representative objects from a spatio-temporal histogram in [4], we try to maximize the representation of the moving process for an object through energy minimization while removing other redundant content as much as possible. Besides, some related objects may be also illustrated in the narrative to present the contextual information when some event happens.

2. OUR APPROACH

To better depict the appearance, behavior or event of a specific object by a static image, we propose an object-centered presentation technique for effective video browsing. First of all, the specific object of interest is extracted and segmented from the video to form a spatio-temporal tube. Then representative object samples are optimally selected from the tube by energy minimization. Finally we map them to the background image

and generate a video narrative. In the rest of this section, we will detail our approach from these three aspects.

2.1. Object Extraction and Segmentation

To generate a narrative, a background image is constructed, to which representative foregrounds will be stitched. Also, the background image is utilized as a prior in the procedure of object extraction and segmentation. Various approaches have been developed to improve the background models. Here, a Gaussian Mixture Model (GMM) [7] is adopted to generate the background.

After modeling the background, we follow [8] using background subtraction together with min-cut to get smooth segmentations of moving objects, considering the problem as an energy minimization problem. And trajectories of objects are obtained by tracking.

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:

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

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 is a contrast term between adjacent pixels s and r. 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.

Like in [5], we define the first term as follows:

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

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

where $d_r = \|I(r) - B(r)\|$ denotes the color differences between the current image and the background , and $k_i, (i = 1, 2)$ are two thresholds set by users. As for the second term, we directly borrow the definition from Sun's work [8]. Then this energy minimization problem is solved by min-cut method [9]. With the trajectory of object of interest, only foregrounds coinciding with the trajectory are preserved in each frame. It is worthy to note that other related objects may be also preserved in this process when foregrounds are joined together with the specific object as one. Finally a spatio-temporal tube of the object is generated.

2.2. Object Sampling by Energy Minimization

2.2.1. Problem formulation

Directly stitching the whole object tube to the background image not only brings a high computation complexity but also is unnecessary. A simple strategy is uniformly sampling object duplications from the tube in temporal. However, this will lead to a loss of information, such as the change of object appearance, behaviors, motion etc. To achieve a visual pleasing presentation for the object appearance, behavior or event, we present three criteria for the sampling of object tube, that is,

1. The sample distribution along the object trajectory should be as spatial uniform as possible;
2. Samples from the object tube should represent the change of appearance and behavior;
3. Samples should represent the change of motion information such as speed, direction etc.

These criteria can be formalized as a series of energy terms, and the sampling procedure is then considered as a labeling problem with minimized energy. Given an object tube containing M total objects, it is denoted as $\Omega = \{O_i\}_{i=1}^{M}$. Let $\lambda = \{R_i\}_{i=1}^{M}$ be one feasible solution, and each R_i has a set of state variables $R_i = (O_i, \ell_i, f_i)$, where ℓ_i is a label indicating whether O_i is selected ($\ell_i = 1$) or not ($\ell_i = 0$) and f_i is a feature vector denoting the appearance of O_i. (In our work, a simple shape context feature is utilized to describe the object appearance.) Then the object sampling problem is formulated to minimize the energy function as follows:

$$\begin{aligned} & E(\lambda) = \omega_1 E_t(\lambda) + \omega_2 E_a(\lambda) + \omega_3 E_v(\lambda) \\ & s.t. \quad \sum_{i=1}^{M} \ell_i = N, \end{aligned} \qquad (2)$$

where $\{\omega_i\}_{i=1}^{3}$ are weighted parameters to balance the effects of three criteria.

The first term $E_t(\lambda)$ measures the deviation of temporal distribution of λ, in the sense of selected objects being uniformly distributed. It is defined as:

$$E_t(\lambda) = \frac{1}{\log N} \sum_{i=1}^{N-1} \ell_i p(R_i) \log p(R_i) \qquad (3)$$

where $p(R_i)$ =(frame interval between O_i and O_{i+1})/(the total duration of tracking). Obviously, the more uniformly λ distributes, the less the cost E_t is.

For the second term $E_a(\lambda)$, we measure the appearance change of object during a constant window δ by $r_i = \|f_i - \bar{f}_i\|$, where the $\bar{f}_i$ is defined as:

$$\bar{f}_i = \sum_{j \in \delta} f_j \exp \frac{-(i-j)^2}{\delta^2} / \sum_{j \in \delta} \exp \frac{-(i-j)^2}{\delta^2} \qquad (4)$$

According to Eq. 4, high local extremum points are likely to be sampled, which stand for the representative appearance changes of this object. So we define the energy cost for the change of object appearance as:

$$E_a(\lambda) = \sum_{i=1}^{M} \ell_i e^{-r_i} \qquad (5)$$

By minimizing the appearance term, neighboring objects occluding with the specific one are preferred to be selected together from the tube, instead of omitting all contextual information as in [4].

For the third term $E_v(\lambda)$, we consider the motion information of the object tube. For two points P_{i-1} and P_i within two consecutive frames I_{i-1} and I_i along the trajectory of object, we denote the displacement by $\Delta d_i = (\Delta x_i, \Delta y_i)$. Then each Δd_i is voted into an 8-bin histogram h_i, according to the orientation of displacement and with the margin as the voting weight. Similar to Eq. 4, a sliding temporal window δ is utilized again to compute an average motion histogram $\bar{h}_i$. Then the energy term can be defined as:

$$E_v(\lambda) = \sum_{i=1}^{M} \ell_i e^{-\|h_i - \bar{h}_i\|} \tag{6}$$

Motion representative objects are likely to be selected by minimize the $E_v(\lambda)$.

Finally, we get the energy function for selecting objects from the tube, which coincides with our sampling criteria.

2.2.2. *Energy minimization*

As described before, the procedure of sampling representative objects is to minimize the energy function Eq. 2. Let $\Omega = \{O_i\}_{i=1}^{M}$ denote all objects in a tube, and let Θ denote a subset of Ω with N objects. The representative object sampling problem is then rewritten as the following energy function,

$$\min_{\Theta} \omega_1 E_t(\Theta) + \omega_2 \sum_{O_i \in \Theta} E_a(R_i) + \omega_3 \sum_{O_i \in \Theta} E_v(R_i) \tag{7}$$

The search space for optimization of energy is a collection of all feasible objects in the tube, which has C_M^N possible solutions totally. In this paper, a heuristic search algorithm is utilized to optimize this problem. We summarize the whole procedure as Algorithm 1. Parameters $\alpha_i > 1 (i = 1, 2)$ are two gain factors to update energy terms in each iteration, and $\gamma = \lfloor M/N \rfloor$, reflecting the effect of $E_t(\lambda)$.

Algorithm 1: Heuristic search for object sampling.

Input: $N, \Omega = \{O_i\}_{i=1}^{M}$
Output: Θ
while $n \leq N$ **do**
  find R_i with $\min(E_a(R_i) + E_v(R_i))$ in Ω;
  $\Omega - = \{O_i\}$;
  for $k = \max(0, i - \gamma)$ **to** $\min(i + \gamma, M)$ **do**
    $E_a(R_k) = \alpha_1 E_a(R_k)$;
    $E_v(R_k) = \alpha_2 E_v(R_k)$;
  $\Theta = \Theta + \{O_i\}$;
  $n++$;

2.3. Object Blending

By this time, representative objects have been selected from the tube, which will be blended with the background image. To achieve a seamless fusion between each object and background, Poisson editing method raised from [10] is adopted in this paper.

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

$$\min_{f} \iint_{\Psi} (\Delta f - \Delta g), \quad s.t. \quad f|_{\partial\Psi} = g|_{\partial\Psi} \tag{8}$$

3. EXPERIMENTS

To demonstrate the performance of the proposed approach, we test our approach on various surveillance video clips collected from both public dataset (PETS2009 [11]) and our own recorded videos.

Fig. 1 shows some results in the procedure of narrative generation. The changes of appearance and motion energy are illustrated in the left column. Note that the motion direction changes at Frame 18 and the appearance changes a lot at Frame 27 due to the occurrence of occlusion, and therefore, the corresponding energies drop lower at these objects. As the right column shows, with energy minimization, the representative objects are selected finally on the premise of spatial uniformity. As discussed before, compared to [4], two neighboring persons are also preserved in our resulting narrative as the contextual information to highlight the occlusion event at Frame 27, which reveals the behavior of object more comprehensively. Other examples are given in Fig. 3 and Fig. 4, generated from different video clips.

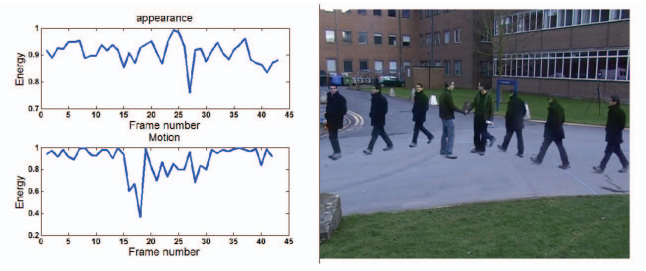

Fig. 1. An illustration of energy of a tube and its narrative.

To the best of our knowledge, few attempts have been to present an object-centered narrative like us. So it is difficult to compare our approach with others quantitatively. However, some objective criteria from [3] are borrowed with slight modifications in our evaluations. With these criteria, we compare our approach with video synopsis [5] and the method described in [4]. Evaluators were asked to give a score (5 is the highest while 1 the lowest) for each following questions:

- Are you "overall satisfied" with the presentation for a specific object in general?
- Do you consider this presentation is "representative" for an object in the video?
- Do you believe this presentation is "visual pleasing"?
- Do you believe this presentation is "compact" enough?
- Can you get to know the "storyline" from this presentation?

Fig. 2 gives the average score for each criterion. Compared with video synopsis [5], narrative presentation exceeds in characteristics of representative, visual pleasing and storyline due to the specificity. In addition, compared with [4], in spite of a slight degeneration in visual pleasing, our approach enhances the user experience of browsing in general by adding some contextual information.

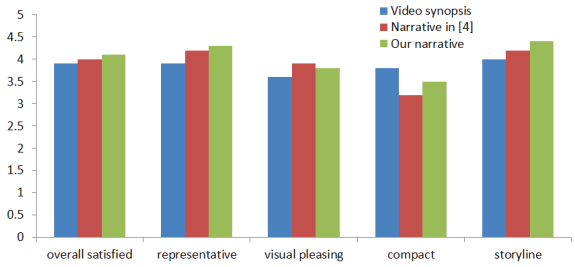

Fig. 2. Comparisons with video synopsis and narrative in [4].

Fig. 3. Six narratives generated from clips of two different scenarios in PETS2009 database.

4. CONCLUSIONS

In this paper, we have proposed a novel video presentation called video narrative to summarize a specific event of object in the source video. Representative objects are selected and illustrated in a single image, to maximally preserve the behavior or event information of the specific object. The experimental results are convincing for a first stage. We believe that this kind of video presentation has large applications in video indexing, fast browsing and video summarization.

Fig. 4. Two narratives from real surveillance videos.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 60833006, 61070104 and 60905008).

6. REFERENCES

[1] B.T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, 2007.

[2] D.B. Goldman, B. Curless, S.M. Seitz, and D. Salesin, "Schematic storyboarding for video visualization and editing," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 862–871, 2006.

[3] T. Mei, B. Yang, S.Q. Yang, and X.S. Hua, "Video collage: presenting a video sequence using a single image," *Vis. Comput.*, vol. 25, no. 1, pp. 39–51, 2008.

[4] W. Fu, J. Wang, X. Zhu, H. Lu, and S. Ma, "Video reshuffling with narratives toward effective video browsing," in *Image and Graphics (ICIG), International Conference on*, 2011, pp. 821 –826.

[5] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1971 –1984, 2008.

[6] C.D. Correa and K.L. Ma, "Dynamic video narratives," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 88:1–88:9, 2010.

[7] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *ICPR*, 2004, vol. 2, pp. 28 – 31.

[8] J. Sun, W. Zhang, X. Tang, and H. Shum, "Background cut," in *In ECCV*, 2006, pp. 628–641.

[9] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 65–81, 2004.

[10] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.

[11] A. Ellis, A. Shahrokni, and J.M. Ferryman, "Pets2009 and winter-pets 2009 results: A combined evaluation," in *PETS, IEEE International Workshop on*, 2009, pp. 1 –8.