21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

A Tracking Based Fast Online Complete Video Synopsis Approach

Lei SUN[1], Junliang XING[1], Haizhou AI[1] and Shihong LAO[2]
[1]*Computer Science and Technology Department, Tsinghua University, Beijing, 100084, China*
[2]*Development Center, OMRON Social Solutions Co., LTD, Kyoto 619-0283, Japan*
E-mail:ahz@mail.tsinghua.edu.cn

Abstract

By segmenting moving objects out and then densely stitching them into background frames, video synopsis provides an efficient way to condense long videos while preserving most activities. Existing video synopsis methods, however, often suffer from either high computation cost due to global energy minimization or unsatisfactory condense rate to avoid loss of important object activities. To address these problems, a tracking based fast online video synopsis approach is proposed in this paper which makes following three main contributions: 1) an online formulation of the video synopsis problem which makes the approach very fast and scalable to endless surveillance videos with reduced chronological disorders, 2) a tracking based schema which can preserve most object activities, and 3) a complete optimization process from both temporal and spatial redundancies of the video which results in much higher condense rate and less object conflict rate. Experimental results demonstrate the effectiveness and efficiency of proposed approach compared to the traditional method on public surveillance videos.

1. Introduction

Millions of surveillance videos are captured to be processed for abnormal incidents and criminal evidences detections. Examining these videos manually, however, is quite time-consuming and wasteful since there exists long periods of inactivity in these videos. Video synopsis, first proposed by Peleg et al. [2, 3], is different from the previous video abstraction approach [7]. It temporally shifts objects of different time intervals and then selectively stitches them into background frames to obtain a much shorter video. By condensing the time axis, it effectively preserves the dynamic aspect of object in video. However, it is an offline method and it cannot guarantee all dynamic objects appear in synopsis video.

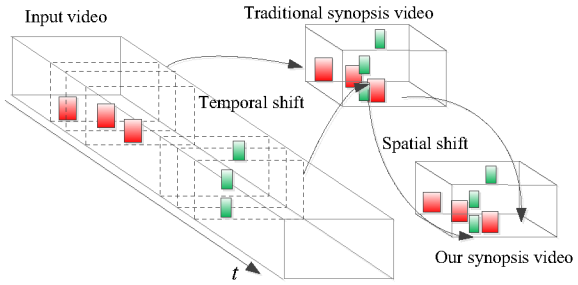


Figure 1: The traditional method only shifts two dynamic objects in the time axis. Our approach furthermore shifts the spatial location to avoid severe collision.

We present a tracking based fast online video synopsis method, which completely optimizes the redundancy of videos in both the temporal domain at the time axis and the spatial domain in the image plane. In this way, the conflict area of objects is decreased greatly and the condense rate thus can be further improved. As illustrated in Figure 1, two dynamic objects are temporally shifted with some spatial collisions, which in the traditional method one dynamic object would be either omitted causing object activity loss or shifted one or two frames behind causing relative longer synopsis video, while in our approach only two spatial shifts are applied to guarantee both low activity loss and shorter synopsis video. The whole optimization process of our synopsis problem is formulated in an online incremental manner which makes it very fast and scalable to videos of arbitrary lengths. Based on object tracking results[1], our approach can preserve most of the high level object activities such as object identities and moving trajectories in videos captured from both fix and moving cameras.

Some other related work includes [4] which applied eye-gaze clues to generate synopsis video. In [6], Yildiz et al. presented a novel approach to realize fast non-linear real-time synopsis by employing the non-linear image scaling method. Although [6] achieved real-time fast video synopsis, it reduced video redundancy at the image level instead of object level, causing it difficult to

incorporate with high level applications such as "who passed this spot" or "select persons going west" .

2. Online Tracking based Video Synopsis

The framework of our approach is illustrated in Figure 2. Objects are stored into tubelet (a short tube segment in the space-time volume) pool once tracked and they are then rearranged into the map (a mapping of tubelets from original video to synopsis video) based on energy minimization. A frame generating condition (FGC) is predefined to determine when to generate synopsis frames. Once it is satisfied, several frames are generated according to the final map and the tubelet pool is cleared afterwards. This procedure is iteratively performed to generate frames which are finally pushed into the stream of synopsis video. As a result, the total synopsis video is the assembling of short synopsis clips, each of which is generated from one certain period of input video.

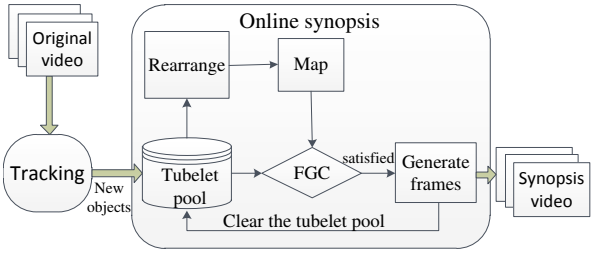


Figure 2: The framework of our approach

2.1. Problem Formulation

The original video with N_o frames is represented in a 3D space-time volume as $I(x, y, t)$, where (x, y) are the spatial coordinates of the pixel and $1 \leq t \leq N_o$. The synopsis video with N_s frames is represented as $S(x, y, t)$ in the same way, where $1 \leq t \leq N_s$. Each activity (tube) A_i is a dynamic object and its corresponding information is obtained from a tracking procedure (a sophisticated tracking algorithm[5] is adopted) to adapt to various high level applications, which produces a sequence of rectangle bounding boxes $\{R_t^i\}$ denoting its spatial locations, moving trajectories $\{T_t^i\}$ and speed vectors $\{S_t^i\}$ over its existing time interval $t_s^i \leq t \leq t_e^i$, formally,

$$A_i = (t_s^i, t_e^i, \{R_t^i\}, \{T_t^i\}, \{S_t^i\}) \tag{1}$$

A tubelet b is a short activity segment with predefined length L. several tubelets can be derived from an activity A_i as follows,

$$A_i = (\{b_j^i\}_{1 \leq j \leq l}, l = \left\lceil \frac{t_e^i - t_s^i}{L} \right\rceil) \tag{2}$$

where $b_j^i = (t_s^{ij}, t_e^{ij}, \{R_t^{ij}\}, \{T_t^{ij}\}, \{S_t^{ij}\})$, $t_s^{ij} \leq t \leq t_e^{ij}$.

Each synopsis clip is generated from a local spatiotemporal mapping M which arranges each tubelet b from original video into one tubelet $\hat{b}$ in synopsis video by shifting in both the temporal domain from its original time $t_{b_j^i} = [t_s^{ij}, t_e^{ij}]$ to the time segment $\hat{t}_{b_j^i} = [\hat{t}_s^{ij}, \hat{t}_e^{ij}]$ in the synopsis video and the spatial domain from its original spatial location $s_{b_j^i} = \{R_t^{ij}\}$ to the location $\hat{s}_{b_j^i} = \{\hat{R}_t^{ij}\}$ in the synopsis video. $M(b) = \hat{b}$ indicates the time and space shifts of the tubelet b , and when b is not mapped, $M(b) = \emptyset$. We optimize this local map M by minimizing an energy function defined on it.

2.2. Energy Definition

Let B represent the tubelet set in one synopsis clip, an energy function is defined over the local map M.

$$E(M) = \alpha \sum_{b \in B} E_a(\hat{b}) + \beta \sum_{b, b' \in B} E_c(\hat{b}, \hat{b'}) + \gamma \sum_{b \in B} E_s(\hat{b}) \tag{3}$$

where $E_a(\hat{b})$ is the activity cost, $E_c(\hat{b}, \hat{b'})$ is the collision cost and $E_s(\hat{b})$ is the spatial location cost. Weights α, β and γ are set by users according to their relative importance for a particular query.

1) Activity cost: we prefer "older" tubelets to be preserved since new tubelet may appear in the next synopsis clip and thus is not likely to be lost. We add one exponential term into the formulation of activity cost in [2], increasing the cost of "older" tubelets.

$$E_a(\hat{b}) = \sum_{x,y,t} \chi_b(x, y, t) \cdot \exp(-\frac{t - t_s}{\sigma_{time}}) \tag{4}$$

where t and t_s are the current frame and the start frame of the tubelet b, respectively. σ_{time} is the parameter to determine the extent of frame interval.

2) Collision cost: the collision cost calculates the weighted intersect area between two tubelets as in [2].

$$E_c(\hat{b}, \hat{b'}) = \sum_{x,y,t} \chi_b(x, y, t) \chi_{b'}(x, y, t) \tag{5}$$

3) Spatial location cost: we penalize the spatial location changing of the tubelet. Since the activity may loss its content if both spatial and temporal locations vary a lot, the range of the spatial shift is restricted.

$$E_s(\hat{b}) = \sum_{x,y,t} ||x_s|| + ||y_s|| \tag{6}$$

where (x_s, y_s) are the coordinate deviations in the spatial domain of center points of the corresponding bounding boxes.

In each synopsis clip generating process, all tubelets are temporally shifted to the beginning frame. And there is no any other temporal shift in one clip, which

largely decreases the high computation cost since each possible temporal shift causes entire re-calculation of overlapped tubelets while each possible spatial shift only results in the re-calculation at current frame. And it also mitigates the chronological disorders. Ignoring other temporal shifts is reasonable and practical as spatial shift already can effectively optimize the location of the overlapped objects and reduce the object conflict rate. As a result, the temporal consistency cost [2] is ignored in our approach.

Although considerable objects are omitted to obtain shorter synopsis video, it would be preferred to show all dynamic objects in the synopsis video. Since each dynamic object is detected by tracking with a unique identity which is attached to each segmented tubelet, no dynamic object loss can be easily realized by adding one hard restrain to energy minimization process that at least one tubelet within each derived dynamic object is included into the map. In other words, each object activity is partially maintained.

This online processing method relatively decreases the chronological disorders. Each synopsis clip consecutively generated in the chronological order can be indexed to a period of the input video. The disorder is existed since the tubelets within one synopsis clip concurrently appear regardless of real order in the input video. But it is relatively smaller and more stable than the energy guided method in [2] which may cause one tube in the earlier time to be shifted to the end of the synopsis video.

2.3. Energy Minimization

In order to realize fast online synopsis, we optimize the map in an incremental manner. Once new objects are added into the tubelets, we rearrange them into the last optimized map. We define a predict cost $P(b)$ for each tubelet b to evaluate the potential of including it into the map. It is updated once one new object is detected.

$$P(b) = \begin{cases} E^{pre}(\hat{b}) + E_{\min}(r), & \text{if } M(b) \neq \emptyset \\ P^{pre}(\hat{b}) + E_{\min}(r), & \text{others} \end{cases} \quad (7)$$

where $E_{\min}(r)=\min(E_a(r), E_c(r)-\lambda e_s(r))$. $E^{pre}(\hat{b})$ and $C^{pre}(\hat{b})$ are the energy cost and the predict cost in previous frame, respectively. $E_a(r)$ is the activity cost and $E_c(r)$ is the collision cost of the new object if it is removed or added into this tubelet. r is the bounding box of the new object. $e_s(r)$ measures the number of objects in the surroundings of this new object. It is also updated once the map is changed,

$$P(b) = \begin{cases} E(b), & \text{if } M(b) \neq \emptyset \\ P^{pre}(\hat{b}) + E_{\delta}(b), & \text{others} \end{cases} \quad (8)$$

where $E(b)$ is the current energy cost and $E_{\Delta}(b)$ is the increased or decreased energy cost under the new map.

The tubelets b of which $P(b)$ is lower than one threshold (it can be either predetermined or dynamic obtained by setting one certain proportion of selected tubelets) is mapped. Once the $P(b)$ of the tubelet b is updated, we reselect the tubelets. And the map is re-optimized only when selection of tubelets alters, otherwise we only simply either add or remove the new added object determined by in which case the energy cost is smaller. The determination of tubelet selection based on the predict costs fast forwards the minimization process. For re-optimization, we keep arbitrarily selecting one tubelet and optimizing it by spatial shifting or removing its objects within to achieve local minima energy until the deviation of total energy is small enough.

2.4. Compact Measurement of Scene

In this paper, the frame generating condition is "the scene is compact enough" , which indicates that the image space is fully utilized. We sum up both the size of all stitched objects denoted as A_a and overlapped area among objects expressed as A_o. Then a simple compact measurement of scene is denoted as C with the form

$$C = \frac{a \cdot A_a - b \cdot A_o}{S(I)} \quad (9)$$

where weights a and b can be set by the user. $S(I)$ is the image size of input video. A compact threshold C_{thre} is defined. Once C exceeds C_{thre}, one synopsis clip is generated according to the map.

3. Experimental Results and Discussions

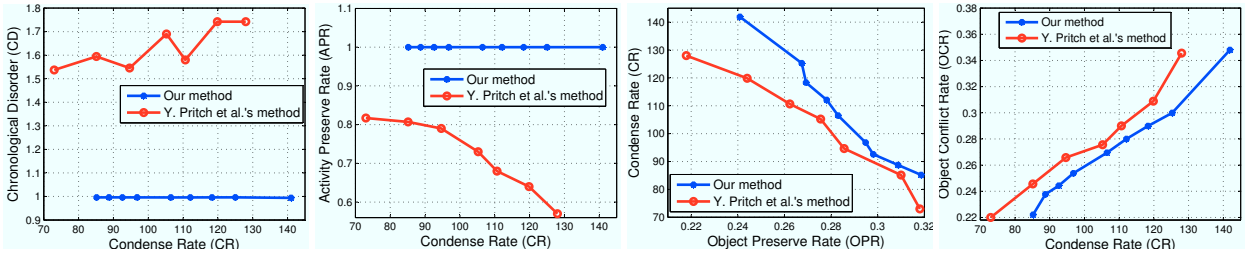
Assuming that the total dynamic object number and the total object number in the original video are D_o and O_o while the corresponding numbers in the synopsis video are D_s and O_s, and the amount of conflicted object in synopsis video is O_c , we adopt several metrics to evaluate our method and compare it with our implementation of the method in [2], which consists of *activity preserve rate* (*APR*), *object preserve rate* (*OPR*), *object conflict rate* (*OCR*), *condense rate* (*CR*) and *chronological disorder* (*CD*).

$$\begin{cases} APR=\frac{D_s}{D_o}, OPR=\frac{O_s}{O_o}, OCR=\frac{O_c}{O_s}, CR=\frac{N_o}{N_s}, \\ CD=\frac{1}{K}\sum_{A_i,A_j} ||1 - \frac{d_s(A_i,A_j)}{d_o(A_i,A_j)} \cdot CR|| \end{cases} \quad (10)$$

where $d_s(A_i, A_j)$ and $d_o(A_i, A_j)$ are the frame distances between activity A_i and A_j in synopsis video and original video. K is the total number of (A_i, A_j).

(a) (b) (c) (d)

Figure 3: The obtained background image ((a)) and three frames of synopsis video ((b)-(d)).



(a) (b) (c) (d)

Figure 4: Comparison of video synopsis results between our method and Pritch et al.'s method[2]

We test on a public surveillance video [7](720×480) with 17039 frames and 385 pedestrians, which can be processed 14-16 frames per second with our approach. Figure 3 shows the obtained background image and three frames of synopsis video containing 168 frames. Figure 4 gives the comparison results from which we can see the *chronological disorder* is much less and more stable with our online approach (Figure 4(a)). And we guarantee the appearing of all dynamic objects while the traditional method loses many especially when the *condense rate* is high (Figure 4(b)). Also, our approach achieves much higher *condense rate* with the same proportions of objects preserved (Figure 4(c)) and the *object conflict rate* is much lower (Figure 4(d)).

Though the spatial location may be altered in synopsis video, we record the trajectory information for video indexing or searching persons who passed a specific spot. Some other information could be as well recorded from tracking for high level applications such as the speed can be recorded to detect over-speeding driving and moving direction can be kept to filter persons who specifically go west. The synopsis video therefore could be even shorter under these specifications.

4. Conclusions

In this paper, we propose a fast online complete video synopsis approach. The online video synopsis is formulated in an incremental manner enabling fast scalable synopsis of endless video, which furthermore alleviates the chronological disorders. The object activities and high level information are effectively preserved and recorded based on tracking results to adapt to various applications. In addition, both spatial and temporal redundancies of video are completely optimized in our approach. The experiment verifies that our algorithm can efficiently perform in real time with a mitigated chronological disorder, a much higher condense rate and a less conflict object rate. In the future work, we will introduce action recognition into current framework to further keep the integrity of object activities in a synopsis video.

5. Acknowledgement

This work is supported by National Science Foundation of China under grant No.61075026.

References

[1] A. Chan, Z. Sheng, J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2008.

[2] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1971–1984, 2008.

[3] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2006.

[4] U. Vural and Y. S. Akgul. Eye-gaze based real-time surveillance video synopsis. *Pattern Recognit. Lett.*, 30(12):1151–1159, 2009.

[5] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2009.

[6] A. Yildiz, A. Ozgur, and Y. S. Akgul. Fast non-linear video synopsis. In *Int. Sym. Comput. Info. Sci.*, 2008.

[7] Y. Zhuangt, Y. Rui, T. S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *IEEE Int. Conf. Image Processing*, 1998.