Neurocomputing 135 (2014) 155–162

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Online video synopsis of structured motion

Wei Fu, Jinqiao Wang *, Liangke Gui, Hanqing Lu, Songde Ma

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, No. 95, Zhongguancun East Road, Beijing 100190, China

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form
11 December 2013
Accepted 19 December 2013
Communicated by Qingshan Liu
Available online 18 January 2014

Keywords:
Video synopsis
Motion structure

ABSTRACT

With the explosive growth of surveillance video data, video synopsis technology is presented for fast browsing a day's worth of video in several minutes. However, for most existing solutions, motion structure in original videos may be destroyed even considering the temporal consistency of related objects. To maintain the important context cues, in this paper, we propose an online motion structure preserved synopsis approach, which can preserve behavior interactions between different objects in the original video while condensing as much content as possible. By measuring sociological proximity of moving objects, we introduce motion structure as a refined term directly added to the problem of energy minimization. A hierarchical fashion is employed to efficiently search an optimal solution for the problem of video synopsis, in which both the spatial collision and the temporal consistency are considered. Experimental results on extensive videos demonstrate the promise of the proposed approach.

1. Introduction

The world witnesses a large amount of video data recorded for security purposes every day, but only a very small percent of them is truly investigated carefully. Browsing and indexing activities in these abundant videos are a time consuming and boring task for viewers. Most methods developed in the literature broadly fall into two categories: activity recognition [1] for specific events of interest predefined by users and video summarization [2] for a sketch of all activities in original videos.

As a prosperous area in video summarization, video synopsis [3] was presented, aiming at condensing content in both the spatial and the temporal dimensions, which allows the viewers to fast browse a day's worth of content in several minutes. Afterwards, various attempts have been made to generate visual pleasing synopsis videos in recent years. However, there still exist some limitations in current research, precluding the effectiveness and practicability of this technology.

Firstly, in the traditional offline synopsis process, all the moving object foregrounds and instantaneous backgrounds must be computed and stored in a large memory before optimization, which is a severe demand for the hardware. Meanwhile, since video synopsis is formulated as a problem of energy minimization, the large solution space often leads to a long time for optimization to search a good solution. Besides, the length of final synopsis video usually needs to be manually set by users as a matter of experience [4]. To address this problem, Feng et al. [5] proposed an online synopsis, in which object foreground sequences were filled into a spatio-temporal video volume one by one like playing a Tetris game. However, for the sake of a high condensation rate, the time consistency between different objects was not taken into account in this method.

The last but not the least, even though the temporal consistency is considered, moving objects in original sequences could be shifted to inappropriate locations where the behavioral interaction information may be sacrificed for the sake of avoiding severe spatial collisions. So it is difficult to explore important context cues directly from the synopsis video. To explain this point, we give a simple toy illustration in Fig. 1. Regarding all the moving objects existing in the original video, we divide the whole video into some segments according to their occupied periods, as shown in the top row of Fig. 1. In the spatio-temporal volume A, objects 2 and 3 cross each other in their moving process, indicating that a behavior interaction may happen between them. The same situation also exists in volume D for objects 8 and 9. In volume B, objects 4 and 5 stand for two persons walking together. To maintain original behavior interactions, objects occluded with each other or sharing a proximity motion structure are preferred to be shifted together in the synopsis video.

To this end, we propose an online motion structure preserved synopsis approach, which can maintain original behavior interactions while condensing as much content as possible. Inspired by the study of sociology, we measure motion structure between different moving objects with their motion proximity and intersection. Embedding this structure information in the final energy minimization problem, a hierarchical optimization method is utilized to efficiently search the optimal shift times for objects one by one in an online fashion.

2. Related work

There has been an increasing interest in video presentation and summarization for a long time, which is critical for video storage,

* Corresponding author.
E-mail address: jqwang@nlpr.ia.ac.cn (J. Wang).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.041

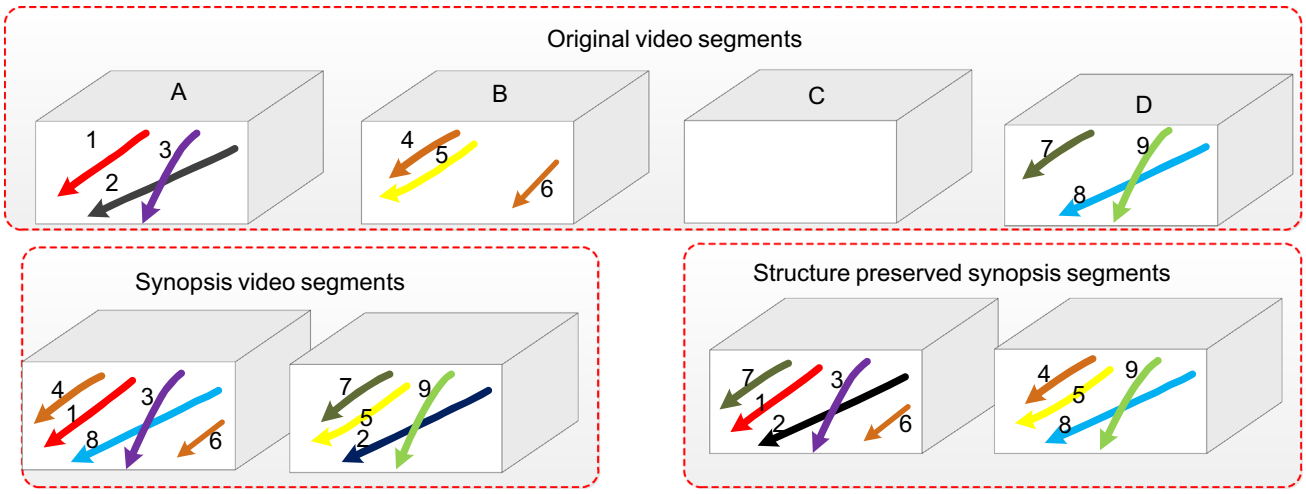


Fig. 1. A schematic illustration for structure preserved synopsis. The top row represents some segments from the original video. The bottom row shows the results by the traditional synopsis and structure preserved approach. In the spatio-temporal volume A, objects 2 and 3 cross each other in their moving process, indicating that a behavior interaction may happen between them. The same situation also exists in volume D for objects 8 and 9. In volume B, objects 4 and 5 stand for two persons walking together. By considering this motion structure information, our approach can maintain these behavior interactions in the synopsis video.

browsing and indexing. In this section, we will give a brief review of state-of-the-art techniques from two aspects: static image based summarization and dynamic content based video abstract.

For the first type of video summarization, key frames are usually selected to form a new representative image. In these methods, key frames are usually selected based on maximum frame discrepancy strategies. Visual features are extracted as the principle to compare the relational degree between frames [6–8]. Beyond a whole frame, some researchers generate new images using regions of interest (ROI). For example, video mosaic [9] is a synthetic representation by stitching successive video frames, covering more comprehensive information than a single key-frame. Another typical work is video collage [10] in which a video sequence was compacted to get a single image by seamlessly arranging ROIs on a given canvas. Storyboards [11] and narratives [12] represent the course of events by a static image with an explicit temporal cue.

The earlier work about dynamic video based summarization can be traced to video fast-forward [13] and video skimming [14], where significant content was extracted to generate a compacted video segment. Following along this route, various attempts have been made in the past few years. For example, in the space–time video montage method [15], spatio-temporal informative portions were extracted from a long video sequence and fused into a new short video volume. In dynamic video narrative [16], all duplications of a specific object are seamlessly stitched into the background video according to its time axis. In terms of a high condense rate, video synopsis [4,3] has made a big success and attracted the attention of many researchers. Feng et al. [5] proposed an online method, in which tubes were filled in a spatio-temporal volume one by one like playing a Tetris game. However, in their method motion structure was not considered, as well as the time consistency of tubes.

With the development of multimedia techniques, information from different media was employed to enhance video summarization in recent researches, both based on static image and dynamic video [17,18]. For example, Huang et al. [19] generated semantically meaningful montage by integrating text information from different media. Li et al. [20] utilized textual information from websites to enhance video summarization under a transferable structured learning framework.

In this paper, we focus on surveillance videos, in the context of which behavior interactions are important cues for effective content understanding. Therefore, we propose a novel motion structure preserved synopsis approach, where behavior interactions in the original video are treated as a refined factor for energy cost, and both the spatial collision and the temporal concordance are considered.

3. Motion structure

As the basic processing unit in video synopsis, a tube is defined as the spatio-temporal sequence of a moving object. For a tube T_i, we use a set of tuples $\{s_i, v_i, t_i\}$ to parameterize the trajectory, where s_i^t denotes the spatial position vector of T_i, and v_i^t is the velocity vector at frame t. While representing as much content as possible within a minimal length, an ideal synopsis video should reflect behavior interactions between different tubes in the original video as well. We measure motion structure from two aspects: motion proximity and intersection.

Proximity measurement: In our method, tubes with proximity movement are preferred to be shifted together to a common segment video. We assume that Γ is the temporal overlap of two tubes T_i and T_j. Informed by sociological models of collective behavior [21], an aggregated pairwise distance combining spatial proximity and velocity cues over time is defined to measure the motion proximity between two tubes

$$\begin{aligned}
&\omega_{ij} = \frac{\sum_t \omega_{ij}^t}{\rho_{ij}|\Gamma|}, \quad t \in \Gamma \\
&\omega_{ij}^t = \alpha \| s_i^t - s_j^t \| + (1-\alpha) \| v_i^t - v_j^t \| \\
&\rho_{ij} = \sum_t \delta_t(i,j)
\end{aligned} \tag{1}$$

where $\delta_t(i,j)$ is set to 1 if $\| s_i^t - s_j^t \| < \tau$ and $\| v_i^t - v_j^t \| < \tau$ and 0 otherwise, α is a weighting parameter. If there is no temporal overlap between T_i and T_j, we simply set $\omega_{ij} = +\infty$.

Intersection measurement: As another aspect manifestation of motion structure, spatial overlap between tubes also indicates the interaction of objects in original videos. Inspired by but different from the sticky tracking method described in [5], we measure the intersection of tubes using an accumulative Euler distance as follows:

$$\begin{aligned}
&I(i,j) = \begin{cases} 1, & \sum_t I_t(i,j) > k_2, \quad t \in \Gamma \\ 0 & \text{otherwise} \end{cases} \\
&I_t(i,j) = \begin{cases} 0, & d_t(i,j) > k_1 \\ k_1 - d_t(i,j) & \text{otherwise} \end{cases}
\end{aligned} \tag{2}$$

where $d_t(i,j) = \| s_i^t - s_j^t \|$ measures the distance of tubes T_i and T_j at time t, and the parameters k_1 and k_2 are two thresholds set by users (10 and 100 in our experiments). Intuitively, tubes with serious occlusions in a long enough time tend to be with a high $I_t(i,j)$, indicating that interaction may exist in the original video.

Motion structure measurement: Now, given two tubes T_i and T_j, we can measure their motion structure by integrating motion proximity and motion intersection. A sigmoid function is employed to depict the intensity of belonging to a group (later we call it a *tubelet*), where a and b are two parameters controlling its shape and central location, respectively (1 and 0 in our experiments):

$$S(i,j) = \frac{1}{1+e^{-a(\omega_{ij}-b)}} \cdot (1 - I(i,j)) \tag{3}$$

Obviously, two tubes with a high interaction tend to bring a low S (even 0), which will guarantee these tubes maintaining their relative interactions in the new synopsis video.

4. Online energy minimization

As a universal strategy of synopsis, in order to achieve a visual pleasing synopsis video, criteria are proposed and formulated as a series of energies. After that, the task becomes solving a problem of energy minimization. Therefore, the definition of energy is critical for the final performance of results.

Let Q denote the set of tubes that need to be processed in the original video. The essence of the synopsis process is to optimally rearrange tubes in Q to fill a new compact and short video. For this purpose, as a general strategy, criteria are proposed and formulated as a series of energy terms. Then the task of video synopsis becomes a problem of energy minimization as below:

$$\begin{aligned} E(\mathscr{\ell}) &= \sum_{i \in Q} E(\mathscr{\ell}_i) \\ E(\mathscr{\ell}_i) &= E_a(\mathscr{\ell}_i) + \sum_{j \neq i, j \in Q} E_p(\mathscr{\ell}_i, \mathscr{\ell}_j) \end{aligned} \tag{4}$$

where $\mathscr{\ell} = \{\mathscr{\ell}_i\}_{i=1}^{|Q|}$ denotes the set of play start times in the final synopsis video for Q. The first term measures the cost of adding tube T_i into a synopsis video at time $\mathscr{\ell}_i$ while the pairwise term penalizes the spatio-temporal collision between each of the two tubes.

Traditional offline strategy minimizes Eq. (4) once at a time after obtaining all tubes in the original video, but often leads to two problems. Firstly, all foregrounds of moving objects and backgrounds must be stored in a huge memory before optimization. Besides, searching in a large solution space takes a long time for optimization.

To bypass these problems, we follow an online stepwise tactic similar to [5], in which tubes are filled in a spatio-temporal volume one by one. Especially, for a new incoming tube T_i its optimal label time can be solved with a greedy algorithm

$$\begin{aligned} \mathscr{\ell}_i^* &= \underset{\mathscr{\ell}_i}{\operatorname{argmin}}\, E(\mathscr{\ell}_i) \\ \text{s.t.} \quad E(\mathscr{\ell}_i) &= E_a(\mathscr{\ell}_i) + \sum_{j \in Q'} E_p(\mathscr{\ell}_i | \mathscr{\ell}_j) \end{aligned} \tag{5}$$

where Q' denotes the set of tubes already processed, a subset of the whole tube set Q. In our approach, we simply regard $E_p(\mathscr{\ell}_i|\mathscr{\ell}_j) = E_p(\mathscr{\ell}_i, \mathscr{\ell}_j)$. In the following, we will present the definition for each term in Eq. (5).

Activity cost: Video synopsis favors a maximum activity presentation for the original video. Therefore, all moving objects in the original video should be shifted to the final synopsis video. In other words, we should avoid the leave out case for any tube. According to this criterion, we define an activity cost for each tube as follows:

$$E_a(\mathscr{\ell}_i) = \sum_{t \notin \tau_p} Area_t(i) \setminus \sum_{t_s \in \tau_s} Area_{t_s}(i) \tag{6}$$

where $Area_t(i)$ stands for the object area of tube T_i at time t, τ_p denotes the duration the result synopsis video lasts for, and τ_s is the tube life in the original video.

Spatial collision cost: This term is utilized to measure the degree of spatial overlapping between two tubes. A visual pleasing synopsis video should avoid serious occlusions, which is contradictory to the requirement of high condense rate. Assuming two tubes T_i and T_j with temporal locations $\mathscr{\ell}_i$ and $\mathscr{\ell}_j$, respectively, a spatial collision cost is defined to penalize the possible of spatial occlusions:

$$E_s(\mathscr{\ell}_i, \mathscr{\ell}_j) = \sum_{t \in \Gamma} \frac{O_{ij}^t}{\min\{Area_t(i), Area_t(j)\}} \tag{7}$$

where Γ denotes the common period of T_i and T_j, and O_{ij}^t stands for the spatial occlusion area at time t. Different from [3], spatial collision cost here is defined in the form of occlusion rate, instead of the absolute object size. It is helpful to guarantee the fair *presence right* for tubes with small sizes.

Temporal consistency cost: This term reflects the temporal relations between tubes in the original video. It is important for the cases with causality intersection. The temporal consistency cost creates a preference for maintaining the temporal relations between objects. Let t_i^s and t_j^s denote the play start times of tubes T_i and T_j in the original video. The temporal consistency cost penalizes cases where original relations are violated:

$$E_t(\mathscr{\ell}_i, \mathscr{\ell}_j) = \begin{cases} 0, & (\mathscr{\ell}_i - \mathscr{\ell}_j) \cdot (t_i^s - t_j^s) > 0 \\ e^{-d(i,j)} & \text{otherwise,} \end{cases} \tag{8}$$

where $d(i,j)$ measures the relative spatio-temporal distance [3] between T_i and T_j.

Embedding motion structure: The pairwise terms E_s and E_t reflect the interactions of tubes to a certain extent. Furthermore, integrating the motion structure information, the pairwise energy term is defined as

$$E_p(\mathscr{\ell}_i, \mathscr{\ell}_j) = (\beta E_s(\mathscr{\ell}_i, \mathscr{\ell}_j) + (1-\beta) E_t(\mathscr{\ell}_i, \mathscr{\ell}_j)) \cdot S(i,j) \tag{9}$$

where β is a weighting parameter balancing the effects of two terms. We can see that $S(i,j)$ plays a role of an attenuation factor, by which tubes with a similar motion structure are preferred to share a less cost.

5. Synopsis procedure

5.1. Hierarchical optimization

Without loss of generality, we assume that all N tubes are processed at one time. Given an incoming tube set $Q = \{T_i\}_{i=1}^N$, we would like to generate a synopsis video that displays all these tubes within a minimal length and at a least cost while preserving original motion structure by optimizing Eq. (5).

To optimize Eq. (5) with our structure preserved energy terms, various existing methods can be employed such as simulated annealing [22], graph cuts [23] and roulette wheel selection [5]. Following the route of stepwise optimization, we present a hierarchical optimization to accelerate this procedure by introducing the concept of *tubelet*. In our method a tubelet stands for a small group of tubes. And the hierarchical optimization for Eq. (5) includes the following three steps:

1. The tube set Q is roughly clustered into M different tubelets according to their motion structure proximity, $Q = \{G_j\}_{j=1}^M$.
2. Within each tubelet, tubes are rearranged in a spatio-temporal volume, i.e., each tube is set a relative play start time.

3. In the tubelet level, each tubelet is optimized to set a global play start time, which will be added to its members as a delay time.

For the first step, many techniques such as graph cut and spectral clustering can be employed to cluster tubes with a similar motion structure. However, note that the term $S(i,j)$ takes effect only when tubes occupy a common temporal overlap. Based on this observation, we can simply divide Q into M parts $\{G_j\}_{j=1}^{M}$ according to their original play times for the sake of efficiency in practice, and we call each part a *tubelet*.

In the second step, for each tube within a tubelet $G_i=\{T_{ik}\}_{k=1}^{K}$ (K is the size of G_i), an optimal time location is determined by solving Eq. (5) using a simplified competitive algorithm [24]. Firstly, the tube with a maximal length is firstly selected as the reference tube (marked as R_i) and filled into an empty spatio-temporal volume V_i. Then other tubes are selected out one by one and filled into the current spatio-temporal volume V_i. To be specific, for a new incoming tube T_{ik}, the competitive force function is defined as $CF(T_{ik})=e^{-E_p(V,\ell_{ik}^*)}$. After that, T_{ik} is placed at its optimal time ℓ_{ik}^* with the probability $p(T_{ik})=CF(T_{ik})/\sum_{ij\in G_i\setminus V} CF(T_{ij})$, where the optimal location ℓ_{ik}^* is determined by solving the following problem:

$$\ell_{ik}^*=\underset{\ell_{ik}}{\operatorname{argmin}}\sum_j E_P(\ell_{ik},\ell_j),\quad \text{s.t.}\, T_j\in V,\quad \forall j \tag{10}$$

For clarify, we summarize the procedure of rearrangement within each tubelet in Algorithm 1.

Algorithm 1. Tube rearrangement within each tubelet.

Input: A tubelet with K tubes $G_i=\{T_{ik}\}_{k=1}^{K}$
Output: $\{\ell_{ik}\}_{k=1}^{K}$, V_i
Initialization:
$\ell_{ik}=0,\quad \forall k\in\{1,\ldots,K\}$;
$V_i\leftarrow\emptyset$;
1. Select the reference tube R_i and update the tubelet:
$G_i\leftarrow G_i\setminus R_i$;
$V_i\leftarrow\{R_i,0\}$;
2. While $G_i\neq\emptyset$ do
Select a tube T_{ik} in G_i using the competitive algorithm;
Determine the located time ℓ_{ik} by Eq. (10);
$G_i\leftarrow G_i\setminus T_{ik}$;
$V_i\leftarrow V_i\cup\{T_{ik},\ell_{ik}\}$;
End while

In the third step, the roulette wheel selection algorithm similar to that in [5] is carried to determine the global play start time for each tubelet. Tubelets are filled into the final synopsis volume S one by one. Note that a tubelet instead of a tube becomes the process unit in this step, energy terms should be changed correspondingly. Especially, let $D=\{d_i\}_{i=1}^{M}$ be the global play start times for $\{G_i\}_{i=1}^{M}$. Then the pairwise energy term can be defined as

$$E_p(d_i,d_j)=\sum_{ik,jm}E_p(\ell_{ik},\ell_{jm}),\quad \text{s.t.}\quad T_{ik}\in G_i,\quad T_{jm}\in G_j \tag{11}$$

And we define the fitness of a new incoming tubelet G_i arranged at time d_i as

$$fit(d_i)=\exp\left\{-\sum_{j,G_j\in S}E_p(d_i,d_j)\right\} \tag{12}$$

Based on this, G_i is placed at d_i with the probability $p(d_i)=fit(d_i)/\sum_{d_j\in[0,len(S))}fit(d_j)$, where $len(\cdot)$ is a function computing the temporal length of a spatio-temporal volume. A roulette wheel is spun to determine the delay times of tubelets one by one.

5.2. Dynamic increment

In all the stepwise approaches, a matter worthy to be considered is the *discontinuity* of motion flow. At the high-layer in the hierarchical optimization, imagine that another new tubeset containing N tubes comes after the preceding one just being filled into a new volume. Note that tubes spatio-temporally extend in the final volume, separately optimizing this incoming tubeset without considering the current volume state will leave a cutoff between two consecutive tubesets, called a *discontinuity* of motion flow. We give a toy illustration in Fig. 2. More proof can be found from the comparison video in our supplementation.

Supplementary material related to this article can be found online at http://dx.doi.org.10.1016/j.neucom.2013.12.041.

In order to address this problem, we employ a dynamic increment method which is simple yet efficient. To be specific, tubes already filled in the final volume are truncated into two parts, i.e., the body parts and tails. When another tubeset comes, the former tails with fixed time stamps are first added to constitute a new tubeset. And then they will be optimized together with new incoming tubes. In this dynamic increment way, the discontinuity could be avoided. In addition, the length of final synopsis video can be determined automatically as well.

6. Experimental evaluation

To evaluate the performance of the proposed approach, we carried experiments on diversified types of videos including both public datasets and videos recorded by ourselves. All these videos

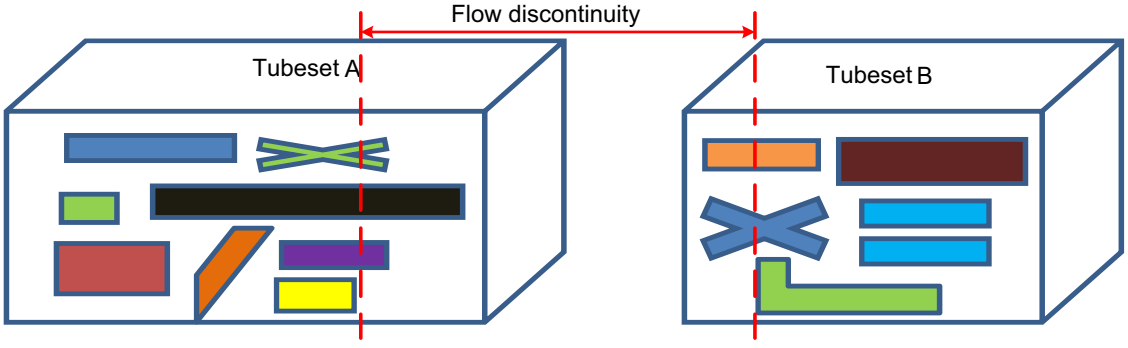


Fig. 2. A toy illustration for flow discontinuity. Separately optimizing tubeset B without considering the current state of the ST volume (filled with A) will lead to a cutoff between the two tubesets.

Table 1
A summary of video description.

Video description	# Frame	# Tube	CR (%)
Square, 320 × 240, 25 fps	4924	40	7.62
Park, 320 × 240, 25 fps	4435	81	18.4
Road1, 320 × 240, 25 fps	2891	18	3.46
Road2, 320 × 240, 25 fps	12 577	289	14.11
IndoorGTTest1 [25], 320 × 240, 14 fps	2659	9	12.64
Institute, 320 × 240, 25 fps	12 328	137	17.85

were captured by surveillance cameras. In order to obtain smooth and accurate foreground segmentations, a method combined the Gaussian mixture model and min-cut [23] is utilized for background modeling. Table 1 presents a summary description of these videos, as well as the condensation rate (CR) by our approach. The condensation rate denotes the frame number ratio between synopsis and original videos, which is related to the original activity density or user settings. Fig. 3 presents an intuitive impression result where a single frame from each synopsis video is selected to exhibit for intuitive impressions.

Fig. 3. Four synopsis frames for Square, Park, Road2, and IndoorGTTest1, respectively.

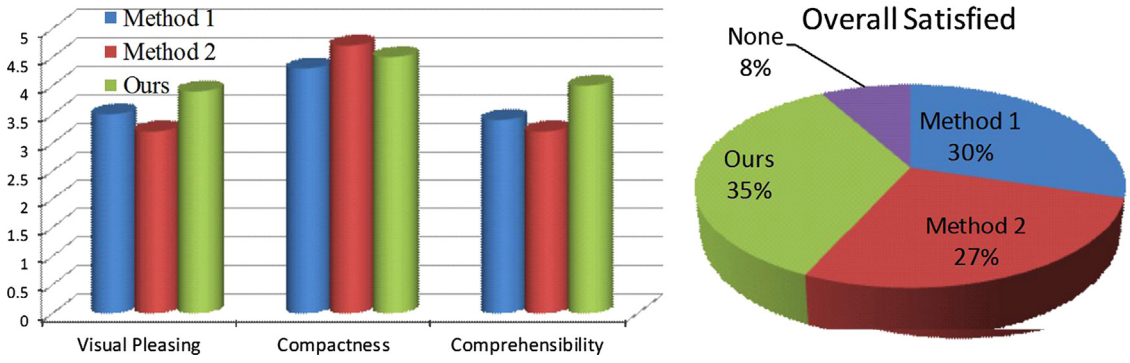


Fig. 4. User study for different synopsis methods.

Fig. 5. The comparison result for Road1 with different selection strategies. SA_1000 stands for offline simulated annealing [22] through 1000 iterations.

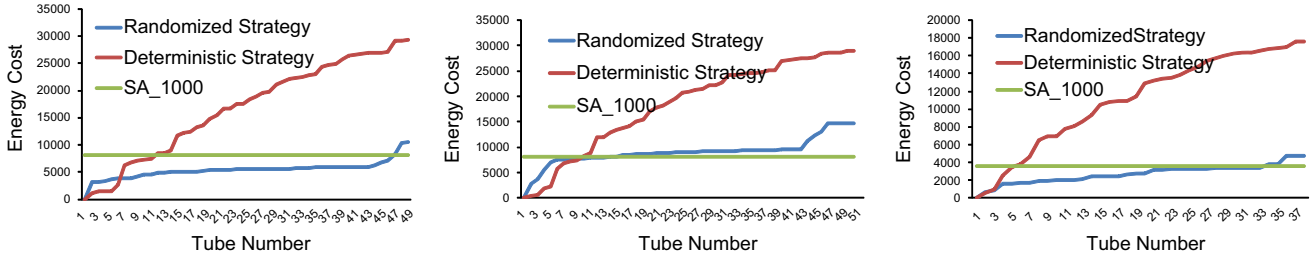


Fig. 6. The energy cost for Institute sequence with different selection strategies.

User study: We reproduced two methods described in [3] (Method 1) and [5] (Method 2) as baselines. Four evaluation criteria are proposed for rating, i.e., visual pleasing, compactness, comprehensibility, and overall satisfied, as follows:

1. *Visual pleasing*: Do you think this synopsis is comfortable for view?
2. *Compactness*: Is this synopsis compact enough?
3. *Comprehensibility*: Can you infer the original behavior interactions from this synopsis?
4. *Overall satisfied*: Which is your most satisfied synopsis overall?

The first criterion reflects the vision comfortable degree, the compactness criterion reflects the object density of the synopsis, and the comprehensibility criterion reflects whether the original movement information can be inferred from the synopsis results. For the first three survey items, the score scale is 1–5, where 1 is the lowest and 5 is the highest. For the last one, evaluators are requested to point out the most satisfied synopsis (give 1 for the corresponding synopsis) or give 0 score if they are unsatisfied with any synopsis videos.

Given the evaluation criteria, we invited 37 participants to score the synopsis results. All the participants have strong background knowledge in video surveillance, aged from 25 to 43. They were requested to watch the original videos first, then watch the synopsis, and give their ratings at the end from 4 aspects: visual pleasing, compactness, comprehensibility and overall satisfied. Finally, classified according to the methods, statistics results of subjective feedbacks are illustrated in Fig. 4.

We can see that, on the premise of guarantee for visual pleasing and compactness, our approach achieves a better performance in comprehensibility by maintaining the original motion structure information.

Different selection strategies: In the process of online synopsis, a key point is how to select a tube or tubelet to fill a spatio-temporal volume. A simple strategy is to keep selecting the current best tube at every turn. As opposed to this deterministic strategy, in our approach, we introduce randomness in our selection process through a competitive or roulette wheel algorithm. This randomized strategy could bring a better solution by considering the future or unknown tubes when making a decision. To give a quantitative evaluation, we examine the tendency of energy cost in the tube filling process, which could reflect the performance of final results in a way. Fig. 5 shows the comparison result for Road1 as an example. More details can be found from videos in the supplementation. Two synopsis frames are also illustrated. For another Institute sequence containing 137 tubes, 50 tubes are treated at one time and the comparison results for each tubelet are illustrated in Fig. 6. We can see that the cumulative energy cost could decrease greatly by introducing randomness in the selection process, which usually indicates improvement of performance.

Evaluation for motion structure: Quantitative evaluation and comparison for the capability of motion structure preservation is difficult by the lack of benchmark datasets with known ground-truth. In order to give a quantitative result, we invited 3 human coders to label the Institute sequence and its corresponding synopsis video, in which collective behaviors of pedestrians appear widely. Human coders were instructed to examine carefully whether structure information of these groups are preserved. Individuals with interactions in small groups are annotated. The final labels determined by coders are translated into a numeric score for each pedestrian according to the group size they belong to: 1 for single pedestrians, 2 for pedestrians in pairs and 3 for triplets or more complex groups with a 87.07% match rate as shown in Fig. 7.

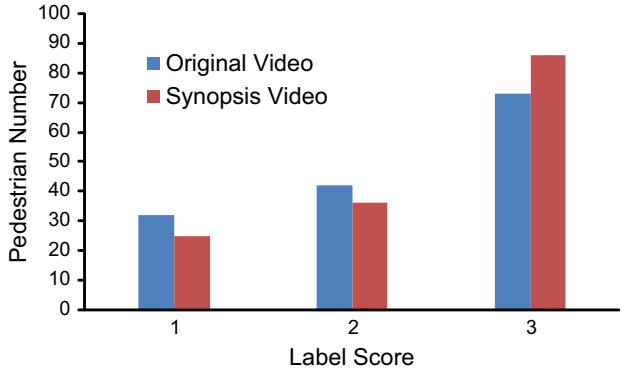


Fig. 7. Label scores from coders for pedestrians in the Institute sequence and its corresponding synopsis video.

Fig. 8. Top: five typical frames from the Institute video. Bottom: three frames selected from the synopsis video.

We would like to present some more intuitive results. In Fig. 8, the top row shows five typical frames from Institute video, and three synopsis frames in the bottom row. In this video, motion structure between pedestrians reflects in terms of their collective behavior. We can see that this group information is preserved in the final synopsis video. More proof can be found from the video in our supplementation.

7. Discussion and future Work

The problem of preserving the motion structure of moving objects which either follow each other or cross each other, while disrupting their chronological order to save time, is worth considering especially in surveillance settings. In this paper, we propose an online motion structure preserved video synopsis method. Measured by motion proximity and intersection, the motion structure is formulated as a refined term to take effect on the final energy minimization. Embedding this information, the final synopsis video could condense as much activities as possible while maintaining their behavior interactions. Experiments on various videos have demonstrated the effectiveness and good potential applications of our approach.

However, due to the introduction of behavior interaction, the optimization problem appears to take more computational time than the previous work. Currently in our approach, a hierarchical fashion optimization is utilized to accelerate the process, which leads to local optimization results. A more suitable optimization method to an approximate minimizer will be studied in our future work. In addition, the synopsis technology still has some limitations within the scope of application by itself. Imagine a video already full of moving objects. It becomes less meaningful to summarize the activity in a significantly short synopsis video. A new automatic synopsis framework, independent of the original activity density, is another future direction.

Acknowledgments

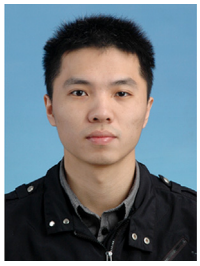
This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61273034, 61070104, and 61202325).

References

[1] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[2] A.G. Money, H. Agius, Video summarisation: a conceptual framework and survey of the state of the art, J. Vis. Commun. Image Represent. 19 (2) (2008) 121–143.
[3] Y. Pritch, A. Rav-Acha, S. Peleg, Nonchronological video synopsis and indexing, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1971–1984.
[4] A. Rav-acha, Y. Pritch, S. Peleg, Making a long video short: dynamic video synopsis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 435–441.
[5] S. Feng, Z. Lei, D. Yi, S. Li, Online content-aware video condensation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2082–2087.
[6] Y. Peng, C.-W. Ngo, Clip-based similarity measure for query-dependent clip retrieval and video summarization, IEEE Trans. Circuits Syst. Video Technol. 16 (5) (2006) 612–627.
[7] J.-M. Odobez, D. Gatica-Perez, M. Guillemot, Video shot clustering using spectral methods, in: Proceedings of 3rd International Workshop on Content-Based MultiMedia Indexing, 2003, pp. 94–102.
[8] Z. Li, G. Schuster, A. Katsaggelos, Minmax optimal video summarization, IEEE Trans. Circuits Syst. Video Technol. 15 (10) (2005) 1245–1256.
[9] M. Irani, P. Anandan, Video indexing based on mosaic representations, Proc. IEEE 86 (5) (1998) 905–921.
[10] X. Liu, T. Mei, X.-S. Hua, B. Yang, H.-Q. Zhou, Video collage, in: Proceedings of the 15th International Conference on Multimedia, 2007, pp. 461–462.
[11] D.B. Goldman, B. Curless, S.M. Seitz, D. Salesin, Schematic storyboarding for video visualization and editing, ACM Trans. Graph. 25 (3) (2006) 862–871.
[12] W. Fu, J. Wang, C. Zhao, H. Lu, S. Ma, Object-centered narratives for video surveillance, in: 19th IEEE International Conference on Image Processing (ICIP), 2012, pp. 29–32.
[13] N. Petrovic, N. Jojic, T.S. Huang, Adaptive video fast forward, Multimed. Tools Appl. 26 (3) (2005) 327–344.
[14] M. Smith, T. Kanade, Video skimming and characterization through the combination of image and language understanding techniques, in: Computer Vision and Pattern Recognition, 1997, pp. 775–781.
[15] H.-W. Kang, X.-Q. Chen, Y. Matsushita, X. Tang, Space–time video montage, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1331–1338.
[16] C.D. Correa, K.-L. Ma, Dynamic video narratives, ACM Trans. Graph. 29 (4) (2010) 88:1–88:9.
[17] A. Bagga, J. Hu, J. Zhong, G. Ramesh, Multi-source combined-media video tracking for summarization, in: ICPR, 2002, pp. 818–821.
[18] B.-W. Chen, J.-C. Wang, J.-F. Wang, A novel video summarization based on mining the story-structure and semantic relations among concept entities, IEEE Trans. Multimed. 11 (2) (2009) 295–312.
[19] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, B. Shahraray, Automated generation of news content hierarchy by integrating audio, video, and text information, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 6, 1999, pp. 3025–3028.
[20] L. Li, K. Zhou, G.-R. Xue, H. Zha, Y. Yu, Video summarization via transferrable structured learning, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 287–296.
[21] C. McPhail, R.T. Wohlstein, Individual and collective behaviors within gatherings, and riots, Annu. Rev. Sociol. 9 (1) (1983) 579–600.
[22] S. Kirkpatrick, C.D. Gelatt , M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.
[23] V. Kolmogorov, R. Zabih, What energy functions can be minimized via graph cuts?, in: Proceedings of the 7th European Conference on Computer Vision-Part III, 2002, pp. 65–81.
[24] M. Manasse, L. McGeoch, D. Sleator, Competitive algorithms for on-line problems, in: Proceedings of the 20th Annual ACM Symposium on Theory of Computing, STOC '88, 1988, pp. 322–333.
[25] L.M. Brown, A.W. Senior, Y.-li Tian, J. Connell, A. Hampapur, C.-fe Shu, H. Merkl, M. Lu, Performance evaluation of surveillance systems under varying conditions, in: Proceedings of IEEE PETS Workshop, 2005, pp. 1–8.

Wei Fu received his B.S. degree from University of Science and Technology of China (USTC), Hefei, China, in 2009. He is currently a Ph.D. candidate at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include pattern recognition and machine learning, image and video processing, and intelligent video surveillance.

Jinqiao Wang received the B.E. degree in 2001 from Hebei University of Technology, China, and the M.S. degree in 2004 from Tianjin University, China. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently an Associate Professor with Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.



Liangke Gui received his B.E. degree from Shandong University, Jinan, China, in 2012. He is currently a Ph.D. candidate at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition and machine learning, image and video processing, and video surveillance.

Hanqing Lu received his B.E. degree in 1982 and his M. E. degree in 1985 from Harbin Institute of Technology, and Ph.D. degree in 1992 from Huazhong University of Sciences and Technology. Currently, he is the deputy director of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include image and video analysis, medical image processing, object recognition, etc. He has published more than 200 papers in these fields.

Songde Ma received his B.S. in Automatic Control from the Tsinghua University in 1968, Ph.D. degree in University of Paris in 1983 and "Doctorat d'Etat es Science" in France in 1986 in image processing and computer vision. He was an invited researcher in Computer Vision Laboratory in the University of Maryland in USA in 1983. From 1984 to 1986, he was a researcher in Robot Vision Laboratory in INRIA, France. Prof. Ma was a member of the Expert Committee of the National High Technology Program and the chief scientist of the Project "Image, Voice and Natural Language Understanding" of the National Fundamental Research Program. His research interests include computer vision, image understanding and searching, robotics and computer graphics, etc.