A Set Theoretical Method for Video Synopsis

Min Xu[1], Stan Z. Li[2], Bin Li[1], Xiao-Tong Yuan[2], Shi-Ming Xiang[2]
[1]Dept. of Electronic Science and Technology, University of Science and Technology of China, Hefei, China
mxu@mail.ustc.edu.cn, binli@ustc.edu.cn
[2]Biometrics and Security Research & National Laboratory of Pattern Recognition, CASIA, Beijing, China
{ szli, xtyuan, smxiang }@cbsr.ia.ac.cn

ABSTRACT

A synopsis video presents a condensed video activities occurring during different periods, based on moving objects extracted in the spatial-temporal domain. How to place different object tubes with least collision in limited video length is crucial to synopsis performance but has not been thoroughly studied in previous work.

In this paper, we address an important problem in video synopsis, that of object start-time programming. We formulate the problem in terms of set theory. An objective is derived to maximize visual information in video synopsis. After relaxing the problem to obtain a continuous one, the problem can be efficiently solved via mean-shift. The resulting algorithm can converge to the local optimum within a few iterations.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Indexing Methods*

General Terms: Algorithms

Keywords: video synopsis, object start-time programming, set theory, mean shift

1. INTRODUCTION

Video synopsis, which generates a short video with the most information of the original video, is a problem of great interest in both research and industry. For example, a 24 hours surveillance video needs to be condensed into a short period without losing any activity to support efficient browsing and retrieval. This representation is significantly different from the traditional video summarization techniques such as key frame representation [2, 11] and video skimming [4, 5, 8]. The synopsis preserves the dynamic characteristic of the original video and changes the relative time between activities to reduce spatial-temporal redundancy.

Recently, there is a trend of using combination of extracted objects from different periods to represent synopsis video [6, 7]. To create the final synopsis video, a temporal mapping of the object start-time from the original video to the synopsis video should be founded with more objects and less spatial-temporal overlap. Here, preserving the chronological order of events isn't considered. We should have in mind that video synopsis is just an index of the original video for the convenience of browsing. One can go back in the original video to know the details according to the synopsis instead of considering the temporal consistency with much computation complexity and spatial-temporal redundancy.

Our work tries to tackle the object start-time programming problem to maximize the visual information in the synopsis video. To enable this, the total spatial-temporal positions in the synopsis video are considered as a universal set, and each object tube can be considered as a set too. Then the problem of visual information maximization is translated to maximize the cardinality of the union set between object tubes in the set theory. According to DeMorgan's law, the equivalent proposition is to minimize the cardinality of the intersection set between the complement sets of object tubes. So it can be expressed by a discrete combinatorial optimization problem which can be solved by continuous relaxation via mean-shift. The proposed algorithm can converge to the local optimum within a few iterations. Experimental results illustrate the validity of our method.

2. RELATED WORK

There are two main kinds of approaches in traditional video summary: keyframe-based methods and videoclips-based methods. In the former kind, a few key frames [2, 11] are selected from the original video. The key frames are the ones that best represent the video, however this representation loses the dynamic aspect of video. In the latter kind, a collection of short video sequences [4, 5, 8] best representing original video's contents are abstracted. The dynamics of the video remains, while the defect is less compact. In both kinds above, each frame in the original video is either shown completely or not shown at all in the synopsis video.

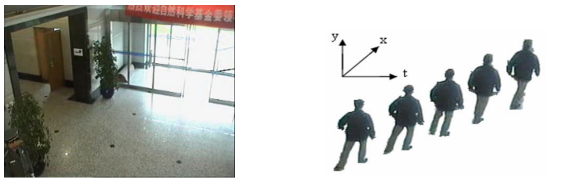
Recently, object-based approaches for video synopsis have been presented in literature [6, 7]. Moving objects, represented in the spatial-temporal domain, are combined to create the synopsis even if they have happened at different periods. Activity cost, collision cost and temporal consistency cost are considered to construct energy function between object tubes for the allocation of the tubes in the synopsis. Simulated annealing is used for solving this problem. This is a comprehensive description of object start-time program-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-312-9/08/10 ...$5.00.

ming, however it suffers from the disadvantage of high computation cost. As video synopsis is just an assistant tool to help people know the outline of a video, you should find the corresponding clip in the original video to get more information. Thus, the temporal consistency cost can be ignored when the computation complexity is reduced significantly without considering this cost.

3. SET COMBINATORICS FOR SYNOPSIS

Suppose that we have got background images and moving object tubes based on mixture Gaussian model [9] and objects tracking [10]. Examples of extracted background and tube are shown in Fig. 1.



(a) background image (b) extracted tube

Figure 1: Examples of extracted background and . tube.

In the following section,we first illustrate the importance of object start-time programming in video synopsis. Then we formulate this problem in set theory.

3.1 Importance of Object Start-time Programming

A simple instance in Fig. 2 is given to show the importance of Object Start-time Programming. Fig. 2(a) shows the original video. There are five object tubes. Each tube is in one color. Here, for simplicity, we ignore tubes' changes in axis x. So the available information of the original video is in 13 slices of time.

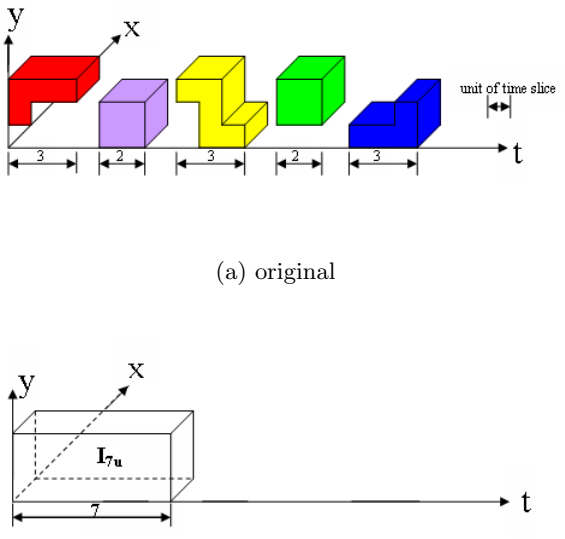
Now the task is to compact the tubes in 7 slices of time with the most preserving information. A cuboid container with the length of 7 units in Fig. 2(b) is a vivid representation of the synopsis' boundaries. Object tubes are placed in this container as many as possible with least overlapping.

Without any order exchange of the tubes, we can make the synopsis like Fig. 2(c). As the yellow tube conflicts with the red and blue tubes, the losing information is 9.5%. But if we do some changes like Fig. 2(d), the visual information reaches the maximum.

The importance of object start-time programming is illustrated from the above example. So the next problem is how to formulate it.

3.2 Optimal Combinations of Set Elements

Suppose that the synopsis video has T frames with the size of $W \times H$. Therefore there are $W \times H \times T$ spatial-temporal positions marked by universal set I_T, such as I_{7u} in Fig. 2(b). Here, u denotes one slice of time. Further, N object tubes are extracted from the original video, and set $O_{n,t}$ denotes tube n with start-time in frame t, for example the red tube can be denoted as $O_{0,0}$ in Fig. 2(c) and $O_{0,4u}$ in Fig. 2(d). $(x, y, z) \in O_{n,t}$ means that tube n is visible in spatial-temporal position (x, y, z), vice versa. Here, (x, y)



(a) original

(b) cuboid container for synopsis

(c) without exchanging (d) after exchanging

Figure 2: Visualization for importance of object . start-time programming

are the spatial coordinates, and $0 \leqslant z \leqslant T-1$ is the frame ID.

Then the problem of maximizing visual information can be mathematically expressed as maximizing the cardinality:

$$\max_{t_0,t_1,\ldots,t_{N-1}} |\bigcup_{n=0}^{N-1}(O_{n,t_n} \cap I_T)| \quad (1)$$

where$|\cdot|$denotes the cardinality of the set, and set-intersection operation to I_T because only the tube's information appearing in the synopsis video is valuable. Using De-Morgan's law, maximizing the visual information is just equivalent to minimizing the rest space without any visual objects, so through set-complement operation the equivalent proposition is

$$\min_{t_0,t_1,\ldots,t_{N-1}} |\bigcap_{n=0}^{N-1}\overline{O_{n,t_n} \cap I_T}| \quad (2)$$

where $\bar{\cdot}$ denotes the complement of the set.

We use function $\{F_{n,x,y}(z,t)|z,t=0,1,...,T-1\}$ to denote $O_{n,t}$, here (x, y) denotes the spatial location in a frame. If $(x, y, z) \in O_{n,t}$, $F_{n,x,y}(z,t) = 1$, vice versa. So expression(2) can be reformulated as the following objective func-

tion:

$$\widetilde{E}(t_0, t_1, ..., t_{N-1}) = \sum_{x,y} \sum_{z=0}^{T-1} \prod_{n=0}^{N-1} [1 - F_{n,x,y}(z, t_n)] \quad (3)$$

As the tubes can and only can shift along the temporal axis, it satisfies

$$F_{n,x,y}(z, t) = F_{n,x,y}(z - t, 0) \quad (4)$$

Where $F_{n,x,y}(z, 0)$ is defined as:

$$F_{n,x,y}(z, 0) = \begin{cases} 1, & (x, y, z) \in O_{n,0}. \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

In this way, the problem of object start-time programming is translated into a combinatorial optimization problem with T^N feasible solutions in Eq.(3).

4. FINDING OPTIMAL SOLUTIONS

In this section, we present a method for solving the combinatorial optimization. We seek to optimize Eq.(3) by continuous relaxation.

4.1 Continuous relaxation

Function $F_{n,x,y}(z, 0)$ can be optimized by convolution with Gaussian kernel as follows:

$$f_{n,x,y}(z, 0) = \sum_{k=0}^{T-1} F_{n,x,y}(k, 0) e^{-s(z-k)^2} \quad (6)$$

where $s > 0$ is the scale coefficient, and will be interpreted in Sect. 4.2.

Similar to Eq.(4), we define $\widetilde{f}_{n,x,y}(z, t_n)$ with continuous variable z and discrete variable t_n as follows:

$$\widetilde{f}_{n,x,y}(z, t_n) = f_{n,x,y}(z - t_n, 0) \quad (7)$$

In the same way, function $F_{n,x,y}(z, t_n)$ with respect to t_n is continuous, and then followed by replacing $\widetilde{f}$ in Eq.(7) and f in Eq.(6):

$$\begin{aligned} f_{n,x,y}(z, t_n) &= \sum_{j=0}^{T-1} \widetilde{f}_{n,x,y}(z, j) e^{-s(t_n-j)^2} \quad (8) \\ &= \sum_{j=0}^{T-1} f_{n,x,y}(z - j, 0) e^{-s(t_n-j)^2} \\ &= \sum_{j=0}^{T-1} \sum_{k=0}^{T-1} F_{n,x,y}(k, 0) e^{-s(z-j-k)^2} e^{-s(t_n-j)^2} \end{aligned}$$

Continuous objective function similar to discrete function Eq.(3) can thus be defined as:

$$E(t_0, t_1, ..., t_{N-1}) = \sum_{x,y} \sum_{z=0}^{T-1} \prod_{n=0}^{N-1} [M - f_{n,x,y}(z, t_n)] \quad (9)$$

where $M = \max_{0 \leqslant z, t_n \leqslant T-1} \sum_{j=0}^{T-1} \sum_{k=0}^{T-1} e^{-s(z-j-k)^2} e^{-s(t_n-j)^2}$ is the maximum of $f_{n,x,y}(z, t_n)$.

In the above continued process of Function $F_{n,x,y}(z, t_n)$, variable z and t_n except x or y are continuous. The reason is that object tubes can only shift along the temporal axis, and only variable z and t_n are defined in the temporal axis. On the other hand, object tubes can't shift along the spatial axes, so variable x and y defined in the spatial axes can't be continuous.

4.2 Algorithm

Fixing $t_0, t_1, ..., t_{i-1}, t_{i+1}, ..., t_{N-1}$, the start-time t_i of tube i that minimizes Eq.(9) is obtained by solving the gradient equation of $E(t_0, t_1, ..., t_{N-1})$ as the following fixed point iteration:

$$t_i^{m+1} = \frac{\sum_{j=0}^{T-1} C_i(j) e^{-s(t_i^m - j)^2} j}{\sum_{j=0}^{T-1} C_i(j) e^{-s(t_i^m - j)^2}} \quad (10)$$

where

$$C_i(j) = \sum_{x,y,z} \prod_{n \neq i} [M - f_{n,x,y}(z, t_n)] \sum_{k=0}^{T-1} F_{i,x,y}(k, 0) e^{-s(z-j-k)^2}. \quad (11)$$

Notice that the iteration (10) is essentially a mean-shift optimization algorithm [1]. Denote by $\{t_i^m\}_{m=1,2,...}$ the sequence of successive start-time locations of the tube i. As the kernel $e^{-s(t_i-j)^2}, s > 0$ has a convex and monotonically decreasing profile, the sequences $\{t_i^m\}_{m=1,2,...}$ converges, as proved in [1]. $1/\sqrt{s}$ is the bandwidth of the kernel.

While this deterministic algorithm is fast, it finds a local optimum. To improve this, an annealing type of algorithm could be incorporated into mean-shift, such as using an adaptive annealing robust estimator [3]. This algorithm performs robust estimation (a mean-shift like algorithm) of the peak of a distribution by varying or annealing the kernel parameter and approximately finds the global peak.

5. EXPERIMENTAL RESULTS

We tested our method using two video streams. As the frame rate is not constant from different video streams, we use the number of frames rather than the absolute time in the presentation.

5.1 Video Data

The first video as in Fig. 4 is taken in a hall under constant lighting condition by a static camera with 320×240 pixels in size. 35 tubes with the total length of 1961 frames are contained in the original video and condensed into 100 frames in the synopsis video. Fig. 4(a) shows the original images with sparse objects. In the synopsis process, each tube's initial start-time is scaled from the original start-time in order to preserve the chronological order of events as much as possible.

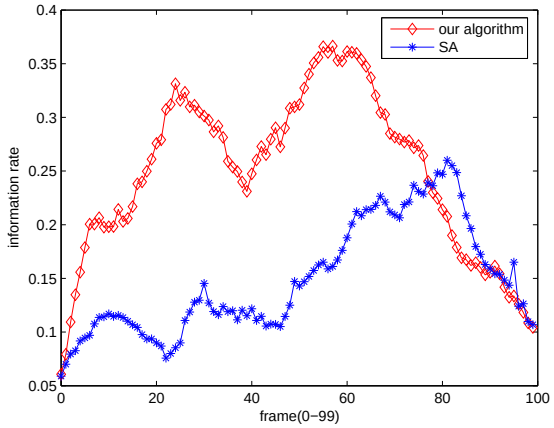
The second video as in Fig. 5 is taken in a car park under varying lighting condition with 352×288 pixels in size. 20 tubes with the total length of 327 frames are contained in the original video and condensed into 40 frames in the synopsis video. Fig. 5(a) shows the original images with sparse objects.

5.2 Results and Performance

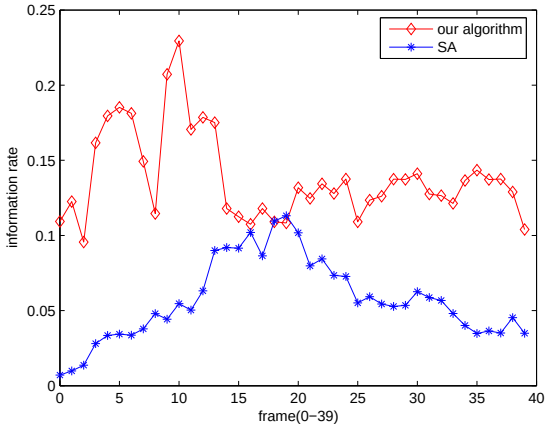
Some synopsis images are shown in Fig. 4(b) and Fig. 5(b).

The performance is evaluated in terms of *information rate* and computing time. We define the information rate(IR) to measure the synopsis efficiency in each frame, as follows:

$$IR = \frac{\textit{the number of pixels occupied by objects}}{\textit{the total number of pixels}} \quad (12)$$



(a) information rate in Experiment 1



(b) information rate in Experiment 2

. **Figure 3: information rate between SA and our algorithm**

Higher information rate means more visual information can be seen in the synopsis.

We compare our algorithm with the simulated annealing method used in [6] (using our own implementation). There, the temporal consistency cost together with the activity and collision cost are considered. However, in our experiment only the activity and collision cost without temporal consistency are computed. This is because the latter one disagrees to the former ones in most cases.

The results are shown in Fig. 3. All the information rate curves are below 0.5 for the reason of all the objects' absence in some regions in the original video. The results show that our method has higher information rate than the simulated annealing based method in most frames.

Our algorithm converges after two cycles in 4 minutes for the first experiment and two cycles in 2 minutes for the second one. In contrast, simulated annealing is well know to be slow (but not reported in [6]).

6. DISCUSSIONS

This work presents a set theoretical formulation for video synopsis, and provided an efficient algorithm which can find a local solution within a few iterations. Currently, the neighborhood information among pixels in each tube has not been fully utilized, as the tube can only move along the temporal axis and so each pixel itself rather than its neighborhood is considered. So in the future work, how to make use of tube's neighborhood information for the purpose of the search speed will be taken into account.

7. ACKNOWLEDGMENTS

This work was performed in Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. It is supported by the following funds: Chinese National Natural Science Foundation Project #60518002, Chinese National 863 Program Projects #2006AA01Z192, #2006AA01Z193, and #2006AA780201-4, Chinese National Science and Technology Support Platform Project #2006BAK08B06, and Chinese Academy of Sciences 100 people project, and Authen-Metric R&D Funds.

8. REFERENCES

[1] D. Cimaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[2] C. Kim and J. Hwang. An integrated scheme for object-based video abstraction. In *Proc. of the 8th ACM Int'l Conf. on Multimedia*, pages 303–311, 2000.

[3] S. Z. Li. Robustizing robust M-estimation using deterministic annealing. *Pattern Recognition*, 29(1):159–166, 1996.

[4] J. Nam and A.Tewfik. Video abstract of video. In *Proc. of 3rd IEEE Workshop on Multimedia Signal Processing*, pages 117–122, September 1999.

[5] N. Petrovic, N. Jojic, and T. Huang. Adaptive video fast forward. *Multimedia Tools and Applications*, 26(3):327–344, 2005.

[6] A. Rav-Acha, Y. Pritch, A.Gutman, and S. Peleg. Web synopsis: Peeking around the world. In *Proc. of IEEE on International Conference on Computer Vision*, October 2007.

[7] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *Proc. of 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 435–441, June 2006.

[8] A. M. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. In *Proc. of IEEE Workshop on Content-Based Access of Image and Video Database*, pages 61–70, 1998.

[9] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 8(22):747–757, 2000.

[10] T.Yang, S. Z. Li, Q. Pan, and J. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *Proc. of 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 970–975, June 2005.

[11] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Syst.*, 10(2):98–115, 2004.

(a) images from the original video

(b) images from the synopsis video

. **Figure 4: Experiment with the hall video.**

(a) images from the original video

(b) images from the synopsis video

. **Figure 5: Experiment with the car park video.**