Surveillance Video Synopsis via Scaling Down Moving Objects

Surafi PS  
PG Student  
Department of Electronics & Communication Engineering  
Marian Engineering College, Trivandrum, India

Hema S Mahesh  
Assistant Professor  
Department of Electronics & Communication Engineering  
Marian Engineering College, Trivandrum, India

Abstract

This paper produces an approach for video synopsis through which shorter and condensed form of video can be obtained. Most of the spatiotemporal redundancies in surveillance video can be avoided and essential activities in the video can be preserved here. In most of the existing approaches collisions are caused while video is condensed. Most of the collisions can be avoided by reducing size of moving objects. Mainly three steps are used. Interesting objects are identified and detected in first step. An optimization framework technique is used in second step in order to find energy mapping values. Finally scaled down objects are added back to the video of user defined length video.

Keywords: Video Synopsis, Surveillance, Collision, Optimization Framework, Reduce Size

I. INTRODUCTION

With the development of our society number of surveillance cameras is increasing nowadays. Most of these work 24 hours a day and this makes the video analyzing time consuming [1]. So most of these videos are never reviewed after they are made. These videos contain too much redundant information and important activities are less. Video synopsis give shorter condensed version of the original video with important events preserved. This is the efficient index of the original video by calculating the time of the interesting object. According to the time label interesting objects can be find out from the original video easily. In addition to this video synopsis can also reduce memory storage for surveillance video so that memory storage can be reduced. So far there is no specified method to determine if the synopsis is good or bad. Because the quality of the synopsis video depends on the content of the original video and the requirement of the application. Besides different users have different views which makes the criteria more difficult. Generally there are some qualitative criteria for video synopsis, they are as follows. Output video should preserve the activities present in the original video by reducing the most of the spatiotemporal redundancies. Then the collisions among the active objects should be reduced as much as possible. Next the consistency of the moving objects should be kept as far as possible in the synopsis video.

There are many approaches proposed for video synopsis. Some of them are based on frames in which the entire video frames are considered as the fundamental building blocks that cannot be decomposed. Commonly in frame based approaches the video is compressed along the time axis which allows the user to see the summary of the entire video. One of these methods is fast forward [2], in which several video frames are skipped. Unfortunately there is a chance of missing fast moving objects while skipping video frames. To overcome this some methods are proposed [3]-[7] where according to some criteria key frames are extracted or some video clips without activities are skipped. However frame based approaches tends to lose dynamic aspect of the original video and retain empty spaces of the background in the synopsis [1], [8]. To reduce most of these empty spaces of the original video a number of different approaches have been proposed to retain informative video portion from the original video together [9]-[12]. To avoid these problems [1], [13], [14] proposed an object based approach which is the mainstream for surveillance video synopsis field. In object based approach objects are first extracted from the original video then they are converted to temporal domain. It can avoid most of the spatiotemporal redundancies but causes unwanted collisions among active objects in the synopsis video. To avoid further collisions and improve compression ratio some spatiotemporal approaches are proposed. These methods are capable of avoiding most of the collisions but both temporal and location information of objects is violated.

II. RELATED WORKS

There are mainly three approaches in video synopsis field which are frame based video synopsis, object based video synopsis and part based video synopsis.

A. Frame based Video Synopsis

It can be subdivided into two classes: key frame selection approaches and video skimming. With respect to key frame selection approach a series of important frames are selected from the original video according to certain criteria. Among the all key frame approach technique uniform temporal sampling is the simplest but it may extract less important frames also. McMillan [15]
proposed a non-uniform sampling approach to create time lapsed video summaries to create users duration. Adaptive fast forward techniques were developed in [3] and [6] to avoid the loss of fast activities when deleting frames. To select key frames more correctly and accurately, some criteria for selection were proposed in [4], [5], and [16]–[18]. These approaches aim at eliminating video clips with no activities or discarding frames of low interest according to user’s definition. In addition to the video frame, multimedia information can also be utilized for video content analysis such as sound to find key frame. In [19]–[21], auxiliary information is taken into consideration to help select key frames for synopsis video.

B. Object based Video Synopsis

An important concept was introduced by [9], [10], [21] which combined activities coming from different times rather than selecting and connecting the whole video frames. These approaches take video portions as building blocks, and first extract informative video portions from the original video. Then, the video portions are shifted and stitched together according to optimization techniques such as simulated annealing and graph cut. Although more activity information in the original video are presented simultaneously, obvious discontinuity usually appear between different portions, which results in unpleasant visual effects. Rav-Acha et al. [13] identified and extracted moving objects that can be shifted along the time axis to create a compact and continuous synopsis. Similar with [10], this approach also disturbed the chronological order of moving objects to get a more condensed video than previous methods. Pritch et al. [3] expanded the previous work [13] to deal with the continuous video streams captured by webcams. A more detailed approach was proposed in [1] which preserved the local chronological order of moving objects present in the frames. It can be seen as the unification and expansion of approaches represented in [13], [33]

C. Part based Video Synopsis

More recently, a novel approach called part-based object movement’s synopsis is proposed. The traditional object-based video synopsis approaches cannot reduce the redundant information in the movements of objects as it is difficult. Different from the previous work, object parts are used as the basic building blocks in [12]. An object is first decomposed into different parts and each part corresponds to a movement sequence. Then, slightly changing subsequence is discarded and the rest are assembled and stitched together optimally to get a compact video. This approach can reduce redundancies in a better level through eliminating several unimportant movement sequences.

III. PROPOSED METHOD

In this section, proposed approach is described in detail. To show the effectiveness of the proposed approach, an example is presented in Fig. 2. Given a source video, if it is condensed using traditional synopsis approaches including [1], [13] and [14], visual collisions will occur when the objects are shifted along the time axis as shown in Fig. 1(a) and (b). It can be observed from Fig. 1(a) that objects move across each other, which results in bad artifacts. In Fig. 2(b), there are collisions among moving objects and the scene is crowded. To decrease collisions and eliminate unwanted crowded visual effects, we attempt to reduce the sizes of objects when collisions occur. The corresponding condensed results produced by proposed approach are presented in Fig. 1(c) and (d). As can be seen from the comparison between Fig. 1(a) and (c), collisions is much fewer objects are scaled down. Besides, the comparative result of Fig. 1(b) and (d) indicates that reducing the sizes of objects can also avoid congested phenomenon in the condensed synopsis.

Fig. 2 presents the main procedure of our approach which can be mainly divided into three parts. First, moving objects are identified and segmented using Background subtraction. Second, an optimization framework is proposed to determine an optimal mapping from the input video to the synopsis video. The moving objects will be shifted in the temporal domain and the reduction coefficient of each object in order to avoid collision will be calculated in this part. Finally, each object is stitched into the reconstructed background according to the results obtained in the second part

![Fig. 1: (a). Collisions occur while shifting objects along the time axis. (b) Crowded conditions occur. (c) Collisions are reduced through decreasing the sizes of objects. (d) Collisions are fewer, and the scene is not so crowded as before](Image)
A. Object detection and segmentation

In object-based approaches, object detection and segmentation are most essential steps. Recently, many detection and segmentation approaches have been proposed [15]–[20]. But, they are not appropriate to address our problem. For video synopsis, a qualitative criterion is to preserve the most interesting activities. Generally, similar with [1] and [8] we define a moving object as an interesting one for synopsis. However, it must be also noted that not all the moving objects in the frames should draw the user’s attention and not all static objects can be ignored. For example, the swing of leaves is not the significant information for us, while a motionless man may be important according to user’s definition. Hence, these exceptions must be considered in object detection stage.

Similar with [1], [8], [10], and [13], a background subtraction method is employed here to extract the moving objects. Before segmenting objects from the original video, we have to reconstruct the background. In most cases, the surveillance camera is static. So that the background changes very slightly and slowly due to the variation of illumination or the vibration of camera. Based on this observation, a temporal median over several video clips which is divided into a fixed number of frames is used to represent the background of the current frame. In this method, the video clips within one minute (30 seconds before and after the current frame) are used to calculate the corresponding median value. In this case, the objects which are remaining stationary for a long time will become a part of the background. Then moving objects are identified according to the difference between the current frame and its corresponding background extracted. For segmentation of the moving objects, first construct a mask of foreground pixels, and apply 2D morphological dilation and erosion on all the mask frames to get a more precise result.

Usually, some noise, such as the motion of leaves, which is unwanted, is included in the foreground pixels by background subtraction approaches. Therefore, Aggregated Channel Features (ACF) detection is also employed to extract the moving objects. The ACF detector can produce a precise bounding box for each moving object. After that, a large area of adjacent foreground pixels inside the bounding box is treated as a moving object.

B. Optimization Framework

In this section, we will introduce the detail about optimization framework in detail which penalizes the loss of activity information, the collision artifacts, and the size reduction of object segmented and the violation of relationships between moving objects. By using the proposed framework, an appropriate mapping result can be obtained to indicate the new time positions of detected objects in the synopsis video and determine the corresponding reduction coefficients for the synopsis video. Video synopsis based on this mapping result will maximize activity information and preserve the temporal relationships between moving objects while avoiding
most collisions that can occur. Let E represent the energy of a function and M represent a mapping from the source video to the synopsis video. The objective function of the proposed approach can be defined as follows:

\[ E(M) = E_a(M) + E_c(M) + E_r(M) + E_o(M) \]  \hspace{1cm} (1)

Where \( E_a(M) \) indicates the activity cost of the current mapping, \( E_c(M) \) is the collision cost, \( E_r(M) \) represents the reduction cost, and \( E_o(M) \) denotes the temporal consistency cost. The activity cost term \( E_a(M) \) is prone to preserve most observations when an object is mapped to the synopsis video. It can be defined as

\[ E_a(M) = \sum o^* \]  \hspace{1cm} (2)

In Eq. (2), O represents the set of all the tubes, and \( o^* \) is a tube with new time positions and new size of object, which corresponds to the mapping result of tube \( o \). As moving objects are represented by tubes, these two concepts can be interchangeable in later formulas.

The second term of Eq. (1) is the product of \( E_c(M) \) and \( E_r(M) \) as shown. \( E_c(M) \) restricts collisions generated by shifting moving objects along the time axis, and \( E_r(M) \) prevents the reduction operations on the sizes of objects. Reducing the sizes of moving objects is done to decreasing collisions, but the size cannot be reduced indefinitely. When reducing the size of an object, the collision cost will be decreased and the collisions between moving objects can be avoided to a great extent. But there is a chance that the reduction cost will be increased and the objects in synopsis video will be difficult to identify. In this case, the multiplied form of these two cost terms \( (E_c(M) \text{ and } E_r(M)) \) is constructed to achieve a compromise between alleviating collisions and obtaining identifiable frames from the condensed synopsis video. These two cost terms can be defined as follows, respectively:

\[ E_c(M) = \sum \gamma \]  \hspace{1cm} (3)
\[ E_r(M) = \sum \alpha \]  \hspace{1cm} (4)

In Eq. (3), \( o^* \) and \( p^* \) represents the mapping results of moving object \( o \) and \( p \), respectively. \( \alpha \) is a weight which can be changed according to the needs of users. For example, if the collision cannot be tolerated, we can set a larger value to \( \alpha \) to reduce the collision. Then, the importance of \( E_c(M) \) will be increased, collision phenomenon in the optimization process will reduced to a greater extent. Consequently, the final mapping result contains only fewer collisions. In Eq. (4), \( \gamma \) is a set of reduction coefficients, \( \gamma_o \) is one of them, which corresponds to the object \( o \). \( \gamma_o \) denotes the area of the bounding box of object \( o \). \( E_r(M) \) prevents the serious violation of temporal relationships between moving objects, and it can be defined as follows:

\[ E_r(M) = \sum \gamma \]  \hspace{1cm} (5)

Where \( \gamma \) is also a weight. If we need to preserve the temporal relationships among objects as much as possible, then \( \gamma \) should be set a larger value to increase the importance of \( E_r(M) \) in the objective function. We can also say that, reducing the value of \( \gamma \) can also decrease the collisions to a certain extent. Because disorder sequence of objects may be more conductive to reduce overlapping areas.

On the whole, the undesired observations such as collisions will increase the energy of the objective function in the optimization process. Therefore, the mapping with minimum energy value will preserve activity information to the large extent and also the temporal relations among objects are preserved. It can also reduce collisions and avoid scaling down the objects to a much smaller size. Gaussian mixture model (GMM) can be used in order to classify the detected moving objects and to estimate its new position. So that the collisions caused by the moving objects can be avoided to a greater extent.

C. Stitching Objects into Background

After the optimization process, objects that are scaled down are stitched back into the background to create the synopsis video. For stitching the moving objects, a background video with user-defined length is generated and objects are added into the synopsis. In previous section we have mentioned that for each frame in the original video, a corresponding background image is reconstructed using a temporal median over a few video clips. Among these reconstructed background images, a uniform temporal sampling of active periods is employed in order to produce the background video. This approach is simple but only efficient for short videos. For more complicated background video generation approach, the time lapse background in [1] can be included. Since the perfect segmentation for all the objects is not possible, direct stitching without any process will produce visible seams.

In this case, Poisson Image Editing [12] can be employed. Poisson Image Editing is an effective image fusion tool which can reduce the change of gradient and fuse two images together smoothly. Thus, this approach can reduce the seams, and it is widely
used in many object based video synopsis techniques [1], [8]. Furthermore, we can also calculate the time that an object appears in the original video by utilizing the frame rate and the frame original number. Then, we can mark the time for the corresponding object in the synopsis, which can be an effective index of the original video and make up for some temporal inconsistency that can occur.

**ACKNOWLEDGEMENT**

I would like to express my gratitude to all other faculty members of department of electronics and communication

**REFERENCES**


