A Quick Browsing System for Surveillance Videos

Cheng-Chieh Chiang
Dept. of Info. Tech., Takming Univ. of Sci. and Tech.
Taipei, Taiwan
kevin@csie.ntnu.edu.tw

Ming-Nan Tsai
Taipei Municipal Nei-Hu High School
Taipei, Taiwan
edwen@nhsh.tp.edu.tw

Huei-Fang Yang
Dept. of Computer Sci. and Eng., Texas A&M Univ.
College Station, TX, USA
hfyang@cse.tamu.edu

Abstract

Surveillance cameras have been widely installed in large cities to monitor and record human activities for different applications. Since surveillance cameras often record all events 24 hours/day, it necessarily takes huge workforce watching surveillance videos to search for specific targets, thus a system that helps the user quickly look for targets of interest is highly demanded. To this end, we propose a quick surveillance video browsing system. Our basic idea is to collect all of moving objects which carry the most significant information in surveillance videos to construct a corresponding compact video by tuning positions of these moving objects. The compact video rearranges the spatiotemporal coordinates of moving objects to enhance the compression, but the temporal relationships among moving objects are still kept. The compact video can preserve the essential activities involved in the original surveillance video. This paper presents the details of our browsing system and the approach to producing the compact video from a source surveillance video.

1. Introduction

Surveillance cameras are widely installed in large cities to monitor and record human activities either in inside or outside environments. To efficiently utilize surveillance videos, how to extract valuable information from hundreds-of-hours videos becomes an important task. An intuitive method is to retrieve relevant segments according to the user's queries in surveillance videos. Unfortunately, it is still difficult to automatically understand the user's intentions and the video contents. Video retrieval [2][7] based on the semantic level is still a challenging task.

Video understanding and analysis [8][12] is a key technology when we try to automatically extract valuable information from hundreds-of-hours surveillance videos. A basic method is to extract video segments that can represent most of informative contents of the source video. These informative segments are also known as key frames [1][13] in a video. The collection of key frames is the simplest way to compactly represent a video. Other possible methods for video understanding and analysis are called video summarization [3][10] or video abstraction [11] that extracts key frames based on a semantic level from a video and fuses them to form a shorter one. However, it is difficult to define proper criteria to determine what a key frame is. Moreover, some information may be lost if it is not involved in the selected key frames. In a real application for the surveillance video, all of moving objects may not be negligible because it may be necessary to preserve them for the witness while crimes occur.

In general, the most informative parts involved in surveillance videos are the foreground, i.e., the moving parts appearing in video frames. If we can collect all, or at least most, of moving objects and compact them in a very short video, the users may be more easily look for what they want. It is similar to a movie trailer: the user can roughly understand the contents of a movie by watching the trailer and then determine whether to watch the movie or not. A. Rav-Acha et al. [9] proposed the dynamic video synopsis to shorten videos by defining an energy function that describes activities of moving objects in a video. The energy function is minimized to optimally compress the corresponding behaviors of the moving objects to form the video synopsis. Their method can achieve a very large compression ratio in video representation, with destroying temporal relationship among objects. It may be difficult to focus on correct targets when the user looks for subjects of interest in surveillance videos. P. DeCamp et al. [6] designed an interactive browsing and visualization system called HouseFly for home-surveillance applications. HouseFly synthesizes audio and visual data/metadata from multiple sensors to generate an immersive and virtual world.

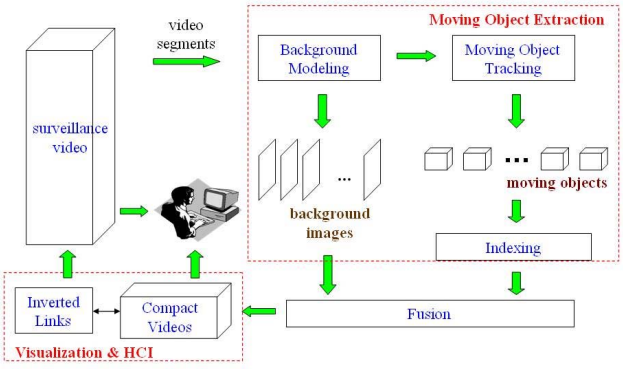


Figure 1. The flowchart of our quick browsing system for surveillance videos.

In this paper, we design a quick browsing system to help the user easily locate the subjects of interest in surveillance videos, showing the flowchart in Figure 1. For each short-time segment from surveillance videos, a background model can be constructed under the assumptions of the fixed camera view and the unchanged lighting, thus the corresponding background images are generated. We employ a tracking process, based on the MeanShift method [5], to extract all of moving objects associated with the background model. A compact frame is generated by first tuning the appearing location and time of objects and then pasting them on the background

image. The compact video is the collection of all compact frames. The compact video not only compactly represents for a copious surveillance video but also preserves all essential components of moving objects appeared in the source video. Also, the temporal relationship among objects can be preserved. Moreover, inverted links from the moving objects in the compact video to those in the source video are also constructed. The user can browse the compact video to fast locate targets of interest and then the system can directly display the corresponding video clip selected by the user. Using our system, the user can spend only several minutes watching the compact video instead of hours monitoring a large number of surveillance videos.

This paper is organized as the follows. Section 2 and 3 introduce the details of moving object extraction and indexing, respectively, in our design. Section 4 describes the generation of the compact video. The details of the system design are presented in Section 5. We finally draw conclusions and future works Section 6.

2. Moving Object Extraction

To well extract objects from a video is an important but challenging task in an outside environment of the real world due to possible changes in lighting, weather, etc. Therefore, we assume that there is only a slight change in the environment in a short period of time (e.g., 10 minutes in our system). To simplify our work, it is also assumed that the image view of the surveillance camera is fixed. Based on our assumptions of the fixed camera view and the unchanged lighting, we construct the background model for each of video segment and employ the tracking process to identify moving objects appearing in the surveillance video.

The background model is constructed for each of video segments. We set the length as 10 minutes for each segment and employ the openCV functions [14] to build a Gaussian background model in the implementation. Certainly, other complex models such as [4] can be applied. All pixels not belonged to the background parts in a video frame are assigned to the foreground. In order to be more robust, the erosion and dilation processes are applied to the foreground regions. Moving objects can be considered as the foreground parts when the background model is dynamically constructed. To avoid the influence of noises, we discard small connected components and perform the MeanShift method [5] to track each of moving objects.

The foreground parts that are extracted according to the background modeling can form moving objects in video frames, and tracking these moving objects can improve the robustness. However, it is still difficult to avoid mistakes in the extraction. For example, one object may be split under the influence of the lighting, or two objects may be merged due to their being too close. All unexpected factors make it hard to exactly extract and track moving objects in general situations. However, this problem is not critically concerned in our work. Even though a moving object is fragmental, it finally appears. The user can watch them, either intact or fragmental, in the compact video to determine whether he clicks the object to directly link to the source of the surveillance video or not.

3. Indexing

All moving objects extracted from the source of surveillance videos are indexed and stored in the database. The main advantage is the flexibility: we can easily extend the capabilities of the system if keeping the information of moving objects. Different types of attributes of moving objects can be attached into the index. For each moving object, the stored information in our implementation contains: the start time and the end time of the appearance, the appearing coordinates in video frames, the width and the height of the minimal rectangle covered the object, and the real area of the connected component of the object.

4. Compact Video

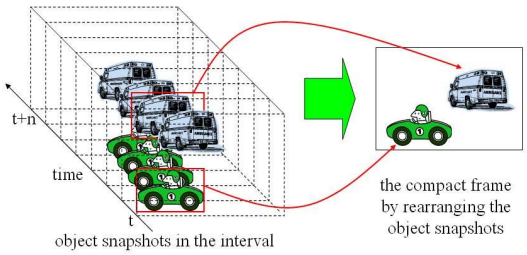


Figure 2. The compact frame is the fusion of the object snapshots in the time interval with n video frames by rearranging their temporal and spatial coordinates.

After all of moving objects are indexed, the next task is to determine their new positions on the background image to generate the corresponding compact frame. Figure 2 depicts the basic idea of generating compact frames. Given continuous video frames in a time interval with length n, the corresponding compact frame is a collection of object snapshots which appear in the interval. We take one snapshot for each appearing moving object and paste it on the background image to form the compact frame.

The most important factor in our design is how to determine the representative snapshots, in the n continuous frames, pasted on the compact frame for all moving objects. An intuitive idea is to compute the average of appearing areas for each object in frames of the time interval. This approach is simple but may produce fragmental objects. Hence, we adopt the median as the representative snapshot for each object in frames of the interval to form the compact frame. The new spatial coordinates of median snapshots in the compact frame are the same as whose coordinates in the original frames. Moreover, since compact frames are generated according to the time order of source video frames, the appearing order of moving objects in the compact videos can be kept. Preserving their time order is an important property for monitoring moving targets in surveillance videos. If not, it may be difficult to focus on correct targets when the user looks for subjects of interest.

The time interval of frames, denoted as n, is a critical parameter to compress the source videos. This parameter can be adjusted by the user. In our design, we set n to 120 as the default value. In order for the user to play the compact video more smoothly, two intervals of frames are overlapped in half. That is to say, the compression

ratio of the compact video is 1/60, or a surveillance video with one hour is reduced to a compact video with one minute when n=120 illustrated as Figure 3.

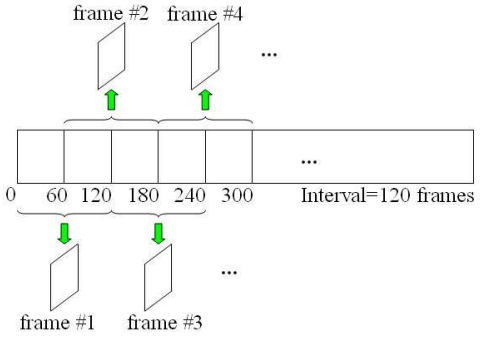


Figure 3. Assume n=120 as the number of video frames in the time interval. Two intervals associated with two consecutive compact frames are overlapped in half.

Occlusion is another possible problem when generating compact frames. In some cases, two representative snapshots, i.e., the medians of two appearing moving objects, may be partially occluded. In practice, the occlusion is not very serious because two representative snapshots of moving objects are not often at the same position in four seconds (suppose 30 frames per second for the source video and n=120=30x4). Most of representative snapshots at least can be partially appeared, and the user can play the compact video slowly and click the correct parts of moving objects to directly link to the source video.

5. Implementation and Experiments

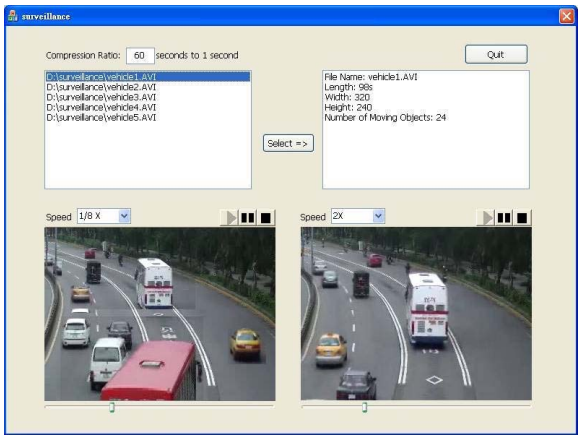


Figure 4. The interface of our quick browsing system for surveillance videos.

Figure 4 presents the interface of our proposed system for browsing surveillance videos. Four main parts are included in the interface: the upper-left part lists all video files that have been indexed, the upper-right shows the information of the video selected by the user, and the lower-left and the lower-right parts display the compact video and the source of the surveillance video respectively. The user needs to first select a surveillance video in the list part. Then, the system can automatically generate the corresponding compact video. When the compact video is playing, the user can select and click a moving object of interest and the system directly links to the source clip of the surveillance video that contains those targets of interest. The user can also change the speed of the video playing for conveniently watching the videos. Since the objects often move very fast in the compact video, it is important to slow down the playing to help the user easily select the subjects of interest. Similarly, the user can also speed up the playing of the source video to assist her/him in the video browsing.

The key parameter for the efficiency of our proposed system is the length n of the time interval, which is presented by the compression ratio of the compact video in the top area of Figure 4. The default value of the field is set to 60, i.e., 60 seconds of the source video is compressed to one second of the compact video. The value 60 for the field is equals to n=120 frames in each time interval. The user can set this parameter to determine the length of the compact video.

An example of compact frames is illustrated in Figure 5. Figure 5(a) shows two consecutive compact frames with the parameter n=60, and Figure 5(b) presents these with n=120. The corresponding source frames are listed in Figure 5(c) by showing one per 10 frames to save the space. The left and the right compact frames in Figure 5(a) are generated according to the 60 source frames marked by a blue rectangle and a green star, respectively in Figure 5(c); the left and right ones in Figure 5(b) are based on the 120 frames in the red and the purple areas, respectively, in Figure 5(c). The influence of parameter n to the compact frames is clearly presented. Moreover, since any two consecutive compact frames cover n/2 overlapped source frames, the compact video can play more smoothly.

We designed a simple but practical evaluation for the utility of the compact video. We set three specific targets in a one-hour video. The positions of the three targets in the source video are: A-21'37", B-29'08", and C-43'18". Five people sequential looked for the three targets by use of the compact video (n=120, i.e., 1-minute compact video), and the results are listed in Table 1. For example, people 1 spent 20s from the end of A to getting the clip (in the source video) of B. In average, people can spend about 70s to find the three targets in a one-hour video.

Table 1. The search time for three targets using the compact video in a one-hour video.

	A	**B**	**C**	**Total**
People 1	36	20	21	77
People 2	41	25	23	89
People 3	30	17	19	66
People 4	32	15	18	65
People 5	28	10	15	53
Average	**33.4**	**17.4**	**19.2**	**70**

6. Conclusion and Future Work

This paper presents a system of quick browsing, instead of retrieving relevant video segments, to help the user easily look for the subjects of interest in surveillance videos. We collected all of moving objects which carry the most significant information in a surveillance video to construct a corresponding compact video. The compact video cannot only preserve the essential activities but also keep temporal relationships among objects in the original surveillance video. Our system links up each moving object in the compact video with its source clip of the

surveillance video. The user can browse the compact video to fast locate targets of interest and then the system can directly display the corresponding video clip selected by the user. Future work includes several directions to extend this work. The generation of the compact video may be improved to achieve more compression and to work on a camera network. Another direction includes attaching semantic-level attributes to the indexing to provide partial retrieval functions.

Acknowledgement

This work was in part supported by National Science Council, Taiwan, under Grant No. NSC 99-2221-E-147 -005 and by Ministry of Economic Affairs, Taiwan, under Grant No. 99-EC-17-A-02-S1-032.

References

[1] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo. "Semantic Annotation of Sports Videos," IEEE Multimedia, vol. 9, no. 2, pp. 52-60, 2002.

[2] S.-F. Chang, L. Kennedy, and E. Zavesky, "Columbia University's Semantic Video Search Engine," Proceedings of ACM International Conference on Image and Video Retrieval, 2007.

[3] B.-W. Chen, J.-C. Wang, and J.-F. Wang, "A Novel Video Summarization Based on Mining the Story-Structure and Semantic Relations among Concept Entities," IEEE Transactions on Multimedia, vol. 11, no. 2, pp. 295-312, 2009.

[4] Y.-T. Chen, C.-S. Chen, C.-R. Huang, and Y.-P. Hung, "Efficient Hierarchical Method for Background Subtraction," Pattern Recognition, vol. 40, no. 10, pp. 2706-2715, 2007.

[5] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603-619, 2002.

[6] P. DeCamp, G. Shaw, R. Kubat, and D. Roy, An Immersive System for Browsing and Visualizing Surveillance Video, in Proceedings of ACM International Conference on Multimedia, 2010.

[7] W.-M. Hu, D. Xie, Z.-Y. Fu, and W.-R. Zeng, "Maybank S.: Semantic-Based Surveillance Video Retrieval," IEEE Transactions on Image Processing, vol. 16, no. 4, pp. 1168-1181, 2007.

[8] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-Based Multimedia Information Retrieval: State of the Art and Challenges," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 2, no. 1, pp. 1-19, 2006.

[9] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological Video Synopsis and Indexing," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1971-1984, 2008.

[10] Y. Takahashi, N. Nitta, N. Babaguchi, "Video Summarization for Large Sports Video Archives," Proceedings of IEEE International Conference on Multimedia and Expo, 2005.

[11] B. T. Truong and S. Venkatesh, "Video Abstraction: A Systematic Review and Classification," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 3, no. 1, 2007.

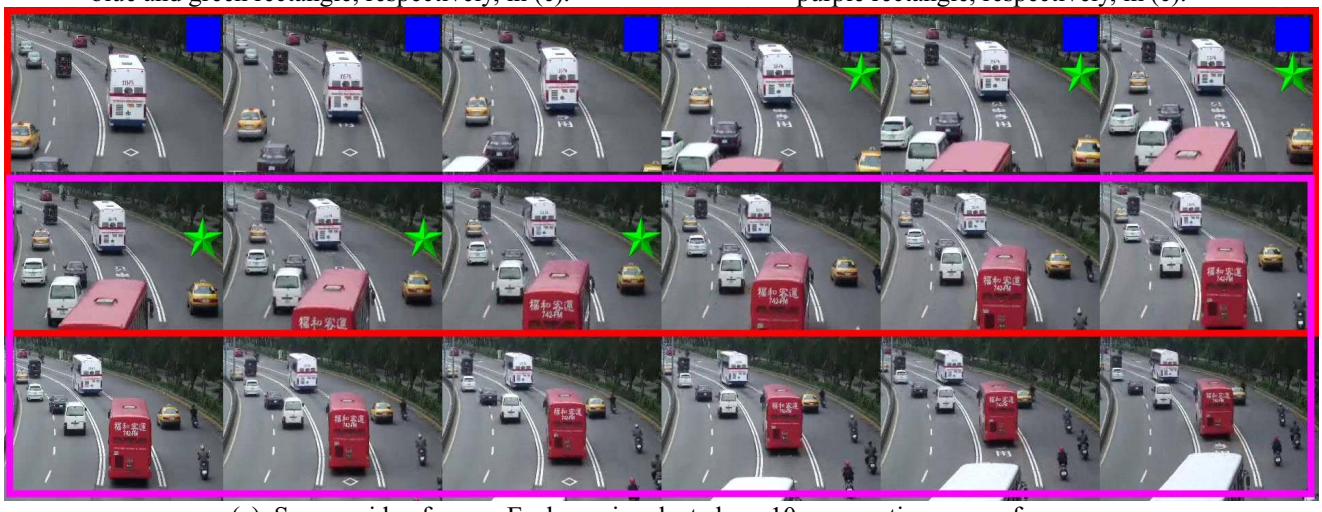
[12] Z. Xiong, X. Zhou, Q. Tian, R. Yong, and T. S. Huang, "Semantic Retrieval of Video - Review of Research on Video Retrieval in Meetings, Movies and Broadcast News, and Sports," IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 18-27, 2006.

[13] X.-D. Yu, L. Wang, Q. Tian, and P. Xue, "Multi-Level Video Representation with Application to Keyframe Extraction," Proceedings of 10th International Multimedia Modeling Conference, 2004.

[14] Open Computer Vision Library, http://sourceforge.net/projects/opencvlibrary/.

(a). Two consecutive compact frames with n=60. The left and the right are generated by the source frames marked by a blue and green rectangle, respectively, in (c).

(b). Two consecutive compact frames with n=120. The left and the right are generated by the source frames in a red and purple rectangle, respectively, in (c).



(c). Source video frames. Each one is selected per 10 consecutive source frames.

Figure 5. An illustration of compact frames. (a) and (b): two compact frames with parameter n=60 and n=120, respectively, (c): the corresponding source video frames.