LIVE VIDEO SYNOPSIS FOR MULTIPLE CAMERAS

Yedid Hoshen *Shmuel Peleg*

School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

ABSTRACT

Video surveillance cameras generate most of recorded video, and there is far more recorded video than operators can watch. Much progress has recently been made using summarization of recorded video, but such techniques do not have much impact on live video surveillance.

We assume a camera hierarchy where a Master camera observes the decision-critical region, and one or more Slave cameras observe regions where past activity is important for making the current decision. We propose that when people appear in the live Master camera, the Slave cameras will display their past activities, and the operator could use past information for real-time decision making.

The basic units of our method are action tubes, representing objects and their trajectories over time. Our object-based method has advantages over frame based methods, as it can handle multiple people, multiple activities for each person, and can address re-identification uncertainty.

Index Terms— Video Surveillance, Video Synopsis, Multi Camera Synopsis

1. INTRODUCTION

Surveillance cameras are installed everywhere, and are becoming even more popular due to lower costs of cameras, networking, and storage. The increase in the number of cameras is not being offset by a proportional increase in the number of operators available to monitor the video, and in practice most surveillance video is not being viewed. The large gap between the availability of human operators and the need to extract the information in the recorded video has attracted much interest from the computer vision community.

Surveillance video has two main purposes: real-time remote sensing and forensic historical analysis, where historical video is rarely used for real-time decision making. In this paper we suggest a novel object-based method for using past surveillance video for live decision making.

One of the great challenges for human operators is being able to exploit relations between the video streams across time and across cameras. Let us consider a library with several cameras. Some cameras view the bookshelves while one camera views the lending desk. Viewing each stream independently may not reveal suspicious behavior. The librarian can not remember all activity of all library visitors, which occur at different times in different cameras. However, if all the bookshelf cameras delayed showing the activities of each reader until he reaches the lending desk, the librarian can easily grasp all the reader's activity before he leaves the library.

The camera synchronization paradigm is quite general. Other cases where cross camera relationship is important are:

- Effective Business intelligence: What items did customers look at before purchasing?
- Anomalous behavior detection: Does a traveler change his pace before going through customs?
- Checkpoint: A guard at the exit from a secure facility can observe if visitors behaved suspiciously during their visit before being allowed to leave.

The relations in these cases are all object-based, which benefit from comparing the behavior of people across different locations and times. All such systems are hierarchical, where one camera is viewed in real-time (Master) while other cameras are of forensic significance (Slaves). The Master camera need not be static, and could even be a body mounted camera worn by the operator. In this case the Master video need not be viewed, as its view is the same as the operator's.

In standard camera networks the analyst needs to remember all objects during a few hours of video, which is unreasonable. However, in a hierarchical camera system, if we display in the Slave videos the previous actions of all persons currently observed in the Master camera, the operator will need to remember only a few seconds of video from the Slave cameras for understanding the activity in the Master camera. This motivates Live Video Synopsis (LVS). In LVS activity tubes are initially extracted from all Slave video streams, and persons are identified and labeled. Tubes are then shifted in time in the Slave videos to be displayed only when the person is observed in the Master view.

Notably, we shift Slave tubes in time but do not attempt to bring tubes from different cameras onto the same screen or shift tubes spatially as object tubes might be placed on semantically unrelated backgrounds sometimes with absurd results (e.g. people floating in mid-air). Also changes in geometry between cameras can cause some tubes to look unnatural and out of place (e.g. front and side views).

Live Video Synopsis has the following benefits: 1) The

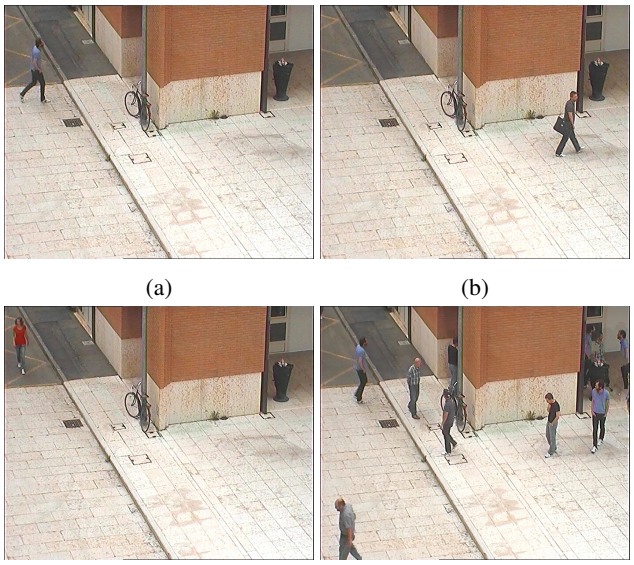

Fig. 1: Video Synopsis: (a-c) Original frames. (d) Synopsis frame. Objects from different times appear simultaneously.

relations between persons observed at multiple cameras are clearly visible to the operator. 2) Multiple persons and histories can be observed on the same screen. 3) In cases of re-identification uncertainty, multiple possibilities can be displayed. 4) The information can aid live decision making.

2. RELATED WORK

Much work has been done on understanding surveillance video. Popular approaches include the classification of activity as normal/anomalous [1, 2], or using activity recognition to transcribe surveillance video into words [3, 4]. High-level activity understanding is a very promising research direction, but current performance has room for improvements. Realizing that the need for human inspection of video will remain for some time, many methods create visual summaries for faster viewing.

One approach for visual summarization is the generation of a storyboard by selecting some key frames [5, 6]. Another approach is adaptive fast forward [7], dropping frames at different rates depending on how interesting the video is. Video synopsis [8, 9, 10, 11], shifting activities in time so that as many activities can be presented simultaneously, presents all activities of a video in a much shorter video. See Fig 1.

Single camera approaches for summarization do not generalize well to multiple cameras, as they do not take into account the relationship between the different cameras. Some work addressed video captured by several overlapping cameras [12]. But this work can not be used with most cameras which are mostly non-overlapping.

Representation of the video from non-overlapping cameras has received little attention, a notable exception is [13], which projects multiple video cameras on a 3D model of the environment. But such a 3D model is not generally available. Another interesting work has been done by [14], who have recognized the importance of using objects for highlighting relationships between video streams from multiple cameras. Their work however has concentrated on the extraction and indexing of objects rather than on visual representation.

A somehow related approach is Multi-Video Browsing and Summarization [15], which attempts to synchronize video streams by shifting frames in time, so that visually similar frames are observed in all videos at the same time. This scheme measures similarity by a set of trained visual similarity descriptors among frames, in contrast to our work which is object based.

3. LIVE VIDEO SYNOPSIS (LVS)

The generation of LVS consists of three stages: Preprocessing (Sec. 3.1), Optimization (Sec. 3.2), and Display (Sec. 3.3).

3.1. Video Preprocessing

Before selecting the Slave action tubes corresponding to the persons observed by the Master camera, several preprocessing steps are required:

1. People are detected and tracked in all slave video streams. Each person is represented as a space-time "tube", which is the union of all pixels of this object in each frame. Relevant literature on object detection using background subtraction appears in [16, 17], and tracking objects across frames appears in [10, 14]. The extraction of video tubes is depicted in Fig. 2
2. People are detected in the current frame of the Master stream. There has been much work on human detection [18] and in particular on Pedestrian detection [19].
3. Re-identification of people between the Master stream detections and the tubes extracted from the Slave streams is performed [20, 21, 22]. Re-identification scores between two objects are often given probabilistically e.g. [22].

3.2. Slave Action Tube Selection

In this section we assume a camera system consisting of one Master camera of real-time importance and one or more Slave cameras of forensic importance. We propose to detect people in the Master camera stream at fixed time intervals, and play for each Slave camera, the activity tubes from the past that contain the observed people.

Pre-processing is done as described in Sec. 3.1. At fixed intervals of length δT the Slave action tubes to be displayed in each Slave video v are selected. The task is to select a set of tubes S_v to display in Slave video v out of the total set of tubes in the Slave view B_v ($S_v \subseteq B_v$). There are three factors that are taken into account: i) displaying the maximal number

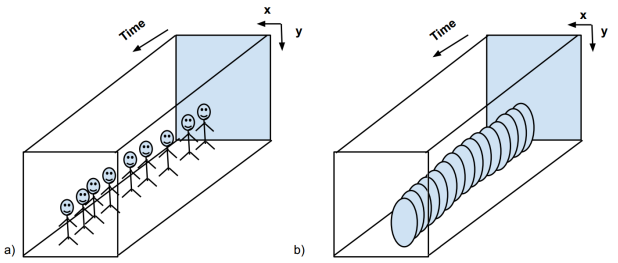

Fig. 2: a) A video showing a single object. b) Tubes are binary masks representing an object, containing all pixels in all frames belonging to the object.

of Slave tubes containing the people observed in the Master camera; ii) minimizing tube collisions; iii) a stable viewing experience: minimizing the number of tube switches at each interval.

This can be formulated using two energy terms, a collision term E_v^C and an identical object overlap term E_v^O:

$$E_v^T(S_v) = \alpha \cdot E_v^C(S_v) - E_v^O(S_v) \mid S_v \subseteq B_v \qquad (1)$$

We do not explicitly take into account the relations between different slave videos. The energy terms E_v^O for Slave videos v are optimized independently of other Slave videos.

3.2.1. Collision Cost

The objective of E^C is to minimize collisions between action tubes placed in the generated videos. A small number of collisions can be tolerated, and it can greatly increase the number of Slave activities displayed simultaneously. The number of collisions that can be tolerated can be modified by adjusting α in Eq. 1.

Let tube b be defined by binary function $\chi_b(x, y, t)$ indicating if the pixel (x, y) in frame t is active for tube b.

Given a slave camera, the collision cost for its generated slave video v is defined in Eq. 2 (similar to [10]): the number of colliding pixels among all pairs of different tubes in the video. We add a discount factor for collisions that are forecast further away in the future as we become increasingly uncertain that the tubes will not be terminated before the forecast collision (due to new persons appearing and old persons disappearing in the Master view). The amount of discounting is determined by factor d.

$$E_v^C(S_v) = \sum_{b,\tilde{b} \in S^v} \sum_{x,y,t} \chi_b(x, y, t) \cdot \chi_{\tilde{b}}(x, y, t) \cdot d^t \qquad (2)$$

where S_v is the set of tubes chosen for display in the output Slave video v at the current time interval from the total set B_v.

3.2.2. Identity Cost

The person identity cost in Eq. 3 encapsulates several requirements: i) displaying the Slave tubes having the highest probability of correspondence to the people detected in the Master stream (this set is labeled O). ii) making the number of tubes corresponding to each object in the Master frame roughly equal. iii) encouraging retention of already playing tubes for smoother viewing. This can be formulated as:

$$E_v^O(S_v) = \sum_{o \in O} \sqrt{\sum_{b \in S_v} (1 + \beta \cdot 1_{b \in S_v^{t-1}}) \cdot P_{b,o}} \qquad (3)$$

Where S_v^{t-1} is the set of Slave tubes selected in the last interval, and β is a constant determining the strength of the preference to retain old tubes. The square root encourages the display of all objects in roughly equal numbers, otherwise most tubes may come from the same most likely object. When the Slave action tube and the person appearing in the Master camera are different persons this term has little effect, as the probability $P_{b,o}$ will be low.

3.2.3. Cost Minimization

The energy for each slave camera as expressed in Eq. 1 can be minimized using standard discrete optimization methods. However the fast greedy approach described below generated good results as well.

- For all Slave videos v
- Set $S_v = \phi$
- Set list $L = B_v$ (all tubes for video v)
- Until no tubes left in L:
  1. For each tube $b \in L$ calculate the approximate decrease in overlap energy $\sum_{o \in O} \frac{(1+\beta \cdot 1_{b \in S^{t-1}}) \cdot p_b^o}{\sqrt{\sum_{\tilde{b} \in S_v} (1+\beta \cdot 1_{\tilde{b} \in S_v^{t-1}}) \cdot p_{\tilde{b}}^o}}$
  2. Select tube b with the largest decrease.
  3. If the sum of collisions between b and the tubes in S_v is smaller than threshold r: if $\sum_{\tilde{b} \in S_v} \sum_{x,y,t} \chi_b(x, y, t) \cdot \chi_{\tilde{b}}(x, y, t) \cdot d^t < r$, add b to S_v
  4. Remove b from L.

We use the binary update rule every δT seconds and display the tubes $\{b | b \in S_v\}$ in slave view v.

3.3. Synopsis Display of Slave Cameras

LVS is now generated for each slave camera v by placing every tube from S_v with the correct temporal offset (in case it has already been playing) over the stationary background of the corresponding Slave video. We emphasize that a synopsis video is created for each Slave camera. Tubes are not transferred between cameras, nor shifted in space. This ensure that all objects remain on their original background and geometries, creating videos that are easy to understand.

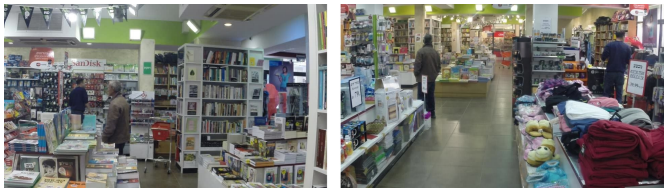

Fig. 3: A sample frame from the output of the MS algorithm on a Store scene. The left image is the Master camera and the right image is the Slave. Tubes corresponding to the two persons in the Master view were rendered simultaneously in the Slave videos. The clip can be seen at: http://www.vision.huji.ac.il/syncvid/

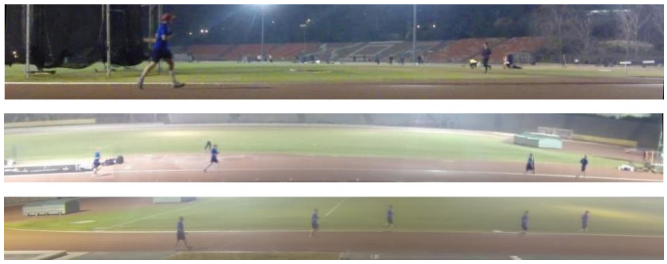

Fig. 4: A sample frame from the output of the Master-Slave algorithm run on a Stadium scene. The top video is the Master camera and the bottom two are Slaves. Multiple tubes corresponding to the person in the Master view are displayed in the slave videos. Many matching tubes were found and are displayed simultaneously. This cannot be achieved by frame-based methods. The clip can be seen at: http://www.vision.huji.ac.il/syncvid/

4. EXPERIMENTS

We present frames from two scenes, demonstrating the output of LVS. The Store Scene was recorded by two non-overlapping cameras in a store (Fig. 3), the Stadium was recorded by three non-overlapping cameras around a stadium (Fig. 4). Tubes were extracted by state of the art background subtraction method such as [23]. Tubes were manually re-identified between Master and Slave tubes. Our method was then run using the following parameters: $\beta = 0.5$, $\delta T = 1\ second$, $r = 15$, $d = 0.978$. The output clips can be seen at http://www.vision.huji.ac.il/syncvid/.

Fig. 5.a) shows a comparison between a frame-based method (showing the whole frames of the highest ranking Slave action tube) and our object-based method - LVS. The frame-based method was able to display only 15-30% of the relevant Slave tubes, whereas our method was able to display 65-85% of Slave tubes with minimal collisions. This performance-gap is expected to increase further when re-identification uncertainty is significant.

The trade-off between collisions and number of relevant Slave tubes can be seen in Fig. 5.b) for the three Slave videos.

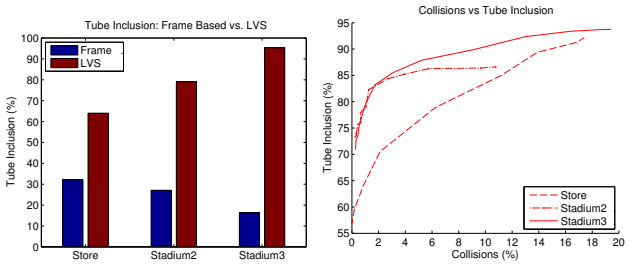

Fig. 5: a) Comparison of tube inclusion rates of LVS vs the frame based method for the three Slave videos. Significant improvements have been obtained. b) The Collision rate vs. the Tube Inclusion rate for the three Slave videos. 65-85% of tubes can be included for a modest collision cost.

A very modest collision rate (2%) is required for displaying 65-85% of relevant tubes.

Several benefits of LVS are apparent:

1) While concentrating on the Master camera, we are able to see much history of the objects in the Slave cameras. This can be of great utility for letting operators make decisions in real-time.

2) In many cases the Master stream is sparse and contains only a small number of objects, it is possible to display several candidate tubes for each object in the Slave nodes. This is helpful as in many scenarios, the top re-identification result has about 30% recall probability, but the top 5 candidates have an accumulated recall probability of above 65% [22]. Showing as many candidates as possible therefore increases the likelihood of seeing the whole history of the object across the scene.

5. CONCLUDING REMARKS

Live video synopsis is a novel object-based method for using summarization of previously recorded video for aiding live decision making. It was shown that our method has many advantages over frame based methods. Although in this paper we have concentrated on people, this method is general and can be used for any type of object that can be detected and re-identified across cameras (animals, cars, etc.). As our method relies on having a reliable object re-identification algorithm, improvements in person re-identification from video will increase the reliability of our method. More interestingly, our method can be used to display object re-identification examples for active learning algorithms. This can be used for obtaining interactive feedback from the operator for refining video re-identification performance.

Acknowledgment: This research was supported by Intel ICRI-CI, by the Israeli Ministry of Science, and by Israel Science Foundation.

6. REFERENCES

[1] Fan Jiang, Junsong Yuan, Sotirios A Tsaftaris, and Aggelos K Katsaggelos, "Anomalous video event detection using spatiotemporal context," *CVIU*, pp. 323–333, 2011.

[2] Bin Zhao, Li Fei-Fei, and Eric P Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *CVPR*, 2011.

[3] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *IJCV*, pp. 171–184, 2002.

[4] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele, "Translating video content to natural language descriptions," in *ICCV*, 2013.

[5] Yihong Gong and Xin Liu, "Video summarization using singular value decomposition," in *CVPR*, 2000.

[6] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan, "Large-scale video summarization using web-image priors," in *CVPR*, 2013.

[7] Nemanja Petrovic, Nebojsa Jojic, and Thomas S Huang, "Adaptive video fast forward," *Multimedia Tools and Applications*, vol. 26, no. 3, pp. 327–344, 2005.

[8] Alex Rav-Acha, Yael Pritch, and Shmuel Peleg, "Making a long video short: Dynamic video synopsis," in *CVPR*, 2006.

[9] Yael Pritch, Alex Rav-Acha, Avital Gutman, and Shmuel Peleg, "Webcam synopsis: Peeking around the world," in *ICCV*, 2007.

[10] Yael Pritch, Alex Rav-Acha, and Shmuel Peleg, "Nonchronological video synopsis and indexing," *IEEE-PAMI*, pp. 1971–1984, 2008.

[11] Shikun Feng, Zhen Lei, Dong Yi, and Stan Z Li, "Online content-aware video condensation," in *CVPR*, 2012.

[12] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou, "Multi-view video summarization," *IEEE Trans. Multimedia*, pp. 717–729, 2010.

[13] Harpreet S Sawhney, Aydin Arpa, Rakesh Kumar, Supun Samarasekera, Manoj Aggarwal, Steve Hsu, David Nister, and K Hanna, "Video flashlights: real time rendering of multiple videos for immersive model visualization," in *Proceedings of the 13th Eurographics workshop on Rendering*, 2002.

[14] Fatih Porikli, "Multi-camera surveillance: object-based summarization approach," *Mitsubishi Research TR-2003-145*, 2004.

[15] Kevin Dale, Eli Shechtman, Shai Avidan, and Hanspeter Pfister, "Multi-video browsing and summarization," in *CVPRW*, 2012.

[16] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999.

[17] Teresa Ko, Stefano Soatto, and Deborah Estrin, "Background subtraction on distributions," in *ECCV'08*, 2008, pp. 276–289.

[18] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[19] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, 2012.

[20] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Transfer re-identification: From person to set-based verification," in *CVPR*, 2012.

[21] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, June 2013.

[22] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, June 2013.

[23] M. Van Droogenbroeck and O. Paquot, "Background subtraction: Experiments and improvements for vibe," in *Change Detection Workshop at CVPR*, June 2012.