A Multi-view Video Synopsis Framework *

Ansuman Mahapatra, Pankaj Kumar Sa, and Banshidhar Majhi †

Abstract

In this paper, a simple framework for multi-view video synopsis is introduced by combining both the benefits of video summarization and video synopsis. While video summarization includes the important actions of an original video, video synopsis, on the other hand, simultaneously displays multiple objects from different intervals of time. To create a multi-view video synopsis, the objects in the field-of-views of all the cameras are shown in a common background plane. All the moving objects are detected, prioritized based on their actions, and only important objects are included in the final synopsis video. The proposed framework is evaluated using two datasets and is compared with existing techniques, which shows a significant reduction in synopsis length with the proper inclusion of important objects.

Key words: Video Synopsis, Multi-view video, Multi-camera Network, Video Summarization

1 Introduction

Creating a synopsis of a long video without missing any important information and keeping the video length shorter is a challenging task. With multiple cameras, the video summarization becomes more challenging, and the existing single video summarization or video synopsis techniques can not directly be applied to multi-view videos. Ou *et al.*, in their work have proposed a multi-view video summarization for wireless video sensor network [1]. Each sensor in the network prepares a summary of its view. View-selection algorithm takes selective frames from each summary to create a multi-view video synopsis. In another summarization scheme proposed by Park *et al.* [2], a user has to input the degree of interest for each event, person, and object, which are used for decision making and ranking them using the fuzzy inference engine. Leo and Manjunath have proposed two methods for summarizing videos by giving importance to actions and presenting the summary stroboscopically [3, 4]. Fu *et al.* have made summarization as a graph labeling task by constructing a

*This paper was presented at *IEEE International Conference on Image Processing*, Quebec City, Canada, 27–30 September 2015.

†The authors are with the Computer Vision Research Laboratory, Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela – 769 008, Odisha, India.

spatiotemporal graph from the input video [5]. An object-based summarization method is proposed by Silva *et al.* [6], which keeps track of the object's entry time, exit time, and important actions performed in the room. A video synopsis technique for single camera is proposed by Rav-Acha *et al.*, in which the objects are shifted along the temporal axis and represented simultaneously, keeping spatial locations unchanged [7].

In the existing video summarization methods, the complexity of object representation increases with the increase in the number of moving objects as well as with the addition of new cameras to the existing multi-camera network infrastructure. The proposed multi-view video synopsis framework is built upon the existing single camera based video synopsis framework, proposed by Rav-Acha [7]. The proposed framework not only creates a synopsis for multi-view videos but also accommodates important actions with optimized video length.

Rest of the paper is organized as follows. The proposed framework is introduced in Section 2. Simulation results are discussed in Section 3. Finally, Section 4 presents the concluding remarks.

2 Proposed Framework

The proposed framework creates a video synopsis by taking multi-view videos as input. The framework incorporates the advantages of both video synopsis and video summarization. It represents objects in the synopsis by shifting them on the temporal axis without changing their spatial location. It includes only those objects that are performing important actions. This leads to a significant reduction in video length. The block diagram of the proposed framework, as shown in Fig. 1, has five stages; common background creation, common plane correspondence, object detection, action recognition, and dynamic video synopsis. The first three stages utilize existing schemes and our proposition thrusts on action recognition and dynamic multi-view video synopsis. The following subsections explains each stage in sequel.

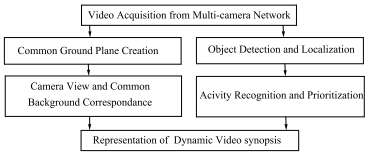

Figure 1: Framework for Multi-view Video Synopsis

2.1 Common Background Creation

Multi-view videos are acquired through multiple cameras. Objects in such videos may appear simultaneously in more than one views. Objects may also exit the field of view (FoV) of one camera and enter the FoV of another camera.

(a) (b)

Figure 2: (a) The top view representation of the surveillance area for PETS 2009 dataset. (b) The common background plane of the indoor surveillance site where the shaded region represents non-walkable area and white represents walkable floor.

In our simulation, static cameras are used and, therefore, each camera has its own FoV with a static background. Hence, a common background is necessary to represent the objects in the synopsis. One of the intuitive ideas is to create a top view of the site from where multi-view videos are acquired. Google Map has been used in our simulation to create a top view of the site for outdoor sequences, and a floor plan has been drawn for indoor videos. Fig. 2(a) shows an example of common background obtained from Google Map for PETS 2009 dataset [8]. Fig. 2(b) depicts the floor plan of our laboratory as common background plane of an indoor surveillance. This creation of the common background plane is one-time process and can easily be modified with the scaling of the multi-camera network.

2.2 View and Common Ground Plane Correspondence

Each camera in a multi-camera network has its coordinate system. These coordinate systems are needed to be mapped to a common coordinate. In other words, they are needed to be mapped to the common background. In this simulation, the homographic technique is used to achieve this mapping [9].

In homographic technique two images of the same planar surface can be related by a homography matrix. Homographic relation between two points X and X' on two images can be written as,

$$(u\ v\ 1)^T = H\,(x\ y\ 1)^T \tag{1}$$

where $(u\ v\ 1)^T$ represents X', $(x\ y\ 1)^T$ represents X and H is the homography matrix. The transformed coordinates are given by,

$$\left.\begin{aligned} u &= \tfrac{h_1x+h_2y+h_3}{h_7x+h_8y+1} \\ v &= \tfrac{h_4x+h_5y+h_6}{h_7x+h_8y+1} \end{aligned}\right\} \tag{2}$$

The above equation set (2) are solved for eight unknown $(h_1, h_2, \cdots, h_8)$ by incorporating four-point correspondences between common background plane and each camera. This process of correspondence matching is done only once

in the setup phase. Unless there is any major change in FoV or position of the camera, the correspondence matching does not require any change.

2.3 Moving Object Detection and Localization

Each object in the video needs a representation in the synopsis. Therefore, objects in the video require detection before they can be processed further. The proposed multi-view video synopsis is restricted to only humans. Hence, the detected objects need further processing to categorize them as human. One of our earlier work has been used for human detection in video [10]. It uses fuzzy inference system to model the background, which is then subtracted from each frame to detect objects. Each detected objects is tested with its contour to verify whether it is human.

2.4 Action Recognition and Prioritization

In video synopsis, action recognition and prioritization give the flexibility of including important actions and thereby making the synopsis short. Six different shape features are extracted from each human silhouette detected from the previous step, height of bounding box, width of bounding box, center of gravity, width of upper half of the body, width of middle part of the body, and width of lower part of the body. Fig. 3 depicts the different shape based features extracted from silhouette.

After the feature extraction phase, the actions performed by all objects are classified using multiple kernel learning (SimpleMKL [11]). Let $L = \{x_i, y_i\}_{i=1}^{l}$ be the learning set where $x_i \in X$(input space) and $y_i \in T$ (target set). The solution of kernel learning takes the form,

$$f(x) = \sum_{i=1}^{l} \alpha_i^* K(x, x_i) + b^* \tag{3}$$

where α_i^*, b^* are some coefficient to be learned from learning set L and K$(\cdot)$ is a positive definite kernel associated with a reproducing kernel Hilbert space (RKHS) H. In MKL, the kernel $K(x, x')$ is a convex combination of basis

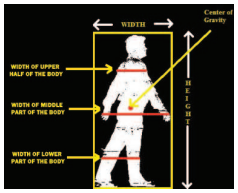

Figure 3: Six different features extracted from a silhouette

kernels.

$$K(x, x') = \sum_{m=1}^{M} d_m K_m(x, x') \tag{4}$$

with $d_m \geq 0, \sum_{m=1}^{M} d_m = 1$,where M is the total number of kernels, K_m is any classical kernels, and d_m is the weight. Learning both coefficients α_i and d_m in a single optimization problem is called MKL. The optimization problem for SimpleMKL is given by,

$$\min_d J(d) \quad \text{such that} \quad \sum_{m=1}^{M} d_m = 1, d_m \geq 0 \tag{5}$$

$$\begin{aligned} \text{where} \quad & J(d) = \min \ \tfrac{1}{2} \textstyle\sum \frac{1}{d_m} ||f_m||^2_{H_m} + C \sum_i \xi_i, \ \forall i \\ \text{such that} \quad & y_i \textstyle\sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \ \forall i \end{aligned}$$

The optimization problem can be solved using general SVM approach. Gaussian and Polynomial kernels are used in multiple kernel learning for classifying seven different actions such as walking, running, bending, jumping, handshake, one hand wave, and both hands wave. The prioritization of actions is purely based on the type of surveillance. The actions that are abnormal are included under the high-priority group.

2.5 Dynamic Video Synopsis Generation

Multi-view video synopsis is a synthesized video obtained by superimposing circles as object identifiers upon a common background plane. Each circle is timestamped to know the appearance of the corresponding object in the original video. Like single-view video synopsis, the proposed multi-view video synopsis shifts all the objects performing important actions along the temporal axis. The number of objects to be shown simultaneously is decided by the user. A sample multi-view video synopsis shot is illustrated in Fig. 4. The action priorities are color coded, for example, red color represents unusual action, and green color represents normal action. The bottom panel shows the preview of the lateral views of each object from the original frame.

To achieve optimized length of synopsis video sequence including all the important objects, an energy minimization optimization technique is formulated as in (6). Simulated Annealing [12] is used to solve the optimization problem. The energy function $E(A, S)$ measures the cost of the selection of a subset of objects S from total object set O given the action priority queue A. The cost function includes E_o the loss incurred due to non-inclusion of some objects in the final synopsis, E_c the penalty of collision or overlapping object representation, and E_l a penalizing term for long synopsis video.

$$\text{Minimize } E(A, S) = w_0 E_o + w_1 E_c + w_2 E_l \tag{6}$$

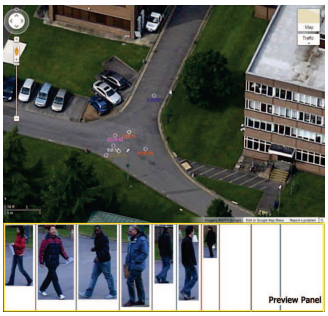

Figure 4: Representation of dynamic video synopsis. The image in the upper part is the top view of the surveillance site. Each object in the upper part can be seen in the preview panel available in the bottom part of the synopsis system. In this illustration, seven objects are identified with unique numbers in the upper panel and their corresponding previews are available in the preview panel.

such that E_l = length of synopsis, $E_c = \sum \#c$, and $E_o = \sum_{o \in O} \#o - \sum_{o \in S} \#o$ where c is the number of collision count obtained from CollisionDetection(P_i, S_{temp}) described in Algorithm 1, P_i denotes the path of object i and S_{temp} denotes temporary solution containing some previously selected paths. The notation $\#o$ denotes the number of objects. If collisions are less, then there are more probability of path i to be included in the final synopsis video.

3 Simulation and Results

To validate our proposed framework for action recognition and synopsis generation, simulations have been carried out on four different datasets namely; KTH dataset [13], WEIZMANN [16], PETS 2009 [8], and Lab video. The KTH dataset comprises 85 videos of each action, WEIZMANN dataset has nine videos of each action, PETS 2009 dataset has four multi-view videos captured with four different cameras each having 794 frames. The Lab video is an indoor sequence captured in the authors' Laboratory. It consists of four cameras, each having 34438 number of frames. Four sample frames from each camera view are shown in Fig. 5.

The proposed action recognition method along with existing schemes like Local SVM approach [13], multi-feature and multi-kernel learning (MF-MKL)[14], and Smart homes [15] are tested on above datasets and the accuracy performance comparison is listed in the Table 1.

The proposed video synopsis framework is evaluated on PETS 2009 dataset, which is an outdoor sequence and Lab video dataset, which is an indoor sequence.

Metrics used for evaluation of the effectiveness of the proposed synopsis

Algorithm 1: Proposed Collision Detection Algorithm

Input : A, P
Output: Number of Collision c

$P = \{P_1, P_2, \cdots, P_N\}$
$P_k = \{(x_k(1), y_k(1)), \cdots, (x_k(m), y_k(m))\}$;
$S_{temp} = P_1$;
$t = T$;
for $k \leftarrow 1$ **to** $length(S_{temp})$ **do**
    for $t \leftarrow 1$ **to** $length(\min(P_{new}, P_k))$ **do**
        $d = \sqrt{(x_{new}(t) - x_k(t))^2 + (y_{new}(t) - y_k(t))^2}$;
        if $d < Th$ **then**
            $c = c + 1$;
        else
            continue;
$T = \min length(P(S_{temp}))$;
$CollisionDetection(P_{new}, S_{temp})$;

Table 1: Accuracy comparison of action recognition methods (in %)

Dataset/Method	Local SVM[13]	MF-MKL [14]	Smart Homes[15]	Proposed
KTH	94.50	95.25	91.00	95.00
WEIZMANN	92.45	90.70	92.25	94.55
PETS 2009	89.45	91.65	87.50	91.40
Lab video	85.25	87.50	76.40	85.50

Figure 5: Sample frames from four cameras of Lab video.

method and for comparison with other existing methods are percentage of video length reduction (R) and quality assessment ratio (Q) which are given in (7) and (8) respectively.

$$R = (1 - \frac{S_L}{O_L}) \times 100 \tag{7}$$

where O_L and S_L signify original video length and synopsis video length in number of frames respectively.

$$Q = \frac{A_S}{A_O} \tag{8}$$

where A_O denotes number of important actions present in the original video, and A_S represents number of important actions included in synopsis video.

Table 2: Results of proposed video synopsis method

Datasets	O_L	S_L	A_O	A_S	R	Q
PETS 2009	794	47	7	6	94.08%	0.90
Lab video	34438	278	21	18	99.19%	0.95

Table 2 shows the result of proposed framework based on length of original (O_L) and synopsis (S_L) video, number of important actions in original (A_O) and synopsis video (A_S), percentage of reduced length (R), and quality assessment ratio (Q). Two other existing methods, fuzzy based summarization [2] and Multi-view video summarization [5] are also tested with both datasets. Their comparison in terms of percentage of video length reduction (R) and quality assessment ratio (Q) are listed in Table 3.

Table 3: Comparison of percentage of length reduction and quality assessment ratio

Methods	R_{PETS}	R_{LAB}	Q_{PETS}	Q_{LAB}
Fu *et al.* [5]	92.50%	97.25%	0.95	0.91
Park and Cho[2]	87.04%	89.02%	0.71	0.47
Proposed Method	94.08%	99.19%	0.90	0.95

4 Conclusion

A simple framework for multi-view video synopsis is presented in this paper. The framework has five stages. Existing methods are used in the first three stages, and contributions are made in the next two stages. Performance comparison with respect to activity recognition accuracy, percentage of synopsis length reduction, and quality assessment ratio has been made with existing schemes. The comparative analysis reveals the superior performance of the proposed framework.

References

[1] S Ou, C LEE, V Somayazulu, Y Chen, and S Chien, "On-line Multi-view Video Summarization for Wireless Video Sensor Network," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, 2014.

[2] Han-Saem Park and Sung-Bae Cho, "A Fuzzy Rule-based System With Ontology for Summarization of Multi-camera Event Sequences," in *Artificial Intelligence and Soft Computing*, pp. 850–860. Springer, 2008.

[3] Carter de Leo and BS Manjunath, "Multicamera Video Summarization and Anomaly Detection from Activity Motifs," *ACM Transactions on Sensor Networks (TOSN)*, vol. 10, no. 2, pp. 27, 2014.

[4] Carter De Leo and BS Manjunath, "Multicamera Video Summarization from Optimal Reconstruction," in *Computer Vision–ACCV 2010 Workshops*. Springer, 2011, pp. 94–103.

[5] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou, "Multi-view Video Summarization," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 717–729, 2010.

[6] Gamhewage C De Silva, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Evaluation of Video Summarization for a Large Number of Cameras in Ubiquitous Home," in *ACM international conference on Multimedia*, 2005, pp. 820–828.

[7] Alex Rav-Acha, Yael Pritch, and Shmuel Peleg, "Making a Long Video Short: Dynamic Video Synopsis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 435–441.

[8] Bo Yang and Ram Nevatia, "Multi-target Tracking by Online Learning of Non-Linear Motion Patterns and Robust Appearance Models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1918–1925.

[9] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

[10] Ansuman Mahapatra, Tusar Kanti Mishra, Pankaj K. Sa, and Banshidhar Majhi, "Human Recognition System for Outdoor Videos using Hidden Markov Model," *AEU - International Journal of Electronics and Communications*, vol. 68, no. 3, pp. 227–236, 2014.

[11] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[12] SP Brooks and BJT Morgan, "Optimization using Simulated Annealing," *The Statistician*, pp. 241–257, 1995.

[13] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing Human Actions: A Local SVM Approach," in *IEEE International Conference on Pattern Recognition*, 2004, vol. 3, pp. 32–36.

[14] Salah Althloothi, Mohammad H. Mahoor, Xiao Zhang, and Richard M. Voyles, "Human Activity Recognition using Multi-features and Multiple Kernel Learning," *Pattern Recognition*, vol. 47, no. 5, pp. 1800 – 1812, 2014.

[15] Iram Fatima, Muhammad Fahim, Young-Koo Lee, and Sungyoung Lee, "A unified framework for activity recognition-based behavior analysis and action prediction in smart homes," *Sensors*, vol. 13, no. 2, pp. 2682–2699, 2013.

[16] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as Space-time Shapes," in *IEEE International Conference on Computer Vision*, 2005, vol. 2, pp. 1395–1402.