Smart Surveillance Based on Video Summarization

Sinnu Susan Thomas
EE Dept
IIT Kanpur
UP, India

Sumana Gupta
EE Dept
IIT Kanpur
UP, India

Venkatesh K. Subramanian
EE Dept
IIT Kanpur
UP, India

***Abstract*—In recent years, video surveillance technology has become ubiquitous in every sphere of our life. But automated video surveillance generates huge quantities of data, which ultimately does rely upon manual inspection at some stage. The present work aims to address this ever increasing gap between the volumes of actual data generated and the volume that can be reasonably inspected manually. It is laborious and time consuming to scrutinize the salient events from the large video databases. We introduce smart surveillance by using video summarization for various applications. Techniques like video summarization epitomizes the vast content of a video in a succinct manner. In this paper, we give an overview how to use an optimal summarization framework for surveillance videos. In addition to reduce the search time we propose to convert content based video retrieval problem into a content based image retrieval problem. We have performed several experiments on different data sets to validate our proposed approach for smart surveillance.**

***Index Terms*—Smart surveillance, video summarization, optimization, content based video retrieval, visual saliency**

I. Introduction

Cameras have always been the eyes of the security industry. As demand continues to grow and costs decline, thousands of cameras are added to surveillance networks every year. With the intention of reducing crime and increasing public safety, an explosive growth in the surveillance industry is seen, leading to an eruption of videos. Video surveillance [1] is expected to generate nearly 24 billion U.S. dollars in revenue worldwide in 2018. Surveillance cameras can be installed at every nook and cranny without much hassle. These cameras work round the clock and capture stack of videos. These large videos pose a serious threat to careful human monitoring. It is highly impractical to scrutinize each and every moment of the video. Techniques such as video summarization have been developed to solve this problem.

A video summary is either a static summary or a dynamic summary. A static summary is a series of key frames and a dynamic summary is a set of short video clips, joined in a sequence and played as a video.

In this paper, we address the following issues.

- The monitoring of surveillance videos round the clock.
- The incorporation of the properties of the Human Visual System (HVS) into the video summarization to make the summary more ergonomic and effective.
- The usage of the summarized output for video retrieval.

Significant research work has occurred in the area of surveillance for various applications. Bilal et al. [2] detected pedestrians in a surveillance video using the discriminating power of the locally significant gradients in building orientation histograms while computing the Histogram Intersection Kernel SVM classifier. Zhou et al. [3] proposed a system for public bus transit system at bus stops. Birnstill et al. [4] anonymized video data to make privacy-aware smart video surveillance.

Video summarization is a crucial technique to analyze surveillance videos. Evangelio et al. [5] proposed a system for the summarization of safety and security surveillance video using low level features and high level events. Wang and Kato [6] presented a system for summarizing nursery school surveillance video. Salehin and Paul [7] used object motion cues to obtain key frames from the surveillance video.

Research work in the area of visual saliency increased after a mathematical model for finding out the salient areas in the image was proposed by Itti [8] using human attention as the basis. Following this work, the research community started using saliency models for summarizing videos. Ma et al. [9] used human attention model for efficient information prioritizing and filtering for video summarization. This was the first work that shows the application of an attention model in video summarization. Following this, Thomas et al. [10] added some more features of human attention for summarizing a video. Along these lines, we propose to use a human attention model to detect events in the surveillance videos. The HVS based attention model helps to find the salient regions from the image and this in turn leads us to extract the key frames from the video. This type of system can be used for cameras installed at any public place including traffic, shopping malls, etc.

In content based video retrieval, the search space is large and cumbersome due to the large size of the video database. To reduce the search space, we propose to combine the extracted key frames from the video shot into a single frame in the database instead of videos. A wide range of experiments carried out shows the superiority of the approach.

The remainder of this paper is organized as follows. Section 2 describes the proposed methodology of summarization and retrieval for surveillance videos. Simulation results are presented in Section 3. Section 4 presents the conclusion of this paper.

978-1-5090-6255-3/17/$31.00

II. Proposed Methodology

We discuss the proposed approach in detail in this section. Fig. 1 presents the flow chart of our approach that can be mainly divided into two parts. First, an optimized framework is proposed for perceptual video summarization. Second, an optimization framework is proposed for video retrieval.

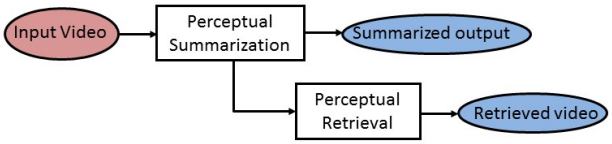

Fig. 1. Flow chart of the proposed approach.

A. Summarization Framework

We propose a new framework —perceptual video summarization where summarization in tune with the properties of human perception play a key role in determining the frames to be selected for the summary. The framework is shown in Fig. 2.

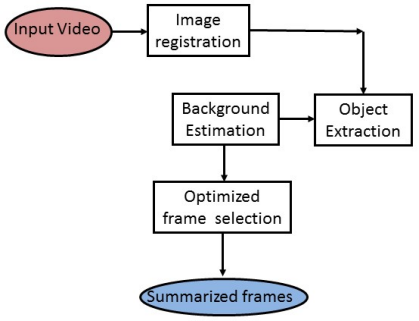

Fig. 2. Summarization Framework.

Let an input video $V = \{I_n | n = 1, 2, ...N\}$. $I_n \in \mathbb{R}^d$ represents the image data in the n^{th} frame with dimension d and (x, y) is the spatial coordinate of a pixel in frame n, satisfying $(1 \leq x \leq W)$, $(1 \leq y \leq H)$. Our primary job is to select the set of frames M to be used for video summarization such that $M \subset N$. $M = \{I_m | m = 1, 2, ...M\}$. For surveillance cameras, we register the input frames with respect to the reference frame [11]. Once all the frames are registered, we need to bucket I_n into relatively static segment and find the median of each bucket and finally, combine them to extract background B. The foreground object is extracted by complementing the background from the registered input image. We use Aggregated Channel Feature (ACF) detection [12] to extract the moving objects and produce a bounding box Bx around the object. An object tube is defined as a tube of object masks.

Let E represent the cost function and λ represent a feasible solution for selecting frames. We set λ_n= 1, if frame 'n' is chosen for the summary and λ_n= 0, if frame 'n' is not chosen for the summary. We define the total cost function $E(\lambda)$ as follows

$$E(\lambda) = E_a(\lambda) + E_c(\lambda) + E_s(\lambda) \quad (1)$$

$E_a(\lambda)$ favors all the significant activities present in the tube to be included in the summarized frame.

We need to control the clutter of object appearance in the summary, and this is done by collision cost $E_c(\lambda)$ which regulates the intra-and-inter-object tube collision. The collision cost is defined as the sum of the overlapping areas between the occurrences of objects and the tubes of distinct objects in the summary. It is given as follows

$$E_c(\lambda_n) = \alpha_1 E_{c_1}(\lambda_n) + \alpha_2 E_{c_2}(\lambda_n) \quad (2)$$

$E_{c_1}(\lambda_n)$ is the collision cost between the objects in a single tube whereas $E_{c_2}(\lambda_n)$ is for different object tubes. The parameters α_1 and α_2 are the weights set by the user in accordance to the intra and inter collisions respectively.

The objects in the tube are selected according to the change or contrast in their various perceptual features. The visual saliency cost $E_s(\lambda)$ of the object tube is measured in terms of change in size, speed, color, acceleration, and direction of an object. The perceptual saliency cost is given as follows

$$E_s(\lambda) = \beta E_{sz}(\lambda) + \chi E_{sp}(\lambda) + \delta E_{cl}(\lambda) + \epsilon E_{acc}(\lambda) + \zeta E_d(\lambda) \quad (3)$$

β, χ, δ, ϵ, and ζ are given by the user as per the requirements of the input video. The cost given in Eq. 3 are the size saliency cost E_{sz}, speed saliency cost E_{sp}, color saliency cost E_{cl}, acceleration saliency cost E_{acc}, and direction saliency cost E_d. The saliency cost are calculated using the weighted average of the parameters of the bounding boxes of the object.

We use a greedy search algorithm [13] to find the key frames as a solution of the optimization problem over the entire video.

B. Retrieval Framework

This framework works well with video shots where the foregrounds in the key frames are stitched together on background to get a single summarized frame. Using our proposed approach, we convert content based video retrieval problem into a content based image retrieval problem. Instead of using the entire video for indexing, content based video retrieval can be carried out with just the summarized frame with reduced system memory requirements as shown in Fig. 3. We create a database with the summarized frames as an index and retrieve an appropriate video using feature extraction and feature matching using NN-classifier.

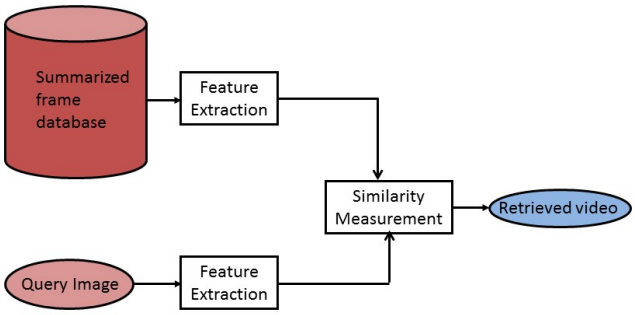

Fig. 3. Retrieval Framework.

III. Results

A rigorous experimental validation is carried out to study the performance of the proposed optimal perceptual video summarization and video retrieval. The simulations were carried out in MATLAB 8.1 Intel Core i-7 3.40 GHz processor. We evaluate our approach on different data sets and evaluate the performance of our approach.

A. Data set

An experimental validation is carried out to study the performance of the proposed smart surveillance using perceptual video summarization. We have used the videos from BBC Motion Gallery, UCF101, OVP, Youtube-8M, UrbanTracker, TRECVID, and other online video archives. We use 1084 video shots from these data sets for our experimentation.

B. Summarization Framework

During preprocessing steps, we find out the object mask of each frame in the video. We fix up the parameter α_1, α_2, β, χ , δ, ϵ, and ζ according to each video and apply greedy search algorithm for finding out the frames. Increasing these parameters signify more weightage on the respective cost and vice-versa. The parameters are fixed as per the needs of the user as shown in Figs. 4-7. In general, though it might appear that a lot of manual parameter setting is required for each individual shot, it is found that a common one-time parameter setting will work for one genre of video such as traffic analysis.

(a) (b)

Fig. 4. (a) Activity Cost, (b) Collision Cost.

(a) (b)

Fig. 5. (a) Size Saliency Cost, (b) Speed Saliency Cost.

We have carried out experiments on road surveillance videos for traffic analysis and accident detection. Apart from color saliency cost, all other saliency cost play a role to summarize traffic surveillance videos. The monitoring of the surveillance videos become easy for the individuals using the proposed approach. We study the smart mobility of the vehicles in our approach. The vehicles can be tracked using the centroid points of the bounding box of the objects as shown in Figs. 6(a) and 6(b). The trajectory of the vehicle maintains its identity before accidents but the trajectories start merging during the accidents. When vehicles collide, they are subsequently enclosed in a single, much larger bounding box after the event whereas if they merely pass by each other, the enlarged combined box disappears instantly. During accidents, the size of the bounding box increases which in turn increases the size saliency cost. In addition, acceleration also changes thereby increasing acceleration cost over a set of frames. After accident, the direction of the object trajectory changes abruptly. We study this situation in traffic surveillance video leading to a better and quicker monitoring of events.

Smart surveillance systems are a powerful tool for real-time detection of potentially suspicious behaviors in shopping malls. We effectively track the trajectories of the people in the malls and study their appearances in the mall. We study the suspicious behavior of a group in a mall. During suspicious event, the individual trajectory merges and the actions are very fast thereby leading to an increase in size saliency cost, speed saliency cost and acceleration saliency cost as shown in Figs. 6(c) and 6(d). The information rate IR conveys the amount of information in the summary that measures the efficiency of the summarized frame. The reduction ratio RR is termed as the ratio of frames in summary to the total frames in the video. We achieve a better reduction ratio and information rate using the proposed approach as shown in Table I.

TABLE I
Comparison for Summarization Results

Videos	IR			RR	
	Shih [14]	Pritch [13]	Ours	Pritch [13]	Ours
Walk1 (Fig. 4(a))	0.23	0.23	0.23	0.38	0.07
Trampouline (Fig. 4(b))	0.31	0.3	0.36	0.42	0.08
Accident1 (Fig. 5(a))	0.3	.28	0.32	0.25	0.08
Walk2 (Fig. 5(b))	0.27	0.28	0.31	0.47	0.1
Accident2 (Fig. 7(a))	0.25	0.26	0.29	0.44	0.08
Accident3 (Fig. 7(b))	0.33	0.33	0.38	0.23	0.07

C. Retrieval Framework

For quick smart surveillance of the videos, we propose to use a single summarized frame instead of a video shot in the database for content based video retrieval [15], [16]. This reduces the search space and computational complexity while retrieving a video. We have created a database with 1306 summarized frames of varying classes. The number of videos in each class varies from 15-30. We use query-by-single-image approach as an input at the query side and single summarized frame at the database side. We retrieve the video based on features such as Graph Based Visual Saliency [17]. To match the feature at the query and database side, we use

(a) (b) (c) (d)

Fig. 6. (a,b) Traffic analysis, (c,d) Suspicious Behavior

(a) (b)

Fig. 7. (a) Acceleration Saliency Cost, (b) Direction Saliency Cost

NN-classifier. We show the top six retrievals for a given query in Fig. 8.

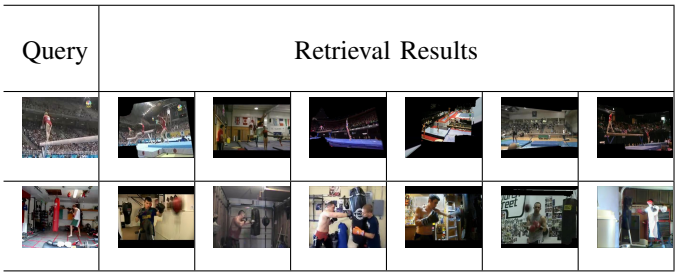

Fig. 8. Examples of retrieval framework. Query image (first column), top six retrieved summarized frames (second column)

We compare the precision values of the retrieval technique using boxplot based visualization. We compare the proposed approach with two other baseline methods Chun et al. [18] and Lai [19] to compute the effectiveness of the proposed approach as shown in Fig. 9. When compared to the motion features of Lai [19] and texture feature of Chun et al. [18], we observe, 50% precision values lie in the third and fourth quartile (that is higher side of precision values) for the proposed features. This implies that the proposed feature work better than the conventional features.

We study this framework for surveillance videos installed indoors and outdoors for different environments. Our work is an efficient approach of smart surveillance.

IV. Conclusion

In this paper, we have presented a smart way of surveillance using video summarization that reduces the surveillance content. We proposed an optimal framework for summarizing gigantic videos using few key frames. In addition, to reduce the search time we perform content based video retrieval using

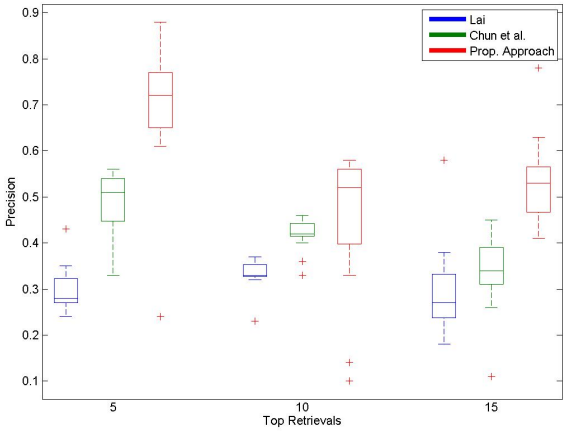

Fig. 9. Boxplot-based visualization of precision values for different retrieval techniques [18], [19].

the summarized frame that is essentially converting a video retrieval problem to a image retrieval one. We proposed a single query by image based video retrieval that reduces the storage space and time complexity. The experiments on various data sets show that it is a promising approach towards the problem of monitoring huge data. This work can be extended with the multi camera scenario and for 3-D videos.

References

[1] "Statistics and Facts About Security and Surveillance Technology," https://www.statista.com/topics/2646/security-and-surveillance-technology/, Oct 2014.

[2] M. Bilal, A. Khan, M. U. K. Khan, and C. M. Kyung, "A Low Complexity Pedestrian Detection Framework for Smart Video Surveillance Systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. 99, pp. 1–1, 2016.

[3] W. Zhou, D. Saha, and S. Rangarajan, "A System Architecture to Aggregate Video Surveillance Data in Smart Cities," in *Proc. IEEE GLOBECOM'15*, Dec 2015, pp. 1–7.

[4] P. Birnstill, D. Ren, and J. Beyerer, "A User Study on Anonymization Techniques for Smart Video Surveillance," in *Proc. IEEE AVSS'15*, Aug 2015, pp. 1–6.

[5] R. H. Evangelio, I. Keller, and T. Sikora, "Multiple Cue Indexing and Summarization of Surveillance Video," in *Proc. IEEE AVSS'13*, Aug 2013, pp. 371–376.

[6] Y. Wang and J. Kato, "A Distance Metric Learning Based Summarization System for Nursery School Surveillance Video," in *Proc. IEEE ICIP'12*, Sep 2012, pp. 37–40.

[7] M. M. Salehin and M. Paul, "Summarizing Surveillance Video by Saliency Transition and Moving Object Information," in *Proc. IEEE DICTA'15*, Nov 2015, pp. 1–8.
[8] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.
[9] Y. F. Ma, X. S. Hua, L. Lu, and H. J. Zhang, "A Generic Framework of User Attention Model and its Application in Video Summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct 2005.
[10] S. Thomas, S. Gupta, and K.S.Venkatesh, "Perceptual Video Summarization - a New Framework for Video Summarization (in press)," *IEEE Trans. Circuits Syst. Video Technol.*, Apr 2016.
[11] S. S. Thomas, S. Gupta, and K. S. Venkatesh, "Perceptual Synoptic View of Pixel, Object and Semantic Based Attributes of Video," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 367–377, Jul 2016.
[12] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug 2014.
[13] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological Video Synopsis and Indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1971–1984, Nov 2008.
[14] H. C. Shih, "A Novel Attention-Based Key-Frame Determination Method," *IEEE Trans. Broadcast*, vol. 59, no. 3, pp. 556–562, Sep 2013.
[15] S. S. Thomas, S. Gupta, and K.S.Venkatesh, "Perceptual Synoptic View based Video Retrieval Using Metadata," *Springer Signal,Image and Video Processing*, vol. 11, pp. 549–555, Mar 2017.
[16] S. S. Thomas, S. Gupta, and K. S. Venkatesh, "Content Based Video Retrieval: A New Perspective (in peer review)," *IEEE Trans. Circuits Syst. Video Technol.*, 2017.
[17] J. Harel, C. Koch, and P. Perona, "Graph-based Visual Saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.
[18] Y. D. Chun, N. C. Kim, and I. Jang, "Content-Based Image Retrieval Using Multiresolution Color and Texture Features," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1073–1084, Oct 2008.
[19] Y. H. Lai and C. K. Yang, "Video Object Retrieval by Trajectory and Appearance," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 6, pp. 1026–1037, Jun 2015.