Time Driven Video Summarization using GMM

Sujatha C
Dept. of CSE
BVBCET Hubli-580031
Email: sujata_c@bvb.edu

Akshay Ravindra Chivate
Dept. of ECE
BVBCET Hubli-580031
Email: akshaychivate@gmail.com

Sayed Altaf Ganihar
Dept. of ECE
BVBCET Hubli-580031
Email: altafganihar@gmail.com

Uma Mudenagudi
Dept. of ECE
BVBCET Hubli-580031
Email: uma@bvb.edu

***Abstract*—In this paper, we propose a method to browse the activities present in the longer videos for the user defined time. Browsing of activities is important for variety of applications and consumes large amount of viewing time for longer videos. The aim is to generate a summary of the video by retaining salient activities in a given time. We propose a method for selection of salient activities using motion of feature points as a key parameter, where the saliency of a frame depends on total motion and specified time for summarization. The motion information in a video is modeled as a Gaussian mixture model (GMM), to estimate the key motion frames in the video. The salient frames are detected depending upon the motion strength of the keyframe and user specified time, which contributes for the summarization keeping the chronology of activities. The proposed method finds applications in summarization of surveillance videos, movies, TV serials *etc.* We demonstrate the proposed method on different types of videos and achieve comparable results with stroboscopic approach and also maintain the chronology with an average retention ratio of 95%.**

I. Introduction

In this paper, we propose a method for quick browsing of activities present in the longer videos for a user defined time. The summarized video is generated by retaining the salient activities, which are useful for viewers to browse over the video to understand the content of the video in a much shorter time than the original video. Since, the video data is rife in the current age, the requirement for viewing time and storage space is huge. Hence, video summarization plays an important role in reducing the viewing time and storage space, and is employed in different areas like surveillance, entertainment, sports *etc.* It also finds applications in video indexing, browsing, retrieval and videos highlights.

A number of methods have been proposed in the literature for the summarization of videos using different features like color histogram [1], text [2], motion in sports video [3], [4] etc. The authors in [2], [3] use clustering-based techniques where the idea is to produce the summary by clustering similar frames/shots and select a limited number of frames per cluster. However, the method is computationally expensive and does not preserve the temporal order. The authors in [5], [6] have proposed object based video summarization methods wherein the moving objects are detected and segmented for summarization. The computational complexity of such methods is high as the number of objects and the length of the video increases. In [6], the summarized video generated contains the compact frames showing multiple activities which have occurred at different times. The method achieves a large condensation ratio, but the temporal relationship among objects is destroyed. However, for better understanding of the content of video and in security related data the chronology is needed.

Most of the summarization techniques extract keyframes that represent the salient content of a video shot. The authors in [7] focus on extracting the key-frames within a given shot depending on the content complexity of the shot. The method is suitable only for scripted videos such as news, movies, TV shows *etc.*, where the content is planned and edited. In [8], highlights are generated based on the semantic concepts of specific sports videos that allows the viewers to customize their choice of interest. Cheng *et al.* [9] proposed a quick browsing system to help the user easily locate the subjects of interest in surveillance videos. They divide the video into segments of equal lengths wherein the size of segment depends on time interval defined by user and assume the background remains same for a segment. The motion tracking algorithm is used to identify the objects, which are then fused with the corresponding background (of the respective segment) to generate a compact frame.

We develop a robust method for video summarization in the user defined time, which is suitable for different types of videos such as surveillance videos, movies, TV shows *etc.* and maintain the chronology of the activities. Towards this we make the following contributions:

1) We propose a method for quick browsing of activities in a video as per the user defined time.
2) We model the motion information in the video using GMM for selecting the set of salient frames.
3) We achieve comparable condensation and retention ratios by maintaining the chronology of the activities.

In Section II, we give the overview of the proposed method for video summarization and selection of the set of salient frames. In Section III, we demonstrate and discuss the results. We give conclusion in Section IV.

II. Time Driven Video Summarization

Motion is an intrinsic property of the visual world which defines the activities present in the video. We are concerned with quick browsing of the activities in a video for user defined time. Thus, motion is used as a key parameter for the summarization of a video. The overview of the proposed summarization method is shown in Fig 1. Initially, the input video is divided into the frames and then feature points present in the frame are detected. The motion information is computed only at the feature points and it is used to compute motion in a frame. In Fig 1., the higher intensity of the gray color represents high amount of motion information present in the frame. The frames with no motion information are eliminated to generate the sub-summarized video. The motion information

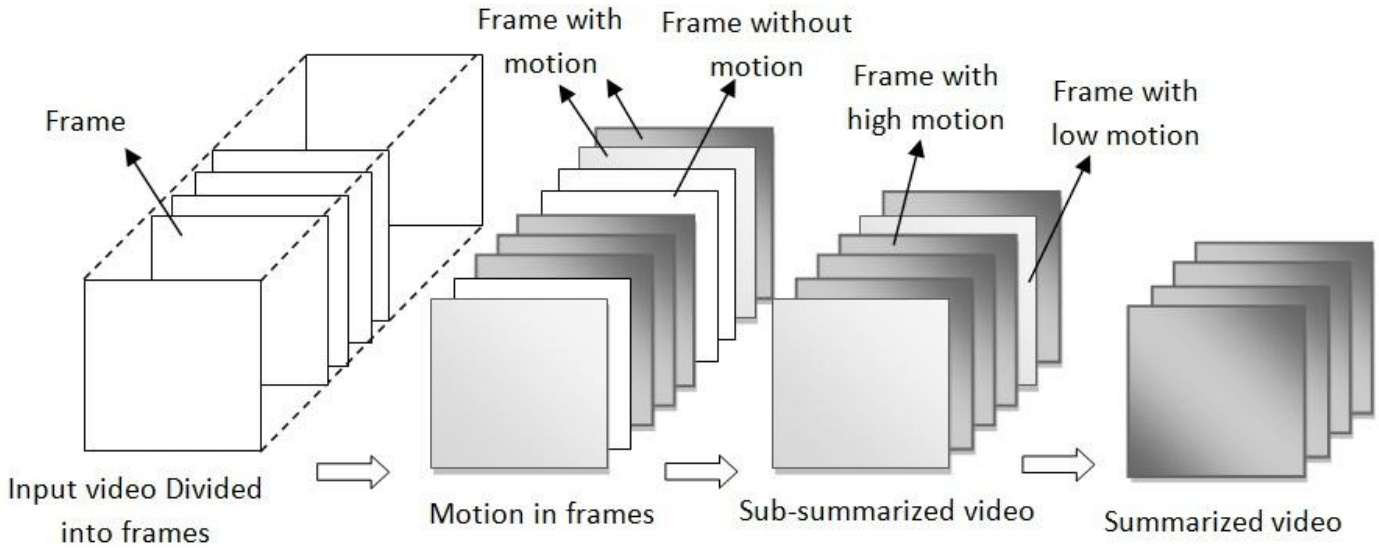

Fig. 1. Overview of the time based video summarization

in the sub-summarized video is modeled as a Gaussian Mixture Model (GMM). The peaks of the individual Gaussian curves (frames with high motion information) are defined as the keyframes which are the indices to construct the building blocks of the summarized video. The bandwidth (set of salient frames) of contributing Gaussian curve at a keyframe is set according to relative strength of motion, number of keyframes and user defined time T. Thus, the final summarized video is constructed using these set of salient frames.

In what follows we present selection of salient frames.

A. Selection of salient frames

Consider $V(x, y, t)$ as a given input video, which needs to be summarized. The motion information in the given video is computed for a set of feature points in every frame using Lukas-Kanade optical flow algorithm [10]. These feature points are selected based on the assumptions of brightness constancy, temporal persistence and spatial coherence.

Let $p(x_r, y_r, t_r)$ be a feature point at a spatial location (x_r, y_r) and at instance t_r. If point p moves to point $q(x_c, y_c)$ during time transition from t_r to t_{r+1}, then the transformation of point p is given by:

$$(x_c, y_c, t_{r+1})^T = \mathcal{A}(x_r, y_r, t_r)^T \tag{1}$$

where, $\mathcal{A}$ defines the motion matrix for the feature point p.

The motion vector $(\vec{v_i})$ for each feature point is calculated from the motion matrix $\mathcal{A}$. Modeling of the motion matrices for each feature point increases the time complexity of the algorithm. To reduce the computational complexity, the motion parameter $\mathcal{A}_T$ is defined for a frame as:

$$\mathcal{A}_T = \sum_{i=1}^{n} mag(\vec{v_i})$$

where, n is the number of feature points in the frame. Based on the motion information $\mathcal{A}_T$ obtained for each frame, the input video is preprocessed to obtain a sub-summarized video by eliminating the frames containing no motion information.

We model the motion parameter $\mathcal{A}_T$ as a weighted combination of Gaussian curves and is given by:

$$\mathcal{A}_T(t|\theta) = \sum_{j=1}^{m} \pi_j \mathcal{N}(t|\mu_j, \Sigma_j) \tag{2}$$

where, j runs from 1 to m and m denotes the number of peaks in the sub-summarized video, π_j is the weight of the j^{th} component, θ is the parameter list for the GMM and $\mathcal{N}(t|\mu_j, \Sigma_j)$ is the probability density function for the individual Gaussian curve [11], [12].

To sample the peaks present in the motion model, the principle of Maximum *aposterior* (MAP) is applied with respect to time to estimate the sample mean μ_j and the variance Σ_j for the individual Gaussian curve. The *aposteriori* probability of an j^{th} individual mixture component is given by [13]:

$$\mathcal{A}_T^j(t_j|t, \theta) = \frac{\mathcal{N}(t|\mu_j, \Sigma_j)\pi_j}{\mathcal{A}_T(t|\theta)} \tag{3}$$

Fig 2 shows Gaussian mixture model for the sub-summarized video, where m is the number of peaks in the motion model signifying high motion information. Each peak is considered to be a keyframe which serves as the index for the construction of building block for the summarized video. The selection of a set of salient frames (block) in the neighborhood of the indexed keyframe depends on the user-specified time T. The size of salient block *i.e.* the bandwidth BW for each Gaussian curve (represented in Fig 2) varies based on the relative intensity of the keyframe η_j, the user defined time T and number of peaks m in the motion model. Thus, we get the bandwidth BW_j for the j^{th} Gaussian distribution as:

$$BW_j \propto \eta_j \cdot \left(\frac{T}{m}\right) \tag{4}$$

TABLE I. CR AND RR FOR DIFFERENT VIDEOS

Method	Video1 (Hostel Parking)		Video2 (Tunnel Video)		Video3 (Street)	
	Duration (CR%)	**No. of Objects (RR)**	**Duration (CR%)**	**No. of Objects (RR)**	**Duration (CR%)**	**No. of Objects (RR)**
Original Video	1614sec	58	480sec	48	2042sec	151
Stroboscopic	153sec (90.52)	58 (1.00)	51sec (89.38)	48 (1.00)	210sec (89.72)	151 (1.00)
Time Based	30sec (98.14)	34 (0.59)	10sec (97.20)	42 (0.88)	120sec (94.12)	124 (0.82)
	60sec (96.30)	41 (0.71)	20sec (95.84)	44 (0.92)	180sec (91.20)	144 (0.95)
	120sec (92.58)	52 (0.90)	30sec (93.75)	45 (0.94)	210sec (89.72)	145 (0.96)
	153sec (90.52)	54 (0.93)	51sec (89.38)	46 (0.96)	360sec (82.38)	147 (0.97)
	300sec (81.42)	58 (1.00)	60sec (87.50)	48 (1.00)	540sec (73.60)	151 (1.00)

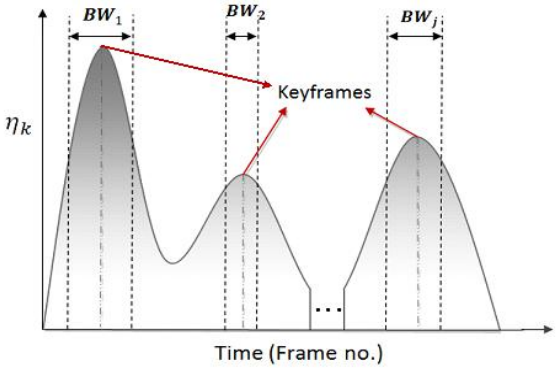

Fig. 2. Gaussian mixture model for motion information in the sub-summarized video

say $\alpha = \beta \cdot (\frac{T}{m})$ be a constant where β is the proportionality constant which is normalized according to the standard deviation of the j^{th} Gaussian distribution. Thus, we get BW_j as:

$$BW_j = \alpha \cdot \eta_j \tag{5}$$

where $j \in \{1, \ldots, m\}$. The i^{th} frame in the neighborhood of j^{th} keyframe is said to be salient if f_i (saliency of the i^{th} frame) is 1, and is defined as:

$$f_i^j = \begin{cases} 1 & \text{if } (\mu_j - \frac{BW_j}{2}) \leq \mathcal{A}_t^i \leq (\mu_j + \frac{BW_j}{2}) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The salient frames which are in chronology are used in the construction of final video summary.

III. RESULTS AND DISCUSSIONS

We evaluate the results of summarized videos using two parameters: Condensation Ratio (CR) and Retention Ratio (RR) [14]. In most video summary algorithms, the CR is used to verify the effectiveness of the algorithms, whereas RR determines the number of objects (amount of information) retained. Let 'V' be the given input video and 'S' be the summarized video, then the CR and RR are defined as:

$$\text{Condensation Ratio (CR)} = \left[1 - \frac{\text{length of S}}{\text{length of V}}\right] \times 100 \tag{7}$$

$$\text{Retention Ratio (RR)} = \left[\frac{\text{\# objects in S}}{\text{\# objects in V}}\right] \tag{8}$$

To test the performance of our proposed time based video summarization method, we have implemented and tested the method on three different types of video clips. The execution time for generating a summarized video of 1 min for a video clip of 30 min took 33 min (approximately), on a Pentium 4 (2.0 GHz) processor with 3GB of RAM with Ubuntu 11.04 O.S.

We compare the results of proposed time driven video summarization method with general stroboscopic approach as shown in Table I for three different surveillance videos. The summarized videos are obtained for different user defined time intervals to analyze the trade off between CR and RR. In case of tunnel video (video 2) of 480 seconds, the summarized video is obtained for different user defined time (10, 20, 30 and 60 sec) *i.e.* the CR is defined by the user and RR is verified. We can see, as the user defined time T increases, the CR decreases and RR increases ($0 \leq RR \leq 1$). Thus, to view all the activities in video, the user defined time must be higher. The activities from different time instances are mapped on to a single time instance in stroboscopic approach. Thus, the summarized video is generated with large CR and RR, but introduces ambiguity due to overlapping of different time instances as shown in Fig 3. The same object is shown multiple times which creates the ambiguity for the viewer, when two or more similar objects (cars, people in same uniform, *etc.*) are present in succession. To overcome this, our method provides relatively lower RR compared to stroboscopic method, but focuses on retaining activities with high motion and also maintain the chronology to reduce the ambiguity. For video 2 (tunnel video) shown in Table I, the stroboscopic method provides RR=1 and our method provides RR=0.96 with same CR. However, our method maintains the chronology and removes the ambiguity at the cost of just 0.04 reduction in RR. The video 1 and video 3 achieves the RR values of 0.93 and 0.96 respectively with same CR as of stroboscopic method. Thus, the method achieves an average of 95% RR as compared to stroboscopic approach.

Fig 4. shows the selection of salient frames for hostel parking video (video 1) to generate the summarized video for the different user defined time intervals. We can see that as the time T is decreased, the the bandwidth BW decreases. For BW_1 with $T = 300sec$, the salient frames ranges from 426 to 932. Similarly, for $T = 120sec$, the salient frames ranges from 194 to 297, thus showing reduction in the size of bandwidth. The size of the bandwidth depends on the user defined time (CR) and is directly proportional to RR and strength of motion in the keyframe.

We also provide a subjective evaluation of our proposed methodology for finding three different specified targets in a 2hr 15min 22sec movie video. The positions of the three targets

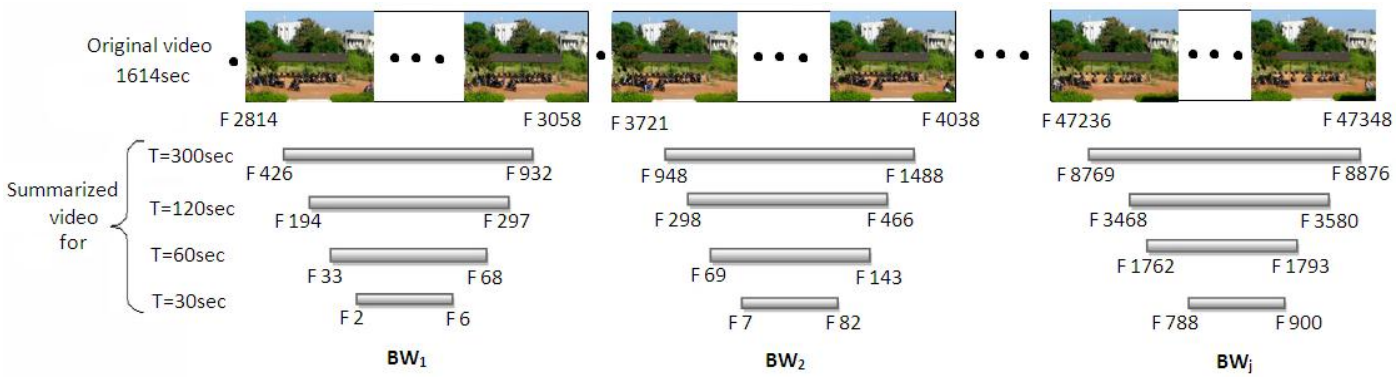

Fig. 4. Results for hostel parking video showing the selection of salient frames using bandwidth based on user defined time ($T = 300, 120, 60, 30sec$)

(a) Stroboscopic method

(b) Stroboscopic method

(c) Time based method

(d) Time based method

Fig. 3. Results for Stroboscopic method (a) & (b) and results for Time based method (c) & (d) for video 2 (Tunnel Video)

are: A-10'54", B-18'33", and C-57'02". The summarized video of 10min is obtained, which is provided to five people to search the specified targets and the respective time taken by each is depicted in Table II. Thus, on an average of 54sec is needed to detect target A which occurred at 10'54".

TABLE II. THE SEARCH TIME FOR THREE TARGETS USING THE SUMMARIZED VIDEO OF 10MIN FOR A 2HR15MIN MOVIE VIDEO

	A	B	C
Person 1	60	88	106
Person 2	45	70	80
Person 3	40	60	75
Person 4	53	75	93
Person 5	72	107	132
Average	54	80	99

IV. CONCLUSION

We have proposed a method to quickly browse through the activities present in the longer videos for the user defined time. We have modeled the motion information in the video as a GMM to determine the keyframes, using which the salient frames are selected. We have achieved an average of 95% retention ratio by maintaining the chronology as compared with stroboscopic approach. We have demonstrated the results on different surveillance videos and movies.

REFERENCES

[1] Y. Zhuang, Y. Rui, H. T. S, and Mehrotra., "Adaptive key frame extraction using unsupervised clustering," *IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 866–870, October 1998.

[2] N. Dimitrova, "Context and memory in multimedia content analysis," *IEEE multimedia*, vol. 11, no. 3, pp. 7–11, 2004.

[3] L. Y. Duan, M. Xu, T. S. Chua, Q. Tian, and C. S. Xu, "A mid-level representation framework for semantic sports video analysis," *ACM Trans.*, vol. 1, pp. 33–44, November 2003.

[4] C. Y. Chen, J. C. Wang, J. F. Wang, and Y. H. Hu, "Motion entropy feature and its applications to event based segmentation of sports video," *EURASIP journal on Advances in Signal Processing*, vol. 2008, no. 152.

[5] Z. Tian, J. Xue, X. Lan, C. Li, and N. Zheng, "Key object based static video summarization," in *MM'11 proceedings of 19th ACM International Conference on Multimedia*, pp. 1301–1304.

[6] A. Rav-Acha, Y. Pritch, and S. Peleg, "Non-chronological video synopsis and indexing.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, pp. 1971–1984, November 2008.

[7] T. Liu, H.-J. Zhang, and F. Qi, "A novel video keyframe extraction algorithm based on perceived motion energy model," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, vol. 13, pp. 1006–1013, October 2003.

[8] M. H. Kolekar and S. Sengupta, "Event-importance based customized and automatic cricket highlight generation," *International Conference on Multimedia and Expo.*, pp. 1617–1620, 2006.

[9] Cheng-Chieh, Chiang, M.-N. Tsai, and H.-F. Yang, "A quick browsing system for surveillance videos," *MVA2011 IAPR Conference on Machine Vision Applications, Nara, Japan*, June 2011.

[10] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2008.

[11] R. Duda, P. Hart, and D. Stork., *Pattern Classification*. John Wiley and Sons, 2 edition.

[12] J. L. Devore, *Probability and statistics for engineering and sciences.* Brooks/Cole Cengage learning, Boston, USA, 8 edition.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[14] C. M. Taskiran, "Evaluation of automatic video summarization systems," in *Proc. SPIE 6073, Multimedia Content Analysis, Management, and Retrieval*, (San Jose, CA), Jan 2006.