Video Summarization using low level Features

Gautam Patel[1]
Bijal Joshi[2]

gautam.patel@bapugkv.ac.in, bijal.joshi@bapugkv.ac.in

Abstract: -** **In this report, an overview of video summarization has been presented. Video summarization as the name implies generates the video summary. A video summary can be generated both manually and automatically, but due to the huge volumes of video data and limited manpower, it's getting more and more important to develop fully automated video analysis. According to proposed method multiple features, obtained from video frames, are combined to describe the frame difference between consecutive frames. It is observed that certain frame difference features have more influence in generating a representative frame difference measure. Moreover, some features are more relevant than others in different video genres. We used three different low level features color histogram, correlation and edge orientation histogram and generate feature vector. Fuzzy C-means and K-means have been effectively used to generate meaningful, enjoyable video summary using generated feature vector.

Keywords: - video, image, summary, frame, Fuzzy logic, C-means, K-means

I INTRODUCTION

The volume of digital video data has been increasing significantly in recent years due to the wide usage of multimedia applications in the areas of education, entertainment, business, and medicine. To handle this huge amount of data efficiently, many techniques about video segmentation, indexing, and abstraction have emerged to catalog, index, and retrieve the stored digital videos. Digital video data refers to the video picture and audio information stored by the computer using digital format [1] [2].

A Video consists of a collection of video frames, where each frame is a picture image. When a video is being played, each frame is being displayed sequentially with a certain frame rate. The typical frame rates are 30 and 25 frames/second as seen in the various video formats. An hour of video has 108,000 or 90,000 frames if it has a 30 or 25 frames/second rates, respectively [1]. A typical video structure is Video is segmented into number of scenes. Scenes are again segmented into Shots. Shots again made up of number of Frames. Figure 1.1 depicts Structure of Video.

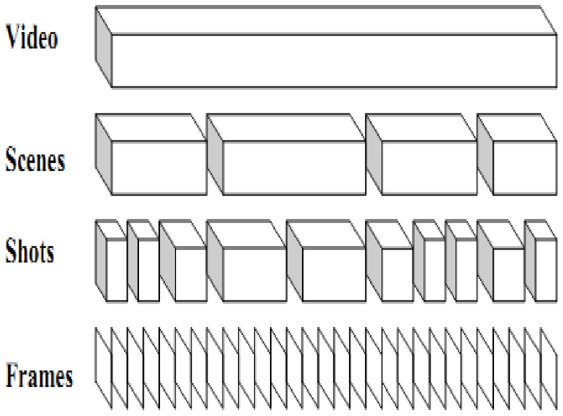


Figure 1.1 Structure of Video [1]

The video summarization methods generate summaries of the videos which are sequence of stationary or moving images. The extraction or generation of these stills (called as key frames/picture image) becomes the main focus of all summary work. The extracted key frames are used to generate the Video summary, which contains the most highlighted scene or the overall story of the video.

A key frame extraction technique must be fully automated in nature and must use the contents of the video to generate summary [3]. The key frames are extracted using some features like color, objects, events, edges, moments, correlation etc. The object, events and moments are considered as high level features that used for the specific applications. But color, edge, correlation are the low level features which can be used to generate the summary for all types of video genre. Here this Report includes how low level features are used to generate the automated summary using unsupervised learning approach.

II RELATED WORK

Various Researchers have used various methods to generate the video summary. Here are different methods will be discussed in the following subsections.

1. Shot Boundary Based Detection

In the shot boundary-based key frame selection, a video is segmented into a number of shots and one or more key frames are selected from each shot. Many techniques have been developed to detect a shot boundary automatically. These schemes mainly differ in the way the inter-frame difference is computed. The difference can be determined by comparing the corresponding pixels of two consecutive frames Color or grayscale histograms can be also used [7]. Alternatively, a technique based on changes in the edges has also been developed [5]. Other schemes use domain knowledge such as predefined models, objects, regions, etc. Hybrids of the above techniques have also been investigated. Once shot detection is completed; key frames are selected from each shot. For example, the first, the middle, or the last frame of each shot can be selected as key frames [3]. If a significant change occurs within a shot, more than one key frame can be selected for the shot [4] [5].

2. Perceptual Feature-based Detection

- **Color-Based Selection**

It is used the color histogram difference between the current frame and the last extracted key frame to draw out key frames from the video. First quantizes the color space into 64 super-cells. Then, a 64-bin color histogram is calculated for each frame where each bin is assigned the normalized count of the number of pixels. The distance (D_{his} (*I*,*Q*)) between two color histograms, *I* and *Q*, each consisting of *N* bins, is quantified by the following metric:

$$\boldsymbol{D_{his}(I,Q) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_{ij} (I_i - Q_i)(I_j - Q_j)}$$

Where the matrix a_{ij} represents the similarity between the colors corresponding to bins *i* and *j*, respectively. This matrix needs to be determined from human visual perception studies. If a_{ij} is an identity matrix, this equation measures the Euclidean distance between two color histograms [4].

- **Motion Based Selection**

Motion is more vulnerable feature to render the visual content of the shot, where more pans and zooms of camera are there. In the algorithm for key frame identification, a motion metric based on optical flow is computed for key frame.

Optical flow is a 3D movement projected on a 2D plane and used to detect the motion. The frame where the motion is very high is selected as a key frame [7]. In the algorithm for key frame identification, a motion metric, M(t) based on optical flow is computed for frame t with a size of r´c using the following formula:

$$\boldsymbol{M(t)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{r}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{c}} |\boldsymbol{O_x(i,j,t)}| + |\boldsymbol{O_y(i,j,t)}|$$

Where $O_x(i,j,t)$the x component of optical flow of a pixel is positioned i and j in frame t and similarly $O_y(i,j,t)$ for the y component. Then, the metric is analyzed as a function of time to select key frames at the minima of motion.

3. Feature Vector Space-based Detection

The feature vector space-based key frame selection represents in [10] considers that the frames in a video sequence are characterized by not just one but multiple features. Several descriptors are first extracted from each video frame by applying a segmentation algorithm to both color and motion domains, which forms the feature vector. Then all frames' feature vectors in the shot are gathered to form a curve in a high-dimensional feature space. Finally, key frames are extracted by estimating appropriate curve points that characterize the feature trajectory, where the curvature is measured based on the magnitude of the second derivative of the feature vectors with respect to time. The curve splitting algorithm is represented in [11].

4. Cluster Based Detection

An unsupervised clustering approach based on color histogram features was presented in [12].In a brief explanation of the cluster-based key frame selection, a given shot, *s* has *N* number of frames, and these *N* number of frames, {f_1, f_2,..., f_n} are clustered into *M* number of clusters, {C_1, C_2, ..., C_m}. This clustering is based on the similarity measures among frames, where the similarity of two frames is defined as the similarity of their features, such as color, texture, shape, or a combination of the above. K-mean algorithm can be used to generate clusters which are represented in [13].Several researchers have stressed on selection of multiple key frames. First they select a key frame that is the closest frame to the centroid of a cluster. The similarity between the key frame and each frame in a cluster is calculated. If this similarity is larger than a predefined similarity threshold, the frame is added to a set of key frames.

5. Motion Activity Descriptor based Detection

In [21] proposed the following strategy for quick video summarization:

First they compute the intensity of motion activity of all the frames and use it to locate regions that will be easy/difficult to summarize. Then,

- Cluster the sequence frames in color feature space.
- Possibly arrange the clusters in easy to access structures such as hierarchical structures.
- Extract a Key frame or a key sequence from each of the clusters.
- The set of all the key-frames/sequences comprises the summary.

6.Sampling based approach

In sampling-based key frame selection approach, key frames are selected by randomly or uniformly sampling the video frames from the original sequence at certain time intervals [1]. Although this is probably the simplest way to extract key frames, the drawback is that such an arrangement may cause some short yet important segments to have no representative frames while longer segments could have multiple frames with similar content, thus failing to capture and represent the actual video content.

7. Segment-based key frame selection approach

One major drawback of using one or more key frames for each shot is that it does not scale up for long videos since scrolling through hundreds of images is still time-consuming, tedious and ineffective. Therefore, recently more and more people begin to work on higher-level video unit, which we call a video segment in this report. A video segment could be a scene, an event, or even the entire sequence. In this context, the segment-based key frame set will surely become more concise than the shot-based key frame set [1].

III PROPOSED WORK

Shot detection is the primary and very important task to identify the boundary. There are various techniques available in the literature and also they are domain dependent. Many of the techniques use domain knowledge to detect shot boundary by the use of threshold applied on the extracted features. Unsupervised learning has the intrinsic characteristic to cluster the data based on their behavior. Hence, we applied unsupervised clustering approach to detect shot transition. However, we used a combination of three features; color histogram, correlation and edge orientation histogram for extracting key frames. Then after we generate feature vector of all three frame difference measure. Unsupervised learning approaches Fuzzy C-means and K-means used to detect cluster. For each cluster there is a representative key frame, which describes the cluster center. Finally to generate video summary we are extracting those frames which are nearer to cluster center. Following are the steps to make video summary.

- **Color Histogram Difference (CHD)**

The color histograms have been commonly used for key frame extraction in both frame difference based techniques and clustering techniques [5][7] .This is because the color is one of the most important visual features to describe an image. Color histograms are easy to compute and are robust in case of small camera motions. For computing FDM (Frame difference Measure), color histogram has been built in HSV color space by performing a quantization step to reduce the number of distinct colors to 64. Instead of computing one histogram for the entire image, we divided image in a total of 10 sections, each of size 35 x 28. This is to effectively measure the level of difference between two frames. Each corresponding section of one frame is compared with the corresponding section of other frame using the histogram intersection mechanism [3]. The histogram difference $HD_{i,j,s}$ between two corresponding section of histogram H_{is} of frame i and histogram H_{js} of frame j is defined as

$$\boldsymbol{HD_{i,j,s} = 1 - \sum_{k=0}^{64} min \quad \left(H_{is}(k) - H_{js}(k)\right)}$$

The Histogram difference (HD) between two frames i and j is then calculated by taking the average of the difference measure between each section.

$$\boldsymbol{HD_{i,j} = \frac{1}{T_s} \sum_{k=1}^{T_s} HD_{i,j,k}}$$

- **Correlation Histogram Difference (CD)**

The correlation coefficients have been very popular scheme to find similarity between two data sets. The correlation coefficients are invariant to brightness and changes in the contrast. Again, for computing correlation measure, we divide frames into 10 sections of size 35 x 28. The correlation values of each section are then averaged.

The correlation is measured for three color channel values red, green and blue [3]. The correlation difference CDp,q,s,c of a color channel 'c' between two corresponding section of frame p and q is calculated. Then after, the correlations of all sections of frame i and j are averaged to obtain the overall correlation CDi,j,c for a color channel.

$$\boldsymbol{CD_{i,j,c} = \frac{1}{T_s}\sum_{k=1}^{T_s} CD_{i,j,k,c}}$$

Where 'c' here denotes color channel (red, green or blue). Then, the overall correlation difference measure CDi,j between frames i and j is obtained by averaging the value of each color channel.

$$\boldsymbol{CD_{i,j} = \frac{CDi_{i,j,Red}+CD_{i,j,Green}+CD_{i,j,Blue}}{3}}$$

- **Edge Histogram Difference (ED)**

The third measure used for computing the histogram of edge orientation. The edges are good under illumination changes. The edges are first computed using horizontal and vertical Sobel operators which are then used to find gradient and angle of edges. Sobel Operators are discussed in following subsection. The angles are then used to build a histogram of edge orientation. For simplicity, we defined only 72 bins for the angles. As in the case of histograms, we compare histograms of corresponding sections of the two frames [3]. The Edge histogram difference (ED) between two frames i and j is calculated by taking the average of the difference measure between each section.

$$\boldsymbol{ED_{i,j} = \frac{1}{T_s}\sum_{k=1}^{T_s}\left|ED_{i,j} - ED_{j,k}\right|}$$

IV Fuzzy C-means Method

Fuzzy C-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade.

Step :1 Read Input Video.

Step :2 Compute Color Histogram Difference (HD), Correlation Histogram Difference (HD), Edge Histogram Difference (ED).

Step :3 Generate Feature Vector of all extracted low level features i.e. HD, CD, ED.

Step :4 Apply Fuzzy C-means on the feature vector that will generate Fuzzy partition matrix having membership grade for every cluster.

Step :5 Calculate distance between highest membership grade and cluster center.

Step :6 If distance is more than given radius then tag that frame as key frame.

V K-means Method

The popular and simplest probabilistic and unsupervised clustering algorithm is K-means algorithm. In K-means algorithm we initially decide the number of clusters let us say K number of clusters and hypothesize the centroid or clusters center point.

Step :1 Read Input Video.

Step :2 Compute Color Histogram Difference (HD), Correlation Histogram Difference (HD), Edge Histogram Difference (ED).

Step :3 Generate Feature Vector of all extracted low level features i.e. HD, CD, ED.

Step :4 Apply K-means on the feature vector that will generate distances from each point to every centroid (D).

Step :5 Calculate Euclidean distance between cluster center and distance from D. If distance is more than given radius then tag that frame as key frame.

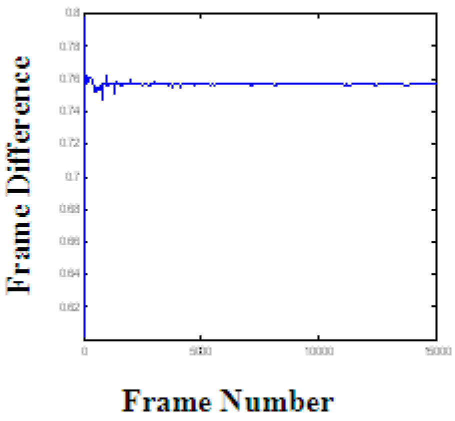
VI RESULT & DISCUSSION

Video summarization techniques have been applied on many different types of videos e.g. home videos, sport, surveillance, news. However, among all these types of videos surveillance and sport genre have their interesting applications related to security and commercial interest. Sport videos are itself in dynamic nature and experience high motion as well as sudden changes. Hence it poses a stiff challenge to generate a good video summary.

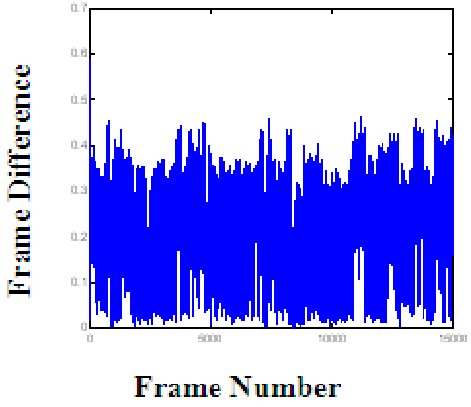
The experiments were conducted on the soccer videos of different conditions. In order to evaluate performance of the proposed algorithm we experimented on 7 different videos. Total duration of each video is about 8 minutes and 30 seconds (15,000 frames) having 25 frame/s. The experiments were conducted on Intel Core i3 processor with 2.53 GHz and 2 GB RAM. Implementation tool used for experimentation is MATLAB 7.8(R2009a).

- **Frame Difference Measure Analysis**

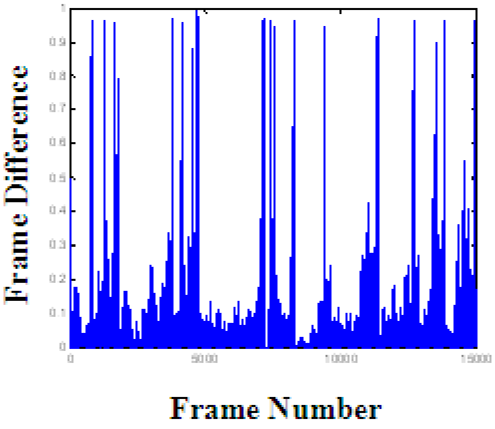
This section starts with a discussion that a single descriptor is not enough to capture all visual features of the image. Here we have represented the three different graphs in Figure 4.1(a) to 4.1(c) of Color, Correlation and Edge Frame Difference respectively for 15000 frames which is generated. It is quite evident from the graph that there is an obvious difference between the FDM (Frame Difference Measure) generated by these three techniques. Here for the color frame difference the graph does not undergo major observable changes where in correlation and Edge the graph undergoes more changes.



(a) Color Frame Difference



(b) Correlation Frame Difference



(c) Edge Frame Difference

VII CONCLUSION

Video summary has been generated by selecting key frames from the created clusters. Well known clustering approaches: Fuzzy C-means and K-means have been used and compared based on the Informativeness of the generated summary. Large numbers of experiments have been conducted on various soccer video datasets. Frames have been chosen as key-frames which are nearer to the centroid of clusters. Experiments have been carried out with different radius to select key frames. Experimental results clearly reflect that Fuzzy c-means achieves better Informativenss than k-means on the soccer video datasets. Increasing the number of clusters definitely boosts the Informativenss but simultaneously it will deteriorate the performance in compression ratio.

VIII REFERENCES

[1]. Jung Hwah Oh, Quan Wen, Sae Hwang, Jeongkyu Lee, "Video Abstraction". *Video Abstraction*, 14, pp. 321-343, 2005.

[2]. Li Y, Zhang T, Tretter D, "An overview of video abstraction techniques". *Proceedings of Tech. Rep., HP-2001-191,* HP Laboratory, 2000.

[3]. Naveed Ejaz and Sung Wook Baik," Weighting low level frame difference features for key frame extraction using Fuzzy Comprehensive Evaluation and indirect feedback relevance mechanism", *Proceedings of IJPS.*, vol.6(14) , pp. 3377-3388,July-2011.

[4]. Ngo, C.W., Pong, T.C., & Zhang, H.J., "On clustering and retrieval of video shots", *Proceedings of ACM Multimedia*, Ottawa, Canada, vol. 1, pp. 51-60, October-2001.

[5]. Girgensohn, A., & Boreczky, J.,"Time-constrained key frame selection technique", *Multimedia Tools and Applications*, vol. 11(3), pp. 347-358, 2000.

[6]. Hanjalic, A., & Zhang, H., "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis". *IEEE Transaction on Circuit and Systems for Video Technology*, vol. 9(8), pp. 1280-1289, 1999.

[7]. Dufaux, F.,"Key frame selection to represent a video". *Proceedings of IEEE 2000 International Conference on Image Processing*, vol. 1, pp. 275-278, 2000.

[8]. Wolf, W.," Key frame selection by motion analysis". *Proceedings of IEEE International. Conference on Acoustics, Speech, and Signal Processing*, pp. 1228-1231, 1996.

[9]. Fatih Poriliki "Multi-Camera Surveillance: Object-Based Summarization Approach" *Proc. of Mitsubishi Electrical Research Laboratory*, 201 Broadway, Cambridge, Massachusetts 02139, 2004.

[10]. Gunsel B, Tekalp AM, "Content-based video abstraction". *Proceedings of IEEE International Conference of Image Processing*, Chicago, USA, pp. 128–132, 1998.

[11]. DeMenthon, D., Kobla, V., & Doermann, D., "Video summarization by curve simplification". *Proceedings of ACM Multimedia*, vol. 3 pp. 211-218, 1998.

[12]. Ramer, U. "An iterative procedure for the polygonal approximation of plane curves". *Computer Graphics and Image Processing*, vol. 1, pp. 244-256, 1972.

[13]. Ciocca G, Schettini R., "Innovative

Algorithm for Key Frame Extraction in Video Summarization”. *J. Real Time Image Process*, vol. 1(1), pp. 69-88, 2006.

[14]. Dim P. Papadopoulos, Savvas A. Chatzichristofis, and Nikos Papamarkos,” Video Summarization Using a Self-Growing and Self-Organized Neural Gas Network”. *Proc of Springer-Verlag Berlin Heidelberg*, Vol.1 (2), pp.216-226, 2011.

[15]. Tomy Chheng “Video Summarization using Clustering”, *Proc. of journal of information and Data Management*, Vol. 1(2), Pages 293-304, June 2010.

[16]. Chen, Tsung, Lee, “Object and color based video Representation for Automated model Free News summarization”. *Proc. IEEE International Conference on Multimedia and Expo*, vol. 3, pp. 209-218, 2003.

[17]. Dian, Yi-ping, Binh Pham, “Sports video summarization using Highlights and play breaks”. *Proc. of MIR*, vol. 30(4), pp. 643-658, 2003.

[18]. Bezdec, J.C.,” Pattern Recognition with Fuzzy Objective Function Algorithms*”, Proc of Plenum Press*, New York, 1981.

[19]. Silvia Corchs, Gianluigi Ciocca, Raimondo Schettini.,” Video Summarization using a Neurodynamical Model of Visual Attention”, *Proceedings of IEEE 2004 International Conference on Image Processing*, vol. 1, pp. 305-309., 2004.

[20]. Ajay Divakaran, Kadir A. Peker , Huifang Sun,” Video Summarization Using Motion Descriptor”. *Proc. of Mitsubishi Electrical Research Laboratory*,Murry Hill,NJ07974,USA pp. 1-6, 2004.