IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p-ISSN: 2278-8735
PP 52-56
www.iosrjournals.org

A Survey on Domain Adaptive Video Summarization Algorithm

Aiswarya.N.R[1], Smitha.P.S[2]
[1](*Department Of Electronics and Communication Engineering, Sree Chitra Thirunnal College of Engineering Trivandrum, Kerala)*
[2](*Department Of Electronics and Communication Engineering, Sree Chitra Thirunnal College of Engineering Trivandrum, Kerala)*
Email:aisu.classmate@gmail.com[1], smitha.krishnendu@gmail.com[2]

Abstract: *Most of the existing methods focus on generating a static summary of the videos based on low level features such as histograms, histogram of oriented gradients (HOG), color histogram of gradients (CHOG) etc. This treats the video as static images thus losing the temporal aspect of the media. Video summarization system can yield good results if the high level features also called the semantic concepts in video frame are modeled accurately by considering the temporal aspects of the frames. The existing system is context aware surveillance video summarization which is a Domain dependent System. It works only on low level features and correlation between them is extracted and updated using dictionary algorithm in an online fashion. Thus dictionary size increases. In contrast to the existing method, the proposed system is a domain adaptive video summarization framework based on high level features in such a way that the summarized video can capture the key contents by assuring minimum number of frames. One of the high level features extracted is Local binary pattern (LBP).Key frames can be extracted after finding the Euclidean distance between the LBP descriptor in different methods. The result is compared with several datasets thus showing the effectiveness of the proposed system. The entire work can be simulated using matlab.*

Keywords - *Euclidean distance, feature extraction, LBP, Video summarization*

I. INTRODUCTION

An electronic medium used for recording, copying, broadcasting and displaying of moving visual media is called a Video. The main characteristics of video streams are: (a) number of frames per second (b) interlaced vs. progressive (c) aspect ratio (d) color space and bits per pixel (e) video quality. One of the particular case of signal processing is video processing which includes video filtering with input and output signals as video files or video streams. This technique is mainly used in television sets, VCRs, DVDs etc.

Thus huge volumes of data are distributed over the web. Large parts of most of the video portions are redundant or non-informative in nature. So watching for hours just to figure out the important or key features of a video is a time consuming process. It's also difficult for people to focus on videos for hours and not to miss the important events in the video. Thus we develop a tool to summarize the most informative video parts. This summarization technique is called Video Summarization.

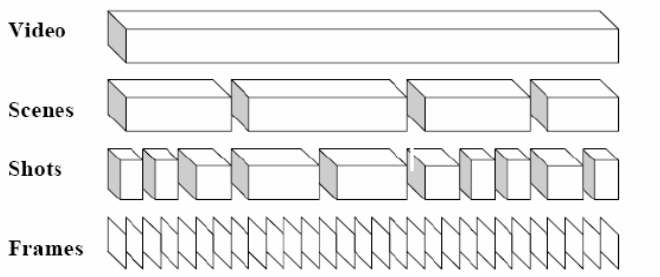


Fig 1: Anatomy of a video

Video Summarization is also called video abstract which gives a brief description about video content. This summary is created by extracting key features or important information of a video into its corresponding storyboard. Video summarization can be classified into two:

a) Static Video Summarization
It is also called static video storyboard, which involves a set of key frames extracted from the original video.
b) Dynamic Video Summarization
It is also called dynamic video skimming, which collects a set of shots by computing the similarity or relationship of each shot.

One of the main advantages of a video skim over a key frame set is its ability to include audio and motion elements which potentially enhance both the expressiveness and the amount of information conveyed by the summary. It is very much entertaining and interesting to watch a skim than a slide show of key frames. While key frame sets are not restricted by any timing issues, so they offer more flexibility in terms of organization for browsing and navigation purposes. Most of the educational and professional fields that are dealing with huge number of videos take a very good advantage of video summarization applications that include image videos, personal videos, sports videos, database management videos.

In this paper, it's proposed a simple approach for video summarization in a domain adaptive framework. This method is based on extraction of high level features from video frames and classified using any unsupervised learning techniques. In addition, a new methodology of evaluating video summarization by comparing with VSUMM dataset and user summaries is considered for comparison. Thus evaluation of VSUMM is performed on different videos e.g.: cartoons, news, sports, tv-shows) etc.

II. LITERATURE REVIEW

Some of the main approach related to video summarization is discussed here. Shu Zhang, Yingying Zhu, and Amit K. Roy-Chowdhury [1], proposed a method called Context-Aware Surveillance Video Summarization. There are two main algorithms used in this method. One is sparse group lasso optimization algorithm and other one is Online updates the dictionary of correlation algorithm. This method is mainly focused on summarizing surveillance videos. Features and correlation that exist among features of individual video frames are also considered. The methodology includes training as well as testing dataset. In training a dataset, the input video segments are first motion boundary detected using background subtraction method. Then this boundary detected video is feature extracted and correlations between these features are also extracted and are stored in the dictionary as feature correlated graph. In testing a video, the input video segments are motion boundary detected; low level features and the correlation between them are also extracted. This correlation is represented using a spatio-temporal graph using sparse group lasso algorithm. Now this graph is compared with correlation graph in the dictionary. The dictionary is updated in an online fashion if a new feature is found. Now reconstruct the extracted features from learned features. If reconstruction error is greater than threshold, then online updating of the dictionary is done using dictionary algorithm and output will be summarized output with new feature. If reconstruction error is less than threshold, then the output will be summarized output. One of the main problems of this method is that it is domain dependent system, and only low level features are extracted using the group lasso algorithm. Another problem is that dictionary size increase each and every time the dictionary is updated in an online fashion. So accuracy of the image is affected.

Zhuang et al. (1998) [2] proposed a method using unsupervised clustering for key frame extraction. Here, the video is segmented into shots and then a color histogram is calculated for every frame. The clustering algorithm uses a threshold that controls the clustering density. Before a new frame is classified, the similarity between the node and the centroid of the cluster is computed first. If this value is less than threshold, then this node is not close enough to be added into the cluster. The key frame selection is employed only to the clusters which are considered as key clusters. In that case, a representative frame is extracted from this cluster as the key frame. The key frame is selected as the frame which is closest, to the key cluster centroid for each key cluster. This proposed technique is efficient and no comparative evaluation is performed for validating such assertions.

Hanjalic and Zhang (1999) [3] proposed a method for producing a summary of an arbitrary video sequence which is based on cluster-validity analysis and is designed to work without any human supervision. This entire video material is first grouped into clusters. Then each frame is represented by color histograms in the YUV color space. Now, a partitional clustering is applied n times to all frames .Then the prespecified number of clusters starts at one and is increased by one each time the clustering is applied. Thus, the system automatically calculates the optimal combination of clusters by applying the cluster-validity analysis. After this optimal number of clusters is found, each cluster is represented by one characteristic frame, which becomes a new key frame. Hanjalic and Zhang (1999) concentrated on the evaluation of the proposed procedure for cluster-validity analysis, rather than on evaluating the produced summaries.

Gong and Liu (2000) [4] proposed a technique for video summarization based on Singular Value Decomposition (SVD). Firstly, a set of frames in the input video is selected. Then, color histograms in the RGB color space are used to represent video frames. Each frame is divided into 3 * 3 blocks, and a 3D-histogram is

created for each of the blocks to incorporate spatial information. Then, these nine histograms are concatenated together to form a feature vector. A feature-frame matrix A (usually sparse) is created for the video sequence using this feature vector extracted from the frames,. Then, SVD is performed on A to obtain the matrix V, in which each column vector represents one frame in the feature space. Then, the cluster closest to the origin of the feature space is found, and then the content value of this cluster is computed . This value is used as the threshold for clustering the remaining frames. Now from each cluster, the system selects the frame that is closest to the cluster center as key frame.

III. PROPOSED METHOD

Fig: shows the steps of our method that produces a domain adaptive static video summary. Initially, the original video is split into frames and then high level features are extracted, i.e. here Local binary pattern (LBP) features are calculated and are classified using an unsupervised method called k-means clustering. In most of the existing methods, domain dependent system is used for summarization (for e.g. only: surveillance videos, sports videos etc.).

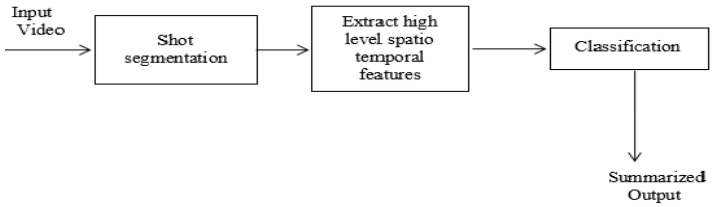


Fig 2: Block diagram for proposed method

VSUMM approach doesn't take all frames. So frame rate is calculated and corresponding frames are taken for feature extraction. Then these frames are classified to extract key-frame. The meaningless frames are removed from video sample. Then the frames are grouped using k-means clustering. By calculating Euclidean distance between frames of each cluster and within each cluster, one frame per cluster is selected. Thus this selected frame is called Key frame. Now similar key frames are eliminated to refine static video summary. Finally, remaining key frames are arranged in temporal order.

Steps involved are:

A. Shot segmentation

Temporal video segmentation is done in summarization approach. Here, the video stream is split into shots or frames of images. In VSUMM approach, not all the frames are considered for feature extraction. Hence, frame rate is calculated for every video. This method is also called pre-sampling approach where sampling rate is fixed on one frame per second. Frame rate is the number of frames extracted in a given duration of each video in seconds. For e.g.: normal frame rates are 24fps, 30fps, 60fps etc.

B. Feature Extraction

In the existing methods, low level spatio temporal features are extracted to detect motion regions and to detect multiple events. For e.g.: spatio-temporal interest point (STIP) detector, histogram of oriented gradients (HOG) and histogram of optical flow (HOF) features are extracted for detecting motion regions. Scale invariant feature transform (SIFT) features are also used for object detection. But this will not give an accurate result after extracting features. So a high level spatio temporal feature is extracted here.

Local binary pattern (LBP), is a visual descriptor used for classification in computer vision. It's mainly used for texture classification. When LBP is combined with Histogram of oriented gradients (HOG) descriptor, it improves the detection performance on datasets.

The LBP feature descriptor is calculated as:

a) First, convert the input colour image after pre-sampling to grayscale image because LBP works only on grayscale image.
b) Now, for each pixel in the grayscale image, select a neighbourhood pixel around the current pixel and then calculate LBP value for pixel using neighbourhood pixel.
c) To calculate LBP values, compare the central pixel value with the neighbouring pixel values. We can either start ordering in clockwise or anticlockwise direction and can start from any neighbouring pixel. But should maintain the same order throughout.
d) When considering 8 neighbouring pixel, perform 8 comparisons for each pixel. The results are stored in an 8- bit binary array.
e) Thus, if the current pixel value is greater than or equal to neighbouring pixel value, then the binary array is set to 1 for corresponding bit. Else if the pixel value is less than neighbouring pixel, then corresponding bit in binary array is set to 0.
f) After calculating LBP values, update the corresponding pixel location in the LBP mask using the LBP value calculate.

An example is shown below:

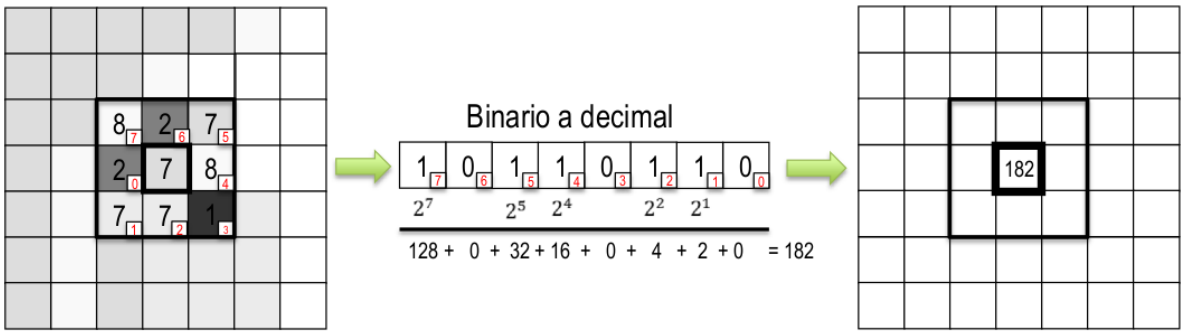


Fig 3: Calculation of LBP values

C. Clustering technique

Clustering is a method of grouping similar frames within a cluster or in between clusters. The different types of clustering algorithms are: a) Hierarchical clustering b) K-means clustering c) Hidden Markov model d) Gaussian mixture models etc. In the existing methods, a High density peak search (HDPS) clustering algorithm and a Video representation based high density peak search (VRHDPS) clustering algorithm is used for integrating some important properties of video.

In this paper, the most efficient clustering method when compared to existing method is K-means clustering algorithm. It is one of the simplest methods of unsupervised learning algorithm. In this work, k-means clustering is applied to frames extracted using LBP feature descriptor. Now, Euclidean distances between LBP features are calculated and then classified using k-means clustering algorithm. In order to find the centroid of each cluster, Euclidean distance between the clusters and within the clusters is calculated. Then for each key cluster, the frame that is closest to the centroid cluster which is measured by Euclidean distance is selected as key frame. The value of k is user defined. When the value of k increases, redundancy will occur. If it decreases, there will be loss of key frames. So threshold value (i.e. k value) is set according to this. But different videos have different k value. This is the main drawback of k-means clustering.

IV. EXPECTED RESULT

To show the effectiveness of video summarization, we perform the experiments on VSUMM dataset which includes VSUMM summaries and user summaries of videos. VSUMM dataset contains several videos with different events for e.g.: cartoons, news, sports, commercials, TV-shows and home videos. These videos differ in colour, length, motion and subject. Surveillance videos are also considered for summarization since they have lot of redundancy. The frames after k-means clustering are compared with VSUMM dataset and user summary. If these frames exist in the summary, then that frame is a key frame else redundant frame. Thus redundant or useless frames are removed. Experimental set up is done for more than 50 videos each with a duration varying from 1 to 10 min.

Fig 4 shows VSUMM summary and one of the user summaries of cartoon video. The extracted frames are compared. But result shows key frames are missing and no: of redundant frames increases when k value

increases. Say for k=10, there are missing frames as well as useless frames. So accuracy is affected. Thus future scope depends on getting accurate result with minimum redundancy and maximum key frames.

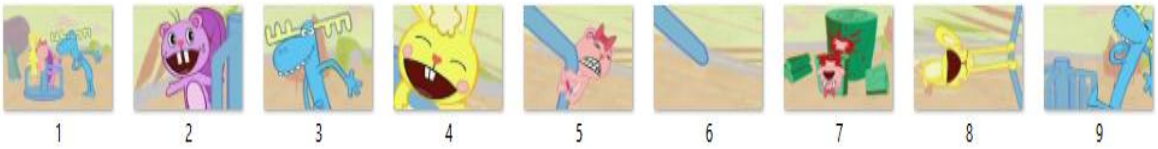
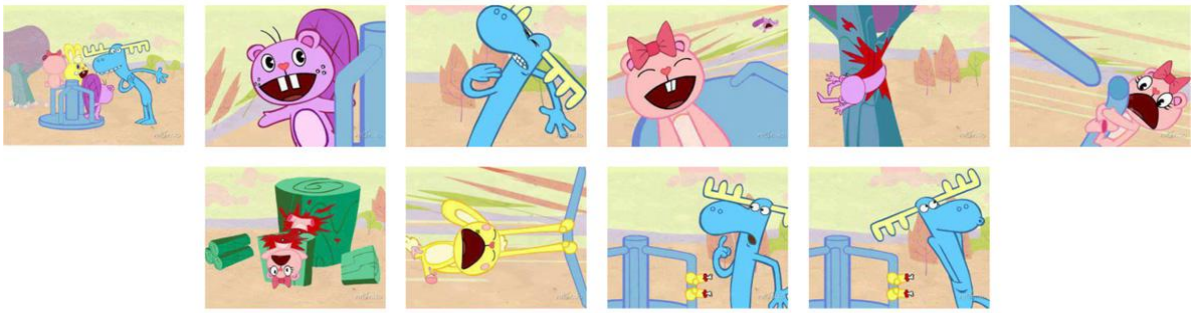


Fig 4: Result analysis of k means clustering for k=10

(a) VSUMM



(b) User #1

Fig 5: VSUMM summary and one user summary of a cartoon video

V. CONCLUSION

Video Summarization has attracted a fast growing attention from researchers and thus various algorithms and techniques have been proposed. In this work, a review on domain adaptive framework using LBP and k-means clustering is carried out. VSUMM dataset is used to produce static video summaries. The evaluation process is based on comparison between VSUMM dataset, user summary and extracted key frames of video. Thus, this technique produces video summaries of high visual quality and also can be used for summarization of different types of compressed videos. Video summarization produces more informative summary if semantic features are combined with visual descriptors. Future work is based on overcoming the drawback of k-means clustering technique and this can be extended using any other semantic feature as well as different clustering algorithm.

REFERENCES

[1] *Shu Zhang, Yingying Zhu, and Amit K. Roy-Chowdhury. Context-Aware Surveillance Video Summarization. IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 25, NO. 11, NOVEMBER 2016.*

[2] *Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S., 1998. Adaptive key frame extraction using unsupervised clustering. In: Proc. IEEE Internat. Conf. on Image Processing (ICIP), vol. 1, pp. 866–870.*

[3] *Hanjalic, A., Zhang, H., 1999. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Trans. Circuits Systems Video Technology 9 (8), 1280–1289*

[4] *Gong, Y., Liu, X., 2000. Video summarization using singular value decomposition. In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition (CVPR).*

[5] *IEEE Computer Society, Los Alamitos, CA, USA, pp. 2174–2180*