International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

A Review on Video Summarization Techniques

Trupti Deshbhakar[1], Simran Meshram[1], Nisha Wakodikar[1], Pranali Wanjari[1], Prof. A. P. Mohod[2]

[1]BE Student, Department of Computer Science of Engineering, Priyadarshini J.L College of Engineering, Nagpur, Maharashtra, India

[2]Assistant Professor, Department of Computer Science of Engineering, Priyadarshini J.L College of Engineering, Nagpur, Maharashtra, India

Article Info
Volume 8, Issue 3
Page Number: 158-165

Publication Issue :
May-June-2021

Article History
Accepted : 15 May 2021
Published: 23 May 2021

ABSTRACT

With the quick development of digital video technology, it is possible to upload large videos to YouTube or any other websites, record huge amount of data as news videos, sports videos, and lecture videos and surveillance videos etc. Storage, transfer and processing of video take considerably large amount of time. The user might not have adequate time to watch enter video before downloading or the user needs the search result of video to be quick and precise. In such cases the highlight or summary of the video makes search and indexing operations fast and user can view the highlight or summary of the video before downloading the video. Video summarization is short summary or highlights of the long video. This work is a detailed study on various video summarization techniques.

Keywords – Component, Static Video Summarization, Video Skimming, Convolutional Neural Networks

I. INTRODUCTION

Nowadays, we have so many electronic devices which are capable to record a huge amount of data like music, videos, sports, news and documents. Security cameras are being installed in all public places, government organizations, and private organizations. All these sources are producing huge amount of data. Storing these huge data is a difficult task. Upon that everyone can upload videos on the internet.

Processing image or video is a huge time consuming processes. The search result is expected to be fast, appropriate and accurate. Video summarization is a very useful technique in each situation. The video summarization was introduced in the 1990's. It is the short summary or the highlights of long video and it should obey the principles: it should contain only the high priority events of the input video, secondly the speed should not be manipulated that is it should contain the speed as the original video, thirdly the series of occurrence of the event should be as same the original one and lastly the summary video should not contain the redundant data.

Video summarization has three steps. In the initial step, video information analyzed to find out the salient factors, structure or methods within the visual, audio and textual component (audio and textual component if exists). The second step, select the meaningful frames which represent the content

of the video and finally output synthesis includes organizing the frames/shots according to which into the original video.

Video summarization is mainly classified into static video summarization or key frame video summarization and video skimming or dynamic video summarization. Static video summarization produces series of images of high priority event as output and video skimming produces short video as output. Comparing both video summarizations, the static video summarization gives an accurate summary but the video skimming summaries are easily understandable. A Static video summarization output consists of video frames only and audio frames are not considered. Dynamic video summarization contains video data contents, audio data contents and/or textual data contents.

Static video summarization, the key frames are extracted by uniformly skipping the frame or randomly selecting the frame. The key frame extraction is the fundamental processes. The size of the key frame [1] can be fixed or unknown. The fixed key frame size is called priori and unknown size is called posteriori. The priori assigns a particular number or proportion over length of the input video. The posteriori determine key frame size internally. Some approaches used Pre-sampling before key frame extraction to find the candidate frames. Finally, the duplicate frames are eliminated and ordered the frames according to the original video.

Video skimming is a short summary video constitute of the interesting scene from the input video is presented to be the user in the form of abstract of video story. Video skimming techniques include single value decomposition, motion model, and semantic analysis. Dynamic video summarization scheme for movies is derived on the progress of the stories. It consists of two steps. The two-dimensional histograms are utilized to distinguish the shots of the input video, the spatio-temporal correlation are applied to shots to extract the semantic scenario among detected shots. At last, some essential guidelines and procedures of motion picture generation are utilized to get a handle on the stream of the story.

Other classification [2] on video summarization techniques is based on features, clustering, trajectory analysis, shot selection and event based. The input video cannot be directly processed so the video is converted into frames/shots. The features like colour, motion, gesture, audio-visual and event- based approach are used to extract key frames. The features are picked out based on what user wants. It is difficult to derive videos summaries from motion only. In motion based

[2] [3], change in the position of the objects or person of perspective frames of the video and motion of camera action are considered as motion. In camera motion, the movement of the camera is filtered from the frames by considering all other frames motion value of the video. The motion based is good when the video contains medium-level motion and fails when video contains no motion or huge motion. The simple technique or algorithm to find the motion is frame difference. The frame difference greater then threshold is taken as the motion. The lesser threshold value result in many motiondetection and higher threshold value result in no motion detection. The other approach is optical flow, mainly applied in to videos. The motion is calculated from the series of images.

The colour based summarization [2] [3] is widely used. The most popular color representations are RGB and HSV. The colours are represented as red, blue and green in RGB and HSV as hue; the wave length of the color, saturation; white light count in the color and value; the intensity of the colour. In this method colour histograms is computed, it provides count of colour distribution of the frame/shot. Shot detection is a difficult task. The abrupt shot detection can easily calculated because it applied only to single frame. The gradual shot detection is applied to many frames and many algorithms fail to detect gradual change. The simple

shot detection method is comparing the histograms of consecutive frames. With the consideration that frames in the same shot have similar values. The histogram value above the threshold is considered as the shots. The problem with method is selecting the threshold value for the video. In static video summarization, the first, middle or last frame is selected as the key frame of a shot. In video skimming, set of frames are taken as key frame of the shot. After frame selection, the clustering algorithms are used to find the key frame from the selected key frames. The k-means clustering algorithm [2] uses Euclidean distance to find the clusters. The frame near the cluster center is taken as the key frame. The key frame count is equivalent to the clusters count. Finally, according to the flow of input video the key frames/ key frame sets are organized.

The online video lectures are summarized on the basics of gestures movements like hand, head and leg etc. Audio-visual based approach [3] is used to summaries the videos that contain audio as well as the video. The audio based video summarization is more often used then text only because unlike from image the audio requires only less space and computation cost is also less. Lecture and sports video consist of audio data that tells what we can see on the screen. These video should contain synchronization between audio and video. So the summary must contain the video segment that corresponds to some audios segment. The audio is sampled with particular frequency and converted in to frames. The key frames are selected from frames. Event based summarization

[4] is the summary of specific event on the input video. In sports video the goal, people applause etc are the events. Event detection is a two-step process; in the first step, energy and absolute value of pixel difference between the reference frame and current frame is calculated. In the final step, the frames which show some events are identified. Then according to the current frame, the reference frame is refreshed. The study of various methods and techniques of video summarization are carried out.

II. LITERATURE SURVEY

Ajmal et al. [2] provide detailed study of various techniques for video summarization. The techniques are mainly classified based on features, clustering, trajectory analysis, shot selection and event based. Feature based summarization are again classified into motion based, color based, gesture based, audio-visual based, object based video summarization etc. The video is converted into shots/frames at first because direct processing of video is impossible. It is difficult to derive videos summaries from motion only. In motion based, change in the position of the objects or person of perspective frames of the video is considered to find the motion. Colour based summarization are widely used. Colour histograms for frames are calculated and key frames are extracted from histogram value. The Gesture based key frame extraction is applicable to summarize video of lecture based on head, hand and leg movements. Clustering is used mostly when we find similar characterize within in frames. Spectral clustering algorithm is utilized to make dynamic summary of short summary videos. In spatio-temporal based video summarization the three dimensional space is used to represent the moving object, along with the spatial dimension of the object, the time is also considered

Srinivas et al. [5] proposed an approach for key frame based summarization. The process contain three steps 1) score the frames, 2) select the key frame and 3) eliminate duplicate frames. The features like quality, representativeness, uniformity, static and dynamic attention are used to assign score for the frames. The score is then normalized to the range of 0 to 1. In the next step the weights are assigned to the frames. The weights are arranged in the way where the higher scored frames get maximum weight. Then the final score is calculated as the dot product of weight and score. The frames are ranked in the descending order

so that the highest scored frame get 1st rank. Highest ranked frame is picked out as the key frame and distance between the key frame and selected frame is calculated. If that distance is greater, then it is elected as key frame. After key frame extraction, another algorithm is used to remove redundancy. The selected key frames are converted into gray scale images. The histograms are computed for those frames and applied Euclidean distance between the frame pair. The distance greater then threshold is chosen as thekey frame.

Song et al. [6] provides an approach of key frame video summarization algorithm for surveillance video. The existing key frame video summarizations are mostly based on shot boundary detection, in surveillance videos there are no explicit shot boundaries, so it is not applicable for surveillance video summarization. This method mainly focuses on the abnormal events of the surveillance video and has highest summarization ratio then the other key frame based summarization methods. Firstly, the pedestrians, and vehicles are extracted from the video. The abnormally event in the frames are calculated using trajectory algorithm and rain forest classifier based on the features like speed of the vehicle, area and direction. Finally abnormal video summarization is done using max-coverage algorithm, to get least number of frame in video summarization.

Deepika and Babu [7] proposed an approach on real time surveillance video with alarm facility. The captured live video is converted in to frames, the preprocessing techniques like brightness, contract etc. are applied on the frames. The background modeling is used to find the abnormal event. If any abnormal event detected a snapshot will taken automatically and alarm module executes alarm based on user settings. JPEG image compression is used to compress the image and stored inthe system.

Mrs and Mr Jadhav [8] proposed a method for efficient key frame extraction from redundant data. This approach consists of two stages, shot boundary detection and key frame extraction. In shot boundary detection the video is converted into m*n block. Then blocks are split into shots on the basis of image histogram, skew and kurtosis values. From each shots the frame with most extreme mean esteem, standard deviation esteem is chosen as key frame. The advantage of this approach is higher level feature image distribution is used to extract the key frames.

Kavitha and Rani [9] proposed approach for slow and fast moving videos. Both Discrete Wavelet Transform (DWT) techniques and static attention model are used to extract the key frames. The original video is divided into frames. Using sobel edge detection algorithm the shots are determined. The sobel edge algorithm is used to extracts shots of video. Using static attention and discrete wavelet transform techniques sets of key frames are extracted separately. The LMS color space is used to extract static features and statistical features are extracted from the wavelet domain. Then the final key frames are extracted by combining two key frames set and eliminatingirrelevant and redundant data.

Wu et al. [10] proposed video representation based high density peaks search (VRHDPS) clustering algorithm for static video summarization. The algorithm contains four steps. In the pre-sampling step, the candidate frames are extracted by using single value decomposition method and removed redundant and unwanted frames. In video representation step, BoW model is to extract the local features like SIFT of candidate frame and each frame are represented with histograms. In clustering step, VRHDPS clustering is used to cluster the candidate frames. The main advantage of this algorithm is the cluster number is not explicitly defined. This approach are considering isolated points they are less similar and far from the cluster center. The cluster center is taken as key frame. The disadvantage of this methods is some frames with visually different frame may have similar SIFT features. So they may group in same cluster.

Cai et al. [11] proposed a 3D CNN model to identify the abnormal behavior in the examination surveillance. This model can also be applied to the abnormal behavior they are not yet designed. The "optical flow" is calculated for the training and testing dataset using farneback's algorithm. The optical flow cannot process directly so it is converted into flow image. The flow image is given to the 3D CNN to build a model. The 3D CNN has the ability to learn the spatial and temporal behavior of the video clip. The main difference between 2D and 3D CNN is the spatial dimensions. The model has the ability to work with number of abnormal behaviors and performs better then motion blob, template matching and skin SVM algorithms.

Salehin and Paul [12] proposed a video summarization method for motion camera. The moving objects in the video are tracked based on the human eye movement. The foveal, parafoveal and perifoveal are the regions around gaze point of the human retina. The motion is identified based on the intensity difference in red, blue and green channel in foveal (identify smooth pursuit). If intensity value is higher than or equivalent to threshold value is considered as motion. The distance between gaze points (frames) are calculated after identifying smooth pursuit. If the distance value is zero then it is considered as no motion. Finally, the frames are sorted descending according to the distance. The defined quantity of frames is picked out as key frame from the top of the order.

Sun et al. [13] provides a semantic attribute based approach on deep convolution networks. These semantic attributes are automatically discovered from a joint image and text corpora. Standard corenlp toolkit is used to automatically discover the caption from the image. The input frames are passed to the trained CNN to extract the visual features of the image. The visual and semantic features are fused by vector concatenation. Bundling center clustering algorithm is used to cluster the frames. The some frames from the center of the cluster are chosen to represent the video shot. The number of the cluster is obtained by dynamic programming approach. The approach will not work well in title-based approach.

Gharbi et al. [14] proposed a static video summarization method with low computational cost. In this approach, initially from the input video the candidate frames are extracted to remove unwanted frames and to increase the process speed using windows rule. The feature vector of candidate set is extracted by using SURF detector and repeatability table is created to find out the similarity between the frames of candidate set. Graph modularity clustering is used to extract the key frame. This method offers resultant video frame with no redundant key frame.

Yao et al. [15] proposed Highlight-driven video summarization framework for first person video. Initially the original video is converted into segments of n frames. Utilizing deep convolutional neural system every video section is isolated into spatial and temporal streams separately. The spatial stream spoke to as the numerous frame appearances and the temporal stream is spoken to by temporal dynamics in a video. The yields of the two streams are consolidated to get the last highlight score of every video fragment. Highlight curve of every video are acquired from the feature score of every video fragment. The highlight scored segment is selected as "highlights" of the video. The video time-lapse summary and video skimming can be easily generated from this method.

Phong and Ribeiro [16] proposed deep learning based offline and online image recognition. The offline image recognition is implemented on convolution neural network upon keras toolkit and the online image by using javascript based convolutional neural network ConvNetJS. They also tired to decrease the error rate by increasing layers in CNN and adding dropout layer to the CNN

Zawbaa et al. [17] proposed machine learning based automatic video summarization technique for soccer videos. This methodology consists of six stages; pre-processing stage, shot process stage, ML-based logo

replay identification stage, ML-based score board recognition stage, excitement event detection stage, and lastly logo-based event detection and summarization stage respectively. In the preprocessing stage, the video is cut into shots utilizing color histogram. Shot-type classification and play/break classification are utilized to process the shots. The SVM classifier and NN classifier are used to detect the replay portion and score board. The K-means clustering algorithms are used to find excitement events. They evaluated the SVM with NN-based classifier; they concluded that SVM is good for soccer video summarization.

Bolanos et al. [18] provided a key frame video summarization based on convolutional neural network to extract visual highlights of the input video. The photostreams are decomposed into events with the help of an unsupervised clustering algorithm. Finally, from the each event the most relevant key frame is extracted. A definitive objective of work is to recover the memory of MCI patients.

Almeida et al. [19] proposed a static video summarization technique video summarization based online application (VISON) for compressed videos. In most of the video summarization techniques compressed videos are decoded and summary video is produced from that. The main drawbacks of those videos are it takes huge memory. This method consists of three stages; the features are extracted in the initial phase followed by grouping the frames of similar content and finally, the unwanted frame are extracted from the summary. The user can customize the time period and threshold value of the summary. The threshold and time period of the video summary depends on the quality of the video. If the time period increases the video summary quality decreases. The threshold depends on the number of key frames.

Li et al. [20] proposed sparse coding based shot boundary based approach for video summarization. The shot boundary algorithms are less sensitive to gradual changes of frames. Sparse coding built dictionary items from the video segments and the shot are extracted from the video using the dictionary items and similarity between the frames. From the extracted shots, one frame with high features is taken as the key frames and post-process is done to remove redundant frames.

Srinivas et al. [21] provides the history of the convolutional neural network. From the feature extraction of the traditional image classification, object classification methods and the beginning of the convolutional network to the recent development of the convolutional neural network. The main drawback of the object detection algorithm and traditional image classification algorithm are the features are explicitly defined. The accuracy of the result depends on the extracted feature. That is when wrong features are selected then the accuracy decreases. The multilayer approach leads to the beginning of deep learning. In the convolutional neural network the feature are not explicitly defined and the weights and bias of neural network are adjusted during training. This paper is detailed study of the convolution neural networks like AlexNet, RNN, multi layer model and hybrid CNN model. It also mentioned some of the current demerits of the CNN: More training data is required; CNN is give false result for artificial images, robust for small geometric changes etc.

Browne et al. [22] describes the architecture of convolutional neural network and some real world examples of convolutional neural networks in Robotics. Land mark detection and sewing pipe crack detection as examples. The sewing pipes crack detection using CNN contains 5 layers. They are input layer, three hidden layers and output layer respectively. The filter sizes 5×5 and the common activation function log-sigmoid are utilized in respective the layers.

TABLE 1: COMPARISON BETWEEN KEY FRAME-BASED SUMMARIZATION AND VIDEO SKIMMING

Key frame based summarization	**Video skimming**
Produce set of key frames as summary	Produce small video as summary
Contains no motion information	Contains motion information
Limited user viewing experience	High User viewing experience
No time bounds and synchronization needed	Time restricted and synchronization is necessary
Only video frames considered	Video, audio and text data are considered
Helps in video indexing and searching	Helps in video indexing and searching but performance is less then key frame-based summarization

In table 1 represents the difference between key frame- based summarization and video skimming. The main advantage of video skimming is it can have motion and audio data. That makes the summary more interesting and informative for users. Anyway these two types of video summarization can be transformed from one to another.

The users and particularly the decision makers in the field of investigation of crime can briefly view through these skimmed video version for further judgments and conclusions. Further the shortened video version are very useful in detecting the anomalies in the critical or danger areas that are observed in particularly the system of surveillance videos for investigation and forensic purpose. Multiple camera sequence tracking system can be developed based on video summarization is one of the promising applications again for the investigation process in the area of crime-detection.

III. CONCLUSION

In the recent development in video summarization, many techniques and algorithms have been proposed. This work is a review on video summarization techniques basically focusing on the static video summary and the video skimming respectively. The evaluation demonstrated that whether the static or dynamic forms are utilized, the proposed techniques and algorithm deliver video summaries of high visual quality. Some methodologies are applicable for real time videos of particular compression type videos. However, video summarization with convolution neural network identifies key frames accurately and efficiently with low processing. The main drawback of the convolution network is it needs largedata for training.

IV. REFERENCES

[1]. Truong, B. T., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. ACM transactions on multimedia computing, communications, and applications (TOMM), 3(1), 3.

[2]. Ajmal, M., Ashraf, M. H., Shakir, M., Abbas, Y., & Shah, F. A. (2012, September). Video summarization: techniques and classification. In International Conference on Computer Vision and Graphics (pp. 1-13). Springer, Berlin, Heidelberg.

[3]. Brezeale, D., & Cook, D. J. (2008). Automatic video classification: A survey of the literature. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(3), 416-430.

[4]. Money, A. G., & Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. Journal of Visual Communication and Image Representation, 19(2), 121-143.

[5]. Srinivas, M., Pai, M. M., & Pai, R. M. (2016). An Improved Algorithm for Video Summarization–A Rank Based Approach. Procedia Computer Science, 89, 812-819.

[6]. Song, X., Sun, L., Lei, J., Tao, D., Yuan, G., & Song, M. (2016). Event- based large scale surveillance video summarization. Neurocomputing, 187, 66-74.

[7]. Deepika, T., & Babu, D. P. S. (2007). Motion Detection In Real-Time Video Surveillance with Movement Frame Capture And Auto Record in International Journal of Innovative Research in Science. Engineering and Technology An ISO, 3297.

[8]. Jadhav, M. P. S., & Jadhav, D. S. (2015). Video Summarization Using Higher Order Color Moments (VSUHCM). Procedia Computer Science, 45, 275-281.

Cite this article as :

Trupti Deshbhakar, Simran Meshram, Nisha Wakodikar, Pranali Wanjari, Prof. A. P. Mohod, "A Review on Video Summarization Techniques", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 3, pp. 158-165, May-June 2021.
Journal URL : https://ijsrset.com/IJSRSET218331