International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Implementation of Methodology for Video Summarization

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

ABSTRACT

Article Info
Volume 8, Issue 3
Page Number : 388-395

Publication Issue
May-June-2021

Article History
Accepted : 15 May 2021
Published : 24 May 2021

Modern era, a massive amount of multimedia data is analysed, browsed, and retrieved, slowing down delivery and increasing computation costs. Video summarization is an aspect of building video and browsing that has been increased to process all video information in the shortest amount of time. This method allows users to browse large amounts of data quickly. It is the method of separating key frames and video skims to create a summarized or abstract view of an entire video in the shortest amount of time while also removing duplication or redundant features. Paper focus on different ways to achieve a sample video: static and dynamic, which are divided into two categories. With both the rapid advancement of digital video technology, it is now possible to upload large videos from Youtube or other websites, as well as record massive amounts of data such as news, sports, lecture, and surveillance videos, among other things. Video storage, transfer, and processing take a significant amount of time. The user may not have enough time to watch the video prior to actually downloading it, or the user requires a quick and precise video search result. In these kind of cases, the video's highlight or summary speeds up search and indexing operations, and the user can view the video's focus or summary before downloading it.

Keywords : static video summarization, dynamic video summarization, key frame, video skim, Static video summarization; video skimming; convolutional neural networks

I. INTRODUCTION

The compressed representation of a video sequence for the user, as well as allowing the user to browse and retrieve a large collection of video data quickly and easily, is now a major topic in video processing.

The popularity of internet videos, particularly Google videos, has greatly increased the availability of videos on the internet, necessitating an automatic process for generating a rationale underlying of moving or still video. This relates to video summaries, which give the user information about the video's contents in a short amount of time. For both the consumer and the production view, the need for automatic video summary generation is pressing [1].

The processing of an image or video takes a long time. The search result should be quick, relevant, and accurate. In every situation, video summarization is a very useful technique. In the 1990s, video summarization became popular. It is a brief description or highlights of a long video that should adhere to the following principles: this should only contain the input video's high priority events, besides which, the speed should not be manipulated, and it should contain the same speed as the original video, thirdly, the sequence of event occurrence must be the same as the original video, and finally, the summary video should not contain the redu.

There are three steps to video summarization. In the first step, video data is analysed to determine the most important factors, structures, or methods in the visual, audio, and textual components (audio and textual component if exists). The second stage is to choose meaningful frames that represent the video's content, and the final step is to output fusion, which involves organising the frames/shots into the original video.

Static video summarization, also known as key frame video summarization, and video skimming, also known as dynamic video summarization, are the two main types of video summarization. As an output, static video summarization produces a series of images of a high-priority event, while video skimming produces a short video. When comparing the two video summarizations, the static video summarization provides a precise summary, whereas the video skimming summaries are simple to comprehend. A static video summarization output only considers video frames and ignores audio frames. Video, audio, and/or textual data contents are all included in dynamic video summarization. The key frames are extracted from static video summarization by uniformly skipping or randomly selecting the frame. The key frame extraction process is the most basic. The key frame's size [2] can be either fixed or unknown. Priori refers to the size of a fixed key frame, while posteriori refers to the size of an unknown key frame. The priori assigns a specific number or proportion to the input video's length. Internally, the posteriori determines key frame size. To find the candidate frames, some approaches used pre-sampling before key frame extraction. Finally, duplicate frames are removed, and the frames are reordered to match the original video.

Other video summarization techniques [3] include features, clustering, trajectory analysis, shot selection, and event-based classification. Because the input video cannot be processed directly, it is converted into frames/shots. To extract key frames, features such as colour, motion, gesture, audio-visual, and event-based approaches are used. The features are chosen based on the user's preferences. It's difficult to create video summaries solely based on motion. Changes in the position of objects or people in perspective frames of the video, as well as motion of camera action, are considered motion in motion based [3] [4]. The camera movement is filtered from the frames in camera motion. Figure 1 depicts the process of extracting features from the original raw video and abstracting the video summary. The extracted features could be a series of still images (key frames) or video skims (moving images).

This paper focuses solely on the static and dynamic video summarization techniques that are used for feature extraction and abstraction in order to summarize video so that the user can easily understand and retrieve the video regardless of time.

Generalized working of video summarization can be shown in the figure 2:

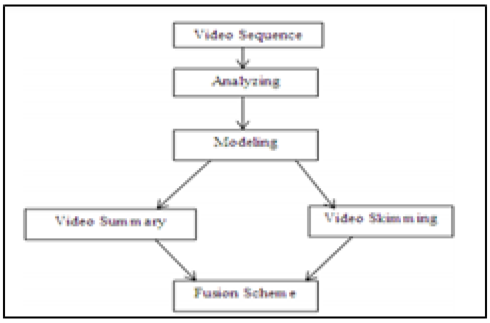

Figure 1 : Block Diagram for Video Summarization

The basic movements of the hand, head, and leg are summarised in the online video lectures. To summarise videos that contain both audio and video, an audio-visual approach [4] is used. Audio-based video summarization is more popular than text summarization because, unlike images, audio takes up less space and costs less to compute. Audio data tells us what we can see on the screen in lectures and sports videos. These videos should have audio and video synchronisation. As a result, the summary must include a video segment that corresponds to one or more audio segments. Audio is sampled at a specific frequency and then converted to frames.

II. LITERATURE REVIEW

An in-depth analysis of various video summarization techniques is given by Ajmal et al. [3]. The most popular classifications feature, clustering, trajectory analysis, shot selection, and event-based techniques. A motion-based, color-based, gesture-based, audio-based, and object-based video summary, as well as a variety of other features-based video summaries, are all examples of feature-based video summaries. In order to be able to apply direct video processing, the video must first be rendered into frames. Creating motion-based video summaries is difficult. Objects or individuals that appear in the different camera positions are used to assess motion in motion-based video. Colour-based summaries are often used.

Researchers Srinivas et al. [5] came up with a way to summarise main frames. The process consists of three stages. Score the frames. Then select the main frame. Finally, you'll want to get rid of any duplicate frames. Performance, representativeness, uniformity, static and dynamic attention are used to rate the frames. Once the score has been normalised to a range of 0 to 1, it will be usual for the following score. In the next step, the weights are allocated to the frames. So that the highest-scoring frames have the greatest amount of weight. The final score is calculated by using the dot product of weight and score. The ranking of the frames is from the highest to the lowest-scoring. The most senior fraternal order

A video surveillance approach with warning features, presented by Deepika and Babu [6], was suggested. The recorded live video is processed with techniques such as brightness, contract, and so on to create frames. Background modelling is used to identify when an unusual occurrence occurs. The alarm activates automatically in the event of an unexpected event. A compressed and processed JPEG image is used in the framework.

Kavitha and Rani [7] suggested an approach for dealing with slow and fast-moving images. Both DWT and static attention models are used to find the main frames. Video has had frames attached to it. The edge detection algorithm is used to decide where the shots are located. The Sobel edge algorithm is used to extract video shots. By separately extracting each set of key frames using static attention and discrete wavelet transform techniques, key frames are extracted. In the LMS space, static features are extracted, while in the wavelet domain, statistical features are extracted. Two sets of key frames are combined to remove obsolete and redundant information and only the final key frames are removed.

Wu et al. [8] proposed the video representation-based high density peaks search (VRHDPS) clustering algorithm as a method for static video summarization. The algorithm consists of four stages. Redundant and unused frames are removed in the pre-sampling phase. In the BoW model, SIFT features are extracted from each frame and are depicted as histograms. VRHDPS clustering is used to cluster the candidate frames in the clustering process. This algorithm has the advantage of the cluster number not being specifically defined. This approach takes into consideration instances where individuals are completely separated. Offline and online image recognition that incorporates deep learning was suggested by Phong and Ribeiro [9]. An offline convolutional neural network, driven by the keras toolkit, is used for offline image recognition, and a convolutional neural network, also called ConvNetJS, is used for online image recognition. Another method they tried was to increase the number of layers in CNN and use a dropout layer to lower the error rate. In a study led by Mrs. and Mr. Jadhav [10], a method for efficiently extracting key frames from redundant data was proposed. The two stages of this process are shot boundary detection and key frame extraction. During shot boundary detection, the video is transformed into m*n blocks. If the picture histogram, skew, and kurtosis values have been obtained, the blocks are classified into shots based on these values. The main frame for each shot is the frame with the most extreme mean and standard deviation value. One benefit of this approach is that main frames are derived from feature image distribution using a higher level feature image.

This motion camera video summarization approach was developed by Salehin and Paul [11]. In order to follow the moving objects in the video, the human eye movement is tracked. In the human retina, the foveal, parafoveal, and perifoveal regions are all situated around the gaze point. a red, blue, and green channel strength comparison is used to detect motion (identify smooth pursuit). Intensity greater than or equal to the threshold value is known as motion. Calculating the distance between gaze points is done by deciding smooth pursuit. If the distance value is zero, there is no motion. In the final step, the frames are sorted in ascending order of distance. According to Song et al. [12], a main frame video summarization algorithm can be used for surveillance video. Surveillance videos don't use key frame video summaries so there are no explicitly defined shot boundaries. This approach mainly concentrates on the outliers that appear in the surveillance footage, and has the highest ratio of the primary frame-based summarization approaches. As a starting point, the pedestrians and vehicles are extracted from the film. A trajectory algorithm and a rain forest classifier are employed to quantify the irregular events in the photos.

Yao et al. [13] created a system that utilises highlight-driven video summaries to comprehensively represent first-person video. Before any of the video segments are created, the original video is first cut into n-frame segments. A deep convolutional neural system is used to break down any video segment into spatial and temporal streams. Temporal dynamics (i.e. how long each event appears on screen) is what you use to describe the temporal stream, while spatial dynamics (how far each frame appears) is what you use to represent the spatial stream. When calculating the last highlight score for each video fragment, the combined yields of the two streams are taken into account. Every video highlight curve is calculated by evaluating each video fragment's function score. We just use the highlight scored section for the video's "highlights." video summarization using an online application named Almeida et al. [14] have developed a static video summarization technique for compressed videos known as video summarization based online application (VISON). The majority of video summarization techniques are capable of decompressing and processing compressed videos and then generate a video description from the results. There are two primary drawbacks associated with such videos: They use a significant amount of memory. This technique is separated into three stages: the first is extracting features, the second is groupings of similar material, and the third is finding frames that do not contain what was intended. You can customise the time and importance of the summary as you'd like. the quality of the video influences the quality of the video description A video summarization approach based on sparse coding and shot boundaries, as suggested by Li et al. [15], has been proposed. Due to frame changes that occur over time, the shot boundary algorithms are less influenced by them. Dictionary items for sparse coding were extracted from video segments, and the footage was selected based on the relevant dictionary items and related frames. Redundant frames are deleted from the final video through post-processing. The research paper "Srinivas et al. [16]" gives an introduction to the history of the convolutional neural network. The beginning of the convolutional neural network and other object classification methods came after conventional image classification methods. Although both object detection and traditional image classification have a major drawback in that the features must be clearly described, object detection has a number of significant advantages over traditional image classification. In order to determine the accuracy of the result, the extracted function must be considered. Choosing the incorrect features would result in less accurate results. This is the first step toward deep learning: A multilayer approach. Convolutional neural networks feature elements that are not clearly specified, and during training, the weights and bias of the neural network are modified. This paper takes a detailed look at the many types of convolutional neural networks, including AlexNet, recurrent neural network, multilayer model, and hybrid CNN model. In addition, it briefly outlined some of the news network's existing shortcomings: CNN creates inaccurate results when processing artificial images, as well as other problems. The researchers named the architecture of convolutional neural networks, as well as some real-world robot examples, in their study by Browne et al. [17]. Examples include, but are not limited to, land mark detection and sewing pipe crack detection. There are five layers to the CNN-based crack detection for sewing pipes. Input layer, three hidden layers, and output layer follow in that order. The 55-pixel-wide filter is used in each layer, while the log-sigmoid activation feature is used in all layers.

III. STATIC VIDEO SUMMARIZATION

This technique is also known as key frame based video summarization, still image abstract, or storyboard. The following are some of the criteria that come up for key frame based techniques:

1. Redundancy: key frames are chosen from frames with minor differences.

2. Clustering is difficult when there are numerous changes in content.

Figure 2 below depicts key frame selection.

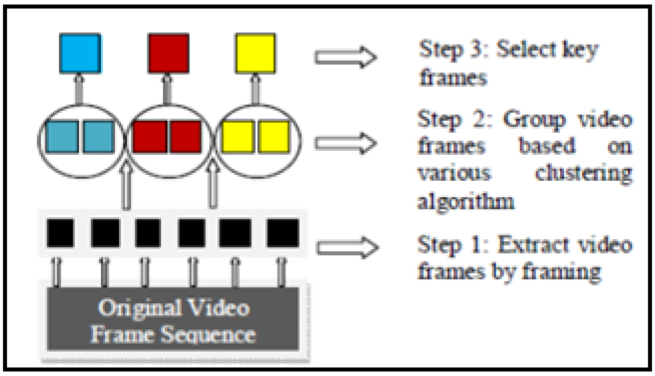
Step 3: Select key frames
Step 2: Group video frames based on various clustering algorithm
Step 1: Extract video frames by framing
Original Video Frame Sequence

Figure 2: Key frame selection[18]

There are three different types of key frame based summarization.

1. Sampling-based classification

It chooses key frames uniformly or at random, regardless of video content.

2. Scene segmentation-based classification

It uses scene detection to extract key frames and includes all semantic links in the video.

3. Shot segmentation-based classification

As a shot key frame, it extracts the first and last images.

Example: Euclidean distance clustering for video summarization.

The detailed explanation for the flowchart, shown in figure 4 is as follows:

1. Video acquisition and framing of a video Video is sampled a constant rate which is divided into set of frames. An ordered set of input digital video sequence V with cardinality N is defined as

V= {f1, f2, f3,……., fN}

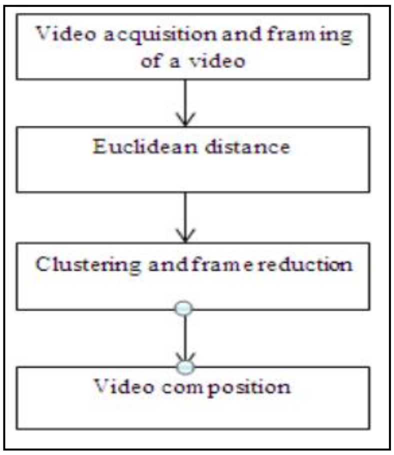
Video acquisition and framing of a video
Euclidean distance
Clustering and frame reduction
Video composition

Figure 3 block diagram of video summarization using Euclidean distance [19]

The detailed explanation for the flowchart, shown in figure 4 is as follows: 1. Video acquisition and framing of a video Video is sampled a constant rate which is divided into set of frames. An ordered set of input digital video sequence V with cardinality N is defined as V= {f1, f2, f3,……., fN}

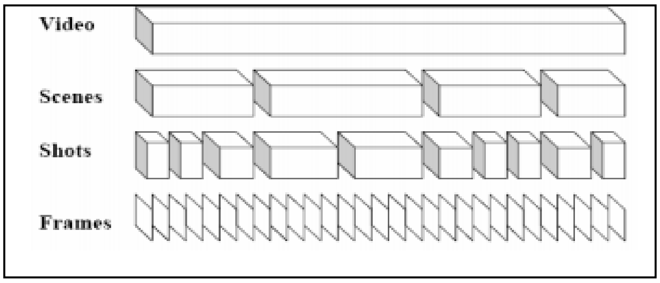

Figure 4: key frame extraction 2.

Euclidean distance calculation The resulting frame is analyzed to obtain the feature frame matrix. Every image has RBG associated with each of its pixel. The first is taken as a reference frame. The word count for each pixel value is taken in a vector form i.e. the vector represents a single frame and the whole video is represented in a set of vectors. The distance between this frame is calculated using Euclidean distance

$$E=\sqrt{\sum(Xj-Yj)^2} \quad \text{……….. equ. (i)}$$

Where x and y represents 2 different frames of an images and j represents columns. The Euclidean distance between two consecutive images is calculated and a threshold value is given, when a distance exceeds a given threshold, a key frame is claimed and that frame serves as a new reference frame. 3. Clustering and frame reduction The extracted features of the frame are clustered and are classified into different classes based on the distance calculated using Euclidean distance measure. Each individual class is classified under same frame name. 4. Video composition The individual frame that has been extracted is considered as a key frame which when combine together will compose a summary of a video.

IV. CONCLUSION

Advances in video summarization have resulted in many new ideas and methods. This paper focuses on static video summarization and video skimming. It was discovered that the proposed techniques and algorithm generate high-quality video summaries, regardless of whether they are being used to create a static image or one that is in motion. With real-time videos with specific compression styles, several methodologies can be used. Since this video summarization method uses a convolutional neural network, it only needs to process a few key frames to extract key concepts from the video.

V. REFERENCES

[1] Sachan Priyavada Ranjendra, Dr. Keshaveni N(2014),"A Survey of Automatic Video Summarization Techniques", International Journal of Electronics, Electrical and computational System , Vol 3,Issue 1,pp.1-5.

[2] 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.

[3] 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.

[4] 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.

[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] 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.

[7] Kavitha, J., & Rani, P. A. J. (2015). Static and Multiresolution Feature Extraction for Video Summarization. Procedia Computer Science, 47, 292-300.

[8] Wu, J., Zhong, S. H., Jiang, J., & Yang, Y. (2017). A novel clustering method for static video summarization. Multimedia Tools and Applications, 76(7), 9625-9641.

[9] Phong, N. H., & Ribeiro, B. (2017, June). Offline and online deep learning for image recognition. In Experiment@ International Conference (exp. at'17), 2017 4th (pp. 171-175). IEEE.

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

[11] Salehin, M. M., & Paul, M. (2017, July). A novel framework for video summarization based on smooth pursuit information from eye tracker data. In Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on (pp. 692-697). IEEE.

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

[13] Yao, T., Mei, T., & Rui, Y. (2016). Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 982- 990).

[14] Almeida, J., Leite, N. J., & Torres, R. D. S. (2012). Vison: Video summarization for online applications. Pattern Recognition Letters, 33(4), 397.

[15] Li, J., Yao, T., Ling, Q., & Mei, T. (2017). Detecting Shot Boundary with Sparse Coding for Video Summarization. Neurocomputing

[16] Srinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S., & Babu, R. V. (2016). A taxonomy of deep convolutional neural nets for computer vision. arXiv preprint arXiv:1601.06615.

[17] Browne, Ghidary, and Mayer, 2008, "Convolutional Neural Networks for Image Processing with Applications in Mobile Robotics", Studies in Computational Intelligence 83, 327–345.

[18] Tanuja Subba1 , Bijoyeta Roy2 , Ashis Pradhan3, "A Study On VIDEO SUMMARIZATION" , International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016, ISSN (Online) 2278-1021 ISSN (Print) 2319 5940

[19] Zaynab El khattabi, et.al., “Video Summarization: Techniques and Applications”, World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:4, 2015.

Cite this article as :

Trupti Deshbhakar, Simran Meshram, Nisha Wakodikar, Pranali Wanjari, Prof. A. P. Mohod, "Implementation of Methodology for Video Summarization", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 3, pp. 388-395, May-June 2021. Journal URL : https://ijsrst.com/IJSRST218383