Neural Computing and Applications (2023) 35:4881–4902
https://doi.org/10.1007/s00521-021-06132-1

S.I. : MACHINE LEARNING APPLICATIONS FOR SECURITY

Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework

Ghazaala Yasmin[1] · Sujit Chowdhury[2] · Janmenjoy Nayak[3] · Priyanka Das[4] · Asit Kumar Das[4]

Received: 4 January 2021 / Accepted: 15 May 2021 / Published online: 9 June 2021

Abstract
Video summarization is the process of refining the original video into a more concise form without losing valuable information. Both efficient storage and extraction of valuable information from a video are the challenging tasks in video analysis. Intelligent video surveillance system has an essential role for ensuring safety and security to the public. Recent intelligent technologies are extensively using the surveillance systems in all areas starting from border security application to street monitoring systems. Now the surveillance camera or motion sensitivity-based cameras produce large volume of data when employed for recording videos. As analysis of videos by humans demands immense manpower, automatic video summarization is an important and growing research topic. Hence, it is necessary to summarize the activities in the scene and eliminate unusual and redundant events recorded in videos. The proposed work has developed a video summarization framework using key moment-based frame selection and clustering of frames to identify only informative frames. The key moment is a simple yet effective characteristic for summarizing a long video shot and motion is the most salient feature in presenting actions or events in video which is used here to extract the key moments of the video frames. The motion is the scene of a video frame which has the most acceleration and deceleration in case of the key moments. Based on the extracted key moments, the frames of the video are partitioned into different groups using a novel similarity-based agglomerative clustering algorithm. The algorithm determines at most K clusters of frames based on Jaccard similarity among the clusters, where K is the user defined parameter set as the 5% to 15% of the size of the video to be summarized. From each cluster, few representative frames are identified based on the centroids of the clusters and arranged according to their original video sequence to generate the summary of the video. The proposed clustering algorithm and the summarization method are evaluated using state-of-the-art video datasets and compared with some related methodologies to demonstrate their effectiveness.

Keywords Motion-based key moments · Key frame extraction · Hierarchical clustering · Cluster Validation Indices · Video summarization

1 Introduction

Video summarization [37] [33] [6] is one of the essential processes in multimedia application. It is the process that deals with the extraction of few frames from each scene in order to create a summary video which explains all course of the entire clip of video within a short duration of time. Recent research is very much focused on big data analysis and summarization including text, image, animation and video. In this advancement in digital technology, video surveillance keeps a very important role for ensuring safety and security. The surveillance systems are being deployed to have a broad range of applications, specially invigilation and forensic purpose, in order to analyze activities in the environment. In security critical regions of the organizations, CCTV (Closed-circuit television) cameras are installed and surveillance videos are being recorded, which provide a massive amount of video data. Cameras are being integrated into recently developed devices, like drones and robots equipping them to record the videos at many places that are impossible to reach by humans. Analyzing these huge volume of video data by human being takes enormous amount of time. Instead of involving tedious human labors, an intelligent video summarization system [25] can be built

Extended author information available on the last page of the article

to provide a summary of the long surveillance videos. A video summary is easy to interpret the actual set of circumstances with respect to safety issues by security personnel, especially when multiple cameras are used to record surveillance videos of a single place. We may generate a summary of surveillance videos which includes specific activities, such as thefts in malls, accidents on roads, abnormal behavior of people in election or examination centers, specific movements like seizures of patients in Intensive Care Units (ICU) of hospitals, etc. These issues can be solved by introducing video summarization. Since large data processing needs more resources, video summarization enables users to manage and browse massive videos effectively and efficiently. Generally, video summarization [19] is a summary representing an abstract view of the original video sequence that can be used either as a video browser or as a retrieval system. Another description of summarization [26] is the highlight of the original video sequence, which is the concatenation of a user-defined and selected video segments and collection of key frames. A video key frame is the frame represents salient content of a video shot that provides a suitable abstraction of video indexing, browsing, and retrieval [2]. Considering the numerous applications of video processing, video summarization is one of the most important topics in the area of multimedia technology advances that aims creating summary of video to quick browsing of a collection of large volume of video dataset and also useful for video indexing and surveillance. A video key moment is the feature that can represent salient content of video shot. Key moments provide a suitable abstraction for video as it represents the important actions and events that occur in video shot. This approach can help a user to get a good overview of overall events occurred in the whole video. In this paper, we focus on extracting motion-based key moments within a segmented time frame of the video shot.

1.1 Preliminary concepts

Before discussing in details about the video summarization methodology, some concepts relevant to the proposed method are briefly described in this section. Figure 1 gives the hierarchical structure in a video sequence by which key frames of the video are extracted.

1. *Frame detection:* The video sequence to be summarized is generally segmented into a sequence of shots with the help of color features, which are extracted using the color histogram computed from HSV (Hue, Saturation and Value). Next, the Principal Component Analysis (PCA) is applied on the one dimensional vector representing the frequency of the color histogram to reduce the dimension of the feature vector. Detecting change of shots automatically is a different problem due to the variety of transitions that can be used between shots. A cut is an abrupt transition between two consecutive shots that occurs between two adjacent frames. A fade is a gradual change in brightness, either starting or ending with a black frame. A dissolve is similar to a fade except that it occurs between two shots. The images of the first shot get dimmer and those of the second shot get brighter until the second shot replaces the first one. Other type of shot transitions includes wipes and computer generated effects such as morphing [36]. Each shot is the collection of frames, which are further consisted of several macroblocks. The motion vector of each macroblock provides the motion information of the block, based on which the frames in the shots may be determined and separated from each other. Finally, the frames are clustered and representative frames are identified. Various popular clustering algorithms are proposed by different researchers.
2. *Clustering of frames:* Video summarization can be broadly classified into two categories, static and dynamic video summarization [37]. Static video summarization consists of keyframes which take into account the visual information without considering the audio message in the video. On the other hand, dynamic video summarization is a video clip which combines image, audio and text information together. In the paper, we have devised a static video summarization technique based on the proposed hierarchical clustering algorithm. The main objective of any clustering algorithm-based video summarization is to reduce the redundancy of video by selecting only the informative frames. There are many clustering algorithms [29] [40] used for this purpose. The main demerit of the existing clustering algorithm-based approaches is that they generally needed predefined number of clusters. The proposed hierarchical

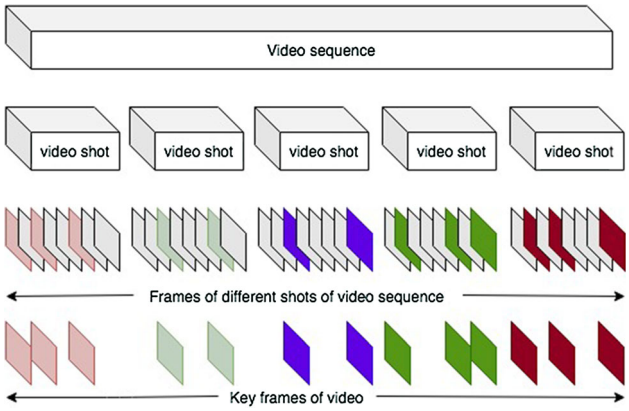

Fig. 1 General steps for key frame extraction-based video summarization

clustering algorithm is an agglomerative approach which does not require to know the number of clusters in advance. As in general, the size of the summary is user input, so we have given the provision of number of clusters in the proposed method, but we can ignore it and only based on the parameter such as overlapping cluster index, the method may be terminated. The quality of clusters obtained by different algorithms is determined by various cluster validation indices [13] [32] [5]. Cluster validation is a process of determining the set of clusters that best fits natural clusters without any class information. There are two types of commonly used cluster validation indices based on internal and external criteria, known as internal cluster validation indices and external cluster validation indices. Internal cluster validation is performed based on the information intrinsic to the data alone, whereas external cluster validation is done based on previous knowledge about data, i.e., the class label of the data is known a priori. These classical cluster validation indices quantify how similar two disjoint clusters are. However, in practical applications, it is quiet natural that an object may have in more than one cluster, such clusters are known as overlapping clusters. Our proposed agglomerative clustering algorithm merges two clusters in each iteration, where initially generated clusters are overlapped in nature. So, we have used the overlapping index defined in [5] as the terminating criteria of our proposed clustering algorithm. It is a new index based on an intuitive probabilistic approach that is applicable to overlapped clusters. After termination of our proposed algorithm based on this overlapping index, we determine non-overlapping clusters based on the fuzzy belongingness of the objects into the clusters, i.e., the frame closest to a cluster centroid is placed into that cluster. The method is compared with some state-of-the-art clustering algorithms [29] [40] with respect to different cluster validation indices. In the literature, there are many internal [27], external [11] [22] [12], and stability-based [3] cluster validity indices. As the proposed clustering algorithm is purely unsupervised in nature and no ground truth class labels are there for the extracted frames, so the method is compared with various related state-of-the-art research works based on widely used internal [27] and stability-based [3] cluster validation indices. The computation of cluster overlapping index defined in [5] is briefly described in next few lines, which have been applied as one of the terminating condition in our algorithm. Let, $U = \{O_1, O_2, \cdots, O_n\}$ is a dataset with n objects, which are partitioned into k number of overlapped clusters, $C = \{C_1, C_2, \cdots, C_k\}$. The probability that any two objects, O_x and O_y, both belong to cluster, say C_i is given by equation (1), where numerator represents number of pairs that can be found in C_i.

$$Prob((O_x, O_y) \in C_i \forall O_x, O_y \in U) = \frac{\binom{|C_i|}{2}}{\binom{n}{2}} \tag{1}$$

So, the probability that any two objects, both belong to any cluster C_i in C is given by equation (2), where the numerator accumulates all the pairs of objects found in each cluster and the denominator is used to normalize the value P.

$$P = \frac{\sum_{i=1}^{k} \binom{|C_i|}{2}}{k\binom{n}{2}} \tag{2}$$

Similarly, if $C' = \{C'_1, C'_2, \cdots, C'_l\}$ be the another set of overlapped clusters, then we can find out the probability of occurrence of any pair of objects in any cluster C'_j in C' using equation (3).

$$P' = \frac{\sum_{j=1}^{l} \binom{|C'_j|}{2}}{l\binom{n}{2}} \tag{3}$$

Therefore, the probability that any pair of objects, O_x and O_y belong to both the clusters C_i of C and C'_j in C' is given by equation (4), where numerator represents number of pairs that can be found in both the clusters C_i in C and C'_j in C'.

$$Prob((O_x, O_y) \in C_i \cap C'_j \forall O_x, O_y \in U) = \frac{\binom{|C_i \cap C'_j|}{2}}{\binom{n}{2}} \tag{4}$$

If the same analysis is done for every possible pair of objects in U considering every pair of clusters C_i of C and C'_j in C', then the probability that any pair of objects lie in pair of clusters in C and C' is given by equation (5).

$$t = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} \binom{|C_i \cap C'_j|}{2}}{\binom{n}{2} \frac{max\{|C_i|,|C'_j|\}}{n} min\{k, l\}} \tag{5}$$

Keeping all these concepts in mind, the overlapped cluster index (OC) is defined as the ratio of the

probability of finding pair of objects in both the clusters of two different cluster sets to the maximum probability of finding them in one of the given clusters, as defined in equation (6).

$$OC = \frac{t}{max\{p, p'\}} \tag{6}$$

In the proposed clustering algorithm, initially overlapping among the clusters is very high, and we merge the pair of highly overlapped clusters in each iteration. Next, we compute OC index using two set of clusters, one set obtained before merging and the other set obtained after merging. If the OC index is less than a predefined threshold then the process is terminated; otherwise the same process of merging is repeated.

1.2 Literature survey

The objective of automatic video summarization is to develop an automatic video monitoring system which provides us a concise summary of what has been captured by a surveillance camera over a long period of time. In some cases, when the scene is usually quiet, a simple motion detection would suffice, but in most of the cases, such simple approach produces a too long summary with many redundant video frames. A video summarization is an important machine learning technique which provides an abstract view of original video sequence and can be used as video browsing and retrieval system. It can be a highlight of original sequence which is the concatenation of a user defined number of selected video segments or can be a collection of key frames. Many of the research [14] [18] [35] have been introduced based on key frame selection. Early research work on key frame selection can be classified based on two different concepts, namely color change [23] and motion activity [25] of the frames. The color change-based approach [23] counts the number of pixels changes between two consecutive frames, whereas in the motion activity-based approach [25], motion vectors of the frames are taken care of. The main objective of both the approaches is to select the frames of the video, which provides the most representative visual content of the shot perceived by the viewer. In fact, these two approaches are inter-related in the sense that larger amount of color change implies a higher motion activity across the frames and vice versa. Zhang et al. [41] computed the difference of color histograms of current frame and previous extracted frame to determine the key frames of a video. The method may lose some global information as only the previous frame is considered to determine whether the current frame is a key frame or not. Gunsel and Tekalp [20] used N (user defined parameter) previous frames and compared the histogram of current frame with that of average of these N frames. Liu et al. [26] proposed the Perceived Motion Energy to determine the number and location of the key frames. They proposed a triangle model of perceived motion energy to determine the motion patterns within the video for extracting the key frames. They have selected the frames containing the turning point of the motion acceleration and deceleration as the key frames. The main advantage of this method is that it is threshold free and computationally efficient. Divakaran et al. [15] used the MPEG-7 motion activity descriptor and fidelity measure for video summarization. They described video summarization techniques based on sampling in the cumulative motion activity space, and then combined the motion activity-based techniques with generalized sound recognition. The method is computationally simple and flexible and generates a summary of any desired length. Wolf et al. [39] used optical flow computations to find out the local minima of motion in a video shot. These motion-based approaches used only the motion vector as the metric of motion activity without considering the macroblock (MB) characteristic of a fast-moving region. The MB type information has been used in Pei and Chou [31] to determine the scene change. The paper [7] proposed a key frame selection approach by the combination of both the MB type characteristic and motion vector activity, which gives more comprehensive measurement of visual content change within a shot. Recently, automatic video processing systems have taken a huge attention of researchers and come up to a great deal of improvements in terms of its performance, scope and acceptance on worldwide basis. Over a long time, triangle model of perceived motion energy (PME) [26] pattern is modeled for video summarization, where the frames at the turning point of the motion acceleration and motion deceleration are selected as key frames. The key-frame selection process is threshold free and fast and the extracted key frames are the representatives of the video. In visual frame descriptors algorithm [14], three visual features, namely color histogram, wavelet statistics and edge direction histogram, are used for selection of key frames. Similarity measures are computed for each descriptor and combined to form a frame difference measure. The key frames are selected by constructing the cumulative graph for the frame difference values. The frames at the sharp slope indicate the significant visual change; hence they are selected and included in the final summary. Fidelity, Shot Reconstruction Degree, Compression Ratio qualities are used to evaluate the video summarization. Jiri et al. [16] proposed a self-attention-based recurrent network to perform the entire sequence to sequence transformation in a single feed forward pass and single backward pass for video summarization. The method is evaluated using two benchmark datasets, namely TvSum and SumMe,

commonly used in this domain to demonstrate its utility. The paper described in [25] proposed a video summarization technique for lane surveillance system. It is a motion focusing method that focused on one constant-speed motion and aligned the video frames by fixing this focused motion into a static situation. A video summary is generated containing all moving objects and embedded with spatial and motional information together with key frames to provide details corresponding to the regions of interest in the summary. Background subtraction and min cut are mainly used in motion focusing. In Camera Motion and Object Motion [30], the video is segmented using camera motion-based classes, such as pan, zoom in, zoom out and fixed. Final key frame selections from each of these segments are extracted based on confidence value formulated for the zoom, pan and steady segments. Zhou et. al [42] have represented the concept of reinforcement learning for video summarization. They formulated video summarization as a sequential decision-making process that develops a deep summarization network. This network initially predicts a probability of how likely a frame of the video is selected in the summary and subsequently, based on the probability distributions, action is taken to select frames for forming video summary. This network is trained by an end-to-end reinforcement learning-based framework by defining a reward function to maintain diversity among the frames in the summary. The concept of generic framework of video summarization is introduced in [28] based on the user attention model which includes both key-frame extraction and video skimming. This framework takes the advantages of computational attention models and eliminates the needs of complex heuristic rules in video summarization. It integrates both visual, audio, and linguistic attentions to generate a user attention curve for a given video sequence. The paper described in [34] extracts key frames based on the rank of the features, such as colorfulness, brightness, contrast, hue count, edge distribution, for summarizing and indexing. The rank of a feature is set by assigning a weight to it based on the standard deviation that allows the feature with maximum variation across the frames as a higher weight feature.

1.3 Motivation

The most challenging task of video summarization is to determine the informative content of the video recording. The important content is generally described by low level features, like texture [1], shape [10] or motion [4]. The texture and shape features are more robust to the noise in the video shots than the motion features. Generally, many video sequences are the combination of slow sequence, fast sequence or an action sequence. The activity feature captures more accurately these type of intuitive notions of intensity of scenes in a video sequence. Video sequence generally spans the whole range of recording from high to low activity, and so a suitable descriptor is essential to accurately demonstrate the activities of different shots of the video. The compressed MPEG-7 video file provides the motion activity descriptor for this purpose. The motion is the more salient and robust to noise feature in describing actions in video, which is popularly used to determine the key moments of the video shot for finding necessary informative frames. The motion features are important in terms of their strong information content and stability over spatio-temporal visual changes. Motion features, like interest points and optical flow, are popularly used for modeling temporal video segments. This motivates us to extract the key moments of different frames for video summarization based on the motion feature. Thus, the frames are described by their motion-based key moments and clustered into different groups and representative frames from each cluster are selected based on their distances from the cluster centroid to generate the summary of the video.

1.4 Contribution

Selection of key frames from a video recording is an important task for video summarization. The recent trend of storing and watching video in portable devices demands for reducing a high volume video to a very small sized video, for which an effective and efficient video summarization algorithm is required. Thus, the main objective of this paper is to propose an effective key frame selection algorithm for reducing the storage size of a video sufficiently without losing the meaningful flow of the original video. The propose algorithm first tries to captured the frames of the video that carry certain information based on the concept of key moments of the frames and then a novel clustering algorithm is applied on these selected frames to determine the key frames of the video, whose order collection gives us the resultant summary of the original video. As there are numerous types of video (such as movies, sports, news, lecture or some speech) present in the real life scenario, the change in event has the prime role to make the video more informative. The safety way to properly differentiate different frames and to select the informative or key frames is based on their key moments. As key moment extraction works very well on all types of videos, so the proposed work first extracts the key moments from the video shot for finding necessary informative frames. The key moment extraction is performed based on the concept of motion vectors of the macroblocks of the frames. The motion is the more salient feature in presenting actions or events in video, which is used to determine the key moments. The motion vector of video often called

MVF (motion vector field) is extracted from the MPEG video using FFMPEG tool and a motion pattern has been built up to compute the key moments. Based on the key moments, the informative frames are extracted. As all the informative frames are not equally important and one may be implied by the other, so informative but redundant frames need not be considered in the video summary. The proposed work thus devises an agglomerative clustering algorithm to partition the frames and selects some representative frames from each partition as the key frames. These key frames are sufficient for generation of the video summary. The algorithm initially considers each frame as a separate cluster, computes Jaccard coefficient-based similarity between every pair of clusters and merges two most similar clusters in each iteration. The process is repeated until the terminating criteria are satisfied. The key contributions of the paper are described as follows:

1. The key moment extraction is performed based on the concept of motion vectors of the macroblocks of the frames. Based on the key moments, the irrelevant frames are removed from the video file.
2. An agglomerative overlapping clustering algorithm is proposed to cluster the frames considering overlapping cluster index as the terminating criteria of the algorithm. Each frame within the overlapping region is compared with the centroids of the overlapping clusters and placed into the cluster to whose centroid it is more similar.
3. The representative frames from each cluster are determined and arranged them in their original order of appearance in the video file, which is considered as the summary of the video file.

The workflow of the proposed video summarization algorithm is described in Fig. 2.

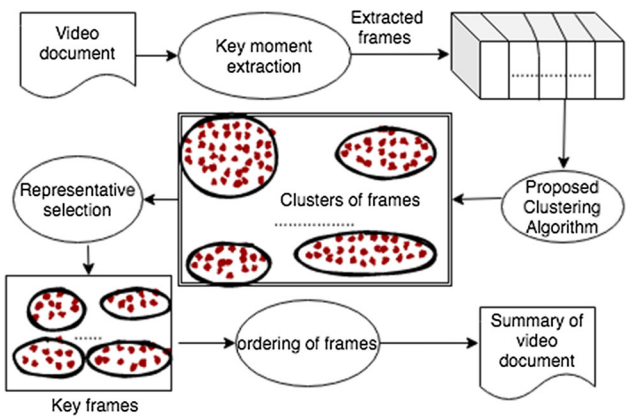

Fig. 2 Workflow of the proposed video summarization algorithm

1.5 Summary

The rest of the paper is organized as follows: Sect. 2 describes the process of extraction of informative frames based on key moments. A novel agglomerative clustering algorithm, proposed for selecting only the necessary and relevant key frames of the video is described in Sect. 3. The experimental results for evaluating both the proposed clustering and summarization methods are discussed in Sect. 4 and finally, the conclusion and future scope of this paper are drawn in Sect. 5.

2 Key moment extraction

There are many different techniques to extract visual, textual, and even audio information for finding the most useful features to understand elements within images. But these are not useful for videos as they are not skimmable like text. Key moments within videos find the important section of the content in faster way. The objective of this section is to determine efficiently the key moments from a video recording, which is a prior step of representing the video in a summarized form. The steps of finding key moments from a video for selecting relevant frames are described in a workflow diagram, as shown in Fig. 3.

Original video file is a recording consists of a collection of scenes where each scene consists of a set of frames. The term frames per second (fps) describes the number of frames of the video recorded in a second. For example, for one minute video recording with 10 fps, we have 600 frames in the whole video. Motion JPEG, popularly known as MJPEG is a video compression format in which each video frame or interlaced field of a digital video sequence is compressed separately as a JPEG image. Each video frame is a collection of pixels, where pixels hold the color intensity of real objects which is visualized in digital media. There are three different frame types of a video used by different video algorithms. These are $I-$frames, $P-$frames and $B-$frames. $I-$frames, also known as Intra coded frames, contain an entire image and are coded without reference to any other frames except themselves. It typically requires more bits to encode than other frame types. Generally, *I*frames are used for random access and are used as references for the decoding of other type of frames. These type of frames are the least compressible but don't require other video frames to decode. *P*frame, also known as Predicted frame, holds only the changes in the image from the previous frame. For example, in a scene of moving a car across a stationary background, only the car's movements need to be encoded. The encoder does not require to store the unchanged background pixels in the

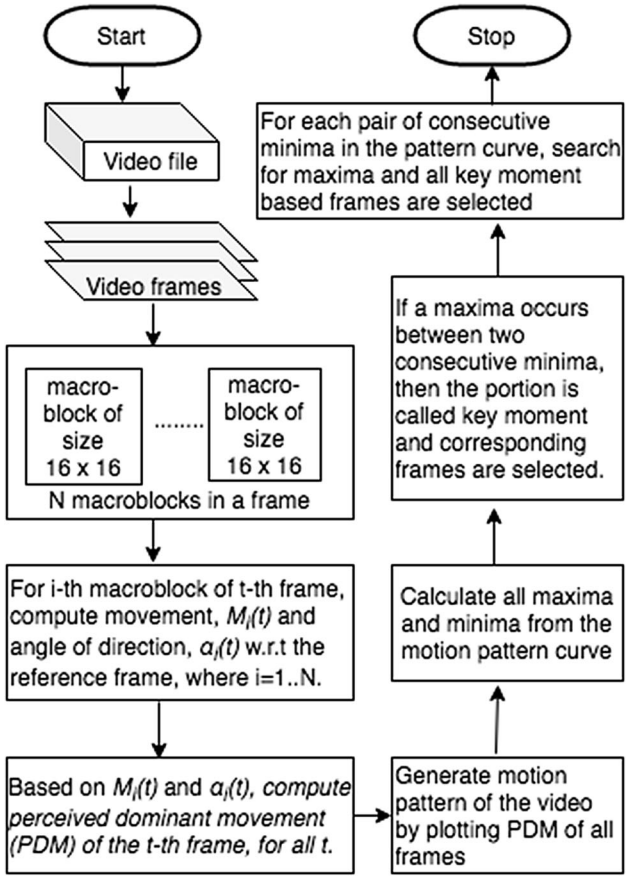

Fig. 3 Work flow of the proposed video summarization framework

*P*frame. Thus, this type of frame saves space. *P*frames can use data from previous frames to decompress and are more compressible than *I*frames. It requires the prior decoding of some other frame(s) in order to be decoded. It may contain both image data and motion vector displacements and combinations of the two. It may refer the previous frames in decoding order. *B*frame, also known as Bidirectional predicted frame, provides the highest amount of data compression by considering both the preceding and following frames together with the current frame to specify its content. It requires the prior decoding of subsequent frame(s) to be displayed and it may contain image data and/or motion vector displacements. Typically, frames are segmented into macroblocks, and individual prediction types are selected on a macroblock basis rather than being the same for the entire frame. For example, $I-$frames contain only intra macroblocks, $P-$frames contain both intra macroblocks and predicted macroblocks and B-frames contain intra, predicted, and bi-predicted macroblocks. In the proposed method, we have used the concept of $P-$frames, as it reduced both the space and the time complexity of the algorithm. We have considered the size of the macroblocks as 16 $\times$ 16. The macroblocks are considered as the processing unit in the frame and the video compression is done based on linear block transform, such as Discrete Cosine Transform.

In our present work, we have measured the visual content complexity of a frame with the help of motion patterns. The motion patterns of a frame are generally composed of a sequence of motion acceleration and motion deceleration. Such motion patterns usually reflect the actions in the events. Thus, the frequency of occurrence of motion patterns is a good indicator of visual content complexity of a frame. The motion patterns of a frame is called Motion vector Field (MVF) that carries the motion information about the macroblocks present in the frame (i.e., the current frame for which we are calculating the motion information). The motion of macroblocks in the source frame is calculated with respect to the reference frame. Motion vector is the more salient feature for presenting actions or events in video, which is basically used to determine the frames that are informative to the viewer. The motion vector is the amount of distance shifted by the macroblocks of the source frame from the reference frame. Let each frame consists of N number of macroblocks, each of dimension, say 16×16. Let, t^{th} frame is the current (or source) frame and previous frame (i.e., $(t-1)^{th}$ frame) is the reference frame. Also let, positions (i.e., top left corner) of i^{th} macroblocks of reference frame and source frame are $(R_{x_i}(t-1), R_{y_i}(t-1))$ and $(S_{x_i}(t), S_{y_i}(t))$, respectively. Then the movement of i^{th} macroblock $(M_i(t))$ of t^{th} frame with respect to its previous frame is obtained using equation (7).

$$M_i(t) = \sqrt{(R_{x_i}(t-1) - S_{x_i}(t))^2 + (R_{y_i}(t-1) - S_{y_i}(t))^2} \tag{7}$$

Thus the average movement $(AM(t))$ of t^{th} frame with respect to $(t-1)^{th}$ frame is given by equation (8).

$$AM(t) = \frac{\sum_{i=1}^{N} M_i(t)}{N} \tag{8}$$

Similarly, using the position of the i^{th} macroblocks of reference and source frames, we calculate the angle of direction $\alpha_i(t)$ of i^{th} macroblock of t^{th} frame, as defined in equation (9).

$$\alpha_i(t) = tan^{-1} \frac{(R_{y_i}(t-1) - S_{y_i}(t))}{(R_{x_i}(t-1) - S_{x_i}(t))} \tag{9}$$

In our proposed methodology, we have divided 360 degree angle of direction of a frame into 8 bins, each of 45 degree. So the bins are B_0, B_{45}, B_{90}, B_{135}, B_{180}, B_{225}, B_{270}, and B_{315} presenting the angular movements, $0^o \leq \alpha < 45^o$, $45^o \leq \alpha < 90^o$, $90^o \leq \alpha < 135^o$, $135^o \leq \alpha < 180^o$, $180^o \leq \alpha < 225^o$, $225^o \leq \alpha < 270^o$, $270^o \leq \alpha < 315^o$, and

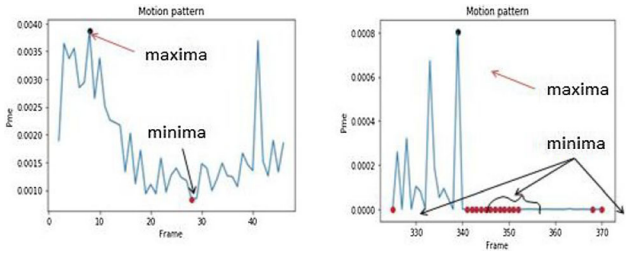

Fig. 4 Maxima and minima point within segment

$315^o \leq \alpha < 360^o$, respectively. The directed movement of a frame is computed using these eight bins. The macroblocks of the frame are placed into the bins based on their angle of directions, i.e., the bin number $B_{45\times j}$ contains all the macroblocks of the frame for which $(45 \times j)^o \leq \alpha < (45 \times (j+1))^o$, $\forall\ j = 0, 1, \cdots, 7$. Thus all the macroblocks of the $t-$th frame are partitioned into eight groups, one for each bin and the directed movement, $DM_j(t)$, of the frame with respect to bin number $B_{45\times j}$ is computed using Eq. (10), where $M_i(t)$ is computed using equation (7), i indicates the macroblock of the frame resides in bin $B_{45\times j}$, and $j = 0, 1, \cdots, 7$.

$$DM_j(t) = \sum_{\forall i \in B_{45\times j}} M_i(t) \tag{10}$$

The perceived dominant movement of $t-$th frame is considered as the maximum directed movement over all eight bins and its normalized value is computed using Eq. (11).

$$PDM(t) = \frac{max_{j=0,1,\cdots,7}\{DM_j(t)\}}{AM(t)} \tag{11}$$

This perceived dominant movement is used to determine the key moments of the frame. For finding the key moments, we use generated motion pattern of the video (i.e., PDM value of all the frames of the video). Here, we segment the motion pattern using fps of video and find the minima and maxima in the specific segment, as shown in Fig. 4. From this figure, it is observed that there may be consecutive maxima or minima along the frames.

After calculating maxima and minima, we again split the pattern using minima and if we find a maxima between two minima, we denote that portion of the segment as a key moments, and corresponding frames are considered as informative frames. If between two minima there is no maxima, we remove that segment (i.e., ignore the corresponding frames), as shown in Fig. 5. Thus, all the informative frames of the video are extracted. The pseudo code of this key moment pattern selection Algorithm is given in Algorithm 1 (i.e., KMPSA).

Algorithm 1: Informative frame selection by Key Moment Extraction (KMPSA)

```
Input: Original MPEG video ;
Output: Informative frames based on extracted key moment
        patterns ;
begin
    Extract the MVF of each macroblock of all frames of a
      MPEG video using FFMPEG tool;
    for each frame t of MPEG video do
        for i=1 to N do
            /*N=No. of macro blocks in a frame*/;
            Compute movement, M_i(t) using eq. (7);
            Compute direction α_i(t) using eq. (9);
        end
        Compute average movement, AM(t) of t−th frame
          using eq. (8);
        Partition macro blocks of t into 8 bins using eq. (10);
        Determine perceived dominant movement (PDM(t))
          of t using eq. (11);
        Key_moment[t]=PDM(t);
    end
    Segment the motion pattern using fps of video;
    for each segment do
        Find maxima and minima of motion;
    end
    Informative_Frames = ∅;
    repeat
        Traverse the pattern from the left;
        if a maxima occurs between two consecutive minima
          then
            Mark the segment between these two minima as
              a key moment;
            Extract all the frames corresponding to this
              segment;
            Merge these frames with Informative_Frames;
        end
    until motion pattern exhausted;
    return Informative_Frames;
end
```

3 Clustering method for key frame extraction

We have extracted the set $F = \{f_1, f_2, \cdots, f_m\}$ of m informative frames, where each frame f_i (which is a 2-D array) is represented by an n dimensional vector $\{f_{i1}, f_{i2}, \cdots, f_{in}\}$ by placing each row one after another starting from the first row of the 2-D array. So, for a video recording, we form a dataset with m rows and n columns where, each row is an extracted informative frame based on key moment and each column is an attribute or feature which describes the extracted frame of the video. As our main objective of this section is to apply a clustering algorithm for partitioning the frames, so first of all we normalize the attribute values within the range [0, 1] using min–max normalization technique [21] to give all the attributes an equal importance during clustering. Let after normalization, we consider the video dataset as an $m \times n$ matrix, $F = (f_{ij})_{m\times n}$. We create a dissimilarity matrix S of size $m \times m$ using Euclidean distance, where each $(i, j) - th$ entry (i.e., S_{ij}) in the matrix

gives the dissimilarity measurement between $i-$th and $j-$th frames. So, the $i-$th row indicates dissimilarity of $i-$th frame with all other frames. The leading diagonal of the matrix S gives the dissimilarity of a frame with itself, which is zero but here ignored it as we are calculating the dissimilarity of a frame with all other frames. The average dissimilarity of $i-$th frame (i.e., δ_i) with all other frames (i.e., m-1 frames) is thus computed using Eq. (12).

$$\delta_i = \frac{1}{m-1} \sum_{j=1, j\neq i}^{m} \sqrt{\sum_{k=1}^{n} (f_{ik} - f_{jk})^2} \tag{12}$$

If the dissimilarity of $i-$th frame to $j-$th frame is less than the average similarity of $i-$th frame to all other frames, i.e., if $S_{ij} < \delta_i$, then we place $j-$th frame in cluster C_i, for all $j = 1, 2, \cdots, m$. Thus, all the frames which are similar to $i-$th frame are placed in a cluster, for all $i = 1, 2, \cdots, m$. This process gives us m number of initial overlapping clusters, $C = \{C_1, C_2, \cdots, C_m\}$. All the clusters are nonempty as $i-$th cluster C_i contains at least $i-$th frame. It may happen that, some clusters are subset of other clusters, so we remove these redundant clusters and get p ($< m$) initial overlapping clusters of frames. If two overlapping clusters have many common frames then they are similar to each other. So cluster similarity is computed between every pair of clusters to decide whether two clusters need to be merged or not. Jaccard coefficient (c_{ij}) is computed between two clusters C_i and C_j using Eq. (13), $\forall i, j(\neq i) = 1, 2, \cdots, p$.

$$c_{ij} = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{13}$$

The similarity value c_{ij} ranges between 0 and 1. The value 0 implies that no common frames are there in between two clusters C_i and C_j and the highest similarity value is less than 1, as there is no two same clusters (because we have already removed a cluster which is subset of some other clusters). Thus a similarity matrix, $C = (c_{ij})_{p\times p}$ is obtained, where $(i, j) - th$ entry in the matrix gives the similarity measurement between $i-$th cluster C_i and $j-$th cluster C_j. Here, also we ignore the leading diagonal as we are interested to merge two different clusters based on their Jaccard similarity. Obviously, the matrix C is a symmetric matrix as the Jaccard similarity relation is symmetric. Thus, we compute only upper triangular matrix of C, i.e., c_{ij} $\forall i, j = 1, 2, \cdots, p$ and $i < j$. That means, we have to compute Jaccard similarity between $\binom{p}{2}$ or $\frac{p(p-1)}{2}$ pair of clusters. Then we apply agglomerative or bottom up approach of merging two clusters to obtain a comparatively larger size cluster. For merging two clusters, we determine which one out of all $\frac{p(p-1)}{2}$ pair of clusters provides the maximum Jaccard similarity. If multiple maximum values occur, we separately merge them. Let, c_{kl} is the maximum value, which implies that two clusters C_k and C_l are the most similar clusters among all p clusters. Merge these two clusters C_k and C_l to get a new cluster C_{kl}, i.e., $C_{kl} = C_k \cup C_l$. After merging all such pair wise maximum similarity-based clusters, we have, say p' (where, $p' < p$) clusters. Thus before merging and after merging, we have two sets of clusters, say $C = \{C_1, C_2, \cdots, C_p\}$ and $C' = \{C'_1, C'_2, \cdots, C'_{p'}\}$. Next, equation (1) to (6) are used on C and C' to compute the overlapping cluster index OC between these two sets of clusters. If OC is less than a predefined threshold (experimentally set as 0.1), then only the process of merging is stopped; otherwise it is continued. After performing merging operations in one iteration, it may again happen that some clusters are subset of newly merged cluster. In this case, the subclusters are removed and treat them as a set of p clusters. Similarly, determine the pair of clusters which have maximum Jaccard similarity and merged them to find the set of clusters which is treated as p'. Then compute the overlapping cluster index OC between them and so on. This process is repeated until either desired number (say, at most K) of clusters are obtained or OC is less than a predefined threshold. After merging, when exactly K clusters are obtain, we terminate the process, but as in this case some clusters may be the subset of the newly merged clusters, so they are removed. Thus we get at most K clusters of frames. The value of K is determined based on the duration of the summary of the

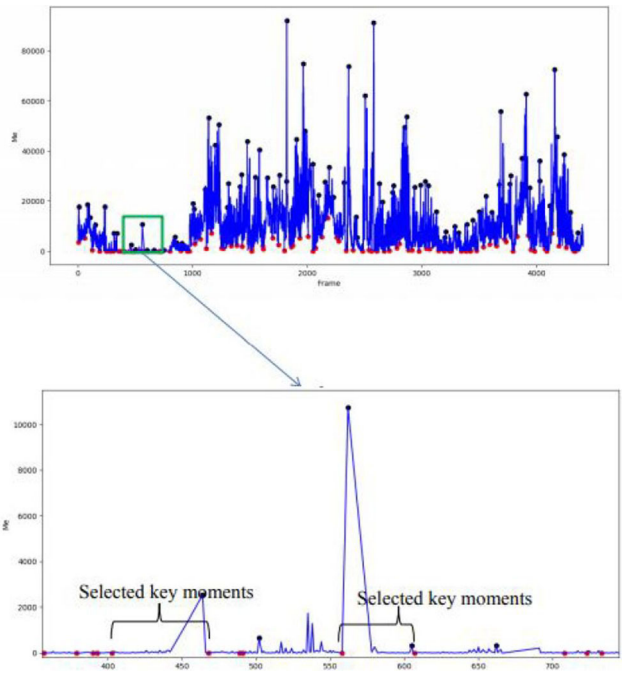

Fig. 5 Selected key moments within minima

video. If each frame is of u seconds and we want summary of a video of at most v seconds, then K is obtained from the equality, $v = K \times u$. But if the summary size is not predefined, then only based on the OC value the process is terminated and an arbitrary set of clusters is obtained. Even if, the summary size is predefined as K, but before achieving K number of clusters, if OC becomes less than t, then also the process terminates. In this case, number of clusters is more than K. Thus the algorithm is not solely dependent on the number of clusters known a priori. In both the cases, from each cluster of frames, centroid frame is considered as the representative frame and order collection of these representative frames into a MPEG video is considered as the summary of the given video file. The proposed clustering algorithm for finding the key frames from a set of frames is described in Algorithm 2 (CAFKF).

Algorithm 2: Clustering Algorithm for Finding Key Frames (CAFKF)

Input: Set of frames $F = \{f_1, f_2, \cdots, f_m\}$, K = the maximum number of clusters, and $t = 0.1$ is the threshold for overlapping cluster index ;
Output: KF = set of Key frames ;
begin
  Apply min-max normalization on each frame f_i in F;
  for *each frame* f_i *in* F */* Form initial m clusters */* **do**
    Compute average dissimilarity δ_i of f_i with all other frames in F using eq. (12);
    for *each frame* f_j *in* F **do**
      if $S_{ij} < \delta_i$ **then**
        Add f_j into cluster C_i /* initially $C_i = \emptyset$*/;
      end
    end
  end
  Let $IC = \{C_1, C_2, \cdots, C_m\}$ be the initial set of clusters;
  Remove all subclusters from IC;
  repeat
    Let $IC = \{C_1, C_2, \cdots, C_p\}$ are the distinct p clusters;
    for *each pair of clusters,* C_i *and* C_j $(i \neq j)$ *in* IC **do**
      Compute similarity c_{ij} using eq. (13);
    end
    $IC' = IC$;
    for *each pair of clusters* C_l *and* C_k *in* IC **do**
      if $C_{lk} = \max_{1 \le i \le |IC|, 1 \le j \le |IC|, i \neq j} \{C_{ij}\}$ **then**
        $IC' = IC' - \{C_l\} - \{C_k\}$;
        $IC' = IC' \cup \{C_l \cup C_k\}$;
      end
    end
    Compute OC between IC and IC' using eq. (1) to (6);
    $IC = IC'$;
    Remove all subclusters from IC;
  until $|IC| \le K$ *OR* $OC \le t$;
  $KF = \phi$;
  for *each cluster* C_i *in* IC **do**
    Determine the centroid frame f_i^C;
    $KF = KF \cup \{f_i^C\}$;
  end
  Return KF;
end

4 Experimental results

For our experiment, we have used the SumMe dataset, which was created to be used as a benchmark for video summarization. The dataset contains 25 videos with the video duration ranging from 1 to 6 minutes. In the paper, we have performed two different types of evaluations, (i) evaluation of clustering algorithm and (ii) evaluation of summarization algorithm.

4.1 Evaluation of proposed CAFKF clustering algorithm

Initially, key moments-based video frame extraction is performed using "KMPSA" algorithm to extract all possible informative frames and remove the irrelevant frames from the video. Next, the proposed agglomerative clustering algorithm, namely CAFKF is applied to cluster the frames into different groups. Two parameters used in the algorithm are number of clusters (K) and a threshold (t) allowing the lower overlapping cluster index (OC) among the generated clusters. If each frame is of duration u seconds and desire duration of the summary of the video is of at most v seconds, then K is obtained from the expression, $v = K \times u$. But before achieving K number of clusters, if OC becomes less than t, then also the process terminates. The value of t is experimentally set as 0.1. Thus, the process terminates based on the values of K and OC. In both the cases, some clusters may be overlapped with each other. We have computed centroid of each cluster and each frame which is in multiple clusters is placed only to its closest-centroid cluster. Thus, all the clusters are made disjoint. After achieving the disjoint set of clusters, they are validated and compared with different clustering algorithms based on cluster validation indices, namely internal indices and stability indices. In order to evaluate the effectiveness of the proposed CAFKF clustering algorithm, we compare it with several very popular and frequently used clustering algorithms such as, K-Means clustering (KM) [8], spectral clustering (SC) [9], clustering by affinity propagation (AP) [17], Delaunay clustering (DC) [29].

1. *Internal cluster validation indices:* Internal cluster validation indices [27], like Silhouette index (SL), Dunn's index (DN), Davies–Bouldin index (DB), Xie–Beni index (XB), Calinski–Harabasz index (CH), I-index (IN) are computed for all the clustering algorithms. The methods are evaluated using 10-fold cross-validation technique and the average values of the performance metrics are listed in Table 1. From Table 1, it is observed that, the proposed CAFKF algorithm gives the best values of Silhouette index (SL) for 18 datasets, Dunn's index (DN) for 20

Table 1 Evaluation of proposed clustering algorithm based on internal indices

Video Data	Clustering Algorithm	Internal cluster validation indices						Video Data	Clustering Algorithm	Internal cluster validation indices					
		SL	DN	DB	XB	CH	IN			SL	DN	DB	XB	CH	IN
Air force one	KM	0.56	0.69	0.47	0.56	468	601	Base jumping	KM	0.68	0.67	0.50	0.61	**452**	579
	SC	0.63	0.76	**0.43**	0.69	375	583		SC	0.72	0.59	0.41	0.58	410	**581**
	AP	0.73	0.84	0.62	**0.37**	451	591		AP	0.59	0.63	0.60	**0.40**	409	572
	DC	0.61	0.88	0.55	0.61	379	**629**		DC	0.72	0.51	0.52	0.50	371	568
	CAFKF	**0.82**	**0.91**	**0.43**	0.46	**474**	627		CAFKF	**0.79**	**0.87**	**0.39**	0.41	**452**	577
Bear park climbing	KM	0.65	0.74	0.40	0.56	450	634	Bike polo	KM	0.62	**0.92**	**0.42**	0.53	437	581
	SC	0.78	0.68	**0.38**	0.61	405	586		SC	0.77	0.88	0.47	0.60	439	579
	AP	0.66	0.64	0.62	0.67	407	599		AP	0.68	0.84	0.50	0.53	425	579
	DC	**0.84**	0.55	0.57	0.53	**461**	643		DC	0.74	0.79	0.56	0.55	417	573
	CAFKF	**0.84**	**0.79**	0.40	**0.47**	436	**652**		CAFKF	**0.80**	**0.92**	0.43	**0.44**	**446**	**584**
Bus in Rock Tunnel	KM	0.61	0.78	0.49	0.57	351	**604**	Car over Camera	KM	0.65	**0.87**	0.51	0.52	410	624
	SC	0.65	0.62	0.37	0.62	389	588		SC	**0.71**	0.73	0.49	0.63	402	604
	AP	0.59	0.76	0.49	0.63	**405**	579		AP	0.69	0.65	0.61	0.50	410	591
	DC	**0.74**	0.64	0.51	0.53	369	594		DC	0.70	0.67	0.50	**0.49**	394	611
	CAFKF	0.70	**0.83**	**0.31**	**0.37**	337	593		CAFKF	0.69	0.85	**0.42**	0.53	**413**	**631**
Car rail crossing	KM	**0.72**	0.73	0.59	**0.53**	411	654	Cockpit Landing	KM	0.79	0.73	0.53	0.63	**430**	597
	SC	0.60	0.67	0.41	0.66	406	586		SC	**0.83**	0.68	0.42	0.58	425	602
	AP	0.65	0.69	0.67	0.72	398	577		AP	0.68	**0.88**	0.61	0.72	400	573
	DC	0.69	0.53	0.59	0.55	368	**659**		DC	0.73	0.57	0.58	0.49	398	**615**
	CAFKF	0.68	**0.85**	**0.40**	0.54	**413**	**659**		CAFKF	**0.83**	0.71	**0.41**	**0.38**	404	611
Cooking	KM	0.66	0.84	0.56	0.70	**436**	612	Eiffel Tower	KM	**0.72**	0.75	0.56	0.61	435	609
	SC	0.64	0.62	**0.48**	0.62	409	596		SC	0.68	**0.88**	0.44	0.57	**447**	618
	AP	0.59	0.74	0.71	0.78	388	587		AP	0.62	0.67	0.53	0.52	411	593
	DC	0.60	0.66	0.56	0.53	371	**615**		DC	0.71	0.63	0.58	**0.44**	392	608
	CAFKF	**0.71**	**0.93**	0.55	**0.50**	**436**	605		CAFKF	0.71	**0.88**	**0.41**	0.49	424	**619**
Excavators river crossing	KM	**0.82**	**0.80**	0.43	0.55	451	604	Fire Domino	KM	0.68	0.92	0.58	0.50	472	599
	SC	0.77	0.68	0.48	0.51	**453**	532		SC	0.71	0.69	**0.45**	0.53	468	578
	AP	0.69	0.77	0.63	0.68	388	597		AP	0.65	0.71	0.61	0.48	434	634
	DC	0.78	0.72	0.50	0.49	371	**611**		DC	0.70	**0.97**	0.51	0.49	418	**681**
	CAFKF	0.81	**0.80**	**0.42**	**0.41**	446	601		CAFKF	**0.75**	0.86	**0.45**	**0.47**	**478**	653
Jumps	KM	0.70	0.67	0.47	0.62	411	650	Kids Playing in leaves	KM	0.77	0.84	**0.40**	0.49	**432**	597
	SC	**0.73**	0.62	**0.44**	0.68	397	585		SC	**0.82**	0.79	0.45	0.51	427	582
	AP	0.69	0.59	0.62	0.70	405	577		AP	0.66	0.63	0.50	0.62	415	591
	DC	0.66	**0.69**	0.51	0.62	389	609		DC	0.71	0.81	0.48	0.53	389	**601**
	CAFKF	0.71	**0.69**	0.45	**0.59**	**444**	**662**		CAFKF	**0.82**	**0.86**	0.45	**0.45**	420	599
Notre Dame	KM	0.77	**0.71**	0.55	0.56	481	598	Paintball	KM	**0.73**	0.91	0.48	0.54	428	598
	SC	0.64	0.68	0.59	0.61	466	**615**		SC	0.68	0.76	**0.44**	0.52	440	542
	AP	0.65	0.61	0.67	0.73	457	581		AP	0.59	0.69	0.62	0.67	432	600
	DC	0.67	0.64	**0.52**	0.55	398	601		DC	**0.73**	0.65	0.57	0.53	**450**	605
	CAFKF	**0.78**	**0.71**	0.57	**0.53**	**490**	590		CAFKF	0.71	**0.93**	**0.44**	**0.50**	412	**609**
Paluma jump	KM	0.79	0.78	0.52	**0.49**	**412**	585	Playing Ball	KM	0.68	**0.89**	0.40	0.70	**491**	606
	SC	0.68	0.72	**0.47**	0.56	395	591		SC	0.70	0.64	0.46	0.66	464	613
	AP	0.74	0.68	0.67	0.60	398	582		AP	0.66	0.60	0.60	0.69	438	592
	DC	0.77	0.52	0.56	0.55	378	594		DC	0.71	0.59	0.58	0.57	410	604
	CAFKF	**0.83**	**0.91**	**0.47**	0.58	410	**605**		CAFKF	**0.80**	0.86	**0.39**	**0.50**	489	**617**

Table 1 (continued)

Video Data	Clustering Algorithm	Internal cluster validation indices						Video Data	Clustering Algorithm	Internal cluster validation indices					
		SL	DN	DB	XB	CH	IN			SL	DN	DB	XB	CH	IN
Playing on water slide	KM	**0.78**	0.84	0.59	0.62	**420**	624	Saving dolphins	KM	0.65	0.75	0.43	0.70	**441**	633
	SC	0.73	0.76	**0.47**	0.61	367	607		SC	0.71	0.77	**0.42**	0.67	431	601
	AP	0.71	0.65	0.60	0.70	388	591		AP	0.64	0.61	0.68	0.72	410	607
	DC	0.68	0.79	0.66	0.59	391	**625**		DC	0.70	0.50	0.58	**0.51**	399	**635**
	CAFKF	**0.78**	**0.92**	0.55	**0.55**	415	620		CAFKF	**0.79**	**0.89**	0.44	0.54	426	**635**
Scuba	KM	0.79	0.78	0.52	**0.49**	412	**618**	Statue of liberty	KM	0.68	**0.89**	0.40	0.70	464	606
	SC	0.68	0.72	**0.47**	0.56	395	591		SC	0.70	0.64	0.46	0.66	**491**	613
	AP	0.74	0.68	0.67	0.60	**415**	582		AP	0.66	0.60	0.60	0.69	438	592
	DC	0.77	0.52	0.56	0.55	378	594		DC	0.71	0.59	0.58	0.57	410	604
	CAFKF	**0.83**	**0.91**	**0.47**	**0.49**	410	605		CAFKF	**0.80**	0.86	**0.39**	**0.50**	489	**617**
St Maarten Landing	KM	0.66	0.91	0.52	**0.43**	407	627	Uncut Evening Flight	KM	0.77	0.94	0.49	0.62	**453**	621
	SC	0.70	**0.93**	0.43	0.50	380	603		SC	0.74	0.69	0.45	0.59	452	610
	AP	0.65	0.67	0.61	0.51	387	598		AP	0.68	0.77	0.67	0.68	401	599
	DC	0.72	0.59	0.58	0.53	400	611		DC	0.75	0.53	0.53	**0.58**	376	617
	CAFKF	**0.79**	**0.93**	**0.42**	0.44	**411**	**631**		CAFKF	**0.82**	**0.96**	**0.43**	**0.58**	410	**623**
Valparaiso downhill	KM	0.72	0.83	0.48	**0.48**	**460**	604	**Average for all Video Data**	KM	0.71	0.81	0.49	0.55	**437**	610
	SC	0.64	0.77	**0.44**	0.52	401	591		SC	0.71	0.72	0.45	0.60	421	591
	AP	0.69	0.64	0.55	0.61	412	555		AP	0.66	0.68	0.61	0.62	411	589
	DC	0.73	0.52	0.51	0.50	399	**617**		DC	0.71	0.62	0.55	0.55	395	613
	CAFKF	**0.76**	**0.86**	0.47	**0.48**	455	612		CAFKF	**0.77**	**0.86**	**0.44**	**0.49**	434	**617**

Bold face is given to indicate that the best result is obtained in that position

datasets, Davies–Bouldin index (DB) for 16 datasets, Xie–Beni index (XB) for 16 datasets, Calinski–Harabasz index (CH) for 10 datasets, and I-index (IN) for 15 datasets out of all 25 datasets. So, on an average, the proposed method provides best result for 64% datasets. Similarly, the best values of all the indices obtained by other clustering algorithms are easily visualized from the table. As we know that, the highest values of SL, DN, CH, IN and the lowest values of DB, XB provide the best clusters of the datasets, so the best values of all the measures obtained by all the clusters are marked by bold faces in the table. Also average index values considering all 25 datasets are separately computed for all clustering algorithms and listed in the last row of the table. These average values also demonstrate that the proposed clustering algorithm provides the best result followed by the KM clustering algorithm. Based on the index values listed in Table 1, the clustering algorithms are ranked and the best ranked algorithm is marked by 1, second best by 2 and so on. This is done separately for each dataset and each internal index, as listed in Table 2. Also the average rank of all clustering algorithms considering all 25 datasets is computed and listed in last row of the table. From this observation, we conclude that the proposed CAFKF clustering algorithm outperforms others with respect to the internal cluster validation indices and thus used for our video summarization work. Also considering all internal indices, K-Means (KM) clustering algorithm provides the second best result.

2. *Stability-based Cluster Validation indices:* Stability-based indices [3], like average distance between measurements (AD), average distance between means (ADM), figure of merit (FOM), global clustering coefficient (GCC), and modularity of the partitions (MOP) are computed for all the clustering algorithms. The methods are evaluated using 10-fold cross-validation technique and the average values of the performance metrics are listed in Table 3. The highest values of all these five measures provide the best clusters and so we have marked the best values by the bold faces in the table. From this table, it is observed that, the

Table 2 Ranking of clustering algorithms based on internal indices

Video	Clustering	Ranking based on Internal indices						Video	Clustering	Ranking based on Internal indices					
Data	Algorithm	SL	DN	DB	XB	CH	IN	Data	Algorithm	SL	DN	DB	XB	CH	IN
Air force one	KM	5	5	2	3	2	3	Base jumping	KM	3	2	3	5	1	2
	SC	3	4	1	5	5	5		SC	2	4	2	4	2	1
	AP	2	3	4	1	3	4		AP	4	3	5	1	3	4
	DC	4	2	3	4	4	1		DC	2	5	4	3	4	5
	CAFKF	1	1	1	2	1	2		CAFKF	1	1	1	2	1	3
Bear park climbing	KM	4	2	2	3	2	3	Bike polo	KM	5	1	1	2	3	2
	SC	2	3	1	4	5	5		SC	2	2	3	4	2	3
	AP	3	4	4	5	4	4		AP	4	3	4	2	4	3
	DC	1	5	3	2	1	2		DC	3	4	5	3	5	4
	CAFKF	1	1	2	1	3	1		CAFKF	1	1	2	1	1	1
Bus in Rock Tunnel	KM	4	2	3	3	4	1	Car over Camera	KM	4	1	4	3	2	2
	SC	3	5	2	4	2	4		SC	1	3	2	5	3	4
	AP	5	3	3	5	1	5		AP	3	5	5	2	2	5
	DC	1	4	4	2	3	2		DC	2	4	3	1	4	3
	CAFKF	2	1	1	1	5	3		CAFKF	3	2	1	4	1	1
Car rail crossing	KM	1	2	3	1	2	2	Cockpit Landing	KM	2	2	3	4	1	4
	SC	5	4	2	4	3	3		SC	1	4	2	3	2	3
	AP	4	3	4	5	4	4		AP	4	1	5	5	4	5
	DC	2	5	3	3	5	1		DC	3	5	4	2	5	1
	CAFKF	3	1	1	2	1	1		CAFKF	1	3	1	1	3	2
Cooking	KM	2	2	3	4	1	2	Eiffel Tower	KM	1	2	4	5	2	3
	SC	3	5	1	3	2	4		SC	3	1	2	4	1	2
	AP	5	3	4	5	3	5		AP	4	3	3	3	4	5
	DC	4	4	2	2	4	1		DC	2	4	5	1	5	4
	CAFKF	1	1	3	1	1	3		CAFKF	2	1	1	2	3	1
Excavators river crossing	KM	1	1	2	4	2	2	Fire Domino	KM	4	2	3	4	2	4
	SC	4	4	3	3	1	5		SC	2	5	1	5	3	5
	AP	5	2	5	5	4	4		AP	5	4	4	2	4	3
	DC	3	3	4	2	5	1		DC	3	1	2	3	5	1
	CAFKF	2	1	1	1	3	3		CAFKF	1	3	1	1	1	2
Jumps	KM	3	2	3	2	2	2	Kids Playing in leaves	KM	2	2	1	2	1	3
	SC	1	3	1	3	4	4		SC	1	4	2	3	2	5
	AP	4	4	5	4	3	5		AP	4	5	4	5	4	4
	DC	5	1	4	2	5	3		DC	3	3	3	4	5	1
	CAFKF	2	1	2	1	1	1		CAFKF	1	1	2	1	3	2
Notre Dame	KM	2	1	2	3	2	3	Paintball	KM	1	2	2	4	4	4
	SC	5	2	4	4	3	1		SC	3	3	1	2	2	5
	AP	4	4	5	5	4	5		AP	4	4	4	5	3	3
	DC	3	3	1	2	5	2		DC	1	5	3	3	1	2
	CAFKF	1	1	3	1	1	4		CAFKF	2	1	1	1	5	1
Paluma jump	KM	2	2	2	1	1	4	Playing Ball	KM	4	1	2	5	1	3
	SC	5	3	1	3	4	3		SC	3	3	3	3	3	2
	AP	4	4	4	5	3	5		AP	5	4	5	4	4	5
	DC	3	5	3	2	5	2		DC	2	5	4	2	5	4
	CAFKF	1	1	1	4	2	1		CAFKF	1	2	1	1	2	1

Table 2 (continued)

Video	Clustering	Ranking based on Internal indices						Video	Clustering	Ranking based on Internal indices					
Data	Algorithm	SL	DN	DB	XB	CH	IN	Data	Algorithm	SL	DN	DB	XB	CH	IN
Playing on water slide	KM	1	2	3	4	1	2	Saving dolphins	KM	4	3	2	4	1	2
	SC	2	4	1	3	5	4		SC	2	2	1	3	2	4
	AP	3	5	4	5	4	5		AP	5	4	5	5	4	3
	DC	4	3	5	2	3	1		DC	3	5	4	1	5	1
	CAFKF	1	1	2	1	2	3		CAFKF	1	1	3	2	3	1
Scuba	KM	2	2	2	1	2	1	Statue of liberty	KM	4	1	2	5	3	3
	SC	5	3	1	3	4	4		SC	3	3	3	3	1	2
	AP	4	4	4	4	1	5		AP	5	4	5	4	4	5
	DC	3	5	3	2	5	3		DC	2	5	4	2	5	4
	CAFKF	1	1	1	1	3	2		CAFKF	1	2	1	1	2	1
St Maarten Landing	KM	4	2	3	1	2	2	Uncut Evening Flight	KM	2	2	3	3	1	2
	SC	3	1	2	3	5	4		SC	4	4	2	2	1	4
	AP	5	3	5	4	4	5		AP	5	3	5	4	4	5
	DC	2	4	4	5	3	3		DC	3	5	4	1	5	3
	CAFKF	1	1	1	2	1	1		CAFKF	1	1	1	1	3	1
Valparaiso downhill	KM	3	2	3	1	1	3	**Average Ranking of Clustering Algorithms**	KM	2.80	1.92	2.52	3.08	1.84	2.56
	SC	5	3	1	3	4	4		SC	2.92	3.28	1.80	3.44	2.76	3.60
	AP	4	4	5	4	3	5		AP	4.16	3.56	4.40	3.96	3.40	4.44
	DC	2	5	4	2	5	1		DC	2.64	4.00	3.36	2.32	4.28	2.24
	CAFKF	1	1	2	1	2	2		CAFKF	1.36	1.28	1.48	1.48	2.16	1.76

proposed CAFKF algorithm gives the best values of GCC for 16 datasets, MOP for 17 datasets, AD and ADM for 14 datasets, and FOM for 13 datasets out of all 25 datasets. So, on an average, the proposed method provides best result for 59.2% datasets. Similarly, the best values of all the stability indices obtained by other clustering algorithms are easily visualized from the table. Also average index values considering all 25 datasets are separately computed for all clustering algorithms and listed in the last row of the table. The average values also express that the proposed clustering algorithm provides the best performance followed by the KM clustering algorithm. Based on the index values listed in Table 3, the clustering algorithms are ranked and the best ranked algorithm is marked by 1, second best by 2 and so on. This is also done separately for each dataset and each stability index, as listed in Table 4. Also the averages rank of all clustering algorithms considering all 25 datasets are computed and listed in last row of the table. From this observation, we conclude that the proposed CAFKF clustering algorithm outperforms others with respect to stability indices and thus used for our video summarization work. Also considering all stability indices, K-Means (KM) clustering algorithm provides the second best result.

4.2 Evaluation of proposed summarization technique

Each of these videos is annotated (annotations are collected by crowd sourcing) by at least 15 humans with a total of 390 human summaries. The ground truth scores are given in fractional number with the implication that if the ground truth score is close to zero then the frame is irrelevant for the summary; otherwise, the frame is selected for summary. Higher score indicates that many users while making the summary select the frame. We have used a cutoff score value as 0.05 which signifies that we have selected those frames in ground truth summary which are selected by at least more than 2 users.

The proposed method is applied on each of the videos of the SumMe dataset and the summary is generated. The method is compared with state-of-the-art methodologies [24], such as Uniform Sampling (USAM), Image histogram (IMHS), Scale Invariant Feature Transform (SIFT) and K-

Table 3 Evaluation of proposed clustering algorithm based on stability indices

Video Data	Clustering Algorithm	Cluster stability indices					Video Data	Clustering Algorithm	Cluster stability indices				
		GCC	MOP	AD	ADM	FOM			GCC	MOP	AD	ADM	FOM
Air force one	KM	0.67	0.58	1.57	**1.04**	0.57	Base jumping	KM	0.71	**0.65**	1.34	0.55	0.57
	SC	0.57	0.63	1.03	0.74	0.45		SC	0.75	0.61	**1.40**	0.49	0.46
	AP	0.65	0.54	1.57	0.65	0.61		AP	0.69	0.56	1.31	0.52	0.43
	DC	0.61	0.62	1.21	0.71	0.49		DC	0.57	0.58	1.29	**0.61**	0.49
	CAFKF	**0.73**	**0.71**	**1.87**	0.81	**0.83**		CAFKF	**0.85**	**0.65**	1.38	0.58	**0.61**
Bear park climbing	KM	**0.61**	0.63	**1.62**	0.85	**0.74**	Bike polo	KM	0.62	**0.75**	1.51	0.58	0.56
	SC	0.52	0.51	1.03	0.74	0.60		SC	0.71	0.57	1.43	0.51	0.48
	AP	0.57	0.59	1.54	0.69	0.55		AP	0.63	0.61	1.38	**0.60**	0.45
	DC	0.55	0.62	1.49	0.75	0.59		DC	0.68	0.55	1.33	0.56	0.49
	CAFKF	0.59	**0.65**	1.59	**0.99**	0.68		CAFKF	**0.74**	0.60	**1.57**	0.57	**0.59**
Bus in Rock Tunnel	KM	0.68	0.67	1.50	0.58	0.58	Car over Camera	KM	**0.70**	0.61	**1.49**	0.58	**0.55**
	SC	**0.72**	0.61	1.48	0.52	**0.59**		SC	0.62	**0.64**	1.42	0.46	0.49
	AP	0.67	**0.69**	1.43	0.51	0.55		AP	0.59	0.55	1.38	0.54	0.40
	DC	0.63	0.58	1.39	0.60	0.49		DC	0.59	0.54	1.35	0.62	0.45
	CAFKF	**0.72**	0.65	**1.51**	**0.62**	0.57		CAFKF	0.68	**0.64**	1.40	**0.64**	0.51
Car rail crossing	KM	**0.81**	0.69	**1.81**	0.77	0.70	Cockpit Landing	KM	**0.77**	0.61	1.61	0.98	**0.74**
	SC	0.70	0.59	1.54	0.58	0.52		SC	0.73	0.55	1.51	0.78	0.61
	AP	0.63	0.57	1.51	**0.79**	0.57		AP	0.69	**0.69**	1.29	0.86	0.69
	DC	0.59	0.64	1.48	0.76	0.69		DC	0.64	0.59	1.48	**1.07**	0.65
	CAFKF	0.79	**0.73**	1.80	0.78	**0.75**		CAFKF	**0.77**	0.64	**1.78**	**1.07**	0.71
Cooking	KM	0.79	0.71	1.66	0.68	**0.72**	Eiffel Tower	KM	0.69	**0.72**	1.62	**0.87**	0.68
	SC	0.80	0.61	1.58	0.65	0.68		SC	0.70	0.64	1.54	0.83	0.59
	AP	0.81	0.62	1.55	0.63	0.69		AP	0.69	0.68	1.44	0.81	0.63
	DC	0.69	0.65	1.50	**0.74**	0.67		DC	0.63	0.59	1.58	0.79	0.70
	CAFKF	**0.83**	**0.75**	**1.72**	0.71	0.68		CAFKF	**0.75**	0.70	**1.77**	0.85	**0.78**
Excavators river crossing	KM	**0.81**	0.59	1.50	**0.74**	0.68	Fire Domino	KM	**0.76**	0.73	1.56	**0.75**	**0.69**
	SC	0.73	0.66	1.61	0.57	0.64		SC	0.68	0.62	1.47	0.69	0.60
	AP	0.69	0.62	1.47	0.61	0.53		AP	0.67	0.68	1.39	0.62	0.58
	DC	0.75	0.59	1.43	0.68	0.60		DC	0.72	0.71	1.48	0.64	0.62
	CAFKF	0.78	**0.68**	**1.68**	0.73	**0.72**		CAFKF	**0.76**	**0.78**	**1.61**	0.71	0.64
Jumps	KM	0.80	**0.68**	**1.64**	0.64	0.63	Kids Playing in leaves	KM	0.78	**0.69**	1.51	**0.62**	**0.65**
	SC	0.67	0.60	1.56	0.59	0.57		SC	0.72	0.62	1.44	0.60	0.58
	AP	0.69	0.61	1.45	0.60	0.51		AP	0.68	0.58	1.50	0.58	0.51
	DC	0.71	0.64	1.48	0.58	0.50		DC	0.59	0.63	1.48	0.59	0.55
	CAFKF	**0.82**	0.67	1.57	**0.67**	**0.70**		CAFKF	**0.79**	0.66	**1.57**	**0.62**	0.57
Notre Dame	KM	0.54	0.71	1.29	0.83	0.57	Paintball	KM	**0.79**	0.81	0.98	0.63	0.74
	SC	**0.63**	0.49	1.31	0.79	0.61		SC	0.61	0.65	1.07	0.47	**0.83**
	AP	0.48	0.56	**1.47**	0.65	0.49		AP	0.54	0.74	0.85	0.59	0.69
	DC	0.57	0.69	1.03	0.72	0.61		DC	0.63	0.59	0.91	**0.71**	0.56
	CAFKF	0.54	**0.75**	1.38	**0.91**	**0.73**		CAFKF	0.75	**0.87**	**1.23**	0.67	**0.83**
Paluma jump	KM	**0.75**	0.61	1.07	0.74	**0.69**	Playing Ball	KM	**0.61**	0.79	1.19	0.66	0.54
	SC	0.63	0.53	0.94	0.63	0.57		SC	0.54	0.82	1.25	**0.71**	0.47
	AP	0.71	0.47	1.13	0.71	0.53		AP	0.37	0.57	1.21	0.59	**0.61**
	DC	0.52	0.63	1.01	0.74	0.61		DC	**0.61**	0.48	1.04	0.43	0.41
	CAFKF	**0.75**	**0.74**	**1.54**	**0.82**	0.63		CAFKF	0.59	**0.86**	**1.33**	**0.71**	0.52

Table 3 (continued)

Video Data	Clustering Algorithm	Cluster stability indices					Video Data	Clustering Algorithm	Cluster stability indices				
		GCC	MOP	AD	ADM	FOM			GCC	MOP	AD	ADM	FOM
Playing on water slide	KM	0.65	0.53	**1.37**	0.69	0.58	Saving dolphins	KM	0.71	0.62	**1.23**	0.57	**0.64**
	SC	0.53	0.61	1.09	**0.87**	0.62		SC	0.65	0.57	1.01	0.43	0.59
	AP	0.49	0.66	1.23	0.71	0.47		AP	0.72	**0.71**	1.17	0.62	0.47
	DC	**0.71**	0.47	1.15	0.54	0.53		DC	0.53	0.45	1.15	0.53	0.56
	CAFKF	0.65	**0.69**	1.21	**0.87**	**0.79**		CAFKF	**0.74**	**0.71**	1.20	**0.81**	0.63
Scuba	KM	0.72	**0.67**	1.28	0.75	0.47	Statue of liberty	KM	0.69	0.56	**1.59**	0.82	0.51
	SC	0.68	0.54	1.31	0.64	0.49		SC	0.58	0.49	1.32	0.74	**0.68**
	AP	0.63	0.62	1.17	**0.78**	0.38		AP	0.52	0.41	1.27	0.91	**0.68**
	DC	0.72	0.59	1.34	**0.78**	0.51		DC	0.49	0.53	**1.59**	0.63	0.47
	CAFKF	**0.75**	0.65	**1.42**	0.73	**0.67**		CAFKF	**0.74**	**0.69**	1.32	**0.95**	0.53
St Maarten Landing	KM	0.63	0.49	**1.75**	0.87	0.61	Uncut Evening Flight	KM	**0.77**	0.59	1.61	**0.92**	0.54
	SC	0.59	0.61	1.29	0.73	**0.65**		SC	0.63	0.61	1.43	0.75	0.69
	AP	0.61	0.53	1.37	0.69	0.57		AP	0.52	**0.71**	1.29	0.68	0.73
	DC	**0.72**	0.47	1.18	0.58	0.53		DC	0.69	0.64	1.52	0.53	0.51
	CAFKF	0.63	**0.72**	1.63	**0.96**	**0.65**		CAFKF	**0.77**	0.69	**1.79**	0.89	**0.86**
Valparaiso downhill	KM	0.56	0.43	**1.57**	0.87	0.67	**Average for all Video Data**	KM	0.70	0.64	1.47	0.74	0.62
	SC	0.59	0.51	1.08	0.75	0.61		SC	0.65	0.60	1.33	0.65	0.59
	AP	0.48	0.66	1.32	0.67	**0.79**		AP	0.62	0.60	1.35	0.66	0.56
	DC	0.69	0.58	1.16	0.73	0.59		DC	0.63	0.59	1.34	0.67	0.55
	CAFKF	**0.71**	**0.69**	1.21	**0.89**	0.71		CAFKF	**0.73**	**0.71**	**1.52**	**0.78**	**0.68**

means clustering-based video summarization (KMVS). Though many clustering algorithms are applied for frame partitioning, but K-means clustering (i.e., KM) provides the second best results in terms of cluster validation indices, as discussed in Sect. 4.1. For this reason, out of all clustering algorithms (i.e., KM, SC, AP, DP) used for evaluation of the proposed clustering algorithm CAFKF, we have selected only KM clustering algorithm and corresponding video summarization technique is named as KMVS. Here, the initial frames selected by our proposed KMPSA algorithm, described in Sect. 2, are partitioned using K-means clustering algorithm. The goal of this algorithm is to partition the frames into different clusters, so that the intra-cluster similarity is the maximum and inter-cluster similarity is the minimum. Next, similar to our proposed CAFKF algorithm, centrally located frame in each cluster is selected as representative of the cluster. These representatives are considered as key frames and used as the summary of the video. We have set the value of K in K-means algorithm as same as the value of K in our proposed CAFKF clustering algorithm. Uniform sampling (USAM) is one of the most frequently used methods for key frame extraction. The main logic behind this method is that it selects every $t-$th frame of the video where the value of t is determined based on the length of the video. In the work, we have considered summary size as the 10% of the size of the original video. So every 10-th frame is chosen in this method for our experimental purpose. Image histograms (IMHS) consider the tonal distribution of the video image. It counts the number of pixels of a specific brightness value within the range of 0 to 255. We extract the histograms of all frames of the video and based on the dissimilarity between histograms of two frames, we decide whether the frames are key frames or not. If there is a significant dissimilarity between two consecutive frames then there must be a rapid change of scene in the video, which might contain informative content of the video. For the experiment, we consider a frame as key frame if it is more than 60% dissimilar to its previous frame. Scale Invariant Feature Transform (SIFT) is one of the most useful local features considered for key frame extraction. The main advantage of this method is that it is invariant to scaling, translation, and rotation, which make the method robust. The method first defines important locations using a scale space of smoothed and resized images and then applies difference of Gaussian functions on these images to find the maximum and minimum responses. To select only the distinct collection of important key points, non-maxima suppression is performed. Also the histogram of oriented gradients is computed by dividing the image into patches to

Table 4 Ranking of clustering algorithms based on stability indices

Video Data	Clustering Algorithm	Ranking based on stability indices					Video Data	Clustering Algorithm	Ranking based on stability indices				
		GCC	MOP	AD	ADM	FOM			GCC	MOP	AD	ADM	FOM
Air force one	KM	2	4	2	1	3	Base jumping	KM	3	1	3	3	2
	SC	5	2	4	3	5		SC	2	2	1	5	4
	AP	3	5	2	5	2		AP	4	4	4	4	5
	DC	4	3	3	4	4		DC	5	3	5	1	3
	CAFKF	1	1	1	2	1		CAFKF	1	1	2	2	1
Bear park climbing	KM	1	2	1	2	1	Bike polo	KM	5	1	2	2	2
	SC	5	5	5	4	3		SC	2	4	3	5	4
	AP	3	4	3	5	5		AP	4	2	4	1	5
	DC	4	3	4	3	4		DC	3	5	5	4	3
	CAFKF	2	1	2	1	2		CAFKF	1	3	1	3	1
Bus in Rock Tunnel	KM	2	2	2	3	2	Car over Camera	KM	1	2	1	3	1
	SC	1	4	3	4	1		SC	3	1	2	5	3
	AP	3	1	4	5	4		AP	4	3	4	4	5
	DC	4	5	5	2	5		DC	4	4	5	2	4
	CAFKF	1	3	1	1	3		CAFKF	2	1	3	1	2
Car rail crossing	KM	1	2	1	3	2	Cockpit Landing	KM	1	3	2	2	1
	SC	3	4	3	5	5		SC	2	5	3	4	5
	AP	4	5	4	1	4		AP	3	1	5	3	3
	DC	5	3	5	4	3		DC	4	4	4	1	4
	CAFKF	2	1	2	2	1		CAFKF	1	2	1	1	2
Cooking	KM	4	2	2	3	1	Eiffel Tower	KM	3	1	2	1	3
	SC	3	5	3	4	3		SC	2	4	4	3	5
	AP	2	4	4	5	2		AP	3	3	5	4	4
	DC	5	3	5	1	4		DC	4	5	3	5	2
	CAFKF	1	1	1	2	3		CAFKF	1	2	1	2	1
Excavators river crossing	KM	1	4	3	1	2	Fire Domino	KM	1	2	2	1	1
	SC	4	2	2	5	3		SC	3	5	4	3	4
	AP	5	3	4	4	5		AP	4	4	5	5	5
	DC	3	4	5	3	4		DC	2	3	3	4	3
	CAFKF	2	1	1	2	1		CAFKF	1	1	1	2	2
Jumps	KM	2	1	1	2	2	Kids Playing in leaves	KM	2	1	2	1	1
	SC	5	5	3	4	3		SC	3	4	5	2	2
	AP	4	4	5	3	4		AP	4	5	3	4	5
	DC	3	3	4	5	5		DC	5	3	4	3	4
	CAFKF	1	2	2	1	1		CAFKF	1	2	1	1	3
Notre Dame	KM	3	2	4	2	3	Paintball	KM	1	2	3	3	2
	SC	1	5	3	3	2		SC	4	4	2	5	1
	AP	4	4	1	5	4		AP	5	3	5	4	3
	DC	2	3	5	4	2		DC	3	5	4	1	4
	CAFKF	3	1	2	1	1		CAFKF	2	1	1	2	1
Paluma jump	KM	1	3	3	2	1	Playing Ball	KM	1	3	4	2	2
	SC	3	4	5	4	4		SC	3	2	2	1	4
	AP	2	5	2	3	5		AP	4	4	3	3	1
	DC	4	2	4	2	3		DC	1	5	5	4	5
	CAFKF	1	1	1	1	2		CAFKF	2	1	1	1	3

Table 4 (continued)

Video Data	Clustering Algorithm	Ranking based on stability indices GCC	MOP	AD	ADM	FOM	Video Data	Clustering Algorithm	Ranking based on stability indices GCC	MOP	AD	ADM	FOM
Playing on water slide	KM	2	4	1	3	3	Saving dolphins	KM	3	2	1	3	1
	SC	3	3	5	1	2		SC	4	3	5	5	3
	AP	4	2	2	2	5		AP	2	1	3	2	5
	DC	1	5	4	4	4		DC	5	4	4	4	4
	CAFKF	2	1	3	1	1		CAFKF	1	1	2	1	2
Scuba	KM	2	1	4	2	4	Statue of liberty	KM	2	2	1	3	2
	SC	3	5	3	4	3		SC	3	4	2	4	1
	AP	4	3	5	1	5		AP	4	5	3	2	1
	DC	2	4	2	1	2		DC	5	3	1	5	4
	CAFKF	1	2	1	3	1		CAFKF	1	1	2	1	3
St Maarten Landing	KM	2	4	1	2	2	Uncut Evening Flight	KM	1	5	2	1	4
	SC	4	2	4	3	1		SC	3	4	4	3	3
	AP	3	3	3	4	3		AP	4	1	5	4	2
	DC	1	5	5	5	4		DC	2	3	3	5	5
	CAFKF	2	1	2	1	1		CAFKF	1	2	1	2	1
Valparaiso downhill	KM	4	5	1	2	3	**Average Ranking of Clustering Algorithms**	KM	1.80	2.44	2.04	2.12	2.04
	SC	3	4	5	3	4		SC	3.08	3.68	3.40	3.68	3.12
	AP	5	2	2	5	1		AP	3.64	3.24	3.60	3.52	3.72
	DC	2	3	4	4	5		DC	3.32	3.72	4.04	3.24	3.76
	CAFKF	1	1	3	1	2		CAFKF	1.40	1.40	1.56	1.52	1.68

Bold face is given to indicate that the best result is obtained in that position

find the dominant orientation of the localized key points, which are selected as local features. In the experiment, we have computed HOGs for each frame in video, and set a threshold such that summary will be 10% of the original video.

The SumMe dataset provides ground truth score to each annotated frames. We evaluate the performance of the proposed method and consider state-of-the-art methods by measuring the Precision (P), Recall (R), and F-score (F) using Eq. (14) – (16), where S_p and S_a are the set of frames in the predicted summary (i.e., summary obtained by a summarization technique) and annotated summary (i.e., ground truth summary), respectively, and F-score is the harmonic mean of Precision and Recall.

$$P = \frac{|S_p \cap S_a|}{|S_p|} \tag{14}$$

$$R = \frac{|S_p \cap S_a|}{|S_a|} \tag{15}$$

$$F = \frac{2PR}{P+R} \tag{16}$$

The results obtained by different methods are listed in Table 5. It shows all Precision (P), Recall (R), F-score (F) and F-score-based P value (FPV) obtained by performing statistical analysis based on Wilcoxon Rank Sum Test [38]. All the best results are marked by bold faces. It is observed from the table that in most of the cases, the proposed summarizer gives the maximum values of P, R, and F. Also, from the last row of the table, it is observed that the average values of all three measures considering all datasets are maximum for the proposed method. The Wilcoxon Rank-Sum Test (nonparameterized) has been performed to check if the proposed video summarization method, i.e., CAFKFVS, is statistically significantly different from considered state-of-the-art methods. It is assumed that the samples used in this test are independent. For this test, if the significance level of p value is $\geq$ 0.05 then the proposed summarization method cannot be considered as statistically significantly different of the state-of-the-art method. If the p value is below the critical level (i.e., 0.05), the decision is to reject the null hypothesis. This statistical analysis is performed based on F-score as it considers both precision and recall. The overall pair wise comparisons between proposed method and state-of-the-art method is done for all the datasets and the *FPV* values are shown in the last column of respective state-of-the-art method in Table 5 and it is observed that the calculated p values are

Table 5 Evaluation of proposed video summarization technique

Video	Summarization	Statistical performance measure metrics				Video	Summarization	Statistical performance measure metrics			
Data	Technique	P	R	F	FPV	Data	Technique	P	R	F	FPV
Air force one	USAM	33	35	34	2.3e-03	Base jumping	USAM	35	37	36	1.6e-03
	IMHS	40	42	41	2.4e-03		IMHS	34	32	33	1.7e-03
	SIFT	39	37	38	2.7e-03		SIFT	39	39	39	1.4e-03
	KMVS	43	43	43	3.9e-03		KMVS	37	38	37	1.5e-03
	CAFKFVS	**46**	**48**	**47**	–		CAFKFVS	**42**	**40**	**41**	–
Bear park climbing	USAM	41	43	42	1.7e-03	Bike polo	USAM	31	32	31	1.9e-03
	IMHS	40	42	41	1.7e-03		IMHS	31	28	29	1.4e-03
	SIFT	44	**46**	**45**	2.5e-03		SIFT	36	31	33	2.5e-03
	KMVS	42	40	41	1.9e-03		KMVS	34	**36**	**35**	0.7e-03
	CAFKFVS	**45**	45	**45**	–		CAFKFVS	**37**	35	**35**	–
Bus in Rock Tunnel	USAM	37	37	37	1.7e-03	Car over Camera	USAM	34	32	33	2.3e-03
	IMHS	39	40	39	2.6e-03		IMHS	35	37	36	1.9e-03
	SIFT	42	41	41	1.9e-03		SIFT	41	38	**41**	1.2e-03
	KMVS	47	46	47	2.1e-03		KMVS	38	**41**	39	1.5e-03
	CAFKFVS	**50**	**48**	**49**	–		CAFKFVS	**43**	39	37	–
Car rail crossing	USAM	36	**38**	37	1.4e-03	Cockpit Landing	USAM	34	33	33	1.3e-03
	IMHS	33	31	32	2.1e-03		IMHS	33	33	33	1.3e-03
	SIFT	34	36	35	1.9e-03		SIFT	40	38	39	1.6e-03
	KMVS	32	30	31	1.7e-03		KMVS	36	37	37	2.6e-03
	CAFKFVS	**41**	**38**	**39**	–		CAFKFVS	**44**	**42**	**43**	–
Cooking	USAM	37	37	37	3.7e-03	Eiffel Tower	USAM	32	31	31	1.9e-03
	IMHS	42	40	41	2.9e-03		IMHS	29	30	29	2.3e-03
	SIFT	49	50	49	3.1e-03		SIFT	31	31	31	1.7e-03
	KMVS	47	46	46	3.4e-03		KMVS	33	**34**	33	2.4e-03
	CAFKFVS	**52**	**51**	**51**	–		CAFKFVS	**35**	33	**34**	–
Excavators river crossing	USAM	30	29	29	2.8e-03	Fire Domino	USAM	31	33	31	3.3e-03
	IMHS	31	33	32	1.6e-03		IMHS	30	32	**39**	3.7e-03
	SIFT	40	38	39	2.9e-03		SIFT	33	31	32	2.8e-03
	KMVS	**42**	41	**41**	2.9e-03		KMVS	**38**	36	37	5.1e-03
	CAFKFVS	41	**42**	**41**	–		CAFKFVS	32	**39**	**39**	–
Jumps	USAM	48	46	47	2.7e-03	Kids Playing in leaves	USAM	44	43	43	1.6e-03
	IMHS	43	44	43	1.5e-03		IMHS	40	42	41	1.4e-03
	SIFT	52	50	51	2.1e-03		SIFT	45	45	45	3.4e-03
	KMVS	66	64	65	2.8e-03		KMVS	48	47	47	4.3e-03
	CAFKFVS	**70**	**72**	**71**	–		CAFKFVS	**52**	**50**	**51**	–
Notre Dame	USAM	31	32	31	0.8e-03	Paintball	USAM	27	28	27	1.9e-03
	IMHS	39	38	38	1.9e-03		IMHS	34	33	33	1.9e-03
	SIFT	45	**47**	**46**	2.7e-03		SIFT	42	40	**43**	3.5e-03
	KMVS	41	41	41	2.2e-03		KMVS	**44**	43	**43**	2.7e-03
	CAFKFVS	**47**	45	**46**	–		CAFKFVS	42	**44**	41	–
Paluma jump	USAM	40	38	39	3.1e-03	Playing Ball	USAM	26	27	26	3.1e-03
	IMHS	36	36	36	3.7e-03		IMHS	24	23	23	3.9e-03
	SIFT	38	**40**	39	3.2e-03		SIFT	37	37	37	5.5e-03
	KMVS	37	38	37	3.5e-03		KMVS	38	**40**	38	5.7e-03
	CAFKFVS	**42**	**40**	**41**	–		CAFKFVS	**40**	39	**39**	–
Playing on water slide	USAM	32	31	31	2.7e-03	Saving dolphins	USAM	36	38	37	1.8e-03
	IMHS	34	36	35	2.6e-03		IMHS	41	41	41	1.9e-03
	SIFT	40	40	40	1.4e-03		SIFT	42	41	41	1.7e-03
	KMVS	42	41	41	1.9e-03		KMVS	**43**	43	39	1.5e-03
	CAFKFVS	**43**	**42**	**42**	–		CAFKFVS	**43**	**44**	**43**	–

Table 5 (continued)

Video	Summarization	Statistical performance measure metrics				Video	Summarization	Statistical performance measure metrics			
Data	Technique	P	R	F	FPV	Data	Technique	P	R	F	FPV
Scuba	USAM	44	43	43	2.3e-03	Statue of liberty	USAM	33	33	33	2.4e-03
	IMHS	43	43	43	2.1e-03		IMHS	36	37	36	2.7e-03
	SIFT	48	46	47	2.9e-03		SIFT	43	42	42	2.5e-03
	KMVS	51	52	51	1.8e-03		KMVS	52	50	51	2.6e-03
	CAFKFVS	**56**	**54**	**55**	–		CAFKFVS	**55**	**54**	**54**	–
St Maarten Landing	USAM	39	38	38	1.3e-03	Uncut Evening Flight	USAM	38	**41**	41	2.7e-03
	IMHS	35	34	34	1.4e-03		IMHS	32	33	32	2.9e-03
	SIFT	48	46	47	1.8e-03		SIFT	32	30	31	2.4e-03
	KMVS	49	49	49	1.7e-03		KMVS	**39**	37	38	1.9e-03
	CAFKFVS	**52**	**51**	**51**	–		CAFKFVS	38	**41**	**43**	–
Valparaiso downhill	USAM	34	32	33	2.9e-03	**Average performance of summarizers**	USAM	35.32	35.48	35.20	–
	IMHS	31	32	31	2.3e-03		IMHS	35.40	35.68	35.64	–
	SIFT	35	37	36	1.9e-03		SIFT	40.60	43.88	40.28	–
	KMVS	44	48	**47**	1.1e-03		KMVS	42.52	42.44	42.16	–
	CAFKFVS	**46**	**49**	**47**	–		CAFKFVS	**44.56**	**44.20**	**45.00**	–

Bold face is given to indicate that the best result is obtained in that position

less than 0.05 in most of the cases, which implies that the probability that the proposed method is statistically significant is more than 95%.

5 Conclusion

Video summarization plays an important role in many video applications. In the literature, there are various methods for key frame-based video summarization. But there is no such universally accepted method available for video summarization that gives better output in all kinds of videos. The summarization viewpoint and perspective are often application dependent. The semantic understanding and its representation are the biggest issues to be addressed for incorporating diversities in video and human perception. Depending upon the changes in contents of the video, the key moments are extracted. As the key moment extraction is dependent on motion feature of the frame, it significantly finds the important events because important event is defined as the duration of time where actions are more. During summarization, audio and video of the recording should not be segmented together because it may not provide us the complete information about an event. In our future work, we will separate audio from video and independently make the summaries of audio and video and based on semantic relationship they will be integrated to generate the final summary of the recording, which will make the summary more presentable and help the user for better understanding about the context of the original video. The most challenging part of the work is to extract MVF (motion vector field) from the original video because the extracted MVF file for the original video becomes too large which lead to high computation time. To improve both the accuracy and efficiency, deep neural network model may be used for key frame extraction, which is the future scope of this paper.

Declarations

Conflict of interest The authors declare that this manuscript has no conflict of interest with any other published source and has not been published previously (partly or in full). No data have been fabricated or manipulated to support our conclusions.

References

1. Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: application to face recognition. IEEE Transactions Patt Anal Mach Intell 28(12):2037–2041
2. Aigrain P, Zhang H, Petkovic D (1996) Content-based representation and retrieval of visual media: a state-of-the-art review. Multimedia Tools Appl 3(3):179–202
3. Brock G, Pihur V, Datta S, Datta S, et al. (2011) clValid, an R package for cluster validation Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta Department of Bioinformatics and Biostatistics, University of Louisville
4. Bruhn A, Weickert J, Schnörr C (2005) Lucas/kanade meets horn/schunck: combining local and global optic flow methods. Int J Comput Vision 61(3):211–231
5. Campo DN, Stegmayer G, Milone DH (2016) A new index for clustering validation with overlapped clusters. Expert Syst Appl 64:549–556

6. Chang IC, Chen KY (2007) Content-selection based video summarization. In: 2007 Digest of Technical Papers International Conference on Consumer Electronics, IEEE, pp 1–2
7. Chau WS, Au OC, Chong TS (2004) Key frame selection by macroblock type and motion vector analysis. In: 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), IEEE, vol 1, pp 575–578
8. Chheng T (2007) Video summarization using clustering. Department of Computer Science University of California, Irvine
9. Cirne MVM, Pedrini H (2013) A video summarization method based on spectral clustering. In: Iberoamerican Congress on Pattern Recognition, Springer, pp 479–486
10. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Ieee, vol 1, pp 886–893
11. Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech: Theory Ex 09:P09008
12. Das P, Das AK, Nayak J (2020) Feature selection generating directed rough-spanning tree for crime pattern analysis. Neural Comput Appl 32(12):7623–7639
13. Deborah LJ, Baskaran R, Kannan A (2010) A survey on internal validity measure for cluster validation. Int J Comput Sci Eng Surv 1(2):85–102
14. Dhawale CA, Jain S (2008) A novel approach towards keyframe selection for video summarization. Asian J Information Technol 7(4):133–137
15. Divakaran A, Peker KA, Radhakrishnan R, Xiong Z, Cabasson R (2003) Video summarization using mpeg- motion activity and audio descriptors. Video Mining. Springer, New York, pp 91–121
16. Fajtl J, Sokeh HS, Argyriou V, Monekosso D, Remagnino P (2018) Summarizing videos with attention. In: Asian Conference on Computer Vision, Springer, pp 39–54
17. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science. 315(5814):972–976
18. Gianluigi C, Raimondo S (2006) An innovative algorithm for key frame extraction in video summarization. J Real-Time Image Process 1(1):69–88
19. Gong B, Chao WL, Grauman K, Sha F (2014) Diverse sequential subset selection for supervised video summarization. Adv Neural Information Process Syst 27:2069–2077
20. Gunsel B, Tekalp AM (1998) Content-based video abstraction. In: Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269), IEEE, pp 128–132
21. Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam
22. Hubert L, Arabie P (1985) Comparing partitions. J Classification 2(1):193–218
23. Jadhava P, Jadhav D (2015) Video summarization using higher order color moments. Proceedings of the International Conference on Advanced Computing Technologies and Applications (ICACTA) 45:275–281
24. Jadon S, Jasim M (2019) Video summarization using keyframe extraction and video skimming. arXiv preprint arXiv:191004792
25. Li C, Wu YT, Yu SS, Chen T (2009) Motion-focusing key frame extraction and video summarization for lane surveillance system. In: 2009 16th IEEE International Conference on Image Processing (ICIP), IEEE, pp 4329–4332
26. Liu T, Zhang HJ, Qi F (2003) A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE Transactions Circuit Syst Video Technol 13(10):1006–1013
27. Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. In: 2010 IEEE international conference on data mining, IEEE, pp 911–916
28. Ma YF, Lu L, Zhang HJ, Li M (2002) A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia, pp 533–542
29. Mundur P, Rao Y, Yesha Y (2006) Keyframe-based video summarization using delaunay clustering. Int J Digital Libr 6(2):219–232
30. Okade M, Biswas PK (2016) A novel moving object segmentation framework utilizing camera motion recognition for h. 264 compressed videos. J Visual Commun Image Represent 36:199–212
31. Pei SC, Chou YZ (1999) Efficient mpeg compressed video analysis using macroblock type information. IEEE Transactions Multimedia 1(4):321–333
32. Rendón E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5(1):27–34
33. Sony A, Ajith K, Thomas K, Thomas T, Deepa P (2011) Video summarization by clustering using euclidean distance. 2011 International Conference on Signal Processing. Communication, Computing and Networking Technologies, IEEE, pp 642–646
34. Srinivas M, Pai MM, Pai RM (2016) An improved algorithm for video summarization-a rank based approach. Procedia Comput Sci 89:812–819
35. Sujatha C, Mudenagudi U (2011) A study on keyframe extraction methods for video summary. In: 2011 International Conference on Computational Intelligence and Communication Networks, IEEE, pp 73–77
36. Tabii Y, Thami R (2009) A new method for soccer video summarizing based on shot detection, classification and finite state machine. In: Proceedings of The 5th international conference SETIT
37. Truong BT, Venkatesh S (2007) Video abstraction: a systematic review and classification. ACM transactions on multimedia computing, communications, and applications (TOMM) 3(1):3–es
38. Wilcoxon F, Katti S, Wilcox RA (1970) Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. Sel Tables Math Stat 1:171–259
39. Wolf W (1996) Key frame selection by motion analysis. In: 1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings, IEEE, vol 2, pp 1228–1231
40. Wu J, Zhong Sh, Jiang J, Yang Y (2017) A novel clustering method for static video summarization. Multimedia Tools Appl 76(7):9625–9641
41. Zhang HJ, Wu J, Zhong D, Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Patt Recognit 30(4):643–658
42. Zhou K, Qiao Y, Xiang T (2017) Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. arXiv preprint arXiv:180100054

Authors and Affiliations

Ghazaala Yasmin[1] · Sujit Chowdhury[2] · Janmenjoy Nayak[3] · Priyanka Das[4] · Asit Kumar Das[4]

✉ Asit Kumar Das
akdas@cs.iiests.ac.in

Ghazaala Yasmin
ghazaala.yasmin@stcet.ac.in

Sujit Chowdhury
sujit_2021cs35@iitp.ac.in

Janmenjoy Nayak
mailforjnayak@gmail.com

Priyanka Das
priyanka.rs2016@cs.iiests.ac.in

[1] Department of Computer Science and Engineering, St. Thomas' College of Engineering & Technology, Kidderpore, Kolkata 700023, India

[2] Department of Computer Science and Engineering, Indian Institute of Technology, Patna, India

[3] Department of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh 532201, India

[4] Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, India