Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-14081-7

1203: APPLICATIONS OF ADVANCED ARTIFICIAL INTELLIGENCE IN MULTIMEDIA AND INFORMATION SECURITY

Experimental analysis of clustering based models and proposal of a novel evaluation metric for static video summarization

Deeksha Gupta[1,2] • Akashdeep Sharma[2] • Pavit Kaur[2] • Ritika Gupta[2]

Received: 20 September 2020 / Revised: 14 July 2022 / Accepted: 23 August 2022

Abstract

Video summarization deals with the identification of relevant and important frames from a video for facilitating efficient storage, browsing and indexing of the videos. Automatic video summarization is a challenging task due to the varied genre, structure and domain of videos. Amongst several video summarization mechanisms, clustering-based approaches have become popular because of their independence from the expensive and time-consuming task of collecting user annotated summaries. The work aims to investigate and compare the behaviour of different nature clustering algorithms and frame descriptors for video summarization tasks. The scope of the presented study is twofold. Firstly, 30 clustering-based video summarization models are implemented and analysed. Secondly, a novel quantitative performance metric, CUS-F (Comparison of User Studies F-Score) is proposed to gauge the quality of generated summaries via a straightforward and concise score value. For comparative evaluation, all experiments are performed on a benchmarking static video summarization dataset - OpenVideo (OV) dataset. The study discovered that DBSCAN clustering shows the best performance when used with local features. The experiments also identifie K-means as a robust clustering method and colour as the most consistent frame descriptor for summarization tasks. In addition, the

✉ Akashdeep Sharma
akashdeep@pu.ac.in

Deeksha Gupta
deeksha.784@gmail.com

Pavit Kaur
pavitk1@gmail.com

Ritika Gupta
ritikagupta8734@gmail.com

[1] Mehr Chand Mahajan DAV College for Women, Panjab University, Chandigarh, India

[2] University Institute of Engineering and Technology, Panjab University, Chandigarh, India

Published online: 15 November 2022

study demonstrates the effectiveness of the proposed evaluation metric CUS-F in obtaining the assessment of automatic key-frame based summaries by considering both true positive and false positive keyframes.

Keywords Static video summarization · Keyframe extraction · Video abstraction · Clustering algorithms · Evaluation metric

1 Introduction

The growth of video capturing technologies, storage mechanisms and multimedia techniques resulted in huge volume digital content. According to a statistical report [38], a total of 500-hour duration videos are uploaded on YouTube every minute. Video is the most powerful digital content in terms of expressiveness due to its characteristic of embedding all other types of media (image, audio, text), but because of its unstructured format, video processing like searching, browsing and indexing etc., has become a very challenging task. This generates the need for development of an automatic mechanism to deal with deluge of video content effectively.

Video summarization deals with the aforementioned concerns, imposed by large video collections, by processing the video to extract important and meaningful frames or skims [40]. These extracted video snippets can be used to replace corresponding video which in terms saves navigation time, memory space and transmission bandwidth constraints significantly. Based upon the form of summary, the summarization system can be classified into two categories: dynamic video summarization (DVS) and static video summarization (SVS) [20]. Dynamic video summarization addresses selection of meaningful excerpts along with temporal information [18]. Whereas, static video summarization results in extraction of diverse but important keyframes from the original video. The resultant keyframe summaries can be presented as storyboard [29], panorama [39], mosaic [42], thumbnail [10] etc. The static video summaries facilitate effective browsing, indexing, retrieval of videos [40]. This makes static video summarization a suitable pre-processing step for other video understanding tasks including anomaly detection in videos [24], action detection and recognition [46] etc. The major concern of the current work is to investigate the behaviour of clustering-based static video summarization models. In existing studies, various clustering based approaches have been proposed employing different types of clustering methods with different kinds of features [3, 4, 9, 19, 23, 25, 27–29, 33, 35, 44, 45]. But there is no consensus on the clustering algorithm as well as frame descriptors that are most suitable for video summarization tasks. Keeping in view the versatile application areas of the static video summaries, we have made an attempt to comparatively analyse the performance of various types of clustering algorithms with different types of features. The foremost purpose of this study is to attain an insight into the behaviour of different clustering methods and features in video summarization-based applications.

The paper also proposes a novel objective metric for keyframe based summary evaluation named CUS-F. CUS-F is an improvement over the popular CUS (Comparison of User Summaries) metric -an objective metric for static summaries evaluation. The main limitation of CUS lies in its inadequacy to represent a specific and concise comparison between two summaries. The proposed metric CUS-F shows its proficiency in evaluating automatic summaries by assigning higher scores to a better-quality summary and vice-versa. The proposed metric can be adopted as a reliable quantitative metric in future research for performance evaluation of static video summarization techniques.

1.1 Motivation of the study

The gaining interest of the research community in the field of video summarization resulted in a number of survey studies targeting different feature selection techniques and frameworks for summary generation [4, 27, 40]. This experimental study is carried out with main focus on the behaviour of various clustering techniques with different feature space. Here, we had made an attempt to evaluate the performance of different models consisting of different combinations of clustering techniques with various frame features. In previous comparative study [33] different clustering-based methods were analyzed but each technique with a different environment. Applying different features maps and processing for different methods fails to provide a comparative analysis on the grounds of clustering method behaviour and feature selection decision. Another attempt John et. al. [23] provides comparative analysis of clustering techniques but with major focus on various colour models. The comparison between existing comparative studies and presented work is specified in Table 1.

The limitations of existing comparative studies lead to a prime motivation for this comparative study. For widening the scope and comprehensive coverage of the experimental study, multiple clustering algorithms and diverse frame descriptors are considered. The selection of six clustering methods and five features resulted in 30 different summarization models for extensive comparative analysis. The paper is targeted to give a fair comparison of various features and clustering technique combinations on the basis of their performance in terms of summary quality.

Major contributions of the study are as follows:

1. An empirical study involving extensive experiments and comprehensive analysis to gauge the behaviour of numerous clustering algorithms, belonging to six different clustering families, is carried out.
2. The study also includes assessment of effectiveness of various scope varied features, such as, global features like colour, texture, GIST and local features like SURF and SIFT.

Table 1 Comparison of present study with existing comparative studies

Reference	Clustering Methods Adopted						Features Used
	Partitioning Based	Soft Computing	Density based	Hirarchical	Graph	Model based	
Sebastian et al. [33] (2015)	K-means, Modified FPF	X	DBSCAN, DT	X	X	X	Color- HSV, Texture- Haar DWT Color- HSV
John et al. [23] (2017)	K-Means	FCM SOM	X	X	X	GMM	Color- RGB, HSV, YCbCr
Ours	K-means	FCM	DBSCAN	AHC	Spectral	GMM	Color- HSV, Texture- Harlick GIST, SIFT, SURF

3. A novel quantitative evaluation metric named Comparison of User Summary- Fscore (CUS-F) is proposed for generating a single score value for the generated summary. The effectiveness of the proposed evaluation metric, CUS-F, is also analysed. Along with CUS-F, predefined evaluation measures - Comparisons of User Summaries Accuracy (CUS-A)/ Recall, Comparisons of User Summaries Error (CUS-E), Precision and F-measure are also employed, for efficacy analysis of all the model's understudy.

The structure of the presented manuscript is as given: Section 2 reviews the existing studies related to clustering-based video summarization mechanisms. Section 3 covers the methodology and taxonomy corresponding to the video summarization process pipeline along with various feature descriptors and clustering methods employed for empirical study. Furthermore, this section also provides discussion about various existing evaluation measures and a proposed evaluation metric CUS-F (Comparison of User Summaries F Score). Section 4 presents the experiment specifications along with the score outcomes. Section 5 includes the comparative analysis of the feature descriptors and clustering algorithms, on the basis of results attained. Section 6, the last section of the manuscript, provides conclusion remarks.

2 Related work

On the basis of the learning mechanism adopted, the vast research work in the direction of video summarization, can be categorized into supervised and unsupervised approaches. Supervised approaches exploit ground-truth annotations for the training of models and exhibit good performance especially for domain specific applications. On the contrary, unsupervised approaches are independent of expensive and hard-to collect ground-truth annotations and, so, are most appropriate for general video summarization. Among unsupervised approaches, clustering-based systems have gained popularity over the years because of the high correlation between clustering mechanisms and the inherent nature of summarization tasks. But the performance of a video summarization system is highly affected by the critical decisions about the model's constituent components like clustering algorithm and feature descriptor selection. The diversity in content, camera position and structure of videos, makes it difficult to select one particular method for generating the most suitable summary for different genre videos.

Under clustering-based approaches Mundur et al. [29] proposed Delaunay Triangulation (DT) clustering that includes Delaunay diagrams construction followed by inter-cluster edge elimination. Avila et. al. [4] adopted K-means clustering algorithm for video abstraction and proposed objective evaluation metrics - CUS. Shroff et. al. [35] proposed modified Kmeans clustering approach by exploiting inter-centroid variance as diversity measure and intra-cluster distance as representativeness measure. Mahmoud et al. [28] proposed VGRAPH approach that includes KNN graph clustering of frames represented with texture feature. Another variant of the proposed approach, VGCOLOUR is also specified where keyframes selection is based on colour feature rather than texture feature. Chamasemani et. al. [9] produced static summaries by presenting DENCLUE clustering approach with multiple features including colour, texture, SURF and energy descriptors. Furini et al. [19] presented Modified Furthest Point First clustering for generating still as well as moving excerpts from a video. Wu et al. [44] proposed High Density Peak Search (HDPS), a density-based clustering method, with colour and texture features. Zhao et al. [45] adopted Affinity propagation algorithm with cluster validity index mechanism. The video is temporally segmented by following Fuzzy Petri-Net model with

Table 2 Clustering -based video summarization approaches

Reference	Clustering Method	Feature	Advantages/Disadvantages
Mundur et al. [29] (2006)	Delaunay Triangulation	Colour	• High computational overhead • Suitable for short and edited videos
Shroff et al. [35] (2010)	modified K-means	BOW with 2D interest point [15]	• Generates diverse and representative summary • Suitable for unedited videos • Linear computational complexity
STIMO [19] (2010)	Modified Furthest Point First (FPF)	Colour	• Produces both keyframes and storyboard • Online video summarization method • Method allows adaptive summarization regarding summary duration and production time
Avila et al. [4] (2011)	K-means	Colour	• Introduces new performance metric - CUS • Suffers with local minima problem • Method is sensitive to outliers
Asadi et al. [3] (2012)	Fuzzy C mean	Colour	• Suitable for long consumer videos • Method performance highly depends upon parameter selection
VSCAN [27] (2013)	DBSCAN	Colour, Texture	• Efficient method for abstracting long videos • Method is robust to outliers
VGRAPH [28] (2013)	KNN graph	Texture	• Sophisticated pre-processing (segmentation) and post-processing is required for comparable results • Trajan Algorithm is employed for strongly connected component detection
Ou et al. [31] (2015)	GMM	Colour	• Clustering is adopted for intra-view summarization
HDPS [44] (2017)	DBSCAN-High Density Peak Search	SIFT	• Employs Singular Value Decomposition for candidate keyframe extraction
Chamasemani et al. [9] 2018	DENCLUE	Colour, Texture, SURF and Energy features	• Robust in handling long duration and noisy videos without further tuning of the model parameters • Fast with quadratic time complexity
Kumar et al. [25] (2018)	K-means	Spatial frequency, Spectral residual and Colour	• Performs real-time video summarization • DBI (Davis-Bouldin Index) method is adopted for estimation of no. of clusters
Zhao et al. [45] (2020)	Affinity propagation clustering	fused pixel level, object level and semantic features	• No need to compute cluster count in advance • Video pre-processing is required for candidate keyframe selection

histogram differences. Kumar et. al. [25] proposed real time summarization by following multi stage clustering where Kmeans clustering is applied separately on the index based prime and non-prime frames. Ou et. al. [31] proposed multi-video video summarization by employing GMM clustering for intra-view summarization. The detail of various clustering-based approaches is provided in Table 2.

For summarization methods' performance assessment, numerous evaluation metrics are employed by existing studies. Among quantitative evaluation approaches, F-measure [28, 43] and Comparison of User Studies (CUS-A and CUS-E) [4, 9, 16], include comparison of automatically extracted keyframes with user annotated ground-truth keyframes. The subjective evaluation measure includes a user study method where the user is required to rank the generated summary on the basis of some predefined criteria. Due to its dependence on only user ranking, this method may suffer from user biasness. To avoid shortcomings from the user study method, our study assesses the effectiveness of underlying models using objective evaluation measures- CUS-A, CUS-E, Recall, Precision and F-measure. Also, a novel evaluation metric derived from CUS-A and CUS-E named CUS-F (Comparison of user summaries F-score) is proposed along with its comparative analysis with previously used metrics.

3 Methodology and taxonomy

A video summarization system aims to reduce the temporal redundancy while retaining the important content from a video. If a video 'V' represented as V= (x1, x2, x3,————, xn), where xi represents ith frame and n denotes total no. of frames, then the video summarization system processes video frames to obtain a condensed summary version 'S', such that:

$$S = f\ (V) = \{x_1, x_2, x_3\text{———}, x_m\}$$

Where f(.) denotes keyframe extraction process, S ⊂ V and x$_i$ in S represents extracted keyframes such that i∊n and m<<n.

The video summarization model pipeline, used in current study, is illustrated in Fig. 1. The summarization model contains three components: video pre-processing, feature extraction and

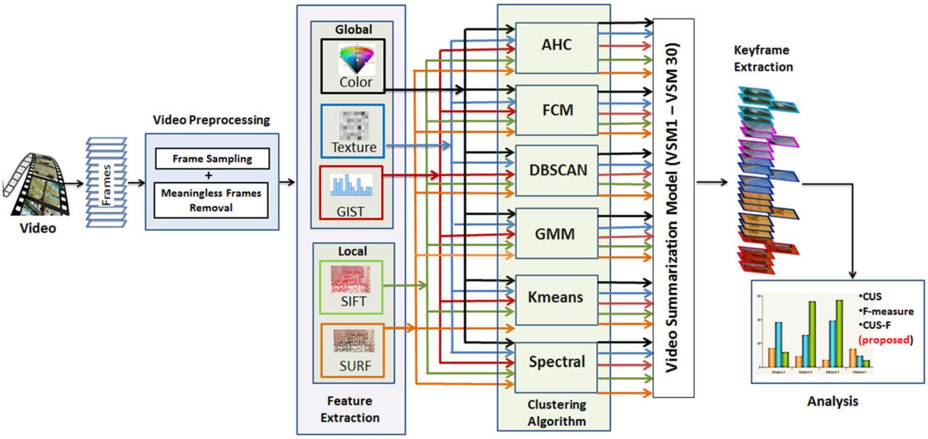

Fig. 1 Video summarization pipeline

clustering algorithm implementation. Major pipeline is followed by keyframe extraction from the cluster generated by the underlying model.

3.1 Video preprocessing

Video preprocessing is the first step of any video summarization model that deals with primitive procedures like video frames' condensation through sampling, temporal segmentation, useless frames removal etc. [11, 17, 34]. In our study, each frame is considered in isolation and down sampling is performed to reduce the number of candidate frames. After downsampling, meaningless monochromatic frames are removed. For this, the frames having standard deviation of colour descriptor frame vectors equal to or close to zero, are identified as monochromatic frames [4].

3.2 Feature extraction

Feature extraction refers to the elicitation of important information, from raw multimedia components, in the form of numeric values. The efficacy of selected feature for frame information representation impacts not only summary quality but also the computational complexity of the model. Here, in total 5 features: colour, texture, GIST, SIFT and SURF are considered for the experiment. The purpose of considering various features is to analyze the effectiveness of global (colour, texture, GIST) and local (SIFT, SURF) scope features in frame representation for video summarization tasks. The global features are used to describe an image as a whole whereas the local features are based on image patches. Global features are utilized in low-level applications, for instance, object detection whereas local level features have shown eminence performance for high-level applications like object recognition. Considering the varied nature and scope of local and global descriptors, both types of features are taken into account in the presented study. The behaviour of features is analyzed on the basis of its frame representation capability and stability, for the video summarization problems. However, explicit feature fusion is not performed. The features-based comparison of various video summarization models is specified in Tables 4, 5, 6, 7 and 8 under Section 5. The aural features, textual features and high-level semantic features are intentionally ignored due to their domain dependence and poor performance for out of specific environmental conditions [29].

3.2.1 Colour descriptor

For experimental study, the user-oriented colour model is preferred from a list of numerous colour models like RGB, YUV, CMYK, HSI etc. Supporting the above said statement, the HSI (Hue- Saturation-Intensity) colour model is selected for frame representation owing to its strong correlation with the human vision perception system [4, 27, 29]. Because of high dominance of Hue component over saturation and intensity components, we have used only Hue components in the summarization approach. To reduce the computational complexity the colour histogram is quantized to 16 bins. Thus, Hue component histogram represented as a normalized 16-D vector is adopted as colour descriptor for experiments.

3.2.2 Texture

Texture feature is one of the fundamental descriptors widely accepted in computer vision applications [21, 27]. This work considers texture features to observe the independent behaviour and efficacy of features in the context of video abstraction while employing different clustering models. Under the presented study, 13-D statistical Haralick texture feature representations are used owing to its fast and intuitive nature [22].

3.2.3 SIFT

Scale Invariant Feature Transform (SIFT) descriptors gained popularity since their inception, due to their robustness against object position, scale, orientation, illumination and minor image artifacts like blur and noise [26]. The Bag of Visual Words (BOVW) model is one of the extensively applied methods in various computer vision problems ranging from image annotation [41] to Content Based Image Retrieval (CBIR) [14].

In the current study, the BOVW approach along with SIFT local features is implemented. From every sampled frame, SIFT local descriptors are extracted and clustered to construct a codebook of key-points. The created codebook is exploited for representation of frames as a normalized frequency distribution of visual words. For abstraction purposes, to balance the performance and computational cost trade-off, codebook size is set equal to 1024D [1].

3.2.4 SURF

Speeded Up Robust Features (SURF) descriptors represent local information inside an image just like SIFT features introduced by [5]. For SURF extraction, Gaussian derivative masks are applied at various scales of integral image, making them significantly faster than SIFT. For SURF descriptor based frame representation also, the BOVW approach with the 1024D codebook, is used. The frames are represented as a normalized frequency histogram of SURF descriptors according to the descriptor distribution of the key-points. Static summaries of videos are generated by using this histogram as input to the clustering algorithm of choice.

3.2.5 GIST

GIST descriptors are global features representing the scene information of the video frame [30]. According to [30], scene recognition features can be defined without extracting individual objects of the scene. GIST contains gradient information simulating the description of the frame as a whole. The extraction of GIST features involves convolution of video frames in RGB colour space with Gabor filters at various scales and orientations followed by concatenation of averaged feature values of the regions obtained from feature maps.

During the experiment, 3 scales, 20 orientations per scale and region size of 4x4 are selected for GIST descriptor extraction, resulting in a 960-Dimensional vector. The 960D GIST descriptor is further processed for dimensionality reduction using the Singular Vector Decomposition (SVD) method. The dimensionality reduction of GIST features is performed to make the evaluation of GIST comparable to other global scope features, colour and text.

3.3 Clustering

Clustering refers to grouping of related data objects into clusters while emphasizing on inter-cluster heterogeneity and intra-cluster homogeneity. Supporting variation in data distribution patterns, different clustering approaches have been devised covering Partition based, Density Based, Soft computing based, Model Based, Hierarchical and Graph Based Clustering [6, 8].

Clustering step in the proposed pipeline includes selection of the clustering method along with its parameters. Various methods under one category consider analogous assumptions about the data points (for example- Kmeans and K-medoid clustering methods), follow similar mathematical models (for example- graph clustering and spectral clustering) or similar expansion procedure (for example- OPTICS density clustering, DBSCAN clustering). This study is an effort towards the analysis of performance and behaviour of different clustering algorithms in the context of static video abstraction. The various clustering families and their corresponding selected algorithm, for experiments, are listed in Table 3. The algorithm selection from each category is influenced by various factors like popularity, citations, applicability, potential of handling high dimensional data and number of research studies where these techniques have been used.

3.3.1 Kmeans clustering

Under partition-based clustering, Stochastic Kmeans clustering algorithm was selected solely for the reasons of its popularity, citations, simplicity and wider acceptance. Kmeans starts with random initialization of cluster centres. The cost function optimized for data points grouping into clusters is mentioned in Eq. (1), which is based upon the Euclidean distance between frames and centroid.

$$\mathit{Cost\ Function}\ (J) = \sum_{j=1}^{k}\sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{1}$$

Here $x_i^{(j)}$ = Frame feature representation, c_j = Cluster centre and ‖. ‖ = Euclidean distance metric.

The implementation of Kmeans algorithm demands no. of clusters parameter (‘k’) to be specified in advance. In presented study, The ‘k’ parameter value depends upon the video content and is computed by using consecutive frame-pair differences. In particular, for the OpenVideo (OV) dataset videos, the ‘k’ value lies within the range 3-26.

For computing ‘k’, pair-wise consecutive frame difference is computed and if the difference exceeds the *threshold* value, then k is incremented. Consecutive frame-pair distance is computed using the Euclidean distance metric. For estimating reasonable no. of clusters, the *threshold* value is computed as follows:

$$\mathit{threshold} = \alpha\ \mu - \beta \tag{2}$$

Where, μ = mean of all consecutive frame-pair difference within a video α =1.6 (constant and empirically driven) and β =0.03 (constant and empirically driven)

Table 3 Clustering types and selected algorithms

Category	Selected Algorithm
Partition Based Clustering	Kmeans
Soft Computing Based Clustering	Fuzzy C Mean (FCM)
Density Based Clustering	Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Hierarchical Clustering	Agglomerative Hierarchical Clustering (AHC)
Model Based Clustering	Gaussian Mixture Models (GMM)
Graph Based Clustering	Spectral Clustering (SC)

The aforementioned procedure is followed in all other clustering techniques that are used in the study and required to specify the number of clusters parameters a priori, for instance, FCM, AHC, GMM and SC.

3.3.2 Fuzzy C mean (FCM) clustering

FCM clustering [37] is one of the famous soft computing based clustering that has proved its successful implication in various applications like image analysis, medical imaging and target recognition. FCM clustering algorithm permits a frame to be associated with multiple clusters based upon their membership bound. This fact results in overlapped clusters and hence the method is also called an overlapping clustering method. The cost function to be minimized is represented in Eq. (3).

$$Cost\ Function\ (J) = \sum_{b=1}^{k}\sum_{a=1}^{n} u_{ab}^{m} \left\|x_{a}^{(b)} - c_{b}\right\|^{2} \quad \text{where } 1.0 < \text{m} < \infty \tag{3}$$

Here u_{ab}^{m} corresponds to the membership scale of frame x_a within cluster b holding fuzziness of membership grade (m). As depicted from Eq. (4), u_{ab}^{m} is inversely correlated to the distance between a frame and the cluster center.

$$u_{ij}^{m} = \frac{1}{\sum_{n=1}^{k}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \tag{4}$$

Where, d_{ij} denotes the distance between i^th and j^th cluster centres. The mechanism adopts greedy search approach for membership matrix construction as well as centroid selection with sensitivity threshold set equal to 1e-3. The cluster count value is approximated using the same approach as specified under Kmeans algorithm.

3.3.3 Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is based upon the two considerations about the cluster centre point [12]. First, among all neighbours, the centre point has higher vicinity density and second, every centre point is adequately far away from the other frames having equivalent higher vicinity density. This algorithm makes use of two parameters: neighbourhood distance (eps), which defines the inter-frame distance assumed as neighbourhood distance and neighbourhood density (Minpts) that describes the minimum no. of neighbours allowed around a cluster centre inside the ‘eps’ radius. In the presented study, ‘eps’ is calculated as follows:

$$\text{eps} = \alpha * \max(\text{dist}) \tag{5}$$

where dist represents the distance between consecutive frames of the video and α is set at 0.45 empirically. The second parameter, Minpts, is set equal to 3 empirically to obtain viable performance. Inter-frame distance is computed by employing the Euclidean distance metric. As last step, the middle frame in sequence from each cluster, is extracted as a keyframe. The independence of DBSCAN algorithm from prior cluster count estimation makes it popular over Kmeans algorithm. Also, the computations of parameters based on the video frames features make the algorithm adaptive to the video content and help to achieve good performance.

3.3.4 Agglomerative hierarchical clustering (AHC)

Agglomerative Hierarchical Clustering is a bottom-up hierarchical clustering technique [13] where data frames are grouped by exploiting inter-frame distance as an affinity matrix. Euclidean distance is adopted to compute inter-frame distance and like Kmeans and FCM clustering algorithms the number of clusters are estimated a priori. Ward's minimum variance method, optimized to minimize the intra-cluster variance, is used for grouping of frames into clusters.

Figure 2 represents the dendrogram diagram obtained for video 69 of the OpenVideo dataset. The x axis represents the truncated clusters for sampled 121 frames. The y axis represents the inter-cluster distance. The black line drawn between the dendrogram specifies the threshold distance considered for generating the desired number of keyframes.

3.3.5 Gaussian mixture model (GMM)

Model-based clustering methods act as a general framework to approximate the most probable parameters of an underlying distribution to the given data. Under model-based approaches, GMM Clustering is selected for an experiment which is based upon the supposition of Gaussian distribution of data points [32]. For each cluster formation, the Expectation–Maximization optimization algorithm is employed for Gaussian distribution mean and variance parameters approximation.

$$P(C_k|f_i) = \frac{P(f_i|C_k)*P(C_k)}{P(f_i)} \tag{6}$$

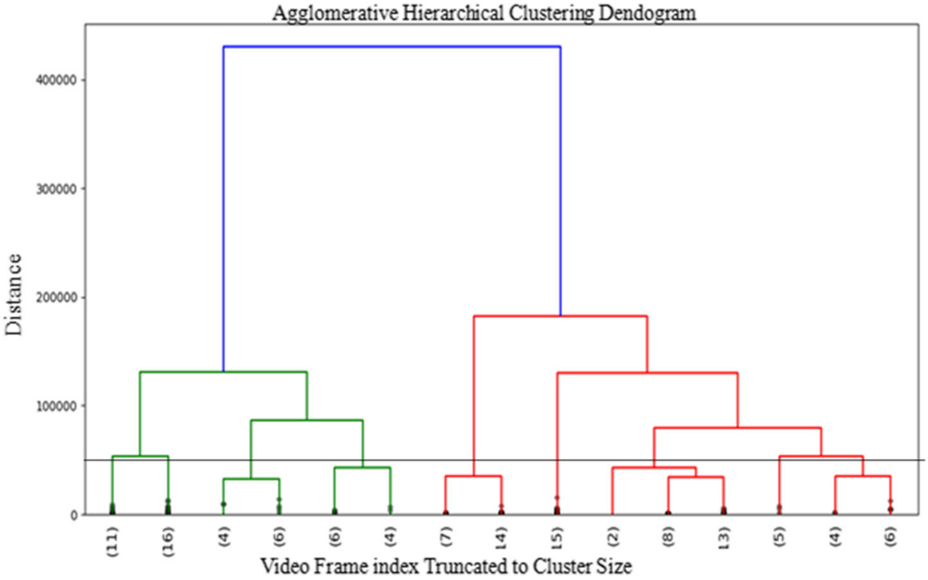

Fig. 2 Agglomerative Hierarchical Clustering generated Dendogram for V69 video of OpenVideo Dataset

Where

$$P(C_k) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left(\frac{-(f_i-\mu_k)^2}{2\sigma_k^2}\right)} \tag{7}$$

Hence,

$$\mathrm{P}(f_i) = \sum_{k=1}^{n} p(f_i| \, C_k \,)^* \, \mathrm{p}(C_k) \tag{8}$$

First, following the common approach as mentioned under 3.3.1, the number of clusters is computed. For each cluster, gaussian parameters are randomly initialized. Finally, the membership probability of video frames is estimated by following Eq. (8).

3.3.6 Spectral clustering (SC)

Graph-based clustering integrates the power of graphs with data points proximity knowledge to cluster similar samples together. Spectral clustering [2], unlike Kmeans, is based upon the computation of adjacency relation ($A \in R^{n \times n}$) between frames for arbitrary shape cluster formulations. The current study employs radial basis function kernel based Laplacian matrix construction for affinity matrix computation. The high dimensional frame feature representations are transformed into low dimensional latent space by exploiting Laplacian matrix eigenvalues and eigenvectors values. The latent space representations are grouped into clusters. This mechanism makes the clustering algorithm free from prior information about the number and density of clusters.

3.4 Evaluation

Evaluation metrics act as a critical tool for analysing the behaviour of models under evaluation. The effectiveness of any evaluation method depends upon its competence to discriminate the model outcomes and provides a clear insight to model behaviour. Current work uses five quantitative evaluation metrics named CUS (CUS-A, CUS-E), Precision and F-measure, along with the proposed metric (CUS-F).

Comparison of User summaries (CUS) [4, 9, 16, 19] and F-measures [28, 43] have been used extensively for the evaluation of static summaries. CUS-F is proposed to cope with the limitation of the CUS metric as discussed in Section 3.4.3. The use of several evaluation methods helps to investigate the models' performance from various facets.

3.4.1 Comparison of user summaries (CUS)

Comparison of User summaries (CUS) encompasses two inherent metrics, named CUS-A (CUS-Accuracy) and CUS-E (CUS-Error) and defined by Eqs. (9) and (10) respectively, for comparing the quality of automatic summary (AS) with respect to user Summary (US).

$$CUS{-}A = \frac{No.of\ \ matched\ \ keyframe\ from\ AS}{Total\ no.of\ \ keyframes\ \ in\ US} \tag{9}$$

$$CUS{-}E = \frac{No.of\ \ non{-}matched\ \ keyframe\ from\ AS}{Total\ no.of\ \ keyframes\ \ in\ US} \tag{10}$$

Where 0 ≤ CUS-A ≤ 1, 0 specifies the worst case, that is, no keyframes of AS match with ground truth US and 1 indicates the best case, that is, all keyframes of US are also present in AS. On the other side, 0 ≤ CUS-E ≤ . Where = AS Keyframes Count/ US Keyframes Count. 0 value of CUS-E represents the best case, that is, all Keyframes of AS are also present in ground truth US, while the CUS-E value equal to specifies that none of the AS Keyframe matches with ground-truth keyframes. CUS-A and CUS-E metrics are complementary in nature, that is, the best performing algorithm should exhibit high value for CUS-A and low value for CUS-E.

Computation of number of matched keyframes for evaluation purpose:

The numbers of matched and unmatched keyframes are computed by matching keyframes from automatic summary and user summary [4]. For this purpose, two keyframes are compared, taking one keyframe from each summary. Manhattan distance is employed for measuring the distance between the keyframes represented in the HSV colour histogram. Keyframes are assumed to be matched if the distance between them is below a prespecified threshold. If two keyframes are found to be similar, they are removed from the next iteration of the comparison process. The threshold value is set equal to 0.5 empirically.

3.4.2 F-measure

Another popular objective evaluation metric is F-measures which comprises two primitive scores Recall and Precision. For static video summarization, Recall and Precision can be obtained by following Eqs. (11) and (12) respectively and F-measure is computed as the harmonic mean of Recall and Precision, as specified in Eqs. (13).

$$Recall = \frac{No.\ of\ matched\ keyframe\ between\ AS\ \ and\ US}{Total\ no.of\ \ keyframes\ in\ US} \tag{11}$$

and

$$Precision = \frac{No.\ of\ matched\ keyframe\ between\ AS\ \ and\ US}{Total\ no.of\ \ keyframes\ in\ AS} \tag{12}$$

So,

$$F\text{-}measure = \frac{2(Recall{*}Precision)}{Recall + Precision} \tag{13}$$

Some works in literature have used the F-measure metric for evaluation of static video summarization methods [28, 43]. So, in this study, we will validate our proposed evaluation metric-CUSF with an F-measure score.

3.4.3 Comparison of user summary – F score (CUS-F)

This paper proposes a new static video summary evaluation metric which is inspired by the need of overcoming the limitations of aforementioned evaluation measures like CUS and F-measure. The main limitation of CUS is that it becomes difficult to quantify any algorithm due to unavailability of a single score value amassing the essence of both accuracy and error measure. CUS provides two metrics and both are complementary in nature. Lack of single condense value makes it quite challenging to gauge the behaviour of different models. While

F-measure does not consider the non-matched keyframes, that is, it ignores the false positive keyframes selected by the model.

Defining CUS-F is an attempt to provide a single condensed score derived from two complementary measures CUS-A and CUS-E. CUS-F is obtained by normalization of CUS-E followed by harmonic mean between CUS-A and normalized CUS-E.

Normalization of CUS-E CUS-E values are normalized to limit the value range between [0,1] and also to complement their nature. Normalization by actuals method [7], represented in Eqs. (14), is followed to drive normalization for CUS-E.

$$Normalized\_V = V^{-c} \tag{14}$$

Where, c is a constant.

So, the normalized value (CUS_E_{norm}) is obtained using the formula mentioned below:

$$CUS\_E_{norm} = (1 + CUS - E)^{-1} \tag{15}$$

After computing normalized CUS-E, CUS-F can be computed as the harmonic mean of CUS-A and normalized CUS-E, as given below:

$$CUS{-}F = \frac{2(CUS\_A * CUS\_E_{norm})}{CUS\_A + CUS\_E_{norm}} \tag{16}$$

The algorithm 1 describes the computation of CUS-F using automatic summary keyframes and Ground truth User Summary keyframes.

Algorithm CUS-F Computation

Algorithm 1 Computation of CUS-F

```
Input: Static Automatic Summary (AS) and Ground-truth User Summary (US)
Output: CUS-F metric value
Let n_AS= no. of keyframes in AS, n_US= no. of keyframes in US
1. Represent keyframes from AS and US, using 16 bin histogram of hue component in HSV
   colour space.
2. Compute the distance between each AS keyframe and US keyframe pair using the
   Manhattan distance metric.
3. If the computed distance is less than predefined threshold (empirically obtained value is
   0.5) then, remove matched keyframes pair for next iteration
4. Repeat steps 2 and 3 until the last pair of AS and US keyframe.
5. Compute Number of match keyframes from AS and no. of non-matched keyframe from
   AS
6. Compute CUS-A= Number of matched keyframes from AS/ n_US
7. Compute CUS-E= Number of non-matched keyframe from AS/ n_US
8. Compute $CUS\text{-}E_{norm}$, using equation (15)
9. Compute CUS-F as harmonic mean of CUS-A and $CUS\text{-}E_{norm}$ as per equation (16)
```

The value of the CUS-F metric lies in the range [0,1] and the high value of CUS-F indicates better summary quality. The comparison and analysis of CUS-F, with respect to other evaluation methods, is provided in the next section for 30 different models operating on Open Video dataset.

4 Experiment and results

This section covers experimental environment specifications like the dataset used, summarization model under consideration as well as performance outcomes.

4.1 Experiments

4.1.1 Dataset

In the study, The OpenVideo (OV) [4] dataset is employed for comparative analysis of various clustering-based models. OV is a benchmark dataset for static video summarization that contains 50 videos from Open Video Project [36]. The included videos cover diverse genres like educational videos, documentary videos, ephemeral videos, historical videos etc. The videos' durations range from 1 to 4 minutes and the total video collection duration of the OV dataset is 75 min. The dataset encompasses five keyframe summaries for each video to address the subjectivity issues of the summarization task.

4.1.2 Video summarization models (VSM)

In presented study, the combinations of each of 6 clustering methods (AHC, FCM, DBSCAN, GMM, Kmeans and SC) and 5 features (colour, texture, GIST, SIFT and SURF) result in 30 clustering method-frame descriptors pairs. These pairs-based models are termed as VSM and hence the paper represents and compares the performance of each of 30 VSM labelled from VSM1 to VSM30. All models are assessed on the basis of five different objective evaluation metrics. The intensive investigation of models on the basis of various aspects like best model and worst model, top performer clustering algorithm and features, consistent clustering model and feature, are provided under Section 5.

4.1.3 Evaluation results

In the presented work, four existing (CUS-A/Recall, CUS-E, Precision, F-measure) and one proposed (CUS-F), total five metrics are exploited for assessment of the summarization models understudy. For evaluation purpose, the automatic summary generated by a model is compared with each of the five ground-truth summaries and the mean of five comparison scores is computed to quantify summary quality. Finally, the average is taken over all the 50 videos of the dataset and the resulting score values are specified in Tables 4, 5, 6, 7 and 8, covering feature-wise performance of all clustering techniques under study.

Table 4, 5, 6, 7 and 8 cover 30 different video summarization models (named VSM1 to VSM 30) categorized into 5 classes on the basis of feature representations taken into consideration. The maximum score obtained for evaluation metric- F-measure and CUS-F, under any class are specified as **bold**.

Table 4 Colour feature based models performance

Model No.	Model (Feature + Clustering)	CUS-A/ Recall	CUS-E	Precision	F-measure	CUS-F (Proposed)
VSM1	Colour + AHC	0.84	0.90	0.59	0.69	0.65
VSM2	Colour + DBSCAN	0.71	0.36	0.71	**0.71**	**0.72**
VSM3	Colour + FCM	0.83	0.91	0.59	0.69	0.64
VSM4	Colour + GMM	0.80	0.94	0.56	0.66	0.63
VSM5	Colour + Kmeans	0.85	0.89	0.60	0.70	0.65
VSM6	Colour + SC	0.83	0..87	0.59	0.69	0.65

Table 5 Texture feature based Models Performance

Model No.	Model (Feature + Clustering)	CUS-A/ Recall	CUS-E	Precision	F-measure	CUS-F (Proposed)
VSM7	Texture + AHC	0.92	1.70	0.43	0.59	0.53
VSM8	Texture + DBSCAN	0.74	0.51	0.66	**0.69**	**0.70**
VSM9	Texture + FCM	0.88	1.73	0.41	0.56	0.52
VSM10	Texture + GMM	0.84	1.78	0.39	0.53	0.50
VSM11	Texture + Kmeans	0.80	1.01	0.52	0.63	0.61
VSM12	Texture + SC	0.91	1.56	0.46	0.61	0.55

Avila et al. [4] proposed two metrics CUS-A and CUS-E which are complementary in nature, that is, for a good quality summary CUS-A should be high while CUS-E should be low. This makes these two metrics inadequate for comparative analysis of summarization approaches when used individually or together. Referring to Tables 4 and 5, the CUS-A and CUS-E metrics pair scores for models VSM1 (Colour + AHC) and VSM7 (Texture + AHC) are (0.84, 0.90) and (0.92, 1.70) respectively. It is quite uncertain to select an absolute winner among aforementioned models only on the basis of CUS-A and CUS-E as though VSM7 exhibits higher CUS-A score but it also possesses higher CUS-E value, which is totally undesirable.

So, the custom CUS-F measure is computed using algorithm 1, for computing a single score value to compare VSM1 and VSM7 and assigns a score value 0.65 and 0.53 respectively, determining VSM1 an apparent winner.

The metric F-measure also shows equivalent behaviour (VSM1 = 0.69, VSM7 = 0.59) while comparing performance and hence validates the proposed condense metric CUS-F. Rather at times, CUS-F metric considers false positive more effectively as compared to F-measure. This fact is justified from performance metric values for VSM1 and VSM3 in

Table 6 GIST feature based Models Performance

Model No.	Model (Feature + Clustering)	CUS-A/ Recall	CUS-E	Precision	F-measure	CUS-F (Proposed)
VSM13	GIST +AHC	0.82	0.92	0.57	0.68	0.64
VSM14	GIST + DBSCAN	0.41	0.19	0.76	0.53	0.55
VSM15	GIST + FCM	0.75	0.78	0.59	0.66	0.64
VSM16	GIST + GMM	0.79	0.96	0.54	0.64	0.62
VSM17	GIST + Kmeans	0.82	0.92	0.57	0.68	0.64
VSM18	GIST + SC	0.83	0.90	0.59	**0.69**	**0.65**

Table 7 SIFT feature based Models Performance

Model No.	Model (Feature + Clustering)	CUS-A/ Recall	CUS-E	Precision	F-measure	CUS-F (Proposed)
VSM19	SIFT + AHC	0.82	0.92	0.58	0.68	0.64
VSM20	SIFT + DBSCAN	0.78	0.41	0.71	**0.74**	**0.74**
VSM21	SIFT + FCM	0.32	0.13	0.76	0.45	0.47
VSM22	SIFT+ GMM	0.71	1.03	0.50	0.59	0.58
VSM23	SIFT + Kmeans	0.76	1.00	0.53	0.63	0.60
VSM24	SIFT + SC	0.53	0.61	0.59	0.56	0.57

Table 4. For VSM1 (Colour + AHC), the F-measure score value is the same 0.69. Even though VSM1 has higher CUS-A and lower CUS-E value in comparison. But our proposed CUS-F assigns a higher score to VSM1 than that of VSM3.

Next section covers the inclusive examination of the obtained experimental outcomes with the help of comparison charts. The visual summary of two videos (V47 and V69) from the Open Video dataset is provided as Appendix A, for justification of obtained results and understanding of readers. The appendix contains keyframe summaries generated by each of 30 models along with their CUS-F score. The appendix also contains figures showcasing standard deviation in CUS-F scores for various clustering methods and features for the mentioned videos.

5 Comparative analysis

The effect of feature representation and clustering algorithm selection is investigated by implementing 30 individual video summarization models. Under this section, the obtained experimental results, as mentioned in Figs. 3, 4, 5 and 6, are carefully assessed on the basis of behaviour analysis of diverse features and clustering algorithms as well as consistency study of features and clustering methods. The results displayed in Tables 4, 5, 6, 7 and 8 are the average metric score over all videos of the dataset. For mitigating the average behaviour of the model with video specific performance, keyframe summaries of two videos V47 and V69 of Open Video dataset [36] are provided in Appendix *A*. *Appendix* contains 5 tables - Table A.I to Table A.V where each table contains a static summary of videos for all clustering methods with corresponding features.

Table 8 SURF feature based Models Performance

Model No.	Model (Feature + Clustering)	CUS-A/ Recall	CUS-E	Precision	F-measure	CUS-F (Proposed)
VSM25	SURF + AHC	0.82	0.92	0.58	0.68	0.64
VSM26	SURF + DBSCAN	0.83	0.49	0.69	**0.75**	**0.74**
VSM27	SURF + FCM	0.37	0.20	0.70	0.49	0.51
VSM28	SURF + GMM	0.77	0.98	0.54	0.63	0.61
VSM29	SURF+ Kmeans	0.80	0.95	0.56	0.66	0.62
VSM30	SURF + SC	0.51	0.57	0.59	0.54	0.56

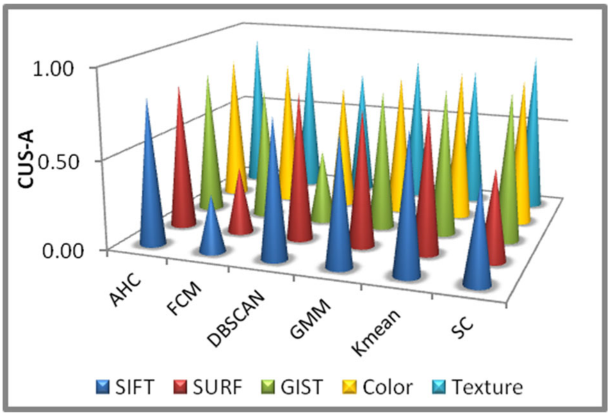

Fig. 3 CUS-A for all models (VSM1-VSM30)

5.1 Best model and worst model

Under this section, two evaluation metrics, CUS-F and F-measure, scores are considered for analysis of model performance. The experiment results discover that, VSM20 (SIFT + DBSCAN) and VSM26 (SURF+DBSCAN) model with CUS-F and F-measure values (0.74, 0.74) and (0.74, 0.75) respectively, exhibits best performance. This outcome witnesses the competency of DBSCAN clustering in managing and handling high-dimensional data effectively. Same behaviour can be verified from the visual keyframe summaries for videos V47 and V69 provided in Appendix A, where from Table A.IV, 3rd row, the SIFT+DBSCAN model is showing best performance for V69 while being the first runner up for V47, following colour + DBSCAN model.

On the other side, from Tables 4, 5, 6, 7 and 8, VSM21 (SIFT+FCM) shows the poorest behaviour followed by VSM27 (SURF + FCM) performance. The visual summary of video V69, presented in Table A.IV, row 2, of Appendix A, indicates the poor performance of SIFT+ FCM with very few and redundant keyframe selection, owing to the sensitivity of FCM clustering towards noise and outliers present in the video frame representations. Although for

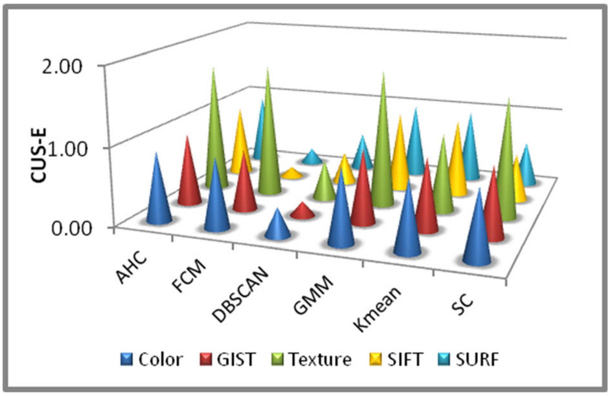

Fig. 4 CUS-E for all models (VSM1-VSM30)

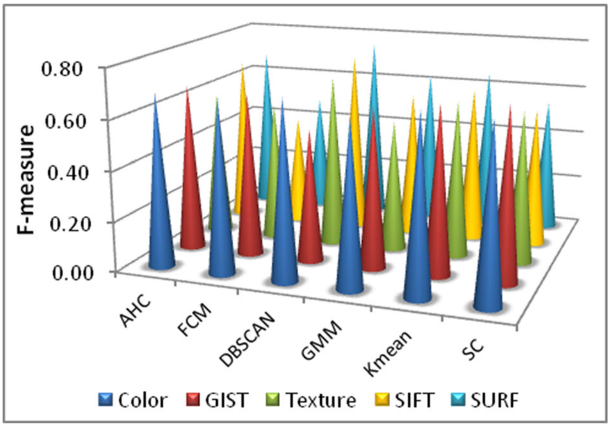

Fig. 5 F-measure for all models (VSM1-VSM30)

V47 visual summary, from Appendix, SIFT+SC model gives the lowest CUS-F value but in general the performance of SIFT+SC is better than SIFT+FCM and SURF+FCM model.

Further, as represented in Table 5, the high values of CUS-E metric for models VSM7 (Texture + AHC), VSM9 (Texture + FCM), and VSM10 (Texture + GMM) indicates the inadequacy of texture features in capturing human perception related frame diversify and hence resulting in large number of redundant and non-matched keyframes extraction, which also reduces the CUS-F score, as shown in Fig. 8. This behaviour of texture features can be verified from keyframes summaries provided for video V47 and V69 in Appendix A, Table A.II, where large numbers of keyframes are selected in the final summary by texture features based models.

Although DBSCAN outperforms other clustering algorithms while used with local features, its performance drops when used with texture features, as shown by metric score in Tables 4, 5, 6, 7 and 8 and visual summaries with CUS-F score of two videos provided in Appendix A, Table A.I to Table A.II. From Fig. 7, observing the overall behaviour of clustering methods, DBSCAN outperforms all other clustering techniques followed by Kmeans and AHC clustering.

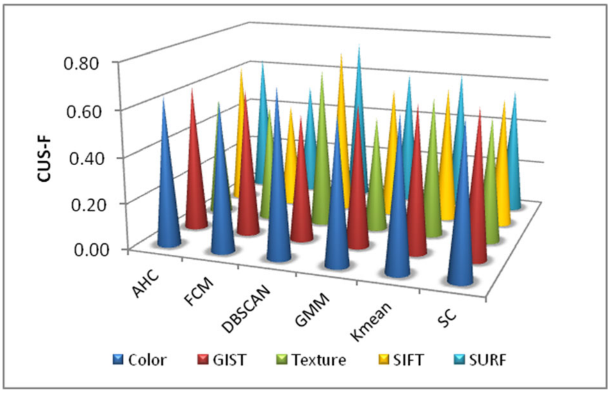

Fig. 6 CUS-F for all models (VSM1-VSM30)

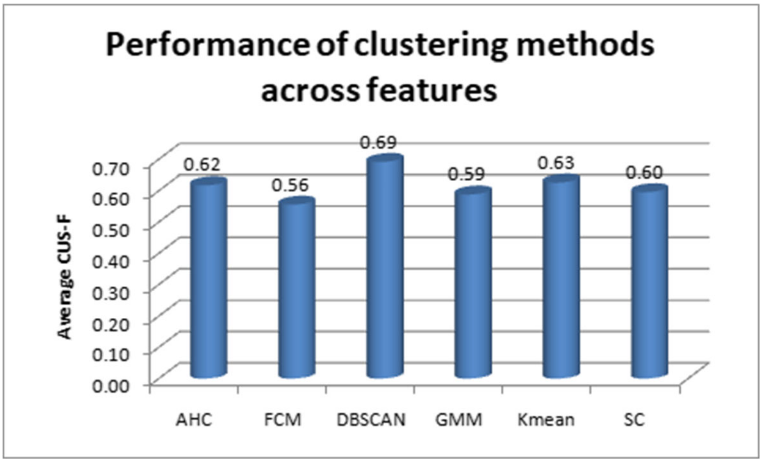

Fig. 7 Average CUS-F score of Clustering Algorithms across Features

5.2 Local v/s global features

Figure 8 depicts the behaviour of various feature descriptors across all clustering algorithms. Colour features give best results despite being the lowest dimensional descriptor in the experiment. This observation manifests the property of high correlation of colour feature representations with human vision perception. Another global feature, GIST, follows colour in performance but this feature acts awkwardly with DBSCAN clustering due to the limitation of DBSCAN to handle varying density clusters. From Figs. 3 and 4, although texture feature shows best and consistent behaviour through CUS-A owing to selection of true positive keyframes but the high value of CUS-E specifies large number of non-matched keyframes in automatically generated summary, hence attenuating the effectiveness of the summarization model.

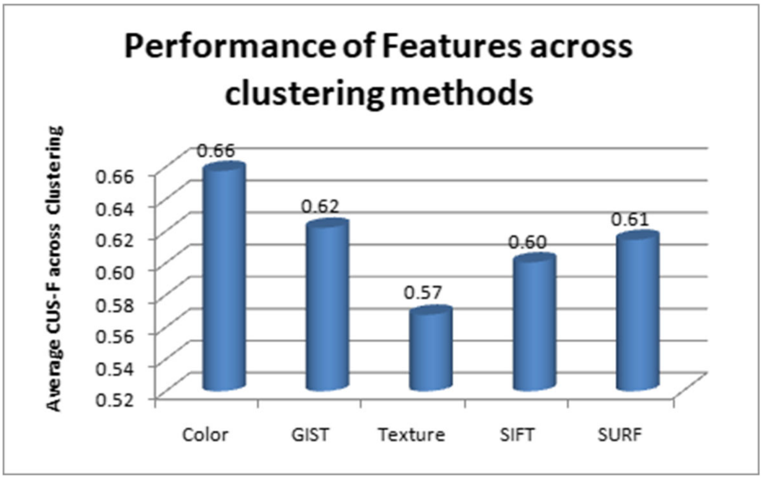

Fig. 8 Average CUS-F score of Features across Clustering Algorithms

This observation implies that texture feature alone is insufficient to capture inter-frames diversity. Further, the local feature, SURF and SIFT, based summarization models exhibit sensitive behaviour towards clustering methods adopted as the local features show moderate performance with GMM and FCM clustering but generate highly quality summaries while employed with DBSCAN algorithm. Although comparing SIFTS and SURF, SURF descriptor is performing slightly better than SIFT. The same observations can be made from Keyframe summaries of videos V47 and V69 provided in Appendix A. Where all colour features based models (Appendix A, Table A.I) provide comparable results. GIST features show analogous behaviour with that of colour feature, with the DBSCAN model as an exception. GIST performance drops with DBSCAN with videos having gradual change in content, as shown by static summaries of V69 in Table A.III, row 3 of Appendix A. Texture based summaries result in a large number of Keyframes for both of the videos (Appendix A, Table A.II). SIFT and SURF based static summaries illustrate drastic change in number and diversity of selected Keyframes with underlying clustering technique (Appendix A, Table A.IV and Table A.V).

5.3 Consistency study of clustering algorithms

Under this section, the consistency in performance of a clustering algorithm is investigated against different frame feature representations. For measuring consistency, the standard deviation of the performance of the clustering method with diverse features is computed. A low standard deviation specifies the least change in performance of the clustering method with a change of frame representation descriptor. This signifies the consistent and robust nature of the method. As interpreted from Fig. 9, Kmeans clustering exhibits the most consistent behaviour followed by Spectral Clustering, While FCM and DBSCAN clustering techniques show the most inconsistent behaviour. Furthermore, AHC, GMM and SC techniques lie somewhere in between the DBSCAN and Kmeans in terms of consistency. The consistent behaviour of Kmean can be verified by standard deviation values of the method for videos V47 and V69, displayed in Fig. A.I of Appendix A. Kmean shows minimum standard deviation among all other clustering methods which verifies its steady performance across various feature maps.

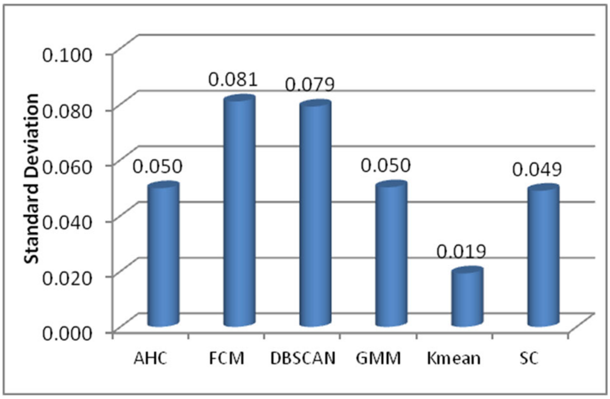

Fig. 9 Standard deviation of clustering method

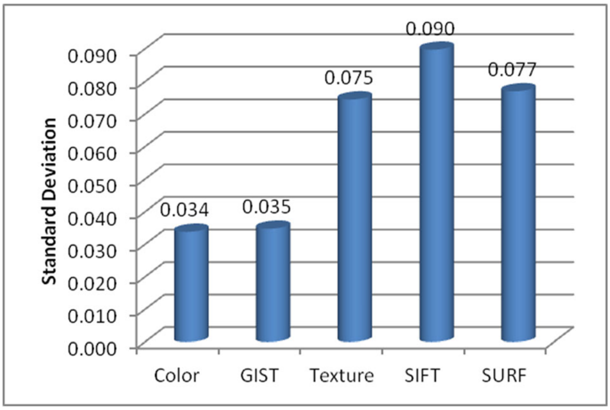

Fig. 10 Standard Deviation of Feature Performance across Clustering Methods

It has been also observed that consistency of a model is also highly dependent on video structure. Models show inconsistent behaviour while processing videos which lack clearly separable shot boundaries and maintain steady performance for videos with visually and semantically separable shots. Figure 9 shows that the feature selection plays an important role while employing FCM and DBSCAN clustering, but Kmeans is quite robust to feature selection decisions.

5.4 Consistency study of features

This section analyses the response of a feature across different clustering algorithms. Figure 10 depicts the standard deviation of performance of a feature corresponding to various clustering approaches. The experimental results signify the consistent behaviour of colour and GIST features. Another global feature texture shows sensitive and unpredictable behaviour across various clustering methods. Further, the local features, SIFT and SURF, present high variation in score value with change in clustering method, which indicates the importance of clustering algorithm selection in a local feature representation-based model.

Same observations are acquired from Fig. A.I provided in Appendix A, where GIST features show minimum standard deviation of 0.06 for V47 video while colour feature depicts most stable behaviour with minimum standard deviation of 0.03 for V69. The standard

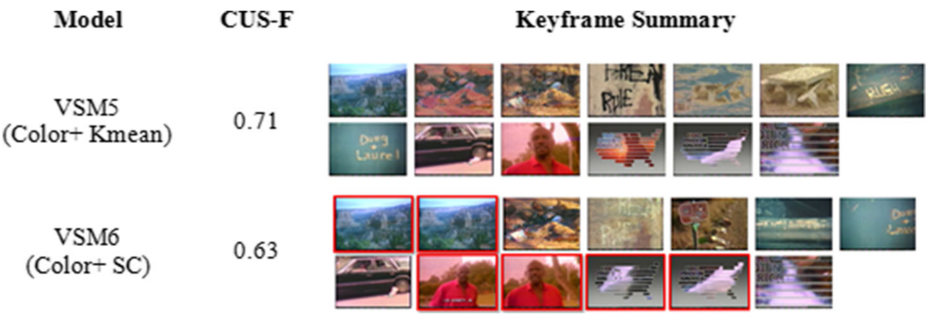

Fig. 11 Model, CUS-F and summary keyframes for video 32 of OV dataset, the red border keyframes represent the redundant frames selected by model

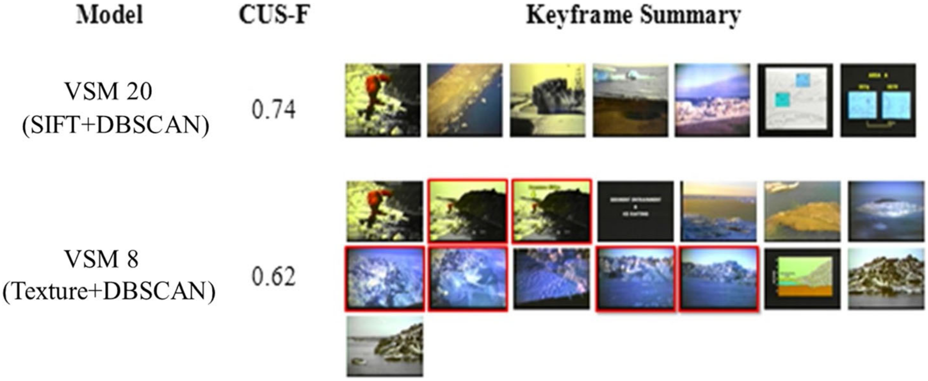

Fig. 12 Model, CUS-F and summary keyframes for video 69 of OV dataset, the red border keyframes represent the redundant frames selected by model

deviation for global feature SIFT illustrates feature sensitivity towards clustering method and video content with highest values 0.17 and 0.15 for video V47 and V69 respectively.

5.5 CUS-F analysis using visual summaries

In this section, the performance of various models along with the proposed evaluation metric CUS-F is discussed with the aid of Keyframe summaries generated.

Figure 11 represents the summary results obtained with model VSM5 (Colour+Kmeans) and VSM6 (Colour+SC). From the figure, it is clear that VSM6 produces a more redundant summary as compared to that of the VSM5 model. This fact is also reflected by the CUS-F score where CUS-F for VSM6 is less than that of VSM5.

Similarly, summaries generated by VSM20 (SIFT+DBSCAN) and VSM 8 (Texture+DBSCAN) are presented in Fig. 12, where texture based summaries contains more redundant frames resulting in lower value of CUS-F than that of VSM20. These comparisons validate the efficiency of the proposed CUS-F metric in considering both true positives as well as false positives effectively. Thus, CUS-F can be used by other researchers in their studies for generating one single score value for Comparison of User evaluation metrics.

6 Conclusion

The wide range of applications of a static video summarization system span over video browsing, indexing, searching, retrieval and human-level decision making about watching a video. In the presented study, the behaviour of six different nature clustering algorithms and five different feature representations, total 30 clustering-based models, are keenly observed and analysed for video abstraction purposes.

The DBSCAN clustering algorithm with local features (SIFT and SURF) outperforms all other models under study. Besides, the performance of DBSCAN drops when used with GIST features. This specifies the sensitivity of the DBSCAN clustering for feature selection. This observation is also supported by the higher standard deviation in the performance of DBSCAN clustering.

The low value of the standard deviation of k-mean clustering indicates its robust performance across various feature representations. This explains the stable performance of k-means based models under different environments.

Spectral clustering and Agglomerative Hierarchical clustering methods exhibited equivalent behaviour when used with global features. Fuzzy C-mean gives a poor performance with local features and shows high vulnerability for different feature spaces. Appropriate pre-processing and post-processing mechanisms are required to obtain reasonable results with GMM clustering based models.

Amongst various local and global feature representations, colour features presented consistent results over different clustering methods. Another global feature, GIST, also manifested comparable performance but performed awfully with a density-based clustering algorithm. Further texture feature failed to represent diversity among frames and resulted in redundant summaries. The performance of local features, SIFT and SURF, is highly sensitive to clustering model selection.

Moreover, this study also proposed a novel evaluation metric CUS-F to facilitate Comparison of User Summaries (CUS) evaluation with better decision capabilities. The experiments demonstrated the efficiency of CUS-F in assigning scores to an automatic summary by considering both true positive and false positive keyframes. Thus, CUS-F can be adopted by the video summarization research community as an effective alternative to the F-measure metric.

The experimental study also concludes that the dependence of clustering-based video summarization methods on finetuning of various hyperparameters makes them difficult to implement for systems dealing with diverse genre videos. More advanced optimization techniques need to be considered for addressing the aforementioned constraint. Also, efforts can be spared to bring machine-learning and deep-learning concepts together for supporting generic video summarization.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11042-022-14081-7.

Funding This research work is supported by the Ministry of Electronics and Information Technology (MeITy), Government of India, New Delhi, India through the Visvesvaraya Research Fellowship.

Data availability In the presented study, all experiments are carried out by utilizing Open Video (OV) dataset [4]. The dataset is publicly available at web link https://sites.google.com/site/vsummsite/download.

Declarations

Conflict of interest The authors declare that they have no conflicts of Interest regarding the work reported in this manuscript.

References

1. Aldavert D, Rusiñol M, Toledo R, Llados J (2015) A study of bag-of-visual-words representations for handwritten keyword spotting. Int J Doc Anal Recognit 18:223–234
2. Arias-Castro E, Chen G, Lerman G (2011) Spectral clustering based on local linear approximations. Electronic journal of statistics, 5: 1537–1587, arXiv:1001.1323. https://doi.org/10.1214/11-ejs651

3. Asadi E, Charkari NM (2012) Video summarization using fuzzy c-means clustering. 20th Iranian conference on electrical engineering (ICEE2012), Tehran, pp. 690-694. https://doi.org/10.1109/IranianCEE.2012.6292442.
4. Avila S, Brandaolopes A, Luz A, Araujo A (2011) VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn Lett 32(1):56–68. https://doi.org/10.1016/j.patrec.2010.08.004
5. Bay H (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110.3:346–359
6. Berkhin P (2006) A survey of clustering data mining techniques. Grouping Multidimensional Data. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-28349-8_2
7. Botchkarev A (2018) Performance metrics (error measures) in machine learning regression. Forecasting and prognostics: properties and typology. ArXiv abs/1809.03006. https://doi.org/10.48550/arXiv.1809.03006
8. Camastra F, Vinciarelli A (2008) Clustering methods. Machine learning for audio, image and video Analysis, pp. 117–148, 978–1–4471-6734-1
9. Chamasemani FF, Affendey LS, Mustapha N, Khalid K (2018) Video abstraction using density-based clustering algorithm. Vis Comput 34:1299–1314. https://doi.org/10.1007/s00371-017-1432-3
10. Choi J, Kim C (2016) A framework for automatic static and dynamic video thumbnail extraction. Multimed Tools Appl 75(23):15975–15991. https://doi.org/10.1007/s11042-015-2909-6
11. Dash A, Albu AB (2017) a domain independent approach to video summarization. Int Conf Adv Concepts Intell Vis Syst, Nov. 2017. https://doi.org/10.1007/978-3-319-70353-4_37
12. Daszykowski M, Walczak B (2009) Density-based clustering methods, In book: Comprehensive chemometrics, vol. 2, pp. 635–654
13. Davidson I, Ravi SS (2005) Agglomerative hierarchical clustering with constraints: theoretical and empirical results. Lecture Notes Comput Sci 3721:59–70 Springer, Heidelberg
14. Dimitrovski V, Kocev D, Loskovska S, Džeroski S (2016) Improving bag-of-visual-words image retrieval with predictive clustering trees. Inf Sci 329:851–865, ISSN 0020-0255. https://doi.org/10.1016/j.ins.2015.05.012
15. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, pp. 65-72. https://doi.org/10.1109/VSPETS.2005.1570899
16. Ejaz N, Bin T, Wook S (2012) Adaptive key frame extraction for video summarization using an aggregation mechanism. J Vis Commun Image Represent 23(7):1031–1040
17. Ejaz N, Baik S, Majeed H, Chang H, Mehmood I (2018) Multi-scale contrast and relative motion-based key frame extraction Journal on Image and Video Processing, 40. https://doi.org/10.1186/s13640-018-0280-z
18. Elharrouss O, Almaadeed N, Al-Maadeed S, Bouridane A, Beghdadi A (2020) A combined multiple action recognition and summarization for surveillance video sequences. Appl Intell 51:690–712. https://doi.org/10.1007/s10489-020-01823-z
19. Furini M, Geraci F, Montangero M, Pellegrini M (2010) STIMO: STIll and MOving video storyboard for the web scenario. Multimed Tools Appl 46:47–69. https://doi.org/10.1007/s11042-009-0307-7
20. Hanjalic A, Zhang HJ (1999) An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Trans Circuits Syst Vid Technol 9(8):1280–1289. https://doi.org/10.1109/76.809162
21. Haralick RM, Shanmugam K, Dinstein I (1973) textural features for image classification. IEEE Trans Syst Man Cybern, vol. SMC-3, no. 6, pp. 610–621. https://doi.org/10.1109/TSMC.1973.4309314.
22. Humeau-Heurtier A (2019) Texture feature extraction methods: a survey. IEEE Access 7:8975–9000. https://doi.org/10.1109/ACCESS.2018.2890743
23. John AA, Nair BB, Kumar PN (2017) Application of clustering techniques for video summarization – an empirical study. Advances in intelligent systems and computing, vol 573. Springer, Cham. https://doi.org/10.1007/978-3-319-57261-1_49
24. Kalita S, Karmakar A, Hazarika SM (2018) Efficient extraction of spatial relations for extended objects Vis-à-Vis human activity recognition in video. Appl Intell 48:204–219. https://doi.org/10.1007/s10489-017-0970-8
25. Kumar K, Shrimankar DD, Singh N (2018) Eratosthenes sieve based key-frame extraction technique for event summarization in videos. Multimed Tools Appl 77:7383–7404. https://doi.org/10.1007/s11042-017-4642-9
26. Low DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis, pp. 91–110, 2004
27. Mahmoud KM, Ismail MA, Ghanem NM (2013) VSCAN: An Enhanced Video Summarization Using Density-Based Spatial Clustering. In: Petrosino A (ed) Image Analysis and Processing – ICIAP 2013. Lecture notes in computer science, vol 8156. Springer, Berlin, Heidelberg

28. Mahmoud KM, Ghanem NM, Ismail MA (2013) VGRAPH: an effective approach for generating static video summaries. IEEE International Conference on Computer Vision Workshops, Sydney, NSW, pp 811–818. https://doi.org/10.1109/ICCVW.2013.111
29. Mundur P, Rao Y, Yesha Y (2006) Keyframe-based video summarization using delaunay clustering. Int J Digit Libr 6:219–232. https://doi.org/10.1007/s00799-005-0129-9
30. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
31. Ou S, Lee C, Somayazulu VS, Chen Y, Chien S (2015) On-line multi-view video summarization for wireless video sensor network. IEEE J Select Top Signal Process 9(1):165–179. https://doi.org/10.1109/JSTSP.2014.2331916
32. Reynolds DA (2009) Gaussian mixture models. Encycl Biom 741:659–663
33. Sebastian T, Puthiyidam JJ (2015) A survey on video summarization techniques. Int J Comput Appl 132(13):30–32
34. Sharghi A, Gong B, Shah M (2016) Query-Focused Extractive Video Summarization. Computer Vision – ECCV 2016. Lecture notes in computer science, vol 9912. Springer, Cham https://doi.org/10.1007/978-3-319-46484-8_1.
35. Shroff N, Turaga SP, Chellappa R (2010) Video Précis: highlighting diverse aspects of videos. IEEE Trans Multimed 12(8):853–868. https://doi.org/10.1109/TMM.2010.2058795
36. The Open Video Project (n.d.) http://www.open-video.org(last accessed on: 9.9.2020)
37. Tilson LV, Excell PS, Green RJ (1988) A Generalisation of The Fuzzy C-means Clustering Algorithm. International Geoscience and Remote Sensing Symposium, 'Remote Sensing: Moving Toward the 21st Century', Edinburgh, UK, pp. 1783–1784. https://doi.org/10.1109/IGARSS.1988.569600.
38. "Track YouTube analytics, future predictions, & live subscriber counts - Social Blade." [Online]. Available: https://socialblade.com/youtube/. Accessed 10 Jul 2020
39. Trinh H, Li J, Miyazawa S, Moreno J, Pankanti S (2012) Efficient UAV video event summarization. Proceedings of the 21st international conference on pattern recognition (ICPR2012), Tsukuba, pp. 2226-2229
40. Truong BT, Venkatesh S (2007) Video abstraction: a systematic review and classification. ACM Trans Multimed Comput Commun Appl 3(1):3:1–3:37
41. Tsai C-F (2012) Bag-of-words representation in image annotation: a review. Int Sch Res Not 2012:1–19. https://doi.org/10.5402/2012/376804
42. Viguier R, Lin CC (2015) Automatic Video Content Summarization Using Geospatial Mosaics of Aerial Imagery. IEEE International Symposium on Multimedia (ISM), Miami, FL, pp. 249–253. https://doi.org/10.1109/ISM.2015.124.
43. Wei H, Ni B, Yan Y, Yu H, Yang X (2018) Video summarization via semantic attended networks. Proceedings of the thirty-second (AAAI) conference on artificial intelligence, New Orleans, Louisiana, USA, pp. 216–223
44. Wu J, Zhong S, Jiang J, Yang Y (2017) A novel clustering method for static video summarization. Multimed Tools Appl 76:9625–9641. https://doi.org/10.1007/s11042-016-3569-x
45. Zhao Y, Guo Y, Sun R, Liu Z, Guo D (2020) Unsupervised video summarization via clustering validity index. Multimed Tools Appl 79(45):33417–33430. https://doi.org/10.1007/s11042-019-7582-8
46. Zhou Y, Cheng Z, Jing L, Hasegawa T (2015) Towards unobtrusive detection and realistic attribute analysis of daily activity sequences using a finger-worn device. Appl Intell 43(2):386–396

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.