Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046185

ARTICLE

Video Summarization Approach Based on Binary Robust Invariant Scalable Keypoints and Bisecting K-Means

Sameh Zarif[1,2,*], Eman Morad[1], Khalid Amin[1], Abdullah Alharbi[3], Wail S. Elkilani[4] and Shouze Tang[5]

[1]Department of Information Technology, Faculty of Computers and Information, Menoufia University, Menoufia, 32511, Egypt

[2]Department of Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Bader, 11829, Egypt

[3]Department of Computer Science, Community College, King Saud University, Riyadh, 11362, Saudi Arabia

[4]College of Applied Computer Science, King Saud University, Al-Muzahmiya, 19676, Saudi Arabia

[5]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

*Corresponding Author: Sameh Zarif. Email: sameh.shenoda@ci.menofia.edu.eg, sameh-zarief@eru.edu.eg

Received: 21 September 2023 Accepted: 12 January 2024 Published: 26 March 2024

ABSTRACT

Due to the exponential growth of video data, aided by rapid advancements in multimedia technologies. It became difficult for the user to obtain information from a large video series. The process of providing an abstract of the entire video that includes the most representative frames is known as static video summarization. This method resulted in rapid exploration, indexing, and retrieval of massive video libraries. We propose a framework for static video summary based on a Binary Robust Invariant Scalable Keypoint (BRISK) and bisecting K-means clustering algorithm. The current method effectively recognizes relevant frames using BRISK by extracting keypoints and the descriptors from video sequences. The video frames' BRISK features are clustered using a bisecting K-means, and the keyframe is determined by selecting the frame that is most near the cluster center. Without applying any clustering parameters, the appropriate clusters number is determined using the silhouette coefficient. Experiments were carried out on a publicly available open video project (OVP) dataset that contained videos of different genres. The proposed method's effectiveness is compared to existing methods using a variety of evaluation metrics, and the proposed method achieves a trade-off between computational cost and quality.

KEYWORDS

BRISK; bisecting K-mean; video summarization; keyframe extraction; shot detection

1 Introduction

The quantity of digital videos has rapidly increased because of multimedia technology advances such as tablets, digital cameras, smartphones, etc. Manipulation of huge amounts of video data faces various challenges, including the time necessary for watching the video to comprehend its contents and the limited memory for video storage. The effective video management system allowed users to browse, retrieve, and index videos quickly while saving storage space. Many researchers are now

interested in summarizing video because of its benefits in several applications such as video-on-demand, video browsing, video indexing, video retrieval, geographic information systems, digital video libraries, and distance education [1]. The objective of the video summary is to give viewers a clear understanding of the video's content through a short summary of the video. The video summarization types are static video summary and dynamic video skimming. A static video summary collects the representative images for the video sequence. Similarly, dynamic video skimming chooses the most important dynamic segments of audio and video to produce the summary. The key benefit of dynamic summary is that it is expressive, and very fun since its summary has both audio and video, but the viewer needs to watch a little snippet of the video to comprehend the video content. On the other hand, static summary is more effective for browsing, retrieval, and indexing. It is also free of timing problems or synchronization. It enables the viewer to have a video overview [2,3].

A static video summary that only includes the bare minimum of keyframes (representative frames) is the focus of this article. The frames are as dissimilar from one another as is possible while still representing the entirety of the video's content. The two different kinds of shot transitions are cut transitions and gradual transitions. A cut transition is an abrupt change in visual content that occurs between two shots within the same frame. A gradual transition is a small variation in dissimilarity and similarity features that occurs between consecutive shots over several frames, whereas these transitions fade-out, fade-in, wipe, and dissolve.

Light variations or flashlights, distortion, noise, camera operation (rotation, tilting, panning, and zooming), camera and object motion, and other factors all have an impact on detecting the transition. These frequently result in false detection, missed detection on shot transition, and the appearance of missed or redundant keyframes [4].

Recent years have seen the emergence of local descriptors, owing to their ability to remain invariant to noise, scale variations, rotation, and illumination changes. These local descriptors can be classified into two primary categories: floating-point descriptors and binary descriptors. The set of floating-point descriptors encompasses the computation of the Euclidean distance and requires a large amount of memory. The primary benefit of binary descriptors is using the calculation of hamming distance. It is extremely fast because it depends on XOR calculation. Binary descriptors are used somewhat limited in video summarization.

In this paper, we are focusing on binary descriptors, due to their invariance to noise, scale, rotation, illumination change, etc. They also distinguish with low computational cost.

In this research, we introduce a technique for static video summary based on bisecting K-means [5] and BRISK [6]. The steps listed below make up the majority of the proposed technique: (i) extract features from the video frames using BRISK, (ii) apply principles component analysis (PCA) [7] to extract the most important features, (iii) use silhouette coefficient [8] for estimating the ideal k-value (clusters number), (iv) cluster BRISK features of the video frames using the bisecting K-means (v) select the keyframes (vi) remove redundant keyframes.

The BRISK has several advantages such as fast keypoint detection, description, matching, rotation invariant, scale-invariant, high quality, and reduce computational cost [9,10]. The bisecting K-means is an extended version of K-means clustering algorithm. It considers all the data as one cluster, then splits it into two subclusters until the required cluster number is obtained. The bisecting K-means algorithm is better than the K-means for the selection of initial center points.

The final video summary of the proposed method is compared with the results of other methods described in the literature that employ the same dataset. A publicly accessible dataset from the Open

Video Project (OVP) [11] that included videos from various genres was used for the experiments. The proposed method outperforms existing methods and has a high F-measure according to the evaluation metrics.

The remaining sections of the paper are structured as follows: Recent methods for summarizing videos are investigated in Section 2. The proposed video summarization method is presented in Section 3. Section 4 presents the datasets, evaluation metrics, and experimental results used to evaluate the video summary. Finally, Section 5 shows the paper's conclusion.

2 Related Work

A significant amount of research effort has been dedicated to the field of video summarization. Numerous approaches aim to extract the utmost crucial details from an extensive video sequence and generate a summary. Static video summarization techniques often rely on both low-level visual characteristics and high-level semantic attributes.

Several approaches based on pixel-wise comparison compute the difference in color or intensity values between two adjacent frames. They compared it to a determined threshold. Some of these methods are in [12,13]. Pixel-based methods are susceptible to false alarms due to their sensitivity to the movement of objects and cameras. These methods depended on the threshold that ignored the temporal relationship of the dissimilarity/similarity signal. Video Summarization by Image Quality Assessment (VSQUAL) [14] summarizes the videos using image quality assessment (IQA) metrics. It did not deal with the video with fast-moving content. The paper in [15] presented a video summarization method called VISCOM. This study employed color co-occurrence matrices as a means of depicting the frames in the video sequence, subsequently generating a summary by selecting the most representative frames. This paper generated rates of false alarms when the image pairs possessed dissimilar contents, but similar color distributions, which consequently yielded low values in the distance function. Pixel-wise techniques might be less efficient if they do not consider the spatial arrangement of neighboring pixels. Therefore, the approach to video summarization proposed in [16] 1is based on key frame extraction through super pixel segmentation (SPSVS). This technique employs a metric for measuring the similarity between image regions. Specifically, it focuses on estimating the spatial relationship similarity using super pixel blocks and selects keyframes with the lowest spatial similarity. resulting in a video summary that exhibits reduced redundancy. The method works best with videos that have bright colors, high contrast, and continuous content. This method did not deal with video content that changed quickly or frequently.

There were also several video summarization approaches that combined the use of global features with clustering techniques. These methods were Delaunay triangulation (DT) algorithm [17] that employed to cluster the video frames and generate the summary by selecting the centroid of each cluster. Another method is called Still and Moving Video Storyboards (STIMO) in [18]. This method utilized Farthest Point-First (FPF) to produce static and dynamic summaries. The two mechanisms VSUMM1 and VSUMM2, as described in reference [19], were characterized by their simplicity, utilizing the K-means clustering method and HSV color histograms. VSUMM1 selected a single keyframe from each cluster, whereas VSUMM2 selected one keyframe from each key cluster. The VISON (Video Summarization for Online Applications) was proposed in [20]. This approach operates in a compressed domain and derives the summary by analyzing visual features. Video summarization technique known as VSCAN, which was discussed in [21], employed the DBSCAN clustering algorithm for its operations. This algorithm enables the clustering of frames by considering the texture and color features. The sparse dictionary method (SD) presented in [22] employed a sparse

constraint based on the l_2 norm. This method was suggested to retain the essential components of the video and eliminate the excess frames in the video sequence. The paper in [23] proposed a color-only clustering-based keyframe extraction approach. They extracted keyframes based on K-means clustering technique. This paper utilized only a single feature, which has the potential to result in false positives within the process of clustering because a lonely feature may not possess the capability to accurately identify data points. The Multi-Featured Cluster based Keyframe Extraction Method (MFCKEM), as proposed in [24], presented a methodology for video summarization by combining both shape and color features.

Some of the methodologies employed for video summarization utilize local descriptors. The local descriptors are divided into Floating-point descriptors such as the Scale Invariant Feature Transform (SIFT) as mentioned in [25] and the Speeded Up Robust Features (SURF) as mentioned in [26]. And another are Binary descriptors which include Oriented Fast and Rotated BRIEF (ORB) in [27], Binary Robust Invariant Scalable Key points (BRISK)in [6], etc.

SIFT and SURF are the foundations of many recent approaches to video summarization in literature. The Keypoints and descriptors were extracted from the video sequence using the employed methods. The consecutive frames descriptors were matched by computing the Euclidean distance, followed by testing the distance against a predetermined threshold to extract the keyframes. Key point-Based Keyframe Selection (KBKS) in [2] extracted SIFT keypoints and descriptors for each frame of the video sequence. After that, through keypoint matching, a global pool of unique keypoints is constructed to represent the entire video taken. Lastly, keyframes are selected as representative frames that best represent the global keypoint pool. Key frame extraction based on the repeatability table (KERT) in [28] employed a window with a size equivalent to the number of frames per second (FPS) to diminish the quantity of frames handled during the selection process from a shot's list of potential frames. By adopting this approach, the count of frames that are processed is effectively reduced. On each candidate frame, interest points were subsequently identified. The candidate sets were solely utilized to calculate the repeatability matrix. Lastly, the relevant keyframes were extracted through the utilization of the shortest path algorithm. The key frame extraction by graph clustering (KEGC) method described in [29] is founded upon the utilization of the Scale-Invariant Feature Transform (SIFT) and the Graph Modularity Clustering technique. This procedure is exclusively employed on a collection of potential frames, which have been selected based on a window, thereby diminishing the quantity of data that necessitates processing. Video summary based on local features (VSLF) is discussed in the study presented in [30]. Initially, the authors employed a leap extraction technique to select a group of potential frames from a video shot. Subsequently, local features on these frames are detected and characterized using SURF. Following this, the FLANN method is employed to eliminate almost identical keyframes and retain a concise set.

Despite the success of the previous video summarization methods, effective video summarization is still challenging because of the challenges imposed by lighting changes, camera operations, object motion, image transformations, and the high cost of computation. As a result, we propose a static video summarization method based on binary descriptors. Due to their invariance to scale, rotation, illumination change, noise, etc., and have a low computational cost.

The proposed method uses BRISK for extracting keypoints and descriptors of the video frames. Then, for each frame, we use PCA to extract the most important features and normalize the feature dimensions of the overall video frames. The video frames' BRISK descriptors are subsequently grouped together employing the bisecting K-means algorithm, and the frame that is closest to the center of the cluster is designated as the keyframe. The methodology pipeline was designed in such

a way as to outperform previous methods while also being simple to incorporate into a variety of applications. Our method attains a good balance between computational cost and quality, due to its reliance on binary descriptors.

3 Proposed Method

In this research, we are focusing on binary descriptors (BRISK), due to their invariance to the factors that effect on video shots transition. These factors are scale, rotation, illumination change, noise, etc., they also distinguish with low computational cost because its dependance of the calculation of hamming distance. For this reason, the proposed method achieves a tradeoff between quality and cost.

The general pipeline of the method that has been suggested is depicted in Fig. 1. The proposed technique extracts the keypoints and the descriptors of the video frames. The number of features in the descriptor for each frame is different from each other. To solve this problem, we apply principal component analysis [7] to extract the most important features for each frame and standardize the dimensions of features for all video frames. The silhouette score [8] is used to define the number of clusters (k-value). The BRISK descriptors of the video frames are then clustered using the bisecting K-means [5]. The keyframe is chosen as the frame that is closest to the center of the cluster. After that, similar keyframes are eliminated. The proposed pipeline is explained in detail in the following subsections.

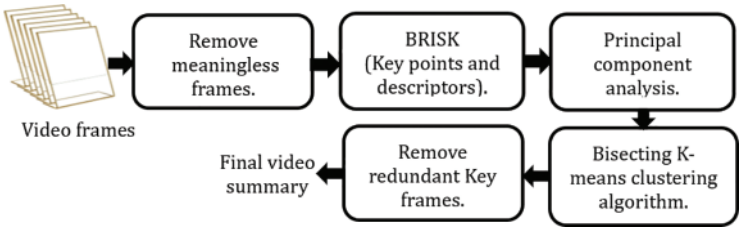

Figure 1: The proposed method framework

3.1 Removing Meaningless Frames

The monochrome frames are meaningless frames that exist in the video due to fade-out/fade-in effects. The proposed method calculates the frames standard deviation to eliminate any potentially meaningless frames. Monochromatic frames have a standard deviation of zero or a small value close to zero. Based on the value of standard deviation the proposed method gets rid of these frames as a pre-processing step. Removing meaningless frames before the clustering process has many benefits on the quality of the proposed method. The monochrome frames are not needed in the clustering step. They also are considered as outlier frames and badly affected on the clustering process. In addition to that, removing meaningless frames reduces the computational time. The standard deviation σ is defined as follows:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{1}$$

where x is each pixel value in the frame, μ is the mean of the frame, and N is the pixels number in the frame.

3.2 BRISK (Binary Robust Invariant Scalable Key Point)

BRISK [6] detects and filters corners using the FAST Corner score. The objective was to locate the maxima in both the image and scale dimensions. Scale-space pyramid levels are made up of n octaves (c_i) and n intra-octaves (d_i) for $i = \{0, 1, \ldots, \mathrm{n}-1\}$ where n is normally equal to 4. The octaves are created by half-sampling the source image gradually (equivalent to c_0). Each intra-octave d_i lies between layers c_i and c_{i+1}. The bit-string binary descriptor of BRISK calculates the typical direction of each feature and the patch orientation by concatenating simple brightness tests. Scale, rotation, affine changes, or simple brightness have no impact on the BRISK algorithm. The main steps of BRISK are Scale-Space keypoint detection and building the descriptor, they are described in detail in [6].

3.3 PCA (Principal Component Analysis)

The main problem is the dimensions of Brisk feature descriptors along the video frames have not the same dimension. For this reason, it uses PCA to extract the main features for each frame and normalize the feature dimensions for all video frames. PCA [7] is a methodology for selecting the most important features of input data. It is used to determine which direction most of the data varies on. It transforms the data into a collection of linearly independent representations of each dimension to extract the primary features of the input data. The following is the specific procedure for using the PCA to select the main components of the traditional 64-dimensional Brisk feature descriptor.

Set n key points and use their feature descriptors as samples. If D represents the sample matrix, then:

$$D = [d_1, d_2, d_3, \ldots . d_n] \tag{2}$$

Find the average feature vector d for n-sample:

$$\overline{d} = \frac{1}{n} \sum\nolimits_{i=1}^{n} d_i \tag{3}$$

where d_i denotes the 64-dimensional feature descriptor of the i^{th} key point.

Identify the difference vector f_i between the average eigenvector and eigenvector of the sample points:

$$f_i = d_i - \overline{d} \tag{4}$$

Create C as a covariance matrix.

$$C = \frac{1}{n} \sum\nolimits_{i=1}^{n} f_i f_i^T = \frac{1}{n} Q Q^T,\ Q = [f_1, f_2, f_3, \ldots, f_n] \tag{5}$$

Find the covariance matrix's 64 eigenvectors λ_i and 64 eigenvalues e_i, arrange the eigenvectors and corresponding 64 eigenvalues in descending order, and then choose the eigenvectors with the Q largest eigenvalues as the principal component's direction.

$$feature\ vector = (h\lambda_1, h\lambda_2, h\lambda_3, \ldots h\lambda_k) \tag{6}$$

where *feature vector* is constructed by taking the eigenvectors ($h\lambda_k$) with the largest eigenvalues as the main feature of the Brisk feature descriptors for each frame along the video frames. k refers to the number of frames in the video sequence.

3.4 Silhouette Coefficient

The role of the silhouette coefficient is to determine the appropriate clusters number, which is used in a bisecting K-means clustering algorithm. As is a common knowledge, the prior methodology is better than the posterior method for determining the keyframe size during the abstraction process. The silhouette coefficient index [8], whose value ranges from -1 to $+1$, it serves as a measure of proximity between a given frame and its respective cluster in relation to alternative clusters. If the silhouette coefficient index is high, the frame matches its cluster but is poorly matched to the other clusters. Throughout the experiment, a series of varying k-values is employed to compute the silhouette coefficient index to ascertain the optimal value. Typically, the optimal cluster number is identified at the maximum point of the silhouette coefficient index. Each sample point's silhouette coefficient index, $p\,(i)$, is defined as follows:

$$p\,(i) = \frac{v(i) - sv(i)}{max\{v(i), sv(i)\}}, -1 \leq p\,(i) \leq +1 \tag{7}$$

The average distance between i and all other data within the same cluster is denoted as $v(i)$, where $sv(i)$ is the smallest average distance between i and all sample points in any other cluster. The silhouette coefficient index is the average $p\,(i)$ of all the points in the dataset, and it shows how well the dataset clustered.

3.5 Bisecting K-Means Clustering Algorithm

The basic concept behind a bisecting K-means clustering algorithm [5] is as follows: choosing two centers, X_0 and X_1, dividing the data set into two different clusters using the standard K-means algorithm, calculating the sum square error (SSE) of two clusters, selecting the largest SSE clusters to split, and repeating this process until the cluster number reaches K. The following are the algorithm stages of a bisecting K-means clustering algorithm:

1. Enter input K and data X.
2. Initialize data and set $n = 1$ as the cluster number.
3. Choose the initial two points x_0, and x_1 at random to serve as the initial cluster center of a first divide. The traditional K-means ($k = 2$) algorithm generates two clusters, $n = n + 1$.
4. Compute the SEE of two clusters, then take the largest SEE and go to step 2.
5. Stages 2 and 3 should be iterative until the cluster number n equals to K value.

3.6 Removing Redundant Key Frames

A redundancy removal process is employed to remove comparable keyframes. Distances are computed between all keyframe descriptor pairs. If the distance between two keyframes exceeds a specific threshold, one is discarded. The threshold value was established through the implementation of empirical tests and meticulous observations, and it was determined to be 0.5. In the end, the remaining frames constitute the final summary, which is subsequently compared to various methodologies described in the literature using metrics such as precision, recall, and F-score.

4 Results and Discussion

4.1 Datasets and Evaluation Metrics

This section encompasses the dataset of the Open Video Project (OVP) [11], along with the assessment metrics. The proposed method is evaluated using Precision, Recall, F-measure, and compression ratio.

The OVP dataset is a very big dataset, which provides video clips from a variety of types like documentary (492 videos), educational (1260 videos), ephemeral (1936 videos), historical (187 videos), lecture (33 videos), other (6 videos), and Public Service (17 videos). These videos are available in color (2084 videos) and black & white (1847 videos). and these are also available with sound (3540 videos) and Silent (391 videos).

50 videos from the OVP dataset have been chosen, these videos include a variety of motion characteristics, colors, and lengths. The overall duration of the video sequences is roughly 75 min (with lengths varying from 1 to 4 min for each movie) and 150,000 frames with 352 $\times$ 240-pixel original dimensions. There are user summaries (ground truth) for these videos, each of them was created by 5 various users. The comparison is made between the summary of the proposed method and the summaries provided by the users, as well as other methods based on the same dataset that are available in the literature.

Open video datasets aim to be representative of real-world scenarios by including diverse and realistic content. There are some common characteristics that contribute to their representativeness such as:

- Diverse Content: OVP dataset includes a diverse range of video content that reflects the complexity and variety of real-world scenarios. This might include scenes from urban and rural environments, indoor and outdoor settings, various lighting conditions, and different types of activities.
- Varied Camera Perspectives: OVP dataset often incorporates videos captured from different camera viewpoints. This includes different angles, heights, and distances.
- Environmental Conditions: OVP dataset may include videos with variations in weather conditions, time of day, and seasonal changes, reflecting the challenges faced in different settings.
- Challenging Scenarios: OVP dataset often include videos with the challenges such as occlusions, crowded scenes, and low-light conditions.
- Temporal Dynamics: OVP dataset incorporates dynamic elements, such as moving objects, changing scenes, and evolving scenarios, to reflect the temporal nature of video data.

The process of evaluation necessitates the utilization of the ground truth, which comprises manually generated summaries by five distinct users for each video. The proposed method's summary of the video is compared with the ground truth. Following this comparison, the average precision, recall, and F-measure are computed for every video. If the semantic content of the two frames is similar, they are supposed to be the same. The terms listed after that are explained:

- True Positive (TP): refers to the Frame that has been selected as a key frame both manually (ground truth) and by the proposed method.
- False Positive (FP): pertains to the frame that has been chosen as a keyframe by the proposed method, not manually (ground truth).
- False Negative (FN): indicates the frame that has been selected as a keyframe manually (ground truth), but not by the proposed method.

These terms are employed to establish the metrics of precision, recall, and F-measure.

$$Precision = \frac{T_P}{T_P + F_P} \tag{8}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{9}$$

$$F - measure = \frac{\sum_{i=1}^{5} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}}{5} \tag{10}$$

The compression ratio quantifies the level of conciseness exhibited by the summary. It is calculated by dividing the total length of the video sequence by the number of key frames present in the summary. It is calculated as follows for a specific video sequence:

$$CR = 1 - \frac{KF}{V} \tag{11}$$

where KF is the keyframe extraction length, and V is the video length.

4.2 Experimental Results

All experiments were conducted on an Intel (R) Core (TM) i7-9750H CPU 2.60 GHz equipped with 16 GB of RAM. The proposed method was implemented using Python and the OpenCV platform. The evaluation of the proposed method utilized the OVP dataset [11] and the assessment metrics outlined in Section 4.1. We conduct a comparison between the proposed approach and various video summarization techniques that also employ the OVP dataset. Certain of these techniques relied on a pixel-wise comparison, wherein the discrepancy in color or intensity values between two consecutive frames was computed and compared against a predetermined threshold. These techniques include VSQUAL [14], VISCOM [15], and SPSVS [16]. Additionally, the proposed approach is evaluated against multiple video summarization techniques that integrate the utilization of global characteristics with clustering methodologies. These methods are DT [17], STIMO [18], VSUMM1 [19], VSUMM2 [19], VISON [20], VSCAN [21], SD [22,23], and MFCKEM [24]. We also compare the proposed method with other methods based on local features such as KBKS [2], KERT [28], KEGC [29], and VSLF [30].

Table 1 illustrates the average precisions, recalls, and F-measures pertaining to the summary of the proposed technique as observed in the OVP dataset. We compare the summary of the proposed technique for each video with the manually created ground truth summaries produced by five distinct users.

Table 1: Average precision, recall, and F-measures of the proposed method's summary for all test videos selected from the OVP dataset [11], as well as their number of frames, and duration (in mm: ss format)

Video	#Frames	Duration	Precision	Recall	F-measure
The Great Web of Water, segment 01	3,279	1:50	0.5	0.58	0.537
The Great Web of Water, segment 02	2,118	1:11	0.85	0.838	0.848

(Continued)

Table 1 (continued)

Video	#Frames	Duration	Precision	Recall	F-measure
The Great Web of Water, segment 07	1,745	0:59	0.64	0.83	0.73
A New Horizon, segment 01	1,806	1:01	0.73	0.793	0.7574
A New Horizon, segment 02	1,797	1:00	0.75	0.775	0.76
A New Horizon, segment 03	6,249	3:29	0.75	0.647	0.7
A New Horizon, segment 04	3,192	1:47	0.8	0.72	0.76
A New Horizon, segment 05	3,561	1:59	0.7	0.7	0.7
A New Horizon, segment 06	1,944	1:05	0.64	0.8	0.71
A New Horizon, segment 08	1,815	1:01	0.8	0.85	0.83
A New Horizon, segment 10	2,517	1:24	0.9	0.833	0.864
Take Pride in America, segment 01	2,691	1:30	0.72	0.657	0.69
Take Pride in America, segment 03	3,261	1:49	0.64	0.576	0.61
Digital Jewelry: Wearable Technology for Every Day Life	4,204	3:00	0.91	1	0.95
HCIL Symposium 2002—Introduction, segment 01	2,336	1:18	0.6	0.596	0.595
Senses and Sensitivity, Introduction to Lecture 1 presenter	4,221	2:20	0.54	0.59	0.57
Senses and Sensitivity, Introduction to Lecture 2	3,411	1:53	1	1	1
Senses and Sensitivity, Introduction to Lecture 3 presenter	4,566	2:32	0.73	0.698	0.71
Senses and Sensitivity, Introduction to Lecture 4 presenter	5,249	2:55	0.81	0.84	0.83
Exotic Terrane, segment 01	2,940	1:38	0.732	0.659	0.69
Exotic Terrane, segment 02	2,776	1:32	0.6	0.53	0.56
Exotic Terrane, segment 03	2,676	1:29	0.72	0.625	0.67
Exotic Terrane, segment 04	4,797	2:40	0.85	0.84	0.845
Exotic Terrane, segment 06	2,425	1:21	0.84	0.93	0.88
Exotic Terrane, segment 08	2,428	1:21	0.7	0.92	0.8
America's New Frontier, segment 01	3,591	1:59	0.66	0.74	0.7
America's New Frontier, segment 03	2,166	1:12	0.8	0.8	0.8

(Continued)

Table 1 (continued)

Video	#Frames	Duration	Precision	Recall	F-measure
America's New Frontier, segment 04	3,705	2:03	1	0.89	0.942
America's New Frontier, segment 07	3,615	2:00	0.77	0.77	0.77
America's New Frontier, segment 10	4,830	2:41	0.8	0.9	0.85
The Future of Energy Gases, segment 03	2,934	1:37	0.73	0.6	0.653
The Future of Energy Gases, segment 05	3,615	2:00	1	0.886	0.94
The Future of Energy Gases, segment 09	1,884	1:02	0.7	0.83	0.76
The Future of Energy Gases, segment 10	2,886	1:36	0.5	0.5	0.5
Oceanfloor Legacy, segment 01	1,740	0:58	0.64	0.87	0.74
Oceanfloor Legacy, segment 02	2,325	1:17	0.7	0.75	0.724
Oceanfloor Legacy, segment 04	3,450	1:55	0.52	0.428	0.47
Oceanfloor Legacy, segment 08	3,186	1:46	0.77	0.896	0.83
Oceanfloor Legacy, segment 09	2,106	1:10	0.722	0.825	0.77
The Voyage of the Lee, segment 05	2,094	1:09	0.72	0.69	0.71
The Voyage of the Lee, segment 15	2,094	1:15	0.69	0.797	0.76
The Voyage of the Lee, segment 16	2,619	1:27	0.92	1	0.96
Hurricane Force—A Coastal Perspective, segment 03	2,310	1:17	0.7	0.75	0.73
Hurricane Force—A Coastal Perspective, segment 04	5,310	2:57	0.67	0.7	0.69
Drift Ice as a Geologic Agent, segment 03	5,310	1:31	0.79	0.735	0.76
Drift Ice as a Geologic Agent, segment 05	2,187	1:12	0.5	0.557	0.53
Drift Ice as a Geologic Agent, segment 06	2,425	1:30	0.86	1	0.93

(Continued)

Table 1 (continued)

Video	#Frames	Duration	Precision	Recall	F-measure
Drift Ice as a Geologic Agent, segment 07	2,425	1:05	0.96	0.88	0.92
Drift Ice as a Geologic Agent, segment 08	2,425	2:00	0.68	0.79	0.731
Drift Ice as a Geologic Agent, segment 10	1,407	0:46	0.88	0.917	0.895

Table 2 presents the average precision, recall, and F-measure of the video summary generated by each individual method for all test videos chosen from the OVP dataset. The data in Table 2 reveal that the distinct values for average precision, recall, and F-measure produced by the proposed approach outperform the values produced by the compared methods, resulting in competitive outcomes. It is evident that the proposed method exhibits a high level of precision and F-measure. Conversely, alternative methods such as VSCAN, KBKS, VISCOM, and SPSVS exhibit a higher level of recall compared to the proposed approach because the proposed method is sensitive to information loss due to PCA use.

Table 2: Average precision, recall, and F-measures of the video summary generated by each approach for all test videos selected from the OVP dataset [11]

Ref. Number	Method	Precision	Recall	F-measure
[17]	DT	0.547	0.433	0.483
[18]	STIMO	0.519	0.621	0.565
[11]	OVP	0.584	0.657	0.618
[20]	VISON	0.595	0.675	0.632
[14]	VSQUAL	0.557	0.743	0.636
[19]	VSUMM1	0.721	0.641	0.679
[19]	VSUMM2	0.42	0.77	0.599
[21]	VSCAN	0.625	0.831	0.713
[22]	SD	0.40	0.61	0.483
[2]	KBKS	0.31	0.89	0.460
[28]	KERT	0.627	0.589	0.607
[15]	VISCOM	0.649	0.811	0.721
[29]	KEGC	0.649	0.798	0.723
[23]	Color feature + K-means	0.5	0.632	0.517
[16]	SPSVS	0.51	0.87	0.626
[24]	MFCKEM	0.547	0.712	0.59
	PROPOSED METHOD	0.75	0.77	0.756

Fig. 2 displays the output from the proposed methodology for a particular video, along with the production from various summary techniques and manual summaries created by five different users. In this case study, which shows the strengths of the proposed method, all the content chosen by the users was covered by the summaries of VSUMM1, KEGC, and the proposed method. The first keyframe of KEGC, OV, and DT is a wipe transition (it is a type of gradual transition), which should not be chosen in the video summary. The other approaches (OV, DT, and VSUMM2) either produced shorter summaries with less satisfying user content or attempted to include the entire content at the cost of a larger final summary (STIMO).

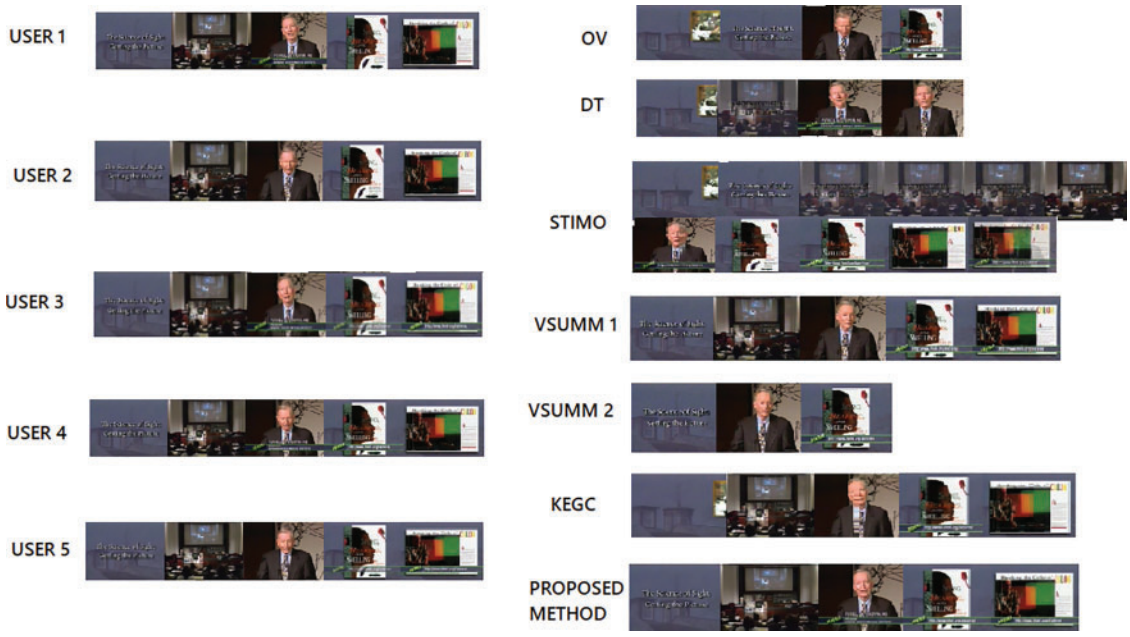

Figure 2: User summaries and keyframes extraction are performed for each method in the video entitled "Senses and Sensitivity, Introduction to Lecture 2"

Similarly, in the example presented in Fig. 3. The proposed method, VISCOM, and VISON produce the best outputs. Whereas the other techniques (STIMO, OV, and VSCAN) received redundancy in their summaries, which led to an increase in the size of their summaries. DT, and VSQUAL miss some keyframes from their summaries. The proposed method gives a high F-score when compared with existing static video summarization methods.

Fig. 4 illustrates the intended methodology summary for a specific video, as well as the output from several summary approaches and manual summaries generated by five distinct users. In this case, all the content chosen by the users was covered by the summary of the proposed method. The method that has been suggested demonstrates a notably elevated F-measure in comparison to alternative methodologies.

All the above results show the effectiveness of the proposed algorithm according to the subjective evaluation, as shown in Tables 1 and 2. The objective evaluation is shown in Figs. 2–4. Bisecting K-means improves clustering results by recursively bisecting clusters until the desired number of clusters is achieved.

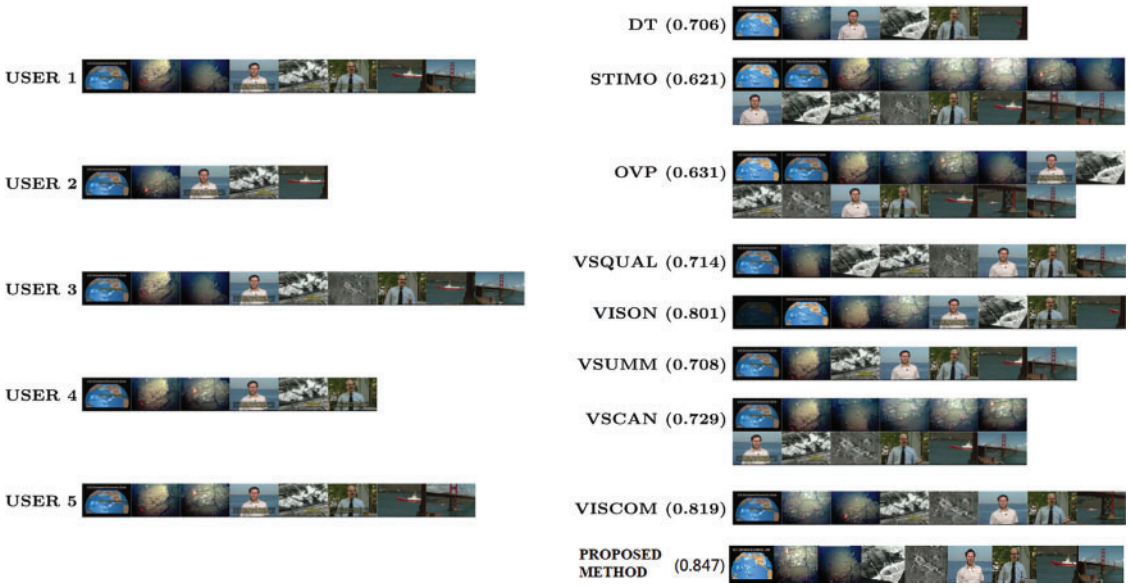

Figure 3: Extraction of user summaries and keyframes from the video titled "America's New Frontier, Segment 10" is performed, alongside the computation of their respective F-measures

Figure 4: User summaries and keyframes extraction for each technique from the video titled "Exotic Terrane, segment 04", alongside the corresponding F-measures

Various parameters can significantly influence the performance and behavior of BRISK. Here are some key parameters associated with BRISK and how they can affect its operation:

- Threshold for Keypoint Detection: Lower thresholds result in more keypoints, while higher thresholds lead to fewer, more salient keypoints.
- Octave Layers: It Determines the number of layers per octave in the image pyramid and affects the scale levels at which BRISK operates. Higher values provide more scale levels but increase computation time.
- Threshold for Orientation Assignment: It affects how keypoints are assigned orientations. Lower values may result in more keypoints having assigned orientations.
- Desired Keypoint Count (MAX_KEYPOINTS): It limits the maximum number of keypoints to be detected, it can be used to control the computational load and focus on the most prominent keypoints.

The quality of the clustering results in Bisecting K-Means can be influenced by various parameters. Here are some key parameters that can affect the quality of Bisecting K-means:

- Number of Clusters (k): It specifies the desired number of clusters; The silhouette coefficient is utilized to find the proper cluster number, which leads to a better result and overcomes the problem of selecting a random k value.
- Maximum Number of Iterations: A higher number of iterations allows the algorithm to refine the clusters further, potentially improving quality. However, setting it too high may lead to an increase in computational costs.
- Initialization Method for K-Means: The choice of initialization method can affect the convergence and quality of the final clustering. Common methods include random initialization and K-means++.
- Convergence Criteria: It Specifies the convergence criteria, such as the change in cluster assignments or centroid positions. Tightening or loosening the convergence criteria can affect the quality of clustering and the algorithm's runtime.
- Initialization Seed: The choice of seed can affect the reproducibility of results. Using a fixed seed is useful for obtaining consistent results in different runs.

The duration of the processing time for specific videos that have been selected as exemplars is delineated in Table 3. All these results indicate that our strategy is fast and achieves a tradeoff between quality and time. This is due to the use of BRISK binary descriptors and bisecting K-means clustering algorithm.

Table 3: Lists some examples of video tests that were taken from the OVP dataset [11], along with the frames number, duration, and time in seconds for the video of the proposed method's summary

Video title	Genre	Duration	#Frames	Execution time (s)
The great of water, segment 01	Documentary	1:50	3279	3.87
Sense and sensitivity introduction to lecture 2	Lecture	1:53	3411	2.70
Sense and sensitivity introduction to lecture presenter 1	Lecture	2:20	4221	3.88
The future of energy gases, segment 09	Documentary	1:02	1884	1.376

(Continued)

Table 3 (continued)

Video title	Genre	Duration	#Frames	Execution time (s)
Digital Jewelry: Wearable Technology for Everyday Life	Educational	3:00	4204	3.446

Fig. 5 illustrates the compression ratio of various experiments conducted on the proposed method in comparison to other existing methods. It is not possible to conclude that the keyframe extraction approach is adequate for video summary based on the compression ratio. Since lossless compression is a significant aspect in the field of video summarization, various evaluation metrics are considered to make comparisons.

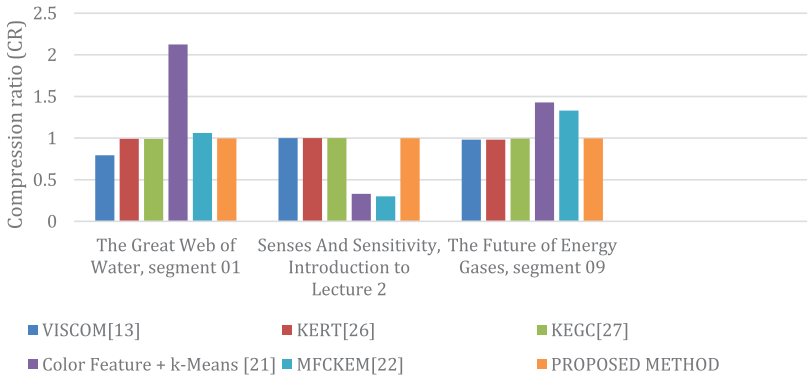

Figure 5: Quality of the produced outputs in terms of the compression ratio (CR) values

4.3 Limitation and Future Work

The proposed method provides good results, but the result is sensitive to the information loss due to PCA use. To avoid this disadvantage, we suggest using another feature selection algorithm like the embedded methods. They combine the advantages of both the wrapper and filter methods, by considering the interactions between features. These methods select the features based on the machine learning procedure, which incorporates feature selection directly into the model training algorithm and optimizes both the model and feature subset simultaneously. This integration helps the model focus on the most relevant features. The most common embedded methods are LASSO (Least Absolute Shrinkage and Selection Operator), Elastic Net, Genetic Algorithms, Decision Trees, etc. After making these changes the quality of the proposed method will increase.

Our intention for the future is to improve the performance of the proposed method for real-world applications by integrating such technologies to support existing multimedia management system requirements for rapid adaptation, storage, processing, retrieval, and reuse of video content.

For Example, for video retrieval application, the query's input will be the set of key frames that are produced by the proposed method. The process of extracting key frames from all the videos in the database will be done offline. The user can then choose his query online. The system will return a list of videos that have comparable content to the query.

5 Conclusion

This study offers a static video summarization method that makes use of BRISK for extracting the keypoints and the descriptors of the video frames. After that, we use PCA to extract the most significant features for each frame and normalize feature dimensions across all video frames. The BRISK descriptors of the video frames are subsequently subjected to clustering utilizing the bisecting K-means algorithm, wherein the frame nearest to the center of the cluster is designated as the keyframe. Experiments with various video types were conducted on the OVP dataset. The video summaries produced by the proposed method are fairly compared with those generated by previous algorithms, employing the same dataset, and evaluating metrics such as average precision, recall, and F-measures. The proposed method's keyframes provide appropriate coverage of the video content. Objective and subjective evaluations have been conducted to prove the efficiency of the work.

The proposed method focuses on improving accuracy and efficiency for video summarization. It reduces computational costs, which leads to an increased chance of using it in real-world applications.

The methodological progression was formulated in a manner that surpasses previous methodologies while simultaneously facilitating integration with diverse applications. Since Our strategy is fast and achieves a tradeoff between quality and time. This is because BRISK binary descriptors are used.

Acknowledgement: None.

Funding Statement: The authors would like to thank Research Supporting Project Number (RSP2024R444), King Saud University, Riyadh, Saudi Arabia.

Author Contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by all authors. The first draft of the manuscript was written by the second author and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. H. Abdulhussain, S. A. R. Al-Haddad, M. I. Saripan, B. M. Mahmmod, and A. Hussien, "Fast temporal video segmentation based on Krawtchouk-Tchebichef moments," *IEEE Access*, vol. 8, pp. 72347–72359, 2020. doi: 10.1109/ACCESS.2020.2987870.

[2] G. L. Guan, Z. Y. Wang, S. Y. Lu, J. D. Deng, and D. D. Feng, "Keypoints-based keyframe selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 729–734, 2013. doi: 10.1109/TCSVT.2012.2214871.

[3] G. Yasmin, S. Chowdhury, J. Nayak, P. Das, and A. K. Das, "Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 7–4902, 2023. doi: 10.1007/s00521-021-06132-1.

[4] T. Kar and P. Kanungo, "A gradient based dual detection model for shot boundary detection," *Multimed. Tools Appl.*, vol. 82, no. 6, pp. 8489–8506, 2023. doi: 10.1007/s11042-022-13547-y.

[5] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, and B. Abuhaija, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, no. 11, pp. 178–210, 2023. doi: 10.1016/j.ins.2022.11.139.

[6] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Int. Conf. Comput. Vis.*, IEEE, 2011. doi: 10.1109/ICCV.2011.6126542.
[7] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 1–30, 2021.
[8] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.
[9] Y. Ou, Z. Cai, J. Lu, J. Dong, and Y. Ling, "Evaluation of image feature detection and matching algorithms," in *5th Int. Conf. Comput. Commun. Syst. (ICCCS)*, IEEE, 2020. doi: 10.1109/ICCCS49078.2020.9118480.
[10] C. H. Lee and E. M. Kim, "Performance comparison and analysis between keypoints extraction algorithms using drone images," *J. Korean Soc. Survey. Geod. Photogramm. Cartogr.*, vol. 40, no. 2, pp. 79–89, 2022. doi: 10.7848/ksgpc.2022.40.2.79.
[11] The Open Video Project 2016. Accessed: Sep. 01, 2023. [Online]. Available: http://www.open-video.org.
[12] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Nature, 2022, doi: 10.1007/978-3-030-34372-9.
[13] S. Lian, "Automatic video temporal segmentation based on multiple features," *Soft Comput.*, vol. 15, no. 3, pp. 469–482, 2011. doi: 10.1007/s00500-009-0527-9.
[14] M. V. Mussel Cirne and H. Pedrini, "Summarization of videos by image quality assessment," in *Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl.: 19th Iberoamerican Congress, CIARP 2014*, Puerto Vallarta, Mexico, Springer International Publishing, 2014, pp. 901–908. doi: 10.1007/978-3-319-12568-8_109.
[15] M. V. Mussel Cirne and H. Pedrini, "VISCOM: A robust video summarization approach using color co-occurrence matrices," *Multimed. Tools Appl.*, vol. 77, no. 1, pp. 857–875, 2018. doi: 10.1007/s11042-016-4300-7.
[16] H. Jin, Y. Yu, Y. Li, and Z. Xiao, "Network video summarization based on key frame extraction via superpixel segmentation," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 6, 2022. doi: 10.1002/ett.3940.
[17] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *Int. J. Digit. Libr.*, vol. 6, no. 2, pp. 219–232, 2006. doi: 10.1007/s00799-005-0129-9.
[18] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," *Multimed. Tools Appl.*, vol. 46, no. 1, pp. 47–69, 2010. doi: 10.1007/s11042-009-0307-7.
[19] S. E. F. de Avila, L. A. P. Bão, and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recogn. Lett.*, vol. 32, no. 1, pp. 1–68, 2011. doi: 10.1016/j.patrec.2010.08.004.
[20] J. Almeida, N. J. Leite, and R. S. Torre, "Vison: Video summarization for online applications," *Pattern Recogn. Lett.*, vol. 33, no. 4, pp. 4–409, 2012. doi: 10.1016/j.patrec.2011.08.007.
[21] K. M. Mahmoud, M. A. Ismail, and N. M. Ghanem, "Vscan: An enhanced video summarization using density-based spatial clustering," in *Image Anal. Process. –ICIAP 2013: 17th Int. Conf.*, Naples, Italy, Berlin Heidelberg, Springer, 2013, pp. 733–742. doi: 10.1007/978-3-642-41181-6_74.
[22] Y. Cong, J. Y. uan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimed.*, vol. 14, no. 1, pp. 1–75, 2011. doi: 10.1109/TMM.2011.2166951.
[23] H. Zhao, W. J. Wang, T. Wang, Z. B. Chang, and X. Y. Zeng, "Key-frame extraction based on HSV histogram and adaptive clustering," *Math. Probl. Eng.*, vol. 2019, pp. 1–10, 2019. doi: 10.1155/2019/5217961.
[24] S. K. Parui, S. K. Biswas, S. Das, and B. Purkayastha, "Multi-featured cluster based keyframe extraction for efficient video processing and management," in *2023 IEEE 3rd Int. Conf. Technol., Eng., Manag. Soc. Impact Using Mark., Entrep. Talent (TEMSMET)*, IEEE, 2023. doi: 10.1109/TEMSMET56707.2023.10150150.
[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.
[26] H. Bay, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.*, vol. 3, no. 2, pp. 346–359, 2008. doi: 10.1007/11744023_32.
[27] E. Rublee, V. Rabaud, and K. Konolige, "ORB: An efficient alternative to SIFT or SURF," in *2011 Int. Conf. Comput. Vis.*, IEEE, 2011. doi: 10.1109/ICCV.2011.6126544.

[28] H. Gharbi, M. Massaoudi, S. Bahroun, and E. Zagrouba, "Key frames extraction based on local features for efficient video summarization," in *Adv. Concepts Intell. Vis. Syst.: 17th Int. Conf., ACIVS 2016*, Lecce, Italy, Springer International Publishing, 2016, pp. 275–285. doi: 10.1007/978-3-319-48680-2_25.

[29] H. Gharbi, S. Bahroun, and E. Zagrouba, "Key frame extraction for video summarization using local description and repeatability graph clustering," *Signal Image Video Process.*, vol. 13, no. 3, pp. 507–515, 2019. doi: 10.1007/s11760-018-1376-8.

[30] M. Massaoudi, S. Bahroun, and E. Zagrouba, "Video summarization based on local features," 2017. doi: 10.4018/ijmdem.2014040103.