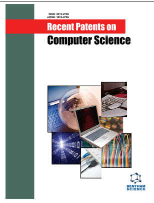
Recent Patents on Computer Science **2018, *11*, 3-16**

REVIEW ARTICLE



Key-Frame Extraction Techniques: A Review

Milan Kumar Asha Paul*, Jeyaraman Kavitha and P. Arockia Jansi Rani

Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, Tamilnadu, India

Abstract: *Background*: The massive database of videos is growing day by day in this era. Analyzing such huge data is always a time-consuming process. The effective use of video content requires a user-friendly access to information. This leads to the evolution of the research area known as video summarization. The effective techniques of video summarization, the videos have let to analyze the content of large volumes of digital video sequences in various categories, such as surveillance, documentaries, movies, sports, lectures, and news. In video summarization, the automatic selection of necessary and informative section from videos using accurate algorithms is essential. The keyframe extraction in video summarization is intended to suffice comprehensive analysis of video by eliminating replications and extraction of keyframes from the video.

***Methods*:** Recent keyframe extraction techniques like clustering, shot, visual content based keyframe extraction methods are discussed for effective keyframe extraction.

***Results*:** First an introduction of various techniques for keyframe extraction pursued by the state-of-the-art review on their properties. Although we have outlined some ideas for effective evaluation of video keyframes, the analytical evaluation of various keyframe extraction techniques is discussed and the approaches based on the methods, dataset and the results are compared.

***Conclusion*:** In the recent years, the use of digital video data has been increasing significantly due to the extensive use of multimedia applications in the areas of education, entertainment, business. So the video has received an incredible attention and research interest in video processing. The use of keyframe extraction has been given incredible attention, in this work, we have carried out a comprehensive survey and review of the research in keyframe extraction techniques. We believe the review paper will provide an update for the reader regarding the progress of keyframe extraction by different keyframe extraction techniques.

ARTICLE HISTORY

Received: February 02, 2018
Revised: May 29, 2018
Accepted: June 28, 2018

DOI:
10.2174/2213275911666180719111118

Keywords: Keyframe, video summarization, clustering based keyframe extraction, shot based keyframe extraction, visual content based keyframe extraction, motion-based keyframe extraction.

1. INTRODUCTION

Today digital video is an emerging force in various multimedia applications. A video is a successive movement of visual images. The rapid growth of the internet has increased the use of video. A video stream is composed of many frames at a frame rate of at least 25 frames per second (fps) that a human can't perceive any discontinuity in the video. The revolution has brought many researchers into new technologies that aim to improve the effective and efficient use of videos.The absolute volume of video data is an obstacle to many applications, and therefore, there is demand for a mechanism, that is used to gain perspective frames from the video without watching the entire video.

Video abstraction is the mechanism for producing the short summary of the entire video. This is very useful to effectively manage and store a huge amount of audiovisual information. A video summary contains the most appropriate information, avoiding any redundancy, but at the same time, preserving the originality of the video [1, 2]. There are two basic forms of video abstract, namely Video skim and Keyframes. Video skim is a type of abstract containing collection of video segments along with their corresponding audio extracted from the original video. This is also called a moving-image abstract, moving storyboard, or summary sequence [3-10].

Keyframe of a video provides the most exact and compact summary of the video. These are also called representative frames, R-frames, still-image abstracts or static storyboard. Keyframes are the basic form for several tasks like video browsing, video summarization, searching, understanding, chapter titles in DVDs and it is also used in many applications such as surveillance video, medical video,

*Address correspondence to this author at the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, Tamilnadu, India; Tel: 009443442978; E-mail: ashanichelson@gmail.com

1874-4796/18 $58.00+.00

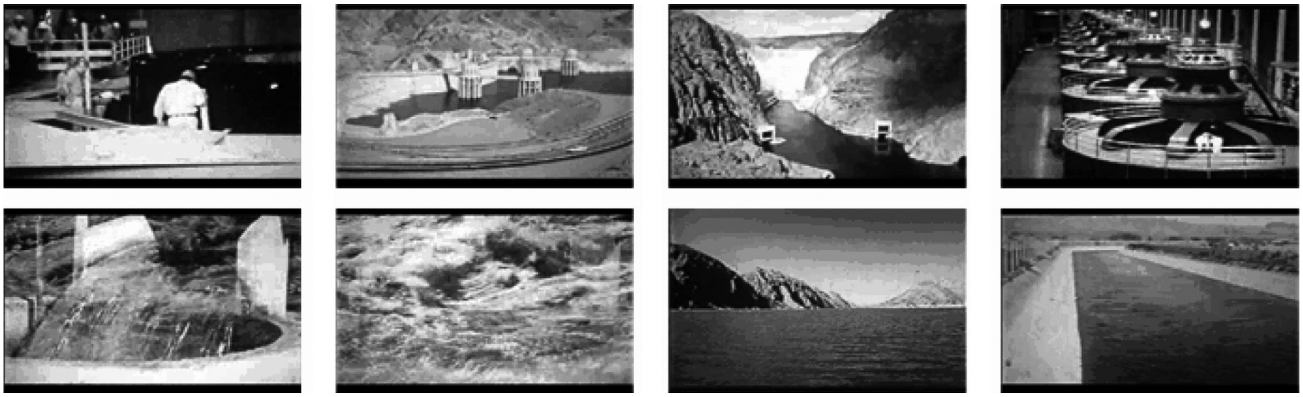


Fig. (1). A sample keyframes from a video.

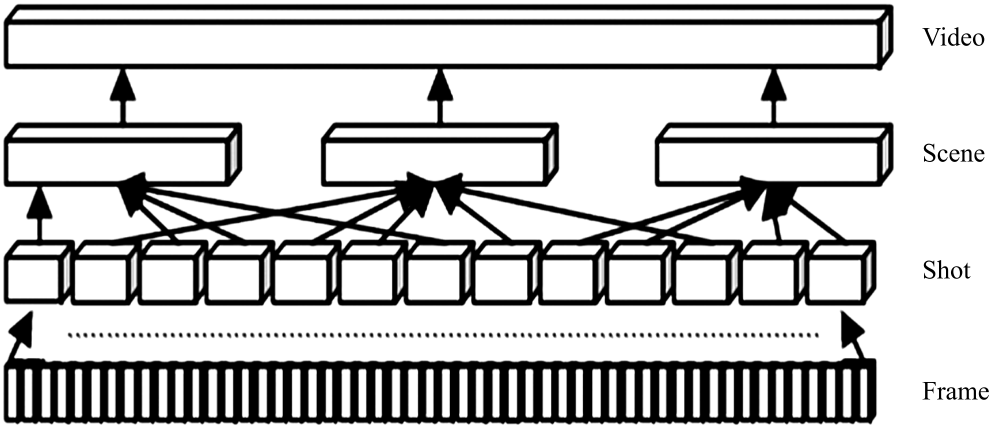


Fig. (2). Schematic representation of a video structure.

underwater video, web browsing video, sports and news programme, indoor and outdoor videos [3, 4], *etc.* Early approaches choose to select keyframes by random or uniform sampling of the video frames at predefined intervals [1].

Traditional summarization method, as illustrated in Fig. (**1**), provides compact static video visualization [11-15] or concentrated video skims [16-20].

In Fig. (**2**), frames are present in the lower level. A frame is a still image and the next higher level frames are grouped into shots and a shot is a continuous camera recording. Then these shots are combined into scenes. Zeeshan Rasheed *et al*., [21] described that a scene is defined as a subdivision of a play, whether the setting is defined, or when it is presented continuously.

There are so many keyframe extraction techniques available. Amiri *et al.,* [3] categorized keyframe extraction into shot based, segment based and motion-based methods. In shot based method, video is segmented into shots and keyframes are selected from each shot. A shot is nothing but a sequence of frames from a video [4, 22]. It has the group of frames, which have constant visual attributes like color, texture and motion. These low-level features are considered to extract keyframes. It is one of the easiest ways to extract keyframes, the main drawback of this method is that they do not scale up efficiently for long video.

In segment based method, based on different semantic levels, video segmentation often refers to two categories as temporal and object-based video segmentations. In temporal video segmentation, a video sequence is partitioned into a set of shots, and some keyframes are extracted to represent a shot. Object-based video segmentation extracts objects for content-based analysis and provide structured representation for many object-oriented video applications [23, 24]. In this method, the video segments are extracted from the frame clusters and keyframes that are closest to the centroid of each calculated cluster. In the hierarchical clustering, the chosen keyframes have been controlled by limiting the number of clusters [5].

In motion based methods, keyframes are extracted based on the optical flow [25], which compute the simple motion metric. Analyze the metric as a function of time to select keyframes at the local minima of motion [7].

In this review article, the main contribution of this work consists of presenting an overview of keyframe extraction in section 2, and various keyframe extraction techniques are discussed in section 3. In section 4, performance evaluation methods for keyframe extraction are discussed and finally, section 5 presents concluding remarks.

2. OVERVIEW OF KEYFRAME EXTRACTION

In Fig. (**3**), a video is given as input. The first step is to convert the video sequence into frames, then in the preprocessing step the video is enhanced by removing the noise, haze and low lighting enhancement are done. Next, a

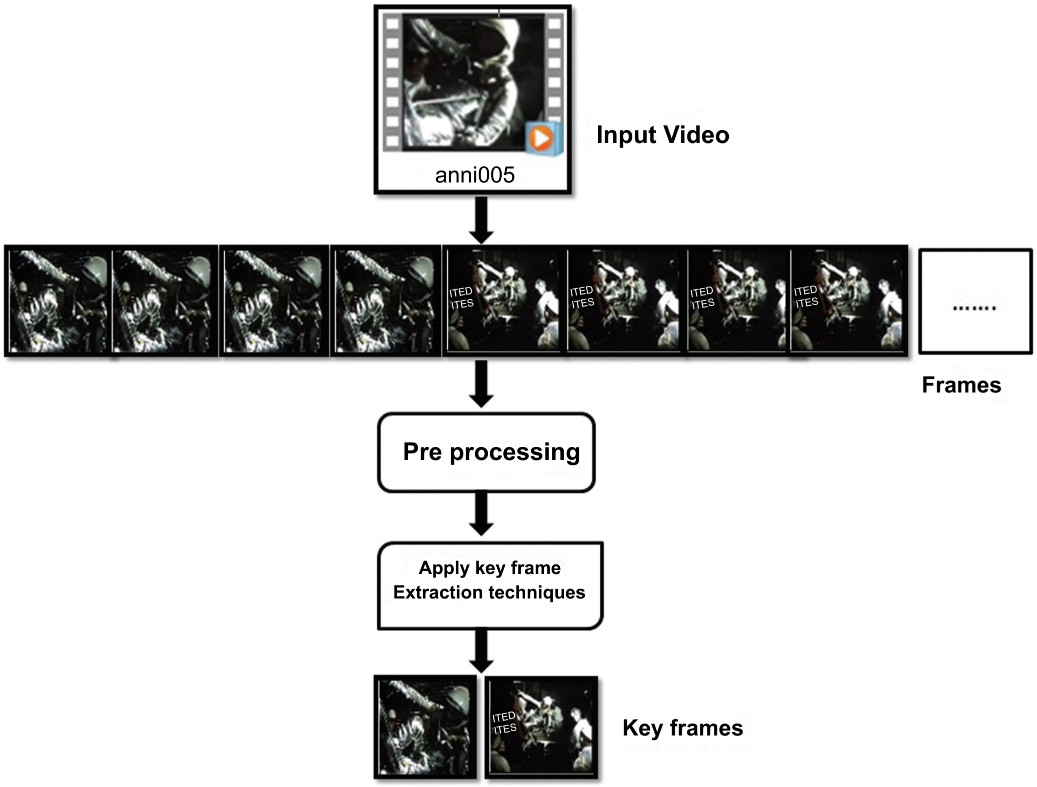


Fig. (3). A flow chart representing the keyframe extraction.

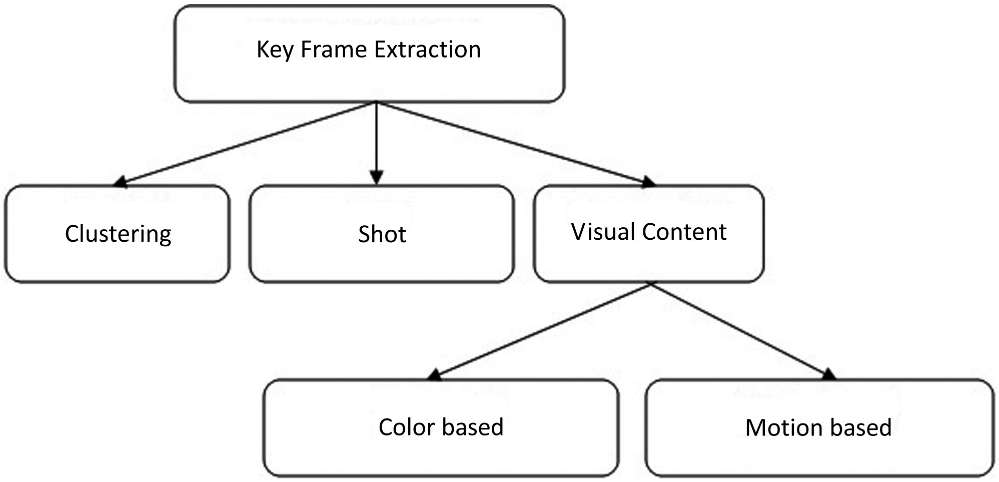


Fig. (4). Hierarchy of keyframe extraction techniques discussed in this review.

keyframe extraction technique is applied and the keyframes are obtained. In this article, keyframe extraction techniques like clustering-based methods, shot based methods, visual content-based methods, and graph modeling based methods are discussed. Fig. (**4**) explains the hierarchy of keyframe extraction techniques that are discussed in this review.

2.1. Keyframe Extraction Using Clustering

Clustering is a technique used to fragment the frames into clusters within a shot and from each candidate clusters a keyframe is selected. Clustering is the powerful mechanism used in various fields such as pattern recognition and information retrieval, *etc.*

In 1998 Zhuang *et al.,* [9] first published keyframe extraction technique based on clustering. Here the keyframe selection is based on the number and size of the clusters. The similar visual frames are grouped into a cluster, where the visual content could be color, texture and shape. This method provides better efficiency, fast to compute and it is easy to implement on online processing. This approach was tested on two movies, namely action movie (movie 1) and a romantic movie (movie 2). The number of keyframes in the movie 1 is greater than movie 2.

A statistical model is used to calculate the clustering threshold. This algorithm can capture the significant content as a keyframe. This technique is tested for sport clip: Golf, a documentary, 'Science eye', TV news program and provided good results, but for strong lighting changes such as flash light on a TV news program, results over segmentation during clustering may lead to an inappropriate output [8].

From the group of frames, a frame which is typically different from their immediate neighbor is selected as a keyframe. The visually similar frames are collected into one group using Fuzzy C- means clustering algorithm. After the formation of clustering, the frames that exhibit a change ratio, which is a measure of the content variation, greater than the average value of the cluster is treated as Keyframes. This method experimented in YouTube sport videos, soccer clips and open video dataset [26]. For analyzing the athletics pose during training of athletes a deep keyframe extraction method use Fully Convolution Networks (FCN) to extract foreground regions that contain athlete and barbell. From the selected foreground, Convolution Neural Networks (CNN) are applied to estimate the pose probability of each frame and extract the keyframes by the maximum probability of each pose [27, 28].

Antonis.I.Ioannidis and his research group [29] worked on weighted multi-view clustering algorithm that applied to produce a single similarity matrix by combining two different image descriptors, that act as an input to a spectral clustering algorithm. For the similarity matrix, a single image descriptor does not contribute equally to the similarity matrix. The weight is associated with each view and learn these weights automatically by the weighted multi-view clustering algorithm. Again the researchers developed segmentation and cluster techniques engage in keyframe extraction for the large variation in the visual content of the video. Using single image descriptors (color, texture, *etc.*,) is not effective for keyframe extraction. To handle this problem, the weighted fusion of several descriptors is used that automatically estimates the weight of each descriptor. For the specific video shot, the weight reflects the relevance of each descriptor. Moreover, they are used to form a composite similarity matrix as the weighted sum of all the similarity matrices corresponds to the individual descriptors. This matrix is then used as input to a spectral clustering algorithm that partitions the shot frames into groups. At last the mediod frame of each group is selected as key-frame. This method is very effective to extract keyframes based on video shots regardless of the characteristics of the visual content of a video [30]. An Eratosthenes Sieve based key-frame extraction can work better for real-time applications. Eratosthenes Sieve is used to produce sets of all Prime number frames and nonprime number frames to total N frames of a video. K-means clustering procedure employs on these sites to extract the key-frames quickly. Davies-Bouldin Index (DBI) is employed to achieve an optimal set of clusters. DBI is a cluster validation technique which allows users with a free parameter to choose the desired number of key-frames without incurring additional computational costs [31].

Jiaxin Wu *et al.,* [32] used a technique by integrating the important properties of the video, the similar frames gather into clusters. In the first step, the redundancy of the video frame is reduced and produced candidate frames by using pre-sampling. BoW (Bag of Word) model is applied to represent the visual content of candidate frames. Finally, candidate frames gather into clusters by the VRHDPS (Video Representation based High Density Peaks Search) clustering approach. The central value of all clusters is collected as a keyframe. This method is tested against many clustering algorithms and produced better results. This experiment is performed on two benchmark databases~~ VSUMM, ~~and 50 videos of the open video project. Mengjuan Fie *et al.* described the syncretic keyframe extraction algorithm by combining Sparse Selection (SS) and Mutual Information based Agglomerative Hierarchical Clustering (MIAHC). Firstly, the optimal keyframes are extracted by applying SS algorithm. Then, using content loss minimization and representativeness ranking, many candidate keyframes are selected and grouped into initial clusters.The improved MIAHC perform further processing to eliminate repeated frames and finally, the keyframes are generated. This work showed better result when compared with conventional methods [33]. The shots are detected from the video that having a correlation. Based on the similarity, the detected shots are clustered. This clustered shot is ranked and a portion of these shots are selected based on cluster ranking as a keyframe [34].

Based on the equipartition problem two keyframe extraction methods are described. In the first case, Iso-content distance principle was used in the video where the keyframes are equidistant. Under Iso-Content Distortion principle, the frame clusters derived by keyframes are equal sized. According to the principles, the selected keyframes have different properties and have the same significance [35].

Normalized cut algorithm engaged to globally and optimally partition a video into clusters. The perceptual quality of the shots and clusters is computed by the motion attention model based on human perception. Then the clusters with computed attention values form a temporal graph similar to 'Markov Chain' which describes the evolution and perceptual importance of the video clusters. This graph is used to group similar clusters into scenes while the appropriate sub shots in scenes [36]. In 2005, the same research team used the same methods that are described above, but at first, the entire video is represented by a complete undirected graph and then apply the same procedure which is described above. The resulting clusters form a directed temporal graph, then the shortest path algorithm is used to efficiently detect video scenes, and then compute the attention value and attach it to the scenes, shots, sub shots and clusters in a temporal graph. It describes the temporal importance of the video [37].

Chamasemani et al proposed DbSva (Density-based Surveillance video abstraction) that integrate the advantages of both the global and local features of video contents by fusion and to employ the DENsity-based CLUstEring algorithm (DENCLUE) to significantly improve the quality of abstract videos. Utilizing fusion and the DENCLUE algorithm resulted in increased robustness of this approach to handle large-scale and noisy videos with no further tuning of the input parameters [38].

突然和渐进
硬切割和软切割

2.2. Shot Based Keyframe Extraction

Shot boundary detection is one of the methods to identify the considerable changes in the content of the video. Video shot segmentation is a significant step in keyframe selection, video summarization, and video indexing for retrieval. The keyframe extraction is done by extracting a keyframe per shot. There are two types of shot boundaries, abrupt and gradual. It is also called hard cuts and soft cuts, respectively. Abrupt shot boundaries occur when the scene changes immediately between two frames, *e.g.*, when the camera focuses changes from one person to another during a conversation. Gradual shot boundaries, on the other hand, involve gradual scene changes over several frames. Gradual shot boundaries often occur at the beginning or end of television shows, advertisements, and movies; the effects include fade in, fade out, and dissolve [39, 40].

The Shot Reconstruction Degree (SRD) is used for keyframe extraction. Based on the degree of retaining motion dynamics of a video shot that examines the representativity of a keyframe set from a viewpoint of its capability to reconstruct the original video shot through frame interpolation. This keyframe selection shows good performance in terms of both fidelity and shot reconstruction degree [41]. From the video frame, the multivariant feature vectors are extracted and arranged in a feature matrix. Then the feature matrix is factorized by singular value decomposition and the significant singular vectors are computed using the sliding window approach. This sliding window approach is used to trace the rank of singular vector. By the rank of the singular vector, the shot boundaries are determined and keyframes are extracted. This work is used in real time videos to detect shot boundaries and to extract keyframes [42].

The feature vector is used to identify cut and gradual transitions. It analyzed each video frame with its immediate left and right neighbor frames for shot detection. For each frame comparison, it considers color, edge, motion, and texture features of the left and right frames. This method is suitable to perform robust shot boundary detection [43]. Estimating a global motion on the video clip that specifies conversion of the scene or scaling of the scene; and labeling each segment forming a plurality of video segment in conformity with a predetermined series of camera motion classes; keyframe candidate are extracted from the labeled segments [44].

Besiris *et al.,* [45] explained that two steps are involved to extract keyframe. The first step is the construction of MST (Minimum Spanning Tree) graph where each node is associated with a single frame of the shot. The second step is to extract the keyframes based on the principle of their maximum speed. An adaptively defined threshold controls the number of selected keyframes.

Kin-Wai and his fellow workers [46] developed an optimal keyframe representation scheme based on global statistics for video shot retrieval. Every pixel in the optimal keyframe is constructed by considering the probability of occurrence of those pixels at the corresponding pixel position among the frames in a shot. This constructed keyframe is called Temporally Maximum Occurrence Frame (TMOF). By considering the k-pixel values with the largest probabilities of occurrence and the highest peaks of the probability distribution of occurrence at each pixel position is used to improve the performance of the representation scheme. These schemes are called k-TMOF and k-pTMOF. The TMOF capture the most important visual contents of a shot.

The Generalized Gaussian Density (GGD) parameters of wavelet transform subbands along Kullback Leibler Distance (KLD) measurement are used to extract the keyframes. Here the GGD parameters are used to construct the frame feature vector and the KLD represents the distance between the GGD feature vector. At first, the KLDs between adjacent frames are used to determine the shot boundaries and segment shot into the cluster. Next, based on similarity and dissimilarity criteria, the keyframes are selected [47].

The shots in the video are further divided into sub-shots. For each sub-shot entropy is calculated. Using the subspace projector, the reconstructed matrix of each frame in a shot is constructed. The sub-shot segmentation is processed according to the difference of singular value between frames and its reconstructed matrix, as well as differences of histogram feature between the frame and the mean value of its previous N frames. Finally, the largest cross-entropy of two adjacent frames obtained and a frame which has large image entropy is selected as a keyframe [48].

The source video contains a plurality of source objects, calculating feature descriptions of some of the source objects. The source objects with similar features or combination of features are placed in clusters to obtain the relevance level of clustered source objects, then generating synopsis objects by sampling respected source objects [49].

2.3. Visual Content Based Keyframe Extraction

Visual content-based keyframe extraction is divided into color based and motion-based keyframe extraction.

2.3.1. Color Based Keyframe Extraction

The keyframe is selected based on the color based features, texture based features and shape based features. Color-based features include color histograms, color moments, color correlograms, a mixture of Gaussian models, *etc.*, [50]. The color-based features depend on color spaces such as RGB, HSV, YCbCr and normalized r-g, YUV, and HVC. Texture-based features are object surface-owned intrinsic visual features that are independent of color or intensity and reflect homogeneous phenomena in images. Shape-based features that describe object shapes in the image can be extracted from object contours or regions. A common approach is to detect edges in images and then describe the distribution of the edges using a histogram [51].

Frames are described as a set of local features. A unified framework compares the local features from the consecutive frames and constructs an auxiliary graph based on the locality of the features. Then the spectral clustering is applied to obtain tentative graph partition. Furthermore, the feature locality represented as a graph allows an independent representation of the sort of features used [52].

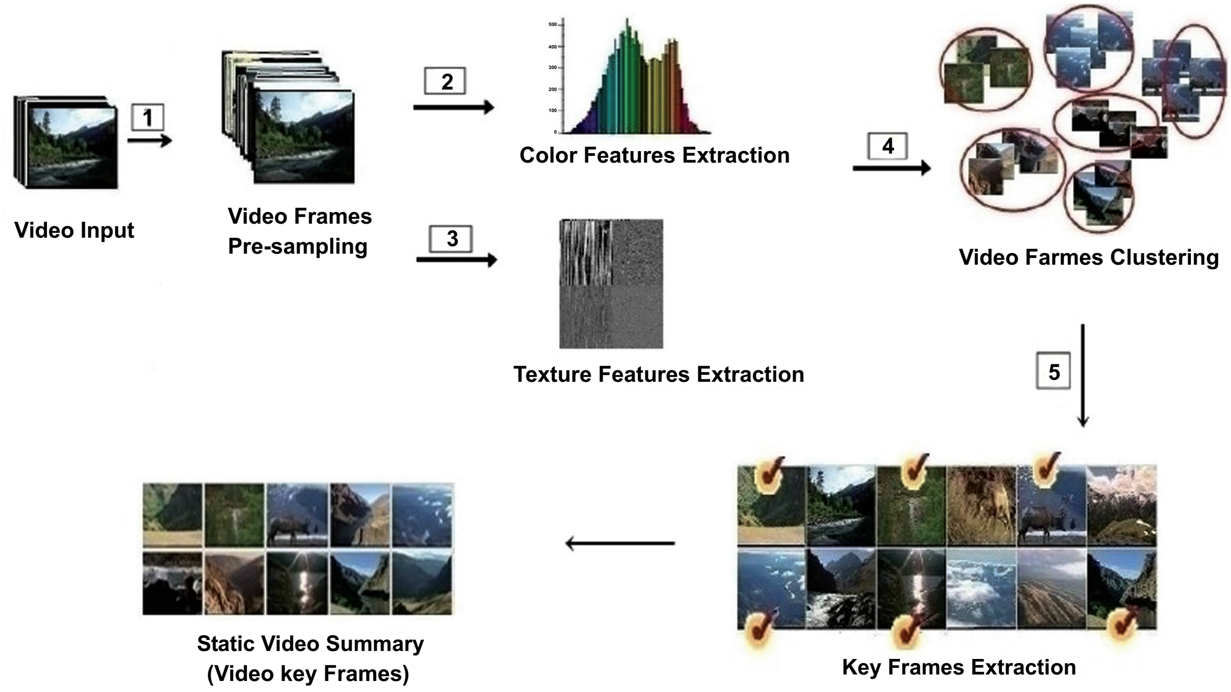


Fig. (5). Steps followed in VSCAN approach [54].

Zhonghua *et al.,* [53] worked on spatial and temporal color distribution based video keyframe extraction. Firstly, the frame which considers the spatial and temporal distribution of the pixels is constructed throughout the video shot. The weighted distance between the color histogram of each frame is calculated in the shot. The frames, which are at the peaks of the distance curve are selected as keyframes.

The VSCAN keyframe selection approach is a modified DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. The color and texture features are used.

In Fig. (**5**). the original video is pre-sampled and color features are extracted using the color histogram method and texture features are extracted using a two-dimensional Haar wavelet transform in the HSV color space. Next video frames are clustered by a modified DBSCAN clustering algorithm. After clustering, the keyframes are selected from the video clusters and arranged in the original order [54].

The video sequence is segmented into sub shots based on detection of gradual and abrupt cuts. Again the long shots are divided into sub shots based on the camera motion and location. One representative keyframe is extracted per subshot. This work is compared with IBM Multimedia Analysis and Retrieval System (IMARS) which is based on visual differences. During the gradual transition, many keyframes selected by IMARS are shaky and blurred due to ignoring motion and visual attention features. But this approach considers the highest amount of visual attention and minimal motion intensity to extract the keyframes [55]. A novel video summarization approach called VIdeo Summarization using Color Co-Occurrence Matrices (VISCOM) is based on color co-occurrence matrices to describe the video frames and generate a synopsis with the most representative frames. Experiments are conducted for three datasets, namely TRECVID, Video Segmentation (VidSeg), Open Video Project (OVP) and the results are reasonable coverage of the video content and still, some enhancement is needed to achieve good esults [56].

Based on visual attention model, instead of traditional optical flow methods, a temporal gradient based dynamic visual saliency detection is used. The static and dynamic visual attention measures are fused by using a nonlinear weighted fusion method and it reduces the computational cost [57]. Spyrou *et al.,* [58] explained that the keyframes are extracted from video shots based on their semantic content. The color and texture features are extracted from keyframe regions. A local region thesaurus is constructed for each frame by using a hierarchical clustering approach. The visual thesaurus was extracted locally from each shot.

Based on motion, color and texture features, the dynamic and static conspicuity maps are constructed. Suppression factor and motion priority schemes are introduced to conspicuity maps that are fused into a slinky map. It includes only true attention regions to produce attention curve. Finally, after time-constraint, cluster algorithm groups frames with similar content. The frames with maximum saliency value are selected as keyframes [59].

A set of features is calculated for each video frame. The features include semantic features and frame-based features. Semantic features identify the likelihoods of semantic concepts of the frame. The video is divided into segments, where each frame is associated with at least one of the semantic features. Based on the semantic value, a score is generated for each subset of frames. Finally, the representative frame is selected based on the score value [60, 61].

The source video containing a plurality of source objects. Calculating feature descriptions of some of the source objects. The source object with similar features or combination of features is placed in clusters. Obtain the relevance level of clustered source objects, then generating synopsis objects by sampling respected source objects [22].

The algorithm differentiates two consecutive frames of a video sequence, by determining the complexity of the sequence in terms of changes in the visual content that expressed by different frame descriptions. By detecting the curvature point within the curve of the cumulative frame differences, the keyframes are extracted as soon as the second high curvature point has been detected [56].

Liping Ren *et al.,* [63] analyzed the keyframe extraction which is based on image information entropy and edge matching rate. For every frame, the information entropy is calculated. Next, the edges of the candidate key are extracted by Prewitt operator. Finally, the edges of the adjacent frames are matched. If the edge matching rate is up to 50% means that keyframe is treated as a redundant keyframe and it should be discarded. The keyframes are extracted by measuring the difference between the neighboring video frames by using the Jensen- Shannon Divergence (JSD), Jensen- Renyi Divergence (JRD) and Jensen-Tsallis Divergence (JTD). Here the video is segmented into shots and further divided into sub shots and keyframes are selected. The visualization tool is used to highlight and remove the possible redundant keyframes. This approach is inexpensive and effective for keyframeextraction [64]. The distance between the neighboring frames is measured by relative entropy and its square root. The extreme Studentized deviate test is employed to identify shot boundaries to segment shots from a video sequence. The video content change is large which means the video shots are divided into sub-shots and keyframes are extracted from these sub-shots. In this approach, the redundancy is reduced [65]. Video bookmark framework is introduced to extract the keyframes where it analyzes the Luminosity. The mean luminosity of the boundary regions of the video frames is calculated and analyzed. It is more accurate than other available frameworks and less consumable regarding time [66].

2.4. Motion Based Keyframe Extraction

Keyframes are selected based on the motion of the video. Motion-based approaches are mostly based on the temporal dynamics of the scene. Motion in the video is used to extract keyframes. This method basically uses the pixel-based image differences [67] or optical flow computation [68]. Wayne wolf [25] described the optical flow computations which are used to identify the local minima of motion in a shot. It is used to identify both gestures which are highlighted by momentary pauses and camera motion. The camera motion links several different images into a single shot. At first, the optical flow is computed for each frame and simple motion metric is also computed. Then the metric was analyzed as a function of time to select the keyframes at the minima of motion.

The motion sequence is clustered into two classes by the similarity distance of the adjacent frames so that the threshold for the next step can be determined adaptively. ISODATA, a dynamic clustering algorithm is used to cluster all frames. Then the frame nearer to the center of each class is automatically selected as a keyframe without any special parameters. It is working well with the motion sequences, including different types of motion like running, jumping, kicking a ball and swordplay [10].

For higher motion video, a number of keyframes are required for summarization. The intensity of motion activity indicated the summarisability of the video segment by using the MPEG-7 motion activity descriptor. The shot is divided into equal parts of cumulative motion activity. Then the frame which is located at the halfway point of each subsegment is selected [69]. Liu *et al.,* [70] explained a triangle model of Perceived Motion Energy (PME) which is modelled motion patterns in the video. Based on this model the keyframes are extracted. It combines motion based temporal segmentation and color based shot detection. The shot is static which means the first frame was selected as a keyframe. The frames at the turning point of the motion acceleration and motion deceleration are selected as keyframes. This approach is threshold free and the extracted key frames are representative.

The difference between the magnitude of motion vector for neighboring frames within a shot to localize frames contains significant content change [71]. Ling Shao *et al.* [72] developed the keyframe extraction based on the intra-frame and inter-frame motion histogram analysis. The frames that enclose complex motion and more significantly than their neighboring frames are extracted as keyframes. It contains more actions and activities of the video. The keyframes are first initialized by finding peaks in the curve of entropy calculated on motion histograms in each video frame. By using inter-frame saliency the peaked entropies are weighted, which use histogram intersection and that produce final keyframes. This algorithm uses motion complexity maxima that the foreground objects contain and the variation of the motion between the consecutive frames to extract keyframe.

Hyun Sung Chang *et al.,* [73] discussed the keyframe extraction algorithm that applied hierarchically and obtained a tree-structured keyframe. It greatly reduces the number of frame comparisons. It generates the multilevel abstract of the video. It enables an efficient content-based retrieval by using the depth-first search algorithm with pruning. Xiaomu Song and his coworkers [74] explained keyframe extraction and object segmentation are jointly constructed by a unified feature space, where the keyframe selection is formulated as a feature selection and the context of Gaussian Mixture Model (GMM) for object segmentation. Here, two divergence criteria are used to extract keyframes. One is to maximize the pairwise inter class divergence between GMM components. Next is maximizing the marginal divergence that finds out the intra-frame variation of the mean density. This scheme extracts the representative keyframes for the object segmentation. This content-based video analysis is performed by combing keyframes and objects. This scheme shows an integrated content-based video analysis that provides a new and flexible frame/ object functionalization.

Markos Mentzelopoulos *et al.,* [75] made an attempt to explore the entropy difference algorithm to perform spatial frame segmentation. The keyframe can be extracted by the entropy that the dominant object contains. This work produces a good result when the object is distinguishable from the background. But the performance drops when the transitory changes like flashes occur.

The content-based video indexing and retrieval was analyzed by using keyframe features like texture, edge and motion. With the help of K-means clustering based methods, the keyframes were extracted. The performance of this method was compared with Volume Local Binary Pattern (VLBP) [76].

Joint Kernel Sparse Representation was introduced to capture human motion data for keyframe extraction, to model the inherent characteristics of human motion capture data. This technique model the sparseness and Riemannian manifold structure of human motion, which has two essential properties of motion data, no matter in what ways motions are recorded. Joint representation enables to obtain the internal structure of motion capture data. As well, the triangle constraint guarantees the local validity of extracting keyframes, especially for periodic motion sequences. The experimental results are good when compared with other state-of-art techniques [77].

2.5. Other Approaches

The keyframes are extracted based on sparse representation from constructing consumer videos, the video frames are projected to a low dimensional random feature space. The spatial temporal information about the video I analyzed by the theory of sparse signal representation and generate keyframes. This method does not require shot detection, segmentation or semantic understanding [78].

Based on key point based framework the local features are considered for Keyframe selection. The suitable keyframes are selected based on the two intuitive metrics of coverage and redundancy. This is one of the promising methods to extract keyframes [79]. Badre *et al.* [80] described Haar wavelet transform with various levels and thepade's sorted pentnary block truncation coding to extract keyframes To measure diversity among consecutive frames, Alias Canberra distance, Sorencen distance, Wavehedge distance, Euclidean distance and mean square error similarity measures are used.

A novel framework for Automatic Summarization of Surveillance Videos (ASOSV) was proposed. It has three properties such as: 1) Unsupervision: It works without any necessities of supervised learning or training; 2) Efficiency: It works fast 3) Scalability: It can achieve a hierarchical analysis/overview of video content. This performance is evaluated and compared with various techniques and demonstrate promising performance [81]. By focusing the analysis on the compressed video features, this paper introduces a real-time algorithm for scene change detection and keyframe extraction that generates the frame difference metrics by analyzing statistics of the macro-block features extracted from the MPEG compressed stream. A discrete Contour evolutionary algorithm is used to extract keyframe by difference metrics curve simplification [82]. The static features and the wavelet features are considered in visual attention that is integrated from the static and wavelet feature set. It is combined by using a prioritized fusion method. This fusion approach is suitable for slow motion videos and fast-moving videos [83].

3. EVALUATION TECHNIQUES FOR KEYFRAMEEXTRACTION

3.1. Evaluation

The performance of the keyframe extraction is analyzed using the metrics such as Recall and Precision, F-measure, compression ratio and processing time [84]. In the input video, detection of all events of interest was done and all redundant frames were eliminated.

3.1.1. Compression Ratio

The Compression Ratio (CR) is used to study the compactness of the shot due to selected keyframes and it depends on the number of key-frames selected. Compression ratio is computed using the equation:

CR = Total number of keyframes in a video shot / number of key-frame selected.

3.1.2. Recall and Precision

Two indexes can be estimated over test series. They are usually employed in the field of image classification, information retrieval and video segmentation.

$$\text{Precision} = \frac{\text{Number of images classified accurately}}{\textit{Total number of images classified}}$$

$$\text{Recall} = \frac{\textit{Number of images classified accurately}}{\textit{Total number of images in the database}}$$

3.1.3. F-Measure

F-measure is a benchmark metric, that consolidates both Precision and Recall values into one value using the harmonic mean [4], and it is defined as:

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

In order to evaluate the automatic video summary, the F-measure is used as a metric. By analyzing [34, 85] the experimental results two databases were used, the first one is VSUMM (VSUMM: A simple and efficient approach for automatic video summarization) database. It is composed of 50 videos selected from open video Project (OVP), which are distributed among several genres (*e.g.*, documentary, educational, ephemeral, historical, lecture). All videos are in MPEG-1 format (30 fps, 352 × 240 pixels). The duration of these videos varies from 1 to 4 min and approximately 75 min of video in total. In Table **1**, the performance of various clustering algorithms is discussed. DT-based clustering based keyframe scheme (Delaunay triangulation) was proposed by Mundur and Yesha [90]. The still and moving (STIMO) video storyboard method was introduced by Furini *et al.* which produces on-the-fly video storyboards. After

Table 1. Performance of some clustering based algorithms [32, 33].

Video	Clustering Algorithm	Precision	Recall	F-measure	Video	Clustering Algorithm	Precision	Recall	F-Measure
Video 40	DT [85]	0.23	0.16	0.19	Video 25	DT	0.45	0.14	0.44
	STIMO [81]	0.51	0.57	0.54		STIMO	0.70	0.34	0.45
	VSUMM1 [82]	0.42	0.66	0.51		VSUMM1	0.82	0.67	0.73
	VSUMM2 [82]	0.36	0.43	0.39		VSUMM2	0.91	0.67	0.77
	OFFMSR [87]	0.45	0.82	0.58		OFFMSR	0.71	0.52	0.60
	SS-MIAHC [33]	0.60	0.83	0.70		SS-MIAHC	0.88	0.88	0.88
23rd video of VSUMM Database	HDPS[32]	0.40	0.79	0.48	23rd video of VSUMM Database	VRHDPS [32]	0.63	0.68	0.63

Table 2. Performance analysis of some shot based techniques [89].

Video	Techniques	Precision	Recall	F-measure	Video	Techniques	Precision	Recall	F-Measure
VSUMM Dataset	OV	64.4	71.6	62.6	Youtube Dataset	OV	-	-	-
	DT [85]	67.7	53.2	57.6		DT	-	-	-
	STIMO [81]	60.3	72.2	63.4		STIMO	-	-	-
	VSUMM1 [82]	68.0	72.2	63.4		VSUMM1	54.6	66.7	54.8
	VSUMM2 [82]	72.8	70.5	68.8		VSUMM2	61.3	53.8	52.0
	SMFR [88]	57.7	80.2	65.0		SMFR	57.7	80.2	65.0
	SFKD [90]	44.2	75.4	53.1		SFKD	51.6	61.6	51.0
	Dictionary Learning [89]	70.0	82.3	73.4		Dictionary Learning	57.2	66.1	57.7

Table 3. Performance of some visual content based keyframe extraction techniques [98].

Method	Precision	Recall	F-Measure
DT [85]	54.7	43.3	0.483
STIMO [81]	51.9	62.1	0.565
VISON [92]	59.5	67.5	0.632
VSQUAL [98]	55.7	74.3	0.636
VSUMM [89]	72.1	64.1	0.679
VSCAN [54]	62.5	83.1	0.713
VISCOM [91]	64.9	81.1	0.721

Table 4. Some key frame extraction techniques and its application.

Reference	Technique	Operating Domain	Methodologies	Major Applications
Zhuang *et al.* [9]	Motion based	Spatio- temporal	Unsupervised Clustering	* Romantic Movie, * Action Movie
Janwe *et al.* [99]	Visual content based	Spatio- Temporal	Unsupervised Clustering	* Fast Camera Motion
Kelm *et al.* [55]	Visual content based	Spatial	*Consider high visual attention *minimal motion intensity	*Unstructured channel "Travel" video
Song *et al.* [74]	Motion based	Spatio- temporal	* GMM-based joint key-frame extraction and object segmentation	* Vehicle, Highway, Truck videos
Jacob *et al.* [100]	Visual attention based	Spatio- temporal	*Visual attention model and computer vision methods (face detection, motion estimation and saliency map computation)	*50 videos selected from the Open Video Project
Verma *et al.* [62]	Shot boundary based	Temporal analysis using histogram Spatial analysis using local features	*Hierarchical clustering	* News video, movie clip and tv- advertisement video
Liu *et al.* [70]	Motion based	Spatial filtering Temporal filtering	*Perceived Motion Energy (PME)	* Home video, sports video, news video and entertainment video
Omidyeganeh *et al.* [47]	Shot based	Spatio- temporal	*Generalized Gaussian Density (GGD) parameters of wavelet transform subbands *Kullback–Leibler Distance (KLD) measurement	* Hollywood-2 Human Actions and Scenes dataset (CVPR09) * The TRECVID 2006 shot boundary detection task dataset * Simon Fraser University (SFU) video Library and Tools dataset * The Open Video Project database

Table 5. Various keyframe extraction techniques with the tested datasets and their results.

Reference	Keyframe Extraction Method	Video Dataset Tested on	Results
[91]	VISON (VIdeo Summarization for ONline applications)	Open Video Project and videos from the YouTube	Good for online application
[92]	Anomaly detection algorithms based on saliency	20 Videos collected from various databases	Good for low quality videos
[93]	Bullet screen comments	Six movies along with the bullet screen comments	Enough for movies of different genres and languages
[94]	Semantic scene content properties	IMPART video dataset	Good for human activity videos
[95]	Unsupervised object-level video summarization	OrangeVille dataset, Base jumping, SumMe and TVSum	Produce a fine-grained keyframes
[96]	Deep Side Semantic Embedding (DSSE)	Thumb1K and TVSum50	Good depict of the video content
[97]	Fusion of the local importance and the global importance	11 Videos from the SumMe dataset and 19 videos which were downloaded from YouKu	Provide good summary with high score

extracting the 256-dimensional HSV color vector of each frame, STIMO uses an improved version of the farthest-point-first clustering algorithm to group similar frames, producing a still (moving) storyboard [81].

De Avila *et al.,* [86] presented VSUMM keyframe extraction that combines color feature extraction and k-means clustering. In VSUMM, after pre-sampling a video, color features are extracted to form a color histogram in the HSV color space. Then, the useless frames are removed and the remaining frames are grouped by k-means clustering, and one frame per cluster is selected as a keyframe. This is the simplest method and has the advantages of utilizing related concepts in the field of keyframe extraction. In SS-MIAHC, a perceptual hashing-based MI calculation is introduced for clustering. MI is an information theory tool that evaluates the similarity between two random values. It is used extensively for shot boundary detection and video summarization [34]. Table **1**, Table **2** and Table **3** explain various keyframe extraction techniques analytical evaluation. In Table **1**, the SS-MIAHC technique produces better results than other clustering algorithms. In Table **2**, dictionary learning technique in shot based keyframe extraction technique provides better results. In Table **3**, the visual attention based keyframe extraction techniques with their results are discussed.

VRHDPS based on HDPS [61], in which the number of clusters doesn’t need to be specified, as the clustering method. HDPS relied on the idea that cluster centers were characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. However, some characteristics of keyframe extraction have not been considered in HDPS. Therefore, the proposed VRHDPS clustering algorithm is more capable than the HDPS. Table **4** and Table **5** show some keyframe extraction techniques and their applications.

CURRENT & FUTURE DEVELOPMENTS

Keyframe extraction is practically an indivisible area for many video applications such as video browsing, retrieval, video abstraction, *etc.*, In the recent years the use of digital video data has been increasing significantly due to the extensive use of multimedia applications in the areas of education, entertainment, business. So the video has received an incredible attention and research interest in video processing. The state of the art of existing keyframe extraction approaches in each major issue has been described.

With the speedy development of computer, electronics, and image processing technology, keyframe extraction will be more innovative. The following aspects will be expected to become serious research topics in the future.

1) Defence

The activities of the unmanned aircraft systems are

2) Underwater Species Classification

The keyframes are extracted
to find out different species

3) Security Alert in Airpo

4) Speech Summarization

1）防御
无人机系统的活动是
2）水下物种分类
从水下视频中提取关键帧以找出不同的物种
3）机场安全警报
4）语音总结

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] L. Ying, T. Zhang and D. Tretter, “An overview of video abstraction techniques. Technical Report HPL-2001-191”, HP Laboratory, 2001.

[2] T. Ba Tu and S. Venkatesh, "Video abstraction: A systematic review and classification", *ACM transactions on multimedia computing, communications, and applications* (TOMM), Vol. 3, 2007.

[3] R. A. Hanjalic and H. J. Zhang, “An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis”, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 9, pp. 1280-1289, 1999.

[4] A. Hanjalic, “Shot-boundary detection: Unraveled and resolved?”, *IEEE Trans. Circuits Syst. Video Technol.,* Vol. 12, pp. 90-105, 2002.

[5] A. Girgensohn and J. Boreczky, “Time-constrained keyframe selection technique”, *In Proceedings of the 6th IEEE International Conference on Multimedia Computing and Systems (ICMCS ’99)*, 1999, pp. 756-776.

[6] Y. Taniguchi, A. Akutsu and Y. Tonomura, “Panorama Excerpts: Extracting and packing panoramas for video browsing”, *In Proceedings of the 5th ACM International Multimedia Conference (Multimedia ’97)*, 1997, pp. 427-436.

[7] L. Rainer, S. Pfeiffer and W. Effelsberg, "Video abstracting" *Comm. ACM,* Vol. 40, pp. 54-62, 1997.

[8] S. Yang and X. Lin, “Key frame extraction using unsupervised clustering based on a statistical model”, *Tsinghua Sci. Technol.*, Vol. 10, pp.169-173, 2005.

[9] Z. Yueting, Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering”, In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, Vol. 1, pp. 866-870, 1998.

[10] N. Ejaz, T. Tariq and S. Baik, “Adaptive key frame extraction for video summarization using an aggregation mechanism”, *Information Syst.,*Vol. 23, pp. 1031-1040, 2012.

[11] C. Bo-Wei, J.-C. Wang and J.-F. Wang, "A novel video summarization based on mining the story-structure and semantic relations among concept entities", *IEEE Trans. Multimedia,* Vol. 11, pp. 295-312, 2009.

[12] Y. Peng and C.-W. Ngo, “Clip-based similarity measure for query-dependentclip retrieval and video summarization”, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 16, pp. 612-627, May, 2006.

[13] S. Lu, I. King and M. R. Lyu, “A novel video summarization framework for document preparation and archival applications”, *In Proc.2005 IEEE Aerospace Conf., Big Sky, MT,* 2005, pp. 1-10.

[14] J. M. Odobez, D. Gatica-Perez and M. Guillemot, “Spectral structuring of home videos”, *In International Conference on Image and Video Retrieval* . Springer, Berlin, Heidelberg, 2003, pp. 310-320.

w, N. Sebe, C. Djeraba and R. Jain, “Content-based mul-
nformation retrieval: State of the art and challenges”, *ACM*
ultimedia Computing, Comm., Applications (TOMM),
.1-19, 2006.

, L. Lu, H.-J. Zhang and M. Li, “A user attention model
summarization”, *In Proc. 10th ACM Int. Conf. Multime-*
les-Pins, France, 2002, pp. 533-542.

[17] J. You, G. Liu, L. Sun and H. Li, “A multiple visual models based perceptive analysis framework for multilevel video summarization*”, IEEE Trans. Circuits Syst. Video Technol.,* Vol. 17*,* pp. 273-285, 2007.

[18] Y.F. Ma, X.-S. Hua, L. Lu and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization”, *IEEE Trans. Multimedia,* Vol. 7, pp. 907-919, 2005.

[19] Y. Li, S.-H. Lee, C.-H.Yeh and C.-C. J. Kuo, “Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques”, *IEEE Signal Process. Mag.,* Vol. 23, pp.79-89, 2006.

[20] Z. Tong, "Key-frame extraction from video”, U.S. Patent 8,379,154, issued February 19, 2013.

[21] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos”, *IEEE Transactions Multimedia,* Vol. 7, pp. 1097-1105, 2005.

[22] C. Feichtenhofer, A. Pinz and A. Zisserman, “Convolutional two-stream network fusion for video action recognition”, In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2016, pp. 1933-1941.

[23] S. Xiaomu and G. Fan, "Joint key-frame extraction and object-based video segmentation", *In Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on,* Vol. 2, pp. 126-131, 2005.

[24] Y. Hadi, M. Rahmati and S. Khadivi, "Content based video retrieval using information theory", *Machine Vision and Image Processing (MVIP), 2013 8th Iranian Conference on,* IEEE, 2013, pp. 214-218.

[25] W. Wolf, "Key frame selection by motion analysis”, *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA, 1996, Vol. 2, pp. 1228-1231.

[26] S. Angadi and V. Naik, "Entropy based fuzzy C means clustering and key frame extraction for sports video summarization”, *2014 Fifth International Conference on Signal and Image Processing*, Bangalore, India, 2014, pp. 271-279.

[27] M. Jian, S. Zhang, X. Wang, Y. He and L. Wu. "Deep key frame extraction for sport training”, In *CCF Chinese Conference on Computer Vision*, Springer, Singapore, 2017, pp. 607-616.

[28] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, "Beyond short snippets: Deep networks for video classification”, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 4694-4702.

[29] A. I. Ioannidis, V. T. Chasanis and A. C. Likas, "Key-frame extraction using weighted multi-view convex mixture models and spectral clustering”, *2014 22nd International Conference on Pattern Recognition*, Stockholm, Sweden, 2014, pp. 3463-3468.

[30] A. Ioannidis, V. Chasanis and A. Likas, "Weighted multi-view key-frame extraction”, *Pattern Recognit. Lett*., Vol. 72, pp. 52-61, 2016.

[31] K. Kumar, D. D. Shrimankar and N. Singh, “Eratosthenes sieve based key-frame extraction technique for event summarization in videos”, *Multimed. Tools Appl.,* Vol. 77, pp. 7383-7404, 2018.

[32] J. Wu, S. -H. Zhong, J. Jiang and Y. Yang, "A novel clustering method for static video summarization”, *Multimed. Tools Appl.,* Vol. 76, pp. 9625-9641, 2017.

[33] M. Fei, W. Jiang, W. Mao and Z. Song, "New fusional framework combining sparse selection and clustering for key frame extraction", In *IET Computer Vision*, Vol. 10, pp. 280-287, 2016.

[34] A. E. Dirik, J. Lai and M. Topkara, "Automatic static video summarization", U. S. Patent 8,687,941, April 1, 2014.

[35] C. Panagiotakis, A. Doulamis and G. Tziritas, "Equivalent key frames selection based on iso-content distance and iso-distortion principles", *Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '07)*, Santorini, Greece, Vol. 2007, pp. 29-29.

[36] C. -W. Ngo, Y. -F. Ma and H. -J. Zhang, "Automatic video summarization by graph modeling", *Proceedings Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003, Vol.1, pp. 104-109.

[37] C. -W. Ngo, Y. -F. Ma and H. -J. Zhang, "Video summarization and scene detection by graph modeling", *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 15, pp. 296-305, Feb, 2005.

[38] F. F. Chamasemani, L. S. Affendey, N. Mustapha and F. Khalid. "Video abstraction using density-based clustering algorithm", *The Visual Computer*, Vol. 34, pp. 1299-1314, October 2018.

[39] J. Baber, N. Afzulpurkar, M. N. Dailey and M. Bakhtyar, "Shot boundary detection from videos using entropy and local descriptor", *2011 17th International Conference on Digital Signal Processing (DSP)*, Corfu, Greece, 2011, pp. 1-6.

[40] R. Hammound and R. Mohr, “A probabilistic framework of selecting effective key frames from video browsing and indexing”, In *Proceedings of International Workshop on Real-Time Image Sequence Analysis*, Oulu, Finland, 2000, pp.79-88.

[41] T. Liu, X. Zhang, J. Feng and K.-T. Lo, "Shot reconstruction degree: A novel criterion for key frame selection" *Pattern Recognit. Lett.,* Vol. 25, pp.1451-1457, 2004.

[42] A.-A. Wael, "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing", *15th International Conference on Image Processing, 2008, ICIP,* 2008, pp. 3200-3203.

[43] J. Kavitha, P. Arockia and J. Rani. "Design of a video summarization scheme in the wavelet domain using statistical feature extraction”, *Intern. J. Image, Graphics Signal Process.,* Vol. 7, pp. 60, 2015.

[44] B. Amit, J. Hu and J. Zhong, "Combined-media scene tracking for audio-video summarization" U.S. Patent 8,872,979, October 28, 2014.

[45] D. Besiris, N. Laskaris, F. Fotopoulou and G. Economou, “Key frame extraction in video sequences: A vantage points approach”, *IEEE 9th Workshop on Multimedia Signal Processing*, pp. 434-437, 2007.

[46] S. Kin-Wai, K.-M. Lam and G. Qiu, "A new key frame representation for video segment retrieval”, *Trans. Circuits Syst. Video Technol.,* Vol. 15, pp.1148-1155, 2005.

[47] O. Mona, S. Ghaemmaghami and S. Shirmohammadi, "Video - keyframe analysis using a segment-based statistical metric in a visually sensitive parametric space", *IEEE Trans. Image Process.,* Vol. 20, pp. 2730-2737, 2011.

[48] P. Lei, X. Wu and X. Shu, "Key frame extraction based on sub-shot segmentation and entropy computing”, *CCPR 2009. Chinese Conference on Pattern Recognition,* pp. 1-5, 2009.

[49] R. Eitan and S. Peleg, "Method and system for producing relevance sorted video summary”, U.S. Patent 9,877,086, January 23, 2018.

[50] Z. Zhikai and Q. Gong. "Key frame extraction based on dynamic color histogram and fast wavelet histogram”, *In International Conference on Information and Automation (ICIA),* pp. 183-188, 2017.

[51] H. Weiming, N. Xie, L. Li, X. Zeng and S. Maybank, “A survey on visual content-based video indexing and retrieval”, *Trans. Syst. Man Cybern., Part C,* Vol. 41, pp. 797-819, 2011.

[52] V.-M. Ricardo and A. Bandera, "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering”, *Pattern Recognit. Lett.,* Vol. 34, pp.770-779, 2013.

[53] S. Zhonghua, K. Jia and H. Chen, "Video key frame extraction based on spatial-temporal color distribution”, *IIHMSP'08 International Conference on Intelligent Information Hiding and Multimedia Signal Processing,* pp. 196-199, 2008.

[54] K. M. Mahmoud, M. A. Ismail and N. M. Ghanem, “VSCAN: an enhanced video summarization using density-based spatial clustering”, In: Petrosino A. (eds), Image Analysis and Processing – ICIAP 2013. Lecture Notes in Computer Science, Vol. 8156. Springer, Berlin, Heidelberg.

[55] K. Pascal, S. Schmiedeke and T. Sikora, “Feature-based video key frame extraction for low quality video sequences”, *In Image Analy-*

sis for Multimedia Interactive Services, WIAMIS'09. 10th Workshop on, IEEE, pp. 25-28, 2009.

[56] S. Sanketh, T. Izo, M-H. Tsai, S. Vijayanarasimhan, A. Natsev, S. Abu-El-Haija and G.D. Toderici, "*Selecting and Presenting Representative Frames for Video Previews*", U.S. Patent 9,953,222, April 24, 2018.

[57] E. Naveed, I. Mehmood and S. WookBaik, "Efficient visual attention based framework for extracting key frames from videos", *Signal Process Image Commun.*, Vol. 28, pp. 34-44, 2013.

[58] S. Evaggelos and Y. Avrithis, "Keyframe extraction using local visual semantics in the form of a region thesaurus", *Semantic Media Adaptation and Personalization, Second International Workshop on*. IEEE, 2007, pp. 98-103.

[60] G. Yihong and X. Liu, "Generating optimal video summaries" *In Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on,* IEEE, 2000, pp. 1559-1562.

[61] L. Jie-Ling and Y. Yi, "Key frame extraction based on visual attention model", *J Vis Commun Image Represent.,* Vol. 23, pp. 114-125, 2012.

[62] V. Manisha and B. Raman, "A hierarchical shot boundary detection algorithm using global and local features", *In Proceedings of International Conference on Computer Vision and Image Processing,* Springer, Singapore, 2017, pp. 389-397.

[63] L. Ren, Z. Qu, W. Niu, C. Niu and Y. Cao, "Key frame extraction based on information entropy and edge matching rate", *Future Computer and Communication (ICFCC), 2010 2nd International Conference on*. IEEE, 2010, pp. V3-91.

[64] Q. Xu,Y. Liu, X. Li, Z. Yang, J. Wang, M. Sbert and R. Scopigno, "Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence", *Inform Sci.,* Vol. 278, pp. 736-756, 2014.

[65] Y. Guo, Q. Xu, S. Sun, X. Luo and M. Sbert, "Selecting video key frames based on relative entropy and the extreme studentized deviate test", *Entropy.*, Vol. 18, pp.73, 2016.

[66] D. Soumik, M. Banerjee and A. Chaudhuri, "An improved video key-frame extraction algorithm leads to video watermarking", *Int. J. Inf. Technol.*, Vol. 10, pp. 21-34, 2018.

[67] R. L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic and E. Persoon, "Visual search in a smash system", *Proc. of ICIP'96*, 1996, pp. 671-674.

[68] B. Shahraray, "Scene change detection and content-based sampling of video sequences", *Proc. of SPIE*, 1995, pp. 2-13.

[69] D. Ajay, R. Radhakrishnan and K. A. Peker, "Motion activity-based extraction of key-frames from video shots" *Image Processing. International Conference on*. Vol. 1. IEEE, 2002, pp. I-I.

[70] L. Tianming, H. -J. Zhang and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model", *IEEE T Circ. Syst. Vid.*, Vol. 13, pp.1006-1013, 2003.

[71] Z. Qiang, Q. Xu, S. Sun and M. Sbert, " Key Frame Extraction Based on Motion Vector", *In Pacific Rim Conference on Multimedia . Springer, Cham,* pp. 387-395, 2016.

[72] L. Shao and L. Ji, "Motion histogram analysis based key frame extraction for human action/activity representation", *Computer and Robot Vision, CRV'09. Canadian Conference on*. IEEE, pp. 25-28 , 2009.

[73] H. S. Chang, S. Sull and S. U. Lee, "Efficient video indexing scheme for content-based retrieval", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 9, pp.1269-1279, 1999.

[74] S. Xiaomu and G. Fan, "Joint key-frame extraction and object segmentation for content-based video analysis", *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 16, pp. 904-914, 2006.

[75] M. Markos and A. Psarrou, "Key-frame extraction algorithm using entropy difference", *Proceedings of the 6th ACM SIGMM international workshop on Multimedia Information Retrieval*. ACM, pp. 39-45, 2004.

[76] M. Ravinder and T. Venugopal, "Content-Based video indexing and retrieval using key frames texture, edge and motion features", *Intl. J. Curr. Engineer. Technol.,* Vol. 6, pp. 672-676, 2016.

[77] G. Xia, H. Sun, X. Niu, G. Zhang and L. Feng, "Keyframe extraction for human motion capture data based on joint kernel sparse representation", *IEEE Transactions on Industrial Electronics*, Vol. 64, pp.1589-1599, 2017.

[78] K. Mrityunjay and C. L. Alexander, "Key frame extraction from consumer videos using sparse representation", *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, pp. 2437-2440, 2011.

[79] G. Guan, Z. Wang, S. Lu, J. D. Deng and D. D. Feng, "Keypoint-based keyframe selection", *IEEE Trans. Circuits Syst. Video Technol.,* Vol. 23, pp. 729-734, 2013.

[80] S. R. Badre and D. T. Sudeep, "Summarization with key frame extraction using thepade's sorted n-ary block truncation coding applied on haar wavelet of video frame", *Advances in Signal Processing (CASP), Conference on*. IEEE, pp. 332-336, 2016.

[81] M. Furini, F. Geraci, M. Montangero and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario", *Multimed. Tools Appl.,* Vol. 46, pp. 47, 2010.

[82] J. Calic and E. Izuierdo, "Efficient key-frame extraction and video analysis", *Information Technology: Coding and Computing, 2002. Proceedings. International Conference on*. IEEE, pp-28-33, 2002.

[83] J. Kavitha and P. A. Rani, "Static and Multiresolution Feature Extraction for Video Summarization", *Procedia Comp. Sci.,*Vol. 47, pp. 292-300, 2015.

[84] C. V. Sheena and N. K. Narayanan, "Key-frame Extraction by Analysis of Histograms of Video Frames Using Statistical Methods", *Procedia Comp. Sci.,*Vol. 70, pp. 36-40, 2015.

[85] P. Mundur, R. Yong and Y. Yelena, "Keyframe-based video summarization using Delaunay clustering", *Intl. J. Digital Libraries,* Vol. 6, pp. 219-232, 2006.

[86] D. Avila, S. E. Fontes, A. P. B. Lopes, A. Luz and A. A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method", *Pattern Recognit. Lett.*, Vol. 32, pp. 56-68, 2011.

[87] M. Vila, A. Bardera, Q. Xu, M. Feixas and M. Sbert, "Tsallis entropy-based information measures for shot boundary detection and keyframe selection", *Signal Image Video P. (SIVIP),* Vol. 7, pp. 507-520, 2013.

[88] E. Elhamifar, G. Sapiro and R. Vidal, "See all by looking at a few: sparse modeling for finding representative objects", *In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* IEEE, pp. 1600-1607, 2012.

[89] L. Jiatong, T. Yao, Q. Ling and T. Mei, "Detecting shot boundary with sparse coding for video summarization", *Neurocomputing,* Vol. 266, pp. 66-78, 2017.

[90] R. Vázquez-Martín and A. Bandera, "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering", *Pattern Recognit. Lett.,* Vol. 34, pp. 770-779 , 2013.

[91] M. V. M. Cirne and H. Pedrini, "VISCOM: A robust video summarization approach using color co-occurrence matrices", *Multimedia Tools Appl.,* Vol. 77, pp. 857-875, 2018.

[92] J. Almeida, N. J. Leite and R. S. Torres, "Vison: Video summarization for online applications", *Pattern Recognition Lett.,*Vol. 33, pp. 397-409, 2012.

[93] C. Kwan, J. Zhou, Z. Wang and B. Li, "Efficient anomaly detection algorithms for summarizing low quality videos", *Pattern Recog. Tracking XXIX,* Vol. 10649, pp. 1064906, 2018.

[94] S. Shan, F. Wang and L. He, "Movie summarization using bullet screen comments", *Multimedia Tools Appl.,* Vol. 77, pp. 9093-9110, 2018.

[95] M. Ioannis, A. Tefas and I. Pitas, "Summarization of human activity videos using a salient dictionary", *In Proceedings of the IEEE International Conference on Image Processing* (ICIP), 2017.

[96] Z. Yujia, X. Liang, D. Zhang, M. Tan and E. P. Xing, "Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder", *arXiv preprint* arXiv:1801.00543, 2018.

[97] Y. Yitian, T. Mei, P. Cui and W. Zhu, "Video summarization by learning deep side semantic embedding", *IEEE Trans. Circuits Syst. Video Technol.*, 2017.

[98] H. Tongling and Z. Li, "Video summarization via exploring the global and local importance", *Multimed. Tools Appl.*, pp. 1-16, 2018.

[99] J. J. Nitin and K. K. Bhoyar, "Video key-frame extraction using unsupervised clustering and mutual comparison", *Intl. J. Image P. (IJIP),* Vol. 10, pp. 73, 2016.

[100] J. Hugo, F. L. C. Pádua, A. Lacerda and A. C. M. Pereira, "A video summarization approach based on the emulation of bottom-up mechanisms of visual attention", *J. Intell. Inform. Syst.,* Vol. 49, pp. 193-211, 2017.