Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Video Structure Analysis: A survey

Talib T. Al-Fatlawi [a], Rafeef M.hamza [b], Adil L. Albukhnefis [c]

[a] *Al-Qadisiyah University/College of Computer Science and Information Technology, Diwaniyah, Iraq, talib.turkey@qu.edu.iq*

[b] *Al-Qadisiyah University/College of Computer Science and Information Technology, Diwaniyah, Iraq, rafeef.hamza@qu.edu.iq*

[c] *Al-Qadisiyah University/College of Computer Science and Information Technology, Diwaniyah, Iraq, adil.lateef@qu.edu.iq*

ARTICLE INFO

Article history:
Received: 20 /04/2023
Rrevised form: 15 /06/2023
Accepted : 21 /06/2023
Available online: 30 /06/2023

Keywords:

Video Structure Analysis (VSA)
Scenes Detection (SD)
Shot Boundary Detection (SBD)
Cut Transition (CT)
Gradual Transition (GT)
Feature Extraction (FE)
Key Frame Extraction (KFE)

ABSTRACT

Due to the great and rapid development in the electronic field, the spread of the Internet, and the diversity of social media, the spread and transmission of data have become an important matter in different areas of life. The video is one of the most important of this information, which includes a lot of information that requires storing and retrieving a large database. Therefore, video structure analysis is the basis for facilitating the video search process based on its contents, indexing, and retrieval. In this research, we will present a survey of the video structure analysis process, the basic concepts and all related processing steps, the procedures used in each method, and the most prominent works used in each step.

MSC..

https://doi.org/10.29304/jqcsm.2023.15.21304

1. Introduction

The big data revolution in multimedia has been sparked by the quick development of computer networks and multimedia technologies, which has resulted in huge daily increases in both the amount of data available and its rate of growth. The most popular type of data on the Internet is video, which can be found on sites like YouTube, Yahoo Video, and social media platforms like Facebook, Twitter, and Instagram. The issue of content management is brought on by the rapid expansion of video content. To evaluate if these videos are relevant or not, people spend their time uploading and browsing enormous videos, which is a demanding and challenging activity for humans[1].

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

Video structure analysis plays a vital role in many fields, such as video summarization, video browsing, compression, analysis, CBVIR, and so on. Generally, video is a massive volume object with high redundancy and insensitive information[2] . Video structure analysis is fairly difficult owing to the following video attributes: (1) videos contain more information than images; (2) videos contain a large volume of raw data; and (3) videos lack or possess a very small prior structure[3].

2. Basic Concepts

- **Video Definition**: A video is defined as a compilation of images, or a signal made up of an audio track and a series of frames with a particular frame rate (F_{rate} measured in frames per second, or f_{ps}). Frames are organized according to a time sequence. A video's frame count depends on the length of the video. These images take up a lot of memory space. There are roughly 20 to 30 frames each second [1], [3].

- **Video Hierarchically**: A video can be divided hierarchically into scenes, shots, and frames. Fig.1 shows the video structure.
- **Frames:** the smallest units of video that make up a shot are frames[1].
- **Video shots:** are the basic units in the video that can be defined as "a sequence of frames taken by a single camera in an uninterrupted run"[4]. In contrast to shots, which have a clear definition, it is much harder to define what a semantically meaningful scene is[5].

- **Scenes:** Different definitions have been proposed to define the scene. According to[6], a scene can be defined as "a collection of shots that form a semantic unity conceptually, a single time and place". The perspective of "semantic" contents in the definition of scenes makes the scene segmentation problem very difficult[7]. It can also be defined as video clips consisting of clusters of shots on the same topic[8]. Another definition of this concept was produced by Sundaram and Chang[9], who defined the scene as "a contiguous segment of visual data with a long-term consistency of chromaticity, lighting, and ambient sound".

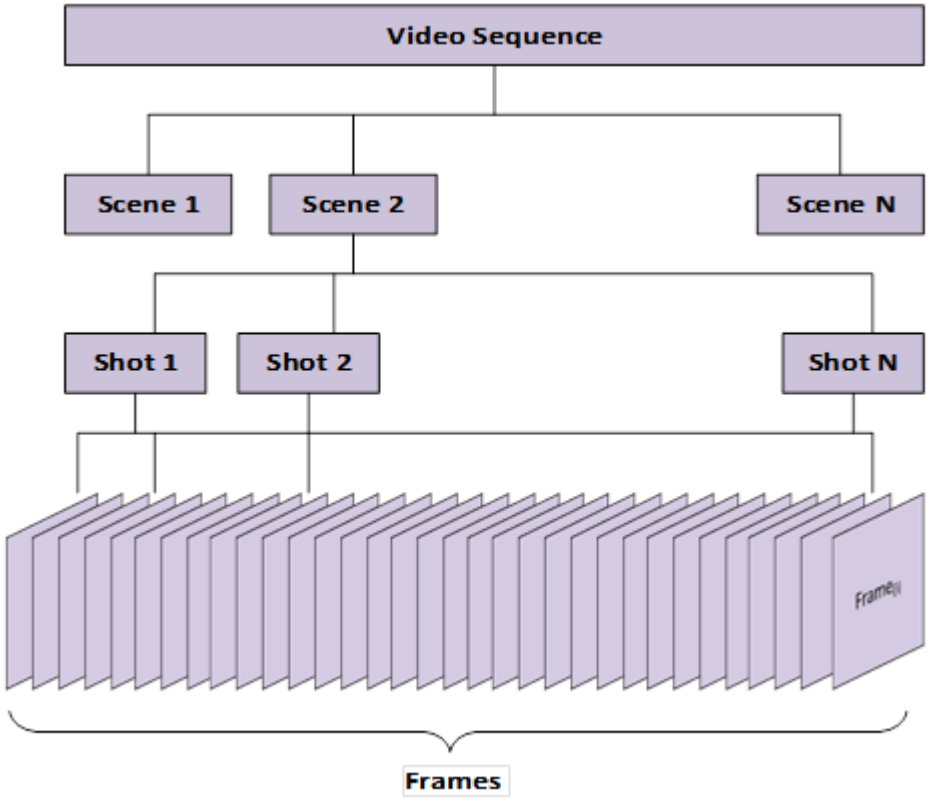

Fig.1- The Structure of Video

3. Video Structure Analysis (VSA)

Video structure analysis aims at segmenting a video into several structural elements that have semantic contents. Video structure analysis includes SBD, KF extraction, and scene segmentation[10], [11]

3.1. Scene Segmentation

Scene segmentation is known as "story unit segmentation". It can be considered as a collection of continuous shots that are consistent with a certain topic. According to[5], different low-level features are used for scene segmentation. These features may belong to visual content, audio content, textual content or may be a combination of these features, as shown in Fig. 2.

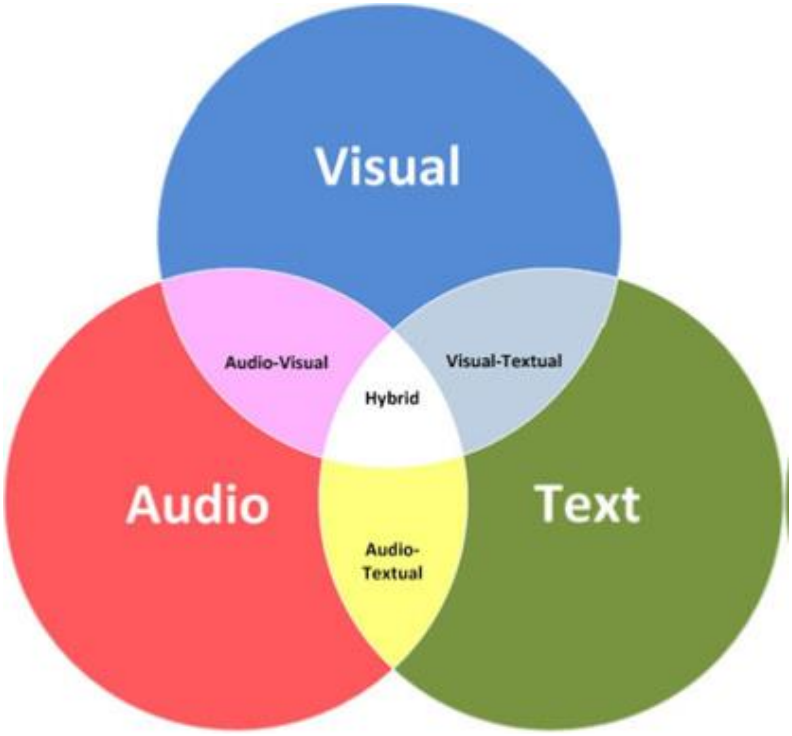

Fig.1 - Different Features Used for Scene Segmentation

In comparison with shots, scenes have a higher level of semantics, Scenes are identified or segmented by grouping successive shots with similar content into a meaningful semantic unit. Different features from texts, images, or audio may be used to group the shots into scenes. According to the representation of the shot, there are three categories of approaches for scene segmentation: methods that are based on keyframes, methods that are based on integration between audio and visual information, and methods that are based on background. These methods are explained as follows[10], [11]:

- **Methods that are based on keyframes:** This category of methods involves the extraction of a set of KFs for each shot of a video; then from this set, the features are extracted, and shots with similar features are grouped into a single scene. For example, a method for scene extraction is proposed; this method uses a block matching between the KFs of the shots, and similar shots are linked. Then the scene is extracted by connecting the overlapping links.
- **Methods that are based on integration between audio and visual information:** In this category, the shot boundary that has changes in both visual and audio content simultaneously is selected as a scene boundary. Sundaram and Chang[12] proposed a method that depends on video and audio features to detect the scene.

They detected the audio and video scenes separately and used the time-constrained nearest neighbor algorithm to find the correspondences between the two sets of scenes.

- **Methods that are based on background:** These methods assume that shots in the same scene must have similar backgrounds. Sometimes, shots within the same scene have different backgrounds, which causes a failed scene segmentation.

The scene segmentation methods can also be classified into four classes based on the processing method: the methods that are based on merging, methods that are based on splitting, methods that are based on statistical models, and finally methods that are based on shot boundary classification[10].

- **Merging-Based Methods:** In this class, similar shots are merged gradually to constitute a scene, these methods are bottom-up [13], [14].
- **Splitting-Based Methods:** These methods are opposite to first-class, they top-down style, it divided the whole video into separate scenes [15], [16].
- **Statistical Model-Based Methods:** This category of methods tries to find a statistical model of shots to segment scenes. The statistical model may be the Gaussian Mixture Model (GMM) [17], stochastic Monte Carlo [18], or the unified energy minimization framework [19].
- **Shot Boundary Classification-Based Methods:** In this category, the feature of boundaries between shots is extracted, and these boundaries are classified either as scene boundaries or not [20].

3.2. Shot Boundary Detection (SBD)

Video shot boundary detection (also known as video temporal segmentation) is a major and an essential step in any video processing applications, such as summarizing, compression, indexing and retrieval, video organization and parsing [2], and so on. It is considered a basic part of CBVIR[21]. SBD is the first fundamental unit in structural analysis and can be defined as "the process of automatically detecting the boundaries between shots in the video"[2], [10], [22].

3.2.1. Shot Transition Types

- **Cut Transitions (CT, or abrupt change, or hard)** are the simple type of transition that is generated by the typical imaging process. In CTs there are no visual effects between shots, and the next shot appears immediately after the last frame of the current shot without any effects. The frames in CT are different, either in the foreground or the background or both. This is a good indicator to exploit, since large changes in features can occur and these can be detected easily, such as in Fig.3[1], [23].

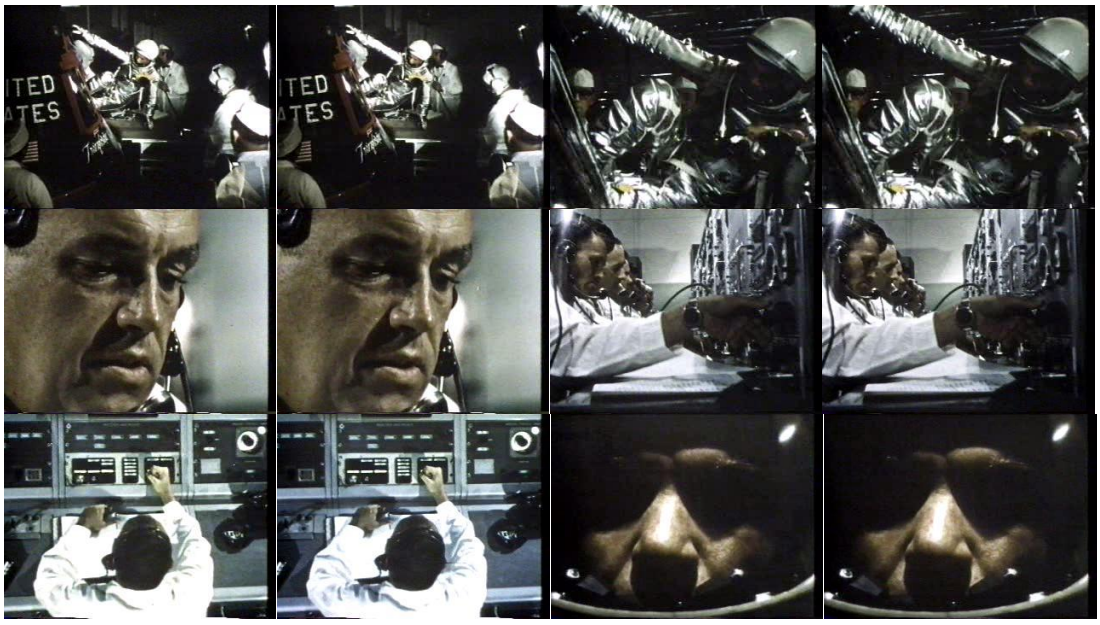

Fig.3- Examples of Cut Transition

- **Gradual Transition (GT, soft):** are considered to be more complex. They are generated by professional video editors by inserting some types of effects (fade in, fade out, dissolve, etc.) between two shots, as shown in Fig.4[1], [23]. In GTs, the neighbouring shots usually contain some interrelated frames, and these frames change slightly so that the adjacent frames during GTs do not exhibit the great differences that hard cuts do [24]. Also, GTs may be confused with undesired factors (motion, camera operation, noise, and so on); therefore, the algorithm must be able to detect all types of GT effects and be able to distinguish them from motion and noise and change in illumination, and any other factors [1][3].

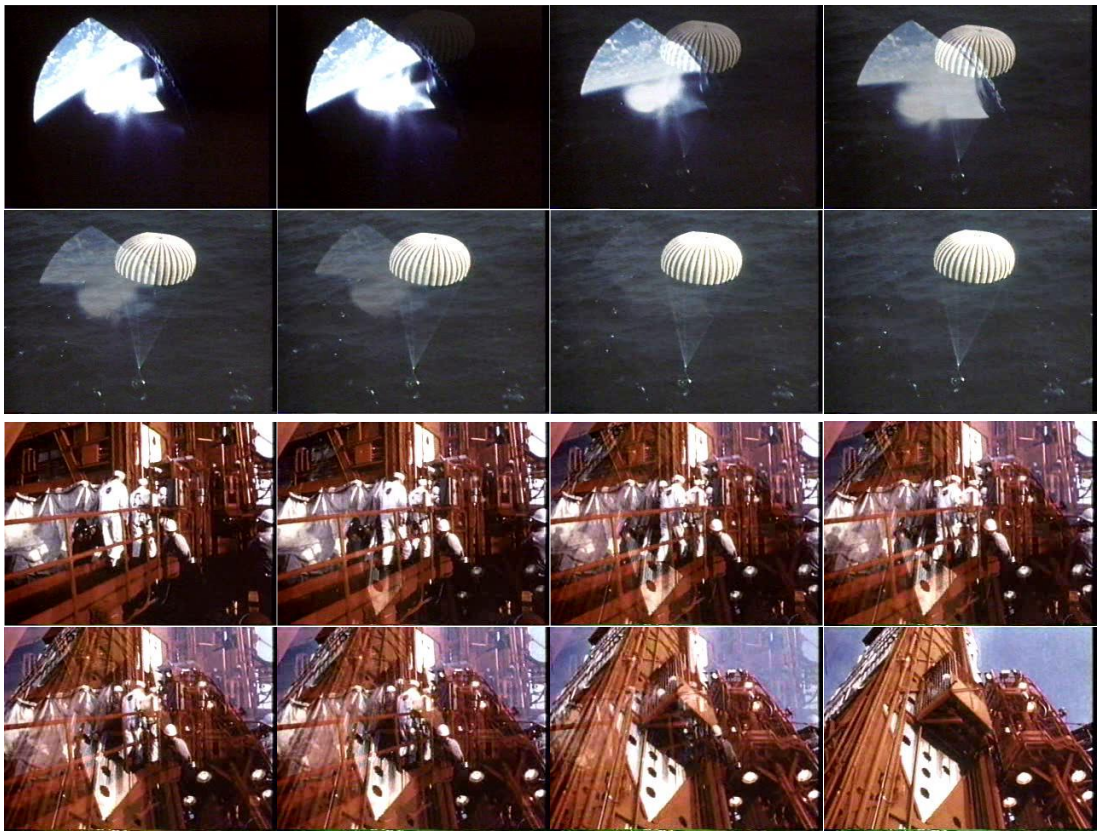

Fig.4 - Examples of Gradual *Transition*

3.2.2. Steps of SBD

There are three main steps in shot boundary detection: feature extraction, similarity measurement, and detection (classification).

- **Features extraction (Representation of Visual Information (ROVI)):** The extracted feature is resilient to temporal frame changes caused by anything moving around the frame, such as the camera or an object. Finding an effective extraction method for traits that must satisfy the requirements of being both sensitive and invariant is the aim of ROVI. An invariant feature is the representation of the frame's visual data [3].

- **Measure similarity between frames:** this is the intermediate stage between ROVI and classification sub-modules [methods]. Current similarity metrics for extracted feature vectors include the 1-norm cosine dissimilarity, the Euclidean distance, the histogram intersection, and the chi-squared similarity, as well as additional special similarity measures like the earth mover's distance and mutual information[10]. In a perfect environment, the dissimilarity signal has high values at shot transitions and low values within the same shots. The converse is true for how the similarity signal works. Due to the erratic nature of video signals, there are numerous disturbances that affect the stability of the dissimilarity/similarity signal, such as object and/or camera motion and flash light occurrence[3].

- **Classification of Dissimilarity/Similarity Signal (CLDS):** The CLDS technique is used to determine transitions and non-transitions between pictures utilizing a dissimilarity/similarity signal. A simple strategy is to group items according to a threshold. This strategy's observed transition depends on one or more fixed parameters. Methods based on a threshold are particularly sensitive to numerous distinct video genres since the threshold is selected based on one or more types of videos[3]. Also, there are several methods used, which will be explained in the section of SBD methods.

3.2.3. SBD Approaches

1) **Pixel-Based Approaches' (PBA):** The intensity of pixels is measured in this method by comparing two consecutive video frames pixel by pixel or by comparing the percentage of pixels that have changed in two consecutive frames. When the intensity of pixels exceeds the threshold, this is referred to as a shot change. The fundamental disadvantage of such techniques (i.e, intensity pixels), regardless of the metric utilized, is that they are susceptible to rapid object and camera movement, camera panning, or magnification. The limitation of this strategy is that the threshold must be established manually. This strategy has been utilized in various studies [1][3].

2) **Histogram-Based Approaches' (HBA):** The difference between the histograms of two successive frames is the most widely used statistic for cut transition detection. Since the histogram depicts the distribution of gray, color, form, and texture without considering their position, we may use it to estimate how similar two photos are. This technique first extracts the histograms from the video frames before figuring out how far apart they are from one another. Shot change is used to describe situations where the distance exceeds the threshold. The histogram distance can be calculated using a variety of techniques, including the Manhattan distance, Euclidean distance, and chi-square distance [1]. In [3], present a summery of the methods based on HBA until 2018. Also, there is research employing Singular Value Decomposition (SVD), with a Hue Saturation Value (HSV) histogram to propose a low computational complexity SBD scheme. It is used to select the candidate segments using an adaptive threshold. While other publications have developed a fresh method for identifying both abrupt (CT) and gradual (GT) transitions. to detect cut transition, these methods based on features are retrieved from the Concatenated Block Based Histograms (CBBH), color histogram, and Histogram of Gradient (HOG) features [1].

3) **Edge-Based Approaches':** The edge information of an image is yet another option for characterizing it. An edge denotes the separation between things that are overlapping and the background of an object. These techniques are used to find HT and ST. When there is a significant change between the locations of the edges in the current frame and those of the previous frame, which have vanished, edge-based techniques announce a transition [1]. Additionally, EBAs are costly and do not outperform HBAs. However, because these methods are more invariant to different lighting changes than HBAs, they can eliminate false positives brought on by flash light occurrences (sudden illumination shifts). Some authors use EBA to detect fade-in and fade-out transitions by using an EBA based on a Robert edge detector, or they use an EBA based on wavelet transform to smooth the image with a Gaussian smoothing filter with radius r and then calculate the gradient value using a Canny operator [3].

4) **Transform-based Approaches (TBAs):** It entails shifting a signal's (frame's) spatial (time) domain into the transform domain. The discrete transform is a helpful tool for signal processing and communication. It offers a significant change in terms of its potent ability to examine the constituent parts of diverse signals and permits the viewing of signals in numerous domains. The capacity to condense energy is one of the characteristics of discrete transforms. Discrete transforms include the discrete Fourier transform (DFT) and the discrete cosine transform (DCT). The base function type of the transform (polynomial) governs how different transforms differ from one another. Signals' important aspects are extracted using basis functions. For instance, DFT relies on a collection of exponential functions that are complex and connected to natural harmonics, whereas DCT is based on a cosine function with real values from -1 to 1[3].

5) **Motion-based approaches (MBAs):** Utilize the block matching algorithm (BMA) to compute motion vectors in order to distinguish between transitions and camera operations like zooming or panning. Video sequences that have been compressed (using MPEG) can have their motion vectors retrieved. However, the BMA used as part of MPEG encoding to select vectors is based on the effectiveness of the compression,

which frequently results in inappropriate vectors. SBD accuracy decreases as a result of using inappropriate motion vectors [3].

6) **Statistical-based approaches (SBAs):** Which compute the statistical characteristics for global or local frame features, are regarded as an extension of the aforementioned methods. Some examples of statistical qualities are mean, median, and standard deviation. Modeling the activity within shots and between shoots is done using these features. SBAs are somewhat noise-tolerant, but because of the complexity of the statistical computation, they are viewed as slow. SBAs also produce a lot of false positives. By calculating the mean and standard deviation of intensity pictures for regions, Jain et al. suggested a method. Also, another method instead of recognizing parabolic forms, SBA detects dissolve transitions by detecting two peaks from the second-order difference, which is seen as a difficulty for continuity signals created by the variance of frames. Hanjalic proposed a statistical model for HT and dissolve detection. Other authers observed that the true dissolve transition with huge spikes is sometimes difficult to see. They believed that a dissolve transition in the first derivative would result in the observation of a monastically growing pattern. They also identified fade-in and fade-out transitions using the same methodology; during a fade-out transition, the frames within a fade exhibit little to no variance, and during a fade-in transition, the opposite is true[3].

7) **Different Aspects of SBD Approaches:** such as Frame-skipping algorithms and algorithms based on video rhythm .Convolution neural networks (CNN) and deep learning techniques were used in [1] to address the problem of shot change detection. These techniques include genetic algorithms, fuzzy logic, SURF matching score and RGB histogram based, information theory using SVM, mixed- method approaches, and more.

3.2.4. SBD Domains

SBD approaches can be divided broadly into two categories based on the feature extraction domain, into:

- Uncompressed Domain-Based (UCD)
- Compressed Domain-Based (CD)

Shot boundary identification can be carried out immediately using features from the compressed domain, such as discrete cosine transform coefficients, DCimage and MBtypes, and motion vectors, to circumvent the time-consuming process of video decompression. However, a number of researchers have tackled the issue of SBD in COD insofar as the speedy processing technique is facilitated by the complexity of the decoding process. Direct access to the used features, like the MPEG stream, is provided by COD. Approaches based on COD, which rely on video compression standards, are less precise than those based on UCD. These techniques have a short computation time because they operate directly in COD. However, due to their dependence on the COD method, they are unable to handle visual input. Due to the shortcomings of COD, researchers are focusing more on UCD because there is much more visual information in the frame under this domain than there is under COD[3], [10].

3.3. Key Frame Extraction

KF extraction is a major topic in a video's structure analysis. It is inspired by the nature of video: as redundancy is a primary characteristic of video, redundant frames can be eliminated to make video more compact. KF extraction is the process of extracting frames or sets of frames that give a good representation of a shot. The frames must preserve the salient feature of the shot while removing most of the repeated frames[2], [25]. The use of KFs is considered a major step in CBVIR and video summarization, since it facilitates the searching of video. If we can get a compact representation of the video, this facilitates the process of indexing and retrieval that depends on the contents.

3.3.1. Size of Key Frames Set

The KFs extraction process is always accompanied by the problem of determining the size of the KFs set. Some methods proposed that for each shot there is only one KF, they proposed that a KF is the frame with maximum entropy in each shot. This proposition is not appropriate for big shots since there are changes in the scene, and some objects may appear while others may be discarded or occluded. Others have proposed that the KFs are the first, middle, or ending frames in the shot. Such methods are simple with low complexity, but the resultant KFs may have a low correlation with visual content. Therefore, the extracted KFs for each shot must give sufficient representation of the shot's frames[2].

According to [2], [10] there are three ways to determine the size of a KFs set. These techniques are described as follows:

- ***A priori*, known as a fixed number**: In this type, the number of KFs is determined a priori as a constant value before starting the extraction process. This is also known as rate constant KF extraction. Assuming that *k* is the number of KFs (the size of the KF set), then the problem of KF extraction can be rewritten as an optimization problem of finding the KF set R, where

$$\boldsymbol{R} = \{\boldsymbol{f}_{r_1}, \boldsymbol{f}_{r_2}, \dots, \boldsymbol{f}_{r_k}\} \quad (1)$$

That differs from the sequence of the video according to a certain summarization perspective:

$$\{\boldsymbol{f}_{r_1}, \boldsymbol{f}_{r_2}, \dots, \boldsymbol{f}_{r_k}\} = \boldsymbol{arg}\ \boldsymbol{\min_{r_i}}\{\boldsymbol{d}(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{\rho}) | \boldsymbol{1} \leq \boldsymbol{f}_{r_i} \leq \boldsymbol{n}\} \quad (2)$$

Where n is the video's frames number, ρ represents the summarization perspective, and D represents the distance (dissimilarity) measure. Most of the recent methods of KF extraction use ρ as "visual coverage," where it aims to cover as much as is possible of the visual content with the lowest number of KFs. P may represent the number of objects, faces, and so forth.

- ***A posteriori* (left unknown)**: In this type, the number of KFs remains unknown until the process of extraction is finished. The level of visual change determines the size of the KFs set. If the shot contains a lot of actions and movement, the required KFs to represent that shot are more than those for a static scene. Therefore, for scenes with dynamic contents, large numbers of KFs are produced, and this causes inconvenience during interactive tasks. The formulation of this type of KF extraction problem can be represented as follows [22]:

$$\{\boldsymbol{f}_{r_1}, \boldsymbol{f}_{r_2}, \dots, \boldsymbol{f}_{r_k}\} = \boldsymbol{arg}\ \mathbf{min}\{\boldsymbol{k} | \boldsymbol{d}(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{\rho}) < \boldsymbol{\varepsilon}, \boldsymbol{1} \leq \boldsymbol{f}_{r_i} \leq \boldsymbol{n}\} \quad (3)$$

Where ε is the dissimilarity tolerance (fidelity level), and n, ρ, and d have the same meaning as the first type.

- ***Determined Internally*:** This is, in essence, *a posteriori*. In this category of KF extraction methods, an appropriate size for KFs is determined before the whole extraction process is executed. For example, this involves methods that depend on clustering techniques.

Also, there exist methods that employ two types of techniques for determining the size of a KF set [26], [27]. For example, the procedure will stop when a particular condition is met or the number of KFs reaches a prior value.

3.3.2. *Key Frame Extraction Methods*

Several methods and several features are employed to extract the KFs from a video sequence. According to [2], [10], [22] the methods can be divided into several categories. The categories that are used to create or extract KFs are:

1: ***Sequential Comparison between frames:*** In these methods, each frame is compared to the KF that has been extracted previously. When the differences between a frame in the video sequence and the extracted KF are great, then this frame is indicated as a new KF. According to [2], some methods present a new KF that is extracted by computing the differences between the colour histogram of the current frame and the previous KF. The advantages of these methods are that they are simple in comparison with other methods, they demand low computational complexity, and the size and content diversity of the shot determine the size of the KF set. The drawbacks of these methods are that the KFs set may contain redundancy, which can occur when the contents appear repeatedly in the same shot, and also that the local properties of the shot are presented in the KF rather than the global properties.

2: ***Global Comparison between frames:*** These methods are based on minimizing a predefined objective function by using the global difference between frames in a shot. The objective function is selected based on the application. According to [2], the objective function can be one of the following:

- ***Minimum correlation***: These methods extract the KFs to minimize the sum of correlations between KFs. They try to make the KFs uncorrelated with each other as much as possible. Porter el al.[28] utilised a directed weighted graph to represent the shot's frames and their correlations. They employed the A* algorithm to find the shortest path in the graph, then the vertices in that path that represent the minimum correlation between frames are selected as KFs.
- **Minimum reconstruction error:** - In this category, the extraction process is done by minimizing the sum of differences between the predicated frame and each frame in the shot. The predicted frame is created by the interpolation process of a set of KFs. According to [2], [10], some researchers used an iterative procedure to select a predetermined number of KFs, and minimize the short reconstruction error. Other researchers suggested a KF selection algorithm based on the extent to which KFs record the motion in the shot. They also used an interpolation algorithm to create interpolation frames.
- **Even temporal variance:** In this type of method, in each shot, the frame with equal temporal variance is selected as a KF for that shot. The objective function can be selected to be the sum of the differences between the temporal variances of all segments. The temporal variance in a segment can be computed by the cumulative change of the contents over the segments.
- **Maximum coverage:** This category of methods attempts to maximize the representational coverage of KFs. The representational coverage can be defined as the number of a video's frames that a specific KF can represent. These algorithms either use a fidelity criterion if the size of KF set is not fixed or maximize the KF's representation if the size of the KF set is fixed.

The advantages of the techniques that are based on global comparison are as follows: the global characteristics are reflected in the extracted KF, the number of extracted KFs is controllable, and, in comparison with sequential methods, the KFs set are considered more compact. The drawback of this type is that it requires more computational complexity in comparison with sequential methods.

3: **Reference frame:** In these methods, a reference frame is generated for the KF extraction process, and each frame in the shot is compared to that reference frame. The advantage of these methods is that they are easy to implement and grasp. The drawback is that some salient content and features in the shot may be missed when the reference frame does not represent the shot adequately[2], [10].

4: **Curve Simplification**: In this category of methods, the shot's frames are represented as points in the feature space. Then the points are linked sequentially to formulate a trajectory curve and try to find the points that give the best representation of the shape of the curve. For example, a real-time method for detecting the change that occurs in the scene is presented. This method extracts the KF, the macro-block features' statistics are analyzed, and then a compressed video stream (especially MPEG) is dealt with to create frame difference metrics. Then a novel discrete contour evolution technique is employed for curve simplification using the frame difference metrics [2], [10].

5: **Clustering**: Methods in this category consider each frame in the video sequences as data points in the feature space. Cluster frames, and then the frames that have the smallest distances from (are closest to) cluster centres are selected to be KFs. Yu et al. [13] present a method for KF extraction process based on fuzzy k-means. [14] proposed a method for the KF extraction based on clustering. They first clustered the motion sequences into two classes based on similarity distances and then used an ISODATA algorithm to cluster all frames, with those closest to the clusters' centres selected as KF. Also Pan et al. in [29] also present an import shot KF extraction method that uses improved fuzzy c-means clustering. In this method colour feature information is employed, shots are clustered into sub-shots, and the frame that has the largest entropy is selected as a KF from each class. In [30] presented a method for extracting KFs and isolating the foreground. They employed a k-means algorithm along with means-squared error. The advantages of methods in this category are that the KFs have the global characteristics of the video, and a generic clustering algorithm can be used. The drawback is that they require a high computation cost, and the KFs set lacks the temporal information of the original video.

6: **Object/Events**: Methods that fall into this category consolidate the KF extraction process with the detection of objects/ events to ensure that the information of objects/ events is involved in the extracted KFs. The advantage of these methods is that the extracted KFs reflect the object or the patterns of object motion. Their disadvantage is that the heuristic rules that appoint according to the application that detects object/event play an important role. Therefore, the efficiency of these algorithms depends on experimental settings, and these settings must be chosen carefully to get good results [2].

7: **Panoramic Frame:** In order to get a KF that is representative of the features and the content of the shot and to avoid the noise in redundant frames, a panoramic frame is the best choice. With panoramic frame, all frames or specific frames in the shot are selected to create a new KF that has a full view of the shot. [31] adopted a method for video registration based on the computation of a homograph matrix between frames. The advantage of these methods is that they provide a KF that has a wide presentation for the shot. The disadvantage is that it has high computational complexity.

7. Conclusion

Due to the developments in the fields of the Internet and electronics, the search for information has become widespread. The most searched - for information is video, as it contains a lot of textual, audio, and visual information. This requires extensive efforts to provide search algorithms for these videos according to content to reduce effort and time for the user. Video structure analysis is considered one of the basics of these algorithms, in addition to being considered the basis for many other processes related to video processing. Therefore, we present in this research a summary of the basic steps and sub-steps in the video structure analysis process, the methods used in it, and the work related to it. The main steps are represented for analyze the structure of the video into the scene segment, the shot segment, and extracting key frames

References

[1] B. A. Halim, T. Faiza, and H. Seridi, “Shot Boundary Detection: Fundamental Concepts and Survey.,” in *CITSC*, 2019, pp. 34–40.

[2] I. H. Ali and T. T. Al-Fatlawi, “Key Frame Extraction Methods,” *International Journal of Pure and Applied Mathematics*, vol. 119, no. 10, pp. 485–490, 2018.

[3] S. H. Abdulhussain, A. R. Ramli, M. I. Saripan, B. M. Mahmmod, S. A. R. Al-Haddad, and W. A. Jassim, “Methods and challenges in shot boundary detection: a review,” *Entropy*, vol. 20, no. 4, p. 214, 2018.

[4] Z. M. Lu and Y. Shi, “Fast video shot boundary detection based on SVD and pattern matching,” *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5136–5145, 2013, doi: 10.1109/TIP.2013.2282081.

[5] M. Del Fabro and L. Böszörmenyi, “State-of-the-art and future challenges in video scene detection: a survey,” *Multimedia Systems*, vol. 19, no. 5, pp. 427–454, 2013, doi: 10.1007/s00530-013-0306-4.

[6] J. Varghese and K. N. Nair, “Detecting Video Shot Boundaries by Modified Tomography,” in *Proceedings of the Third International Symposium on Computer Vision and the Internet*, 2016, pp. 131–135.

[7] J. Son, S. Lee, S. Park, and S. Kim, “Video scene segmentation based on multiview shot representation,” in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 2016, pp. 381–383, doi: 10.1109/ICTC.2016.7763501.

[8] G. Gao and C. H. Liu, “Multimodality Movie Scene Detection,” in *Video Cataloguing Structure Parsing and Content Extraction*, Boca Raton, FL, USA.: CRC Press,Inc., 2015, pp. 49–60.

[9] H. Sundaram and S.-F. Chang, “Computable scenes and structures in films,” *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 482–491, 2002, doi: 10.1109/TMM.2002.802017.

[10] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A survey on visual content-based video indexing and retrieval,” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 41, no. 6. pp. 797–819, 2011, doi: 10.1109/TSMCC.2011.2109710.

[11] J. Yuan *et al.*, “A formal study of shot boundary detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 2, pp. 168–186, 2007, doi: 10.1109/TCSVT.2006.888023.

[12] H. Sundaram and S.-F. Chang, “Video scene segmentation using video and audio features,” in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, 2000, vol. 2, pp. 1145–1148, doi: 10.1109/ICME.2000.871563.

[13] L. Zhao, W. Qi, Y.-J. Wang, S.-Q. Yang, and H. Zhang, “Video shot grouping using best-first model merging,” in *Storage and Retrieval for Media Databases 2001*, 2001, vol. 4315, pp. 262–270.

[14] Z. Rasheed and M. Shah, “Scene detection in Hollywood movies and TV shows,” in *2003 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition, 2003. Proceedings., 2003, vol. 2, pp. II–343, doi: 10.1109/CVPR.2003.1211489.

[15] W. Tavanapong and J. Zhou, "Shot clustering techniques for story browsing," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 517–527, 2004, doi: 10.1109/TMM.2004.830810.

[16] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1097–1105, 2005, doi: 10.1109/TMM.2005.858392.

[17] Y.-P. Tan and H. Lu, "Model-based clustering and analysis of video scenes," in *Proceedings. International Conference on Image Processing*, 2002, vol. 1, pp. I–I, doi: 10.1109/ICIP.2002.1038099.

[18] Y. Zhai and M. Shah, "Video scene segmentation using Markov chain Monte Carlo," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 686–697, 2006, doi: 10.1109/TMM.2006.876299.

[19] Z. Gu, T. Mei, X. Hua, X. Wu, and S. Li, "EMS: Energy Minimization Based Video Scene Segmentation," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 520–523, doi: 10.1109/ICME.2007.4284701.

[20] N. Goela, K. Wilson, F. Niu, A. Divakaran, and I. Otsuka, "An SVM Framework for Genre-Independent Scene Change Detection," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 532–535, doi: 10.1109/ICME.2007.4284704.

[21] Z. Wu and P. Xu, "Shot boundary detection in video retrieval," in *2013 IEEE 4th International Conference on Electronics Information and Emergency Communication*, 2013, pp. 86–89, doi: 10.1109/ICEIEC.2013.6835460.

[22] G. Gao and C. H. Liu, *Video Cataloguing: Structure Parsing and Content Extraction*. Boca Raton, FL, USA.: CRC Press,Inc., 2015.

[23] I. H. Ali and T. T. Al-Fatlawi, "Video's Cut Transitions Detection Based on Multiple Features," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 3, pp. 1203–1211, 2019.

[24] Y. N. Li, Z. M. Lu, and X. M. Niu, "Fast video shot boundary detection framework employing pre-processing techniques," *IET Image Processing*, vol. 3, no. 3, pp. 121–134, 2009, doi: 10.1049/iet-ipr.2007.0193.

[25] S. D. Thepade and A. A. Tonge, "Extraction of key frames from video using discrete cosine transform," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies, ICCICCT 2014*, 2014, pp. 1294–1297, doi: 10.1109/ICCICCT.2014.6993160.

[26] A. Divakaran, R. Radhakrishnan, and K. A. Peker, "Motion activity-based extraction of key-frames from video shots," in *Proceedings. International Conference on Image Processing*, 2002, vol. 1, pp. I–I, doi: 10.1109/ICIP.2002.1038180.

[27] T. Liu, X. Zhang, J. Feng, and K. T. Lo, "Shot reconstruction degree: A novel criterion for key frame selection," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1451–1457, 2004, doi: 10.1016/j.patrec.2004.05.020.

[28] S. V. Porter, M. Mirmehdi, and B. T. Thomas, "A shortest path representation for video summarisation," in *Proceedings - 12th International Conference on Image Analysis and Processing, ICIAP 2003*, 2003, pp. 460–465, doi: 10.1109/ICIAP.2003.1234093.

[29] Rong Pan, Yumin Tian, and Zhong Wang, "Key-frame extraction based on clustering," in *2010 IEEE International Conference on Progress in Informatics and Computing*, 2010, vol. 2, pp. 867–871, doi: 10.1109/PIC.2010.5687901.

[30] A. Nasreen, K. Roy, K. Roy, and G. Shobha, "Key Frame Extraction and Foreground Modelling Using K-Means Clustering," in *Proceedings - 7th International Conference on Computational Intelligence, Communication Systems and Networks, CICSyN 2015*, 2015, pp. 141–145, doi: 10.1109/CICSyN.2015.34.

[31] B. Ghanem, T. Zhang, and A. Narendra, "Robust video registration applied to field-sports video analysis," *Computer Engineering*, pp. 1473–1476, 2012.