Key Frame Extraction from Video Sequence: On-The-Fly A Comparison and Analysis

Swati J. Patel[a,b], Mehul C. Parikh[b]

[a] Information Technology Department, L.D.College of Engineering, Ahmedabad, Gujarat, India
[b] Gujarat Technological University, Ahmedabad, Gujarat, India

Abstract

As a result of the digital revolution, media on the Internet has completely taken over the world. Digital video may now be made available to everyone because of the growing use of internet services, declining prices for digital storage devices, and the advent of 4-G technology. An extensive collection of videos is constantly growing, and it always takes time to analyze such a significant amount of data. Several still images called frames make up the video sequence. A video has a lot of information, and because of this, the frames often contain extraneous and identical data that is pointless if the film's content is a concern. The efficient processing of video content requires a practical and informative presentation. It is crucial to select pertinent and informative content from videos automatically. By removing replications and extracting important frames from the video, keyframe extraction is considered appropriate for thorough video analysis. A key frame is a representative frame that includes the facts of the video collection, representing vital information from the video. It is not only helpful to recognize the whole video but also can reduce the processing time, computational costs, and storage requirements of each video sequence in various applications. Extraction of these frames is one of the essential tasks in video processing. This paper presents different methods of key-frame extraction proposed in the past.

Keywords: Key-frame; Video Processing; Video Key Frame; Key Frame Extraction

1. Introduction

With the development of video recording devices such as smartphones, portable cameras, surveillance equipment, and others, video capturing, sharing, and creating becomes a straightforward process. The amount of video data has been explosively increasing. Due to the tremendous use of digital media over the Internet in information, education, entertainment, business, and surveillance, video processing has become a popular research topic in image processing [1]. It is not advisable to process a whole video sequence as a video composed of many frames at a frame rate of at least 24 frames per second (fps) for high-definition video. So it would be best if the methods could extract important frames in a video sequence that would be sufficient to represent the video and could be used to recognize the whole video sequence [2]. Key-frames provide a quicker view of video content and help reduce the computational complexity for various video analysis and retrieval applications. The video is recreated using extracted Keyframes. [3]. Keyframes are the fundamental building blocks for multiple tasks, including video browsing, summarization, searching, understanding, and chapter titles in DVDs. They are also used in numerous applications, including surveillance, medical, underwater, web browsing videos, sports and news programs, indoor and outdoor videos, and surveillance.

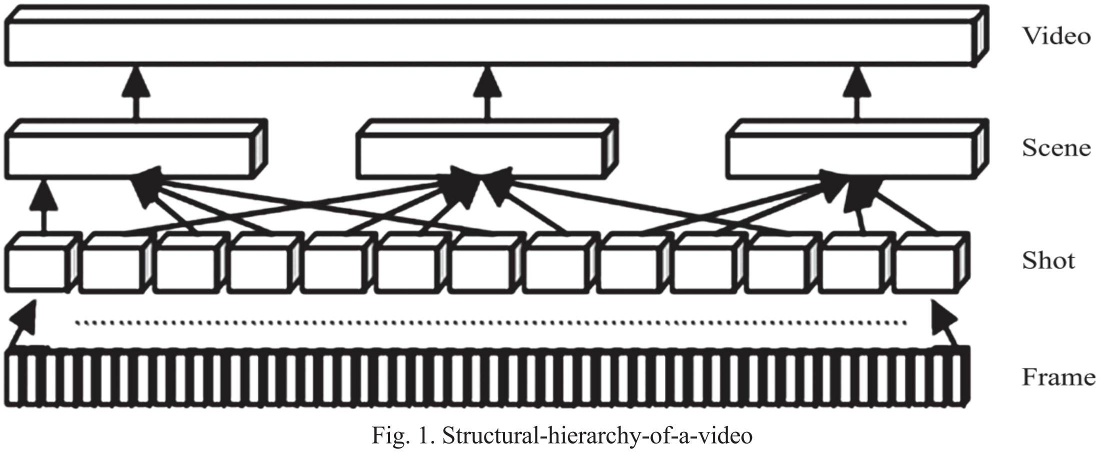

Fig. 1. Structural-hierarchy-of-a-video

Video is an immense volume of data objects containing highly redundant and insignificant information. As shown in Figure 1, the video has a complex structure consisting of scenes, shots, and frames [5]. A shot is a consecutive, adjacent sequence of frames captured by a single camera in continuous action. The key Frame is the video part representing a visual summary and meaningful information about the sequence [6]. The key Frame must contain the high-priority entities and events of the video and be free of repetition and redundancy [7]. Video processing is essential in many applications, including watermarking, various scene segmentation, shot boundary detections from that scenes, and key-frame extraction from that shots. Key-frame is either a frame or a set of frames that represent the entire content of the video clip. It refers to the image frame in the video sequence, which is representative and can reflect the summary of video content. It must contain most of the salient features of the represented video clip [3]. The concept of key-frame extraction focuses on the most specific part of a video sequence and is selected such that the video can be reproduced using the key-frames. Depending on the content complexity of the shot, one or more key-frames can be extracted from one single shot. A shot is defined as an unbroken sequence of frames recorded from a single camera, which forms the building blocks of video. In video data that contains multiple shots, it is necessary to identify individual shots for key-frame extraction [8]. Selecting key-frames from a video is a ranking process of unique frames regarding their representativeness to the video [3]. One can express the main content of video data clearly and reduce the amount of memory needed for video data processing and complexity by using the keyframe. Three properties must be considered when selecting key-frames: continuity, priority, and repetition. Continuity means that the video must be as uninterrupted as possible. Priority means that particular objects or events may be more critical than others based on a given application, and thus the key-frame must contain high-priority items. It is a highly task-dependent property. Repetition means that it is essential not to represent the same events repeatedly. It is often challenging to successfully incorporate these semantic properties [9].

2. Classification of Key Frame extraction method

2.1. Uniform Sampling Method

Uniform sampling is the most common method to extract the key Frame. In this method, every kth Frame from the video sequence is extracted, where the predefined value of k is decided from the length of the video. If the video sequence is more, then the value of k is large, else is kept small. On average, 5% to 15% from the original video, the total number of key-frames should be extracted. This concept is straightforward and does not have semantic relevance [10] [7]. As it is based on the predefined fixed value, these approaches are not content-based and do not consider the dynamics of the visual content, and selected frames are often unstable [11].

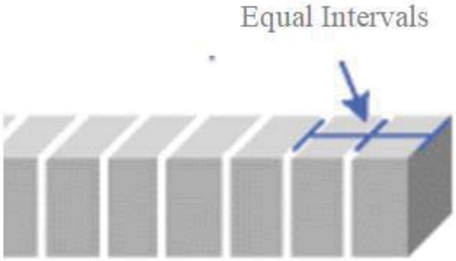

Fig. 2. Uniform sampling

2.2. Pixel Compare Method

In this method, every consecutive Frame is compared pixel-wise, and when the comparison difference crosses a given threshold value, the system identifies that Frame as a keyframe. This method is highly time-consuming and too sensitive to the motion of objects in Frame [12].

2.3. Image Histogram Method

Image histogram gives us the number of pixels for specific brightness values rated from 0 to 256. It contains essential information about an image and hence can be utilized to extract the keyframes. In this method, the histogram of each Frame is calculated and, based on the difference between the two frames, can decide the dissimilarity between them. If the histogram of two consecutive frames is either 50% or more dissimilar, then we can extract that Frame as a keyframe [7].

2.4. Scale-Invariant Feature Transform

Scale-invariant feature transform (SIFT) is used for feature detection to detect and describe the local features in an image. It is the most important method used for local features in computer vision applications. SIFT feature descriptor is invariant to uniform scaling, orientation, illumination changes, translation, and rotation and partially invariant to affine distortion. So we can use SIFT features for key frame extraction. Important locations are first defined using a scale space of smoothed and resized images and applying the difference of Gaussian functions on these images to find the maximum and minimum responses. Non-maxima suppression is performed, and putative matches are discarded to ensure a collection of highly interesting and distinct collection of key points. A histogram of oriented gradients is performed by dividing the image into patches to find the dominant orientation of the localized key points. These key points are extracted as local features [7] [13]

2.5. Cluster-Based Method

Clustering is a popular technique for the keyframe extraction method. Clustering algorithms can automatically classify video data according to their similarity. In this method, key frame clusters are created using the data points and various features of video sequences. The set of keyframes is created with frames that have the closest distance from the center of the cluster. The advantage of this method is that it covers the global characteristics of the scene. Still, it requires a high computational cost for cluster generation and feature extraction from the scene [12]. The main drawback of these methods is that, depending on the number of clusters, keyframes can be either redundant or fail to represent the content of the whole shot efficiently.

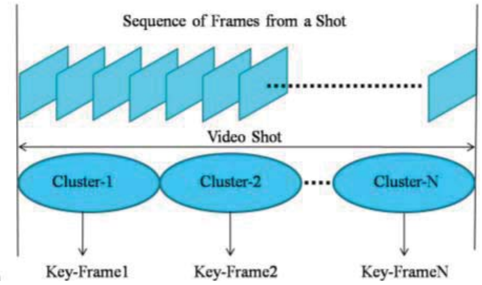

Fig. 3. Cluster-based [14]

2.6. Shot-Based Method

One technique for identifying the significant changes in the video's content is shot boundary detection. The keyframe extraction is done by extracting a keyframe per shot. The number of keyframes used to abstract a shot should be compliant with visual content complexity within the shot, and the placement of keyframes should represent the most salient visual content. The shots in the video are divided into sub-shots. For each sub-shot, entropy is calculated, and the extraction of the key Frame in each shot is based on the maximum entropy value of each shot. However, this method has the drawback of not including the content complexity and is also not appropriate and accurate for a video which is having a big shot [6].

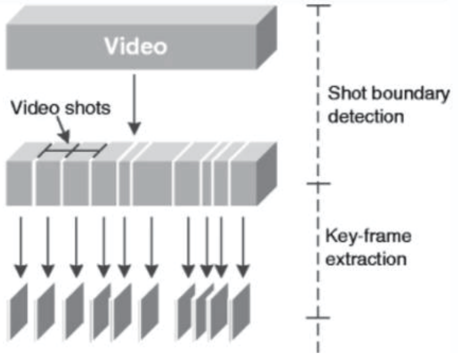

Fig. 4. Shot based

2.7. Content- Analysis-Based Method

In this method, keyframes are selected based on the color, texture, and other valuable visual information of each Frame. All the frames of the video in which this information is changing significantly are considered as the keyframes. First of all, the first Frame is selected as a new frame, and the next subsequent frames are compared with this reference frame. The kth Frame becomes a new reference frame if the distance between kth and the reference frame exceeds the predefined threshold value. This method selects the keyframe based on the degree of change in the content of the Frame. It is very insensitive to camera movement and hence produces unstable and very poor efficient keyframes [12]

2.8. Motion-Based Method

This motion-based method first segments an input video clip into homogeneous parts based on major types of camera motion, e.g., pan, zoom, pause, and steady, and dedicated rules are used to extract keyframes from each segment. Movement in the video shots can easily be detected or analyzed by analyzing the optical flow of the video sequence. In this method, the local minimum in the movement is considered the keyframe. One of the drawbacks of this method is its low robustness, as this method depends on the local information and does not count the global information for the keyframe extraction [15].

2.9. Sparse Representation-Based Method

In this sparse representation-based method, video frames are projected to a low dimensions feature space using a random projection matrix, and sparse representation is exploited in the random feature space to analyze the Spatiotemporal information of the video data and generate keyframes [15]. This approach does not require shot(s) detection, segmentation, or semantic understanding and is computationally efficient.

3. Comparison of Key Frame extraction method

Table 1. compares keyframe extraction methods based on their characteristics, merits, and demerits.

Method	Characteristics	Merits	Demerits
Uniform Sampling	Most common method	Straightforward method	Not content-based and Selected frames are often unstable
Pixel Compare	Pixel-wise comparison	Easy to evaluate	Time-consuming
Image Histogram	Similarity measure between keyframe	High-level segmentation	Don't consider the local similarities
Scale-invariant feature transform	Describe the local features in an image	Most prominent local feature	-
Cluster-based	Clustering similar frames/shots	Covers global characteristics of the scene	High computational cost (Takes 10 times the video length)
Shot based	keyframe in each shot is based on the maximum entropy value of each shot	-	Not appropriate for a big shot
Content-analysis based	Keyframes extraction based on the degree of change in the content of the Frame	Maintain good segmentation results	Insensitive to camera movement.
Motion-based	Adopts advantage of the digital capture device.	Reduce the spatiotemporal effects	High-quality video expected

4. Prior Art of Key frame extraction method

The earlier studies done in the key frame extraction field are briefly summarized in this section.

The first clustering-based keyframe extraction algorithm was published in 1998 by Zhuang et al [16]. The number and size of the clusters are used to determine which keyframes to use. A cluster is formed by comparable visual frames, and the cluster's visual content may include colour, texture, and shape. This method is more efficient, quick to compute, and simple to apply to online processing. This method was tried on two films: a romantic comedy(movie 1) and an action movie (movie 2). The second movie has more keyframes than the first.

A frame from the collection that is frequently distinct from its succeeding neighbor is chosen as the key Frame. The fuzzy C-means clustering algorithm is used to group the visually comparable frames into one group. After clustering, frames with change ratios—a metric for content variation—that is higher than the average value of the cluster are handled as Keyframes. This technique was tested using available video datasets, football footage, and sports videos on YouTube [17].

By fusing the key aspects of the video, Jiaxin Wu et almethod .'s [18] allows comparable frames to cluster together. Pre-sampling is used in the first stage to decrease the video frame's redundancy and generate candidate frames. In order to represent the visual content of candidate frames, the BoW (Bag of Words) model is used. Lastly, the VRHDPS (Video Representation based High-

Density Peaks Search) clustering technique groups candidate frame data into groups. As a keyframe, the central value of each cluster is compiled.

Keyframe extraction requires two phases, according to Besiris et al. [19]. The creation of the MST (Minimum Spanning Tree) graph, in which each node is connected to a single frame of the shot, comes first. The keyframes are extracted in the second stage using the maxim speed approach. The number of selected keyframes is controlled by an adaptively set threshold.

On the basis of spatial and temporal color distribution, Zhonghua et al [20].'s research focused on video keyframe extraction. First, a frame is built during the video shot that takes into account the spatial and temporal distribution of the pixels. The shot calculates the weighted separation between each Frame's color histogram. As keyframes, they choose the frames that are closest to the distance curve's peaks.

According to Spyrou et al. [21], keyframes are taken from video clips based on their semantic context. Keyframe regions are used to extract the color and texture features. Each Frame's local region thesaurus is created using a hierarchical clustering method. Each photo contained a local extraction of the visual thesaurus.

Each video frame is assigned a collection of features. Semantic features and frame-based features are among the features. Semantic characteristics pinpoint the likelihood of the Frame's semantic concepts. Each Frame in each segment of the video is connected to at least one of the semantic attributes. For each collection of frames, a score is generated based on the semantic value. Lastly, the score value is used to choose the representative frame [22].

Keyframe extraction was created by Ling Shao et al. [23] based on intra-frame and inter-frame motion histogram analysis. Keyframes are extracted from the frames that contain complicated motion and are more significant than their neighboring frames. It includes more of the video's actions and activities. Finding peaks in the entropy curve that is produced using the motion histograms in each video frame is the first step in initializing the keyframes. The peaked entropies are weighted using inter-frame saliency, which employs histogram intersection and results in the creation of final keyframes. The foreground objects' motion complexity maxima and the variance in motion between subsequent frames are used by this approach to extract keyframes.

The keyframe extraction approach that was performed hierarchically to produce a keyframe with a tree-structured was discussed by Hyun Sung Chang et al. in their study [24]. There are a lot fewer frame comparisons as a result. It creates the video's multilevel abstract. By utilizing the depth-first search technique with pruning, it offers an effective content-based retrieval.

Keyframe extraction and object segmentation are concurrently built by a unified feature space, according to Xiaomu Song and his colleagues [25]. The keyframe selection is articulated as a feature selection inside the context of the Gaussian Mixture Model (GMM) for object segmentation. In this case, keyframes are extracted using two divergence criteria. Maximizing pairwise interclass divergence between GMM components is one strategy. The next step is to maximize the marginal divergence, which determines how the mean density varies within frames. With this method, the representative keyframes for object segmentation are extracted. Combining keyframes and objects allows for the performance of this content-based video analysis. This scheme demonstrates an integrated content-based video analysis that offers a novel and adaptable functionalization of frames and objects.

The entropy difference approach was investigated by Markos Mentzelopoulos et al., [26] in an effort to segment spatial frames. The entropy that the dominating item possesses can be used to extract the keyframe. When the object can be distinguished from the backdrop, this work produces good results. Yet, when transient changes like flashes happen, performance suffers.

Keyframe attributes like texture, edge, and motion were used to analyze the content-based video indexing and retrieval. Keyframes were retrieved using clustering techniques based on K-means. Comparing this method's effectiveness to the Volume Local Binary Pattern (VLBP)[27].

To modify the fundamental properties of human motion capture data, Joint Kernel Sparse Representation was developed to capture human motion data for keyframe extraction. No matter how motions are captured, this method models the sparseness and Riemannian manifold structure of human motion, which are two crucial characteristics of motion data. The internal structure of the motion capture data can be obtained by joint representation. Moreover, the triangle restriction ensures that keyframe extraction is valid locally, particularly for periodic motion sequences. As compared to other state-of-the-art methods, the experimental results are favorable [28].

The video frames are projected to a low dimensional random feature space, and keyframes are recovered based on sparse representation from creating consumer films. The author used the notion of sparse signal representation to evaluate the spatial and temporal information of the video and produce keyframes. Shot detection, segmentation, or semantic comprehension are not necessary with this technique [15].

The keyframe selection process takes into account local features based on a key point-based architecture. Based on the two obvious parameters of coverage and redundancy, the appropriate keyframes are chosen. One of the promising keyframe extraction techniques is this one [29].

In order to extract keyframes, Badre et al. [30] described the Haar wavelet transform with different levels and thepade's sorted pentnary block truncation coding. The Alias Canberra distance, Sorensen distance, Wavehedge distance, Euclidean distance, and mean square error similarity measurements are used to measure variety among successive frames.

5. Summary and Conclusion

The key frame extraction procedure eliminates the majority of the unnecessary frames from the video and is regarded as a fundamental unit in the structural analysis of the video. An accurate representation of the complete shot is given to the user. It's quite important in a lot of different areas, including video summarization, content-based video indexing and retrieval, video searching, video compression, and many more. This paper provides a thorough analysis of the methods used to locate the Key Frames, their benefits and drawbacks, and the challenges that a user encounters when attempting to extract the Key Frame. Although there are no standard metrics for evaluating Key Frames extraction methods, these approaches should be highly effective, reliable, and computationally simple, and the extracted Key Frames must be as small as feasible and reflective of the full video's sequence frame. Depending on the use for which it is intended, a particular way may be best. An advanced strategy for key frame extraction is the cluster-based approach.

References

1. Ma M, Mei S, Wan S, Hou J, Wang Z, Feng DD. Video summarization via block sparse dictionary selection. *Neurocomputing*. 2020;378:197-209. doi:10.1016/j.neucom.2019.07.108
2. Elahi GMEM, Yang YH. Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition. *arXiv*. 2020.
3. Mei S, Guan G, Wang Z, Wan S, He M, Feng DD. Author ' s Accepted Manuscript reconstruction Video Summarization via Minimum Sparse Reconstruction. *Pattern Recognit*. 2014. doi:10.1016/j.patcog.2014.08.002
4. Asha Paul MK, Kavitha J, Jansi Rani PA. Key-Frame Extraction Techniques: A Review. *Recent Patents Comput Sci*. 2018;11(1):3-16. doi:10.2174/2213275911666180719111118
5. Ali IH, Al-Fatlawi TT. Key Frame Extraction Methods. *Int J Pure Appl Math*. 2018;119(10):485-490.
6. Gawande U, Hajari K, Golhar Y. Deep Learning Approach to Key Frame Detection in Human Action Videos. *Recent Trends Comput Intell*. 2020:1-16. doi:10.5772/intechopen.91188
7. Jadon S, Jasim M. Unsupervised video summarization framework using keyframe extraction and video skimming. *2020 IEEE 5th Int Conf Comput Commun Autom ICCCA 2020*. 2020:140-145. doi:10.1109/ICCCA49541.2020.9250764
8. Sheena C V., Narayanan NK. Key-frame Extraction by Analysis of Histograms of Video Frames Using Statistical Methods. *Procedia Comput Sci*. 2015;70:36-40. doi:10.1016/j.procs.2015.10.021
9. Gianluigi C, Raimondo S. An innovative algorithm for key frame extraction in video summarization. *J Real-Time Image Process*. 2006;1(1):69-88. doi:10.1007/s11554-006-0001-1
10. Jeong D ju, Yoo HJ, Cho NI. A static video summarization method based on the sparse coding of features and representativeness of frames. *Eurasip J Image Video Process*. 2016;2017(1):1-14. doi:10.1186/s13640-016-0122-9
11. Barhoumi W, Zagrouba E. On-the-fly Extraction of Key Frames for Efficient Video Summarization. *AASRI Procedia*. 2013;4:78-84. doi:10.1016/j.aasri.2013.10.013
12. Kavita Sahu MSV. Key Frame Extraction From Video Sequence : A Survey. *Int Res J Eng Technol*. 2017;04(05):1346-1350. https://www.irjet.net/archives/V4/i5/IRJET-V4I5404.pdf.
13. Jin H, Yu Y, Li Y, Xiao Z. Network video summarization based on key frame extraction via superpixel segmentation. *Trans Emerg Telecommun Technol*. 2020;(February):1-11. doi:10.1002/ett.3940
14. Janwe NJ. Video Key-Frame Extraction using Unsupervised Clustering and Mutual Comparison. 2016;(10):73-84.
15. Kumar M, Loui AC. Key frame extraction from consumer videos using sparse representation. *Proc - Int Conf Image Process ICIP*. 2011;(1):2437-2440. doi:10.1109/ICIP.2011.6116136
16. Zhuangt Y, Rui Y, Huang TS, Mehrotra S. ADAPTIVE KEY FRAME EXTRACTION USING UNSUPERVISED CLUSTERING. 1998;(94):866-870.
17. Angadi S, Naik V. Entropy based fuzzy C means clustering and key frame extraction for sports video summarization. *Proc - 2014 5th Int Conf Signal Image Process ICSIP 2014*. 2014:271-279. doi:10.1109/ICSIP.2014.49
18. Wu J, Zhong S, Jiang J. A novel clustering method for static video summarization. *Multimed Tools Appl*. 2016.

doi:10.1007/s11042-016-3569-x

19. Besiris D. Key frame extraction in video sequences : a vantage points approach. 2007:0-3.
20. Chen H. Video Key Frame Extraction Based on Spatial-temporal Color Distribution. 2008:196-199. doi:10.1109/IIH-MSP.2008.245
21. Spyrou E, Avrithis Y. Keyframe Extraction using Local Visual Semantics in the form of a Region Thesaurus. 2007:98-103. doi:10.1109/SMAP.2007.39
22. R JVCI, Lai J, Yi Y. Key frame extraction based on visual attention model. *J Vis Commun Image Represent*. 2012;23(1):114-125. doi:10.1016/j.jvcir.2011.08.005
23. Shao L, Ji L. Motion Histogram Analysis Based Key Frame Extraction for Human Action / Activity Representation. 2009. doi:10.1109/CRV.2009.36
24. Chang HS, Sull S, Lee SU, Member S. for Content-Based Retrieval. 1999;9(8):1269-1279.
25. Song X, Fan G. Joint Key-Frame Extraction and Object Segmentation for Content-Based Video Analysis. 2006;16(7):904-914.
26. Mentzelopoulos M, Psarrou A. Key-Frame Extraction Algorithm using Entropy Difference. 2004:39-45.
27. Ravinder M, Venugopal T. Content-Based Video Indexing and Retrieval using Key frames Texture , Edge and Motion Features. 2016;6(2):672-676.
28. Xia G, Sun H, Niu X, Zhang G, Feng L. Keyframe Extraction for Human Motion Capture Data Based on Joint Kernel Sparse. 2016;2(c). doi:10.1109/TIE.2016.2610946
29. Guan G, Wang Z, Lu S, Deng J Da, Feng DD. Transactions Letters. 2013;23(4):729-734.
30. Badre SR, Coding ABT. Summarization with Key Frame Extraction using Thepade ' s Sorted n-ary Block Truncation Coding Applied on Haar Wavelet of Video Frame. 2016:332-336.