VIDEO SUMMARIZATION

Abstract

Video summarization is one of the primitive tasks in video processing domain that eliminates redundant frame processing and also ensures not losing any crucial information from a video. In the present scenario of digital advancement, due to extensive use of multimedia applications, the digital video repositories are growing in a big scale. Video, being one of the robust sources of information; utilization and consumption of video (both online and offline) is practicing massively in the fields of education, surveillance, business, entertainment, news etc. Video summary plays an important role in searching for a required content of interest from the ocean like video repository and also in finding a crucial moment in video surveillance. Video summarization can be implemented both statically and dynamically. When only the set of key-frames are summarized it is known as static video summarization (or a storyboard) but when a small video clip is summarized by taking small clip collections of some more consecutive frames before and after the key-frames, refers to dynamic video summarization or video skimming (like a highlight of a match). The process of finding the informative frames of interest (key-frames) from a video is called key-frame extraction, which is the most integral part of video summarization. In this Chapter, we have discussed different features in a video, feature-set selection scenarios, feature extraction techniques, modeling and classification techniques for finding a summary from the video. Video summarization has been carried out by different state-of-the-art techniques with different situations, which are discussed in this chapter.

Keywords: Video Features, Feature Extraction, Video summarization

Authors

Mrinal Jyoti Sarma
Department of Computer Science and Engineering
Rajiv Gandhi University
Arunachal Pradesh, India.
mrinaljyotisarma@gmail.com

Marpe Sora
Department of Computer Science and Engineering
Rajiv Gandhi University
Arunachal Pradesh, India.
marpe.sora@rgu.ac.in

Bomken Kamdak Bam
Department of Computer Science and Engineering
Rajiv Gandhi University
Arunachal Pradesh, India.
bomken.kamdak@rgu.ac.in

I. INTRODUCTION

Video summarization refers to the process of condensing a video into a shorter version while preserving its main content and key information. It involves extracting the most important frames, scenes, or segments from the original video and arranging them in a concise and coherent manner.

In the present scenario of digital advancement, due to extensive use of multimedia applications, the digital video repositories are growing in a big scale. Video, being one of the robust sources of information; utilization and consumption of video (both online and offline) is practicing massively in the fields of education, surveillance, business, entertainment, news etc. In this busy world people want only the sufficient information of their interest, for example, a highlight of 30 minutes of an ODI cricket match is sufficient for a general viewer where all the wicket falls, boundaries along with some other high intensive clips of the match are covered. But for a coach, the information about field placements or some other strategies of a game may also be required to be contained by the highlight.

The process of finding only the informative frames of interest (key-frames) from a video is called key-frame extraction and the process of keeping the selected key-frames together is known as video summarization. A good video summarization output must be able to represent the input video in terms of having all the crucial and sufficient information about the video. The key-frames are selected based on the interesting features extracted from the frames of a video. When only the set of key-frames are summarized it is known as static video summarization (or a storyboard) but when a small video clip is summarized by taking small clip collections of some more consecutive frames before and after the key-frames, refers to dynamic video summarization or video skimming (like a highlight of a match).

There are generally two approaches to video summarization: keyframe-based summarization and key-shot-based summarization.

1. **Keyframe-Based Summarization:** This method selects representative frames from the video, typically based on visual or content-based features. These keyframes are chosen to capture the essential information and convey the main ideas of the video. Keyframe-based summarization focuses on reducing the temporal redundancy of the video.

2. **Key-Shot-Based Summarization:** Instead of individual frames, key-shot-based summarization identifies important shots or segments within the video. A shot refers to a continuous sequence of frames taken from the same camera viewpoint. Key-shot-based summarization aims to capture the temporal dynamics and story progression of the video.

II. VIDEO SUMMARIZATION TECHNIQUES OFTEN INVOLVE SEVERAL STEPS, INCLUDING:

1. **Video Segmentation:** Breaking down the video into smaller units, such as shots or scenes, based on visual or temporal cues.

2. **Feature extraction:** Extracting visual, audio, or textual features from the video to represent its content. These features can include color histograms, motion vectors, audio patterns, or text transcripts.

3. **Importance Scoring:** Assigning scores or weights to the video segments or keyframes based on their relevance, importance, or novelty. This step typically involves analyzing various features and applying machine learning algorithms.

4. **Selection and Ordering:** Selecting the most representative or informative segments or keyframes based on the importance scores. These selected units are then arranged in a coherent manner to form the video summary.

   Video summarization finds applications in various domains, such as surveillance video analysis, news summarization, sports highlights, video browsing, and content organization. It helps users quickly grasp the main content of long videos, saving time and enhancing the overall video viewing experience.

III. VIDEO SEGMENTATION

Video segmentation is the process of partitioning a video into different regions or segments based on certain criteria. It involves identifying and separating the foreground objects or regions of interest from the background or other parts of the video.

Video segmentation is a fundamental step in various computer vision tasks, such as object tracking, object recognition, video editing, visual effects, and video compression. It allows for isolating specific objects or regions within a video, enabling further analysis or manipulation.

There are different approaches to video segmentation, including both traditional and deep learning-based methods. Here are some commonly used techniques:

1. **Background Subtraction:** This method assumes that the background of a video sequence remains relatively static. It involves creating a background model from a set of initial frames and then subtracting this model from subsequent frames to extract the foreground objects.

2. **Optical Flow:** Optical flow methods track the motion of pixels between consecutive frames. By analyzing the motion vectors, regions with consistent motion patterns can be identified as separate segments.

3. **Graph Cut:** Graph cut methods treat video segmentation as an energy minimization problem. They model the video as a graph, where nodes represent pixels and edges represent connections between pixels. By minimizing an energy function, the graph cut algorithm separates the video into different segments.

4. **Deep Learning:** With the advancements in deep learning and convolutional neural networks (CNNs), video segmentation can be approached using deep learning architectures. Methods like semantic segmentation and instance segmentation have been extended to video data, where CNN models are trained to directly predict segmentation masks for each frame.

Deep learning-based video segmentation methods often achieve state-of-the-art results due to their ability to learn complex spatiotemporal patterns. They can handle challenging scenarios with occlusions, lighting changes, and complex motion.

It's important to note that video segmentation is an active area of research, and new techniques and approaches are constantly being developed to improve the accuracy and efficiency of the process.

IV. FEATURE EXTRACTION

Feature extraction is a process in which meaningful and informative characteristics, known as features, are extracted from raw data. It is a fundamental step in various fields, including computer vision, natural language processing, signal processing, and machine learning.

For feature extraction, we firstly understand the term “Feature”. But, the term feature is not a well defined entity. It depends on the area of interest and the situation under which it is considered. Digitally an image is displayed in terms of the pixels, which contains the various information like color, contrast, energy, hue etc. or may be information less at that point. These set of pixel level information are termed as local features.

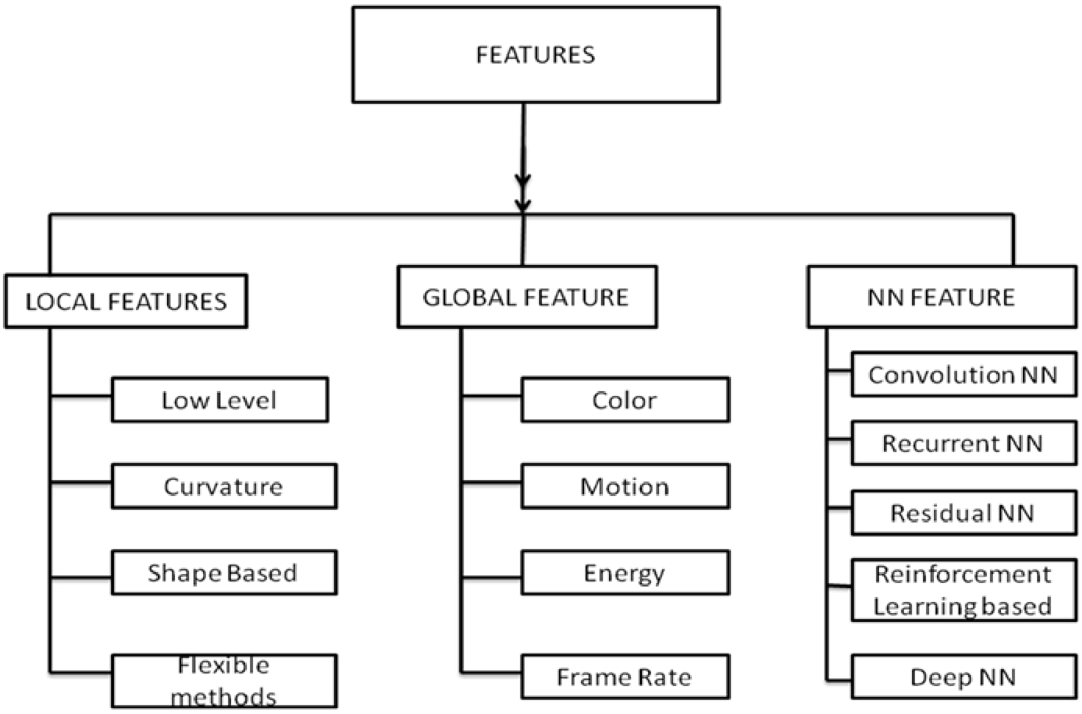

Figure 1: Features

As described in the Figure 1, features can be divided into three categories:

1. Local features,
2. Global features and
3. Neural Network (NN) based features.

Local features are the features computed as a selected part of a frame whereas global features are computed for the whole frame [1][2]. NN based features are defined and computed by the machine itself based on the type of network the system is using. Local features can be classified into:

1. Low-level features,
2. Curvature,
3. Shape based and
4. Flexible methods.

Low level features include Edge detection, Corner detection, Blob detection, Ridge detection and Scale-invariant feature transform (SIFT)[3]. Curvature features contains Edge direction, changing intensity and auto Co-relation. Different Shape based features are-Thresholding, Blob extraction, Template matching, Hough transform and Generalized Hough transform. Another type of local feature called Flexible methods contains Deformable, Parameterized shapes and Active contours.

Global features can also be classified into:

1. Color based (brightness, hue),
2. Motion based (motion detection, optical flow),
3. Energy based and
4. Intensity Based (frame rate).

Similarly, NN based features can be classified into:

1. Convolution NN features (ResNet),
2. Recurrent NN features (LSTM),
3. Residual NN features (ResNet),
4. Reinforcement Learning features (DSNet) and
5. Deep NN features.

A video consists of sequences of frames, so frame rate, sampling rate etc. are some basic features of a video. Though we normally use the term video processing separately, but inside the system it is carried out as a loop of image processing. Image processing uses feature descriptor as its first operation which is guided with higher algorithms for efficient evaluation of dominant features of that area of interest. A good number of different feature descriptors are available, based on the requirements of different interests. Feature descriptors detect the low level features like: edges, corners, blobs, ridges etc. to represent the image. Well-known feature descriptors are listed here along with the features detected by them:

1. Canny(Edge)
2. Sobel (Edge)
3. Harris & Stephens/ Plessey (Edge, Corner),
4. SUSAN(Edge, Corner),
5. Shi & Tomasi (Corner),
6. Level curve curvature (Corner) ,
7. FAST (Corner, Blob),
8. Laplacian of Gaussian (Corner, Blob),
9. Difference of Gaussians (Corner, Blob),
10. Determinant of Hessian (Corner, Blob),
11. Hessian Strength feature measures (Corner, Blob),
12. MSER (Blob),

13. Principal curvature ridges(Ridge) and
14. Grey-level blobs (Blob).

In the context of computer vision, feature extraction involves capturing relevant information from images or videos to enable subsequent analysis, classification, or recognition tasks. Instead of using raw pixel values, features provide a more compact and descriptive representation of the data, making it easier for algorithms to understand and process.

Feature extraction can be performed at different levels, depending on the complexity of the task and the specific requirements. Here are some commonly used techniques for feature extraction in computer vision:

1. **Handcrafted Features:** These are manually designed features that are specifically engineered for a particular task. They are often based on domain knowledge and prior understanding of the problem. Examples of handcrafted features include Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and Local Binary Patterns (LBP).

   A histogram is a pixel wise representation of a frame. A color (RGB) histogram [4] contains the pixel wise color (brightness, hue etc.) information for a frame with 16x16x16 matrix representation. An LBP (Local Binary Patterns) histogram is hierarchy mask to RGB histogram so that each pixel is compared with its eight neighbors in a 3x3 neighborhood matrix space by subtracting the central pixel value. Then the negative values and positive values coming for each neighborhood pixel after subtraction are encoded with 0 and 1 respectively and as a result a binary number is obtained by merging all these binary values in clock-wise direction which represents the central pixel.

2. **Convolutional Neural Networks (CNNs):** CNNs have revolutionized feature extraction in computer vision. Instead of manually designing features, CNNs learn hierarchical representations directly from the data. The convolutional layers of CNNs automatically extract features at different levels of abstraction, starting from low-level features such as edges and textures, and progressing to higher-level features representing complex patterns and objects.

3. **Transfer Learning:** Transfer learning leverages pre-trained CNN models that have been trained on large-scale datasets, such as ImageNet. By using these models as a starting point, the learned features can be transferred or fine-tuned for a specific task or dataset with limited labeled data. This approach is particularly useful when dealing with limited training data.

4. **Deep Feature Extraction:** Deep feature extraction refers to extracting features from intermediate layers of deep neural networks. Instead of using the final output of the network, features are extracted from earlier layers, which capture more detailed and localized information. These features can be used for various tasks, including image retrieval, image clustering, and object detection.

5. **Dimensionality Reduction:** In some cases, feature extraction involves reducing the dimensionality of the data while preserving its most important characteristics. Techniques

such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) can be used to transform high-dimensional feature vectors into lower-dimensional representations, making them easier to visualize and process.

The dimensionality is reduced with some higher level algorithms like PCA (Principal Component Analysis) [5][6] for faster execution. PCA is a process of computing the principal features under a real world co-ordinate by calculating the eigen-decomposition of feature set's covariance matrix or singular value decomposition of feature matrix. Normally first few principal components are considered and rests are discarded which actually reduces the dimensionality of a feature set. Robust PCA (RPCA) L1-norm-based PCA are two improved variations of PCA [7] to achieve more robustness and less computing complexity of the process. In paper [8], different RPCA implementations have been introduced:

- RPCA via PCP(Principal Component Pursuit),
- RPCA via Outlier Pursuit,
- RPCA via Iteratively Reweighted Least Squares,
- Bayesian RPCA(BRPCA),
- Variational BRPCA and
- Approximated RPCA.

6. **Transform based Feature Extraction:** Transformation based feature extraction techniques uses different transformation methods like Fourier, Gabor and Wavelet based transformations. They represent the texture by transforming into their own domain space with relative co-ordinates and original information. Fourier transformation uses frequency domain space. Gabor transformation is a special case of the short time Fourier transformation and used to determine the sinusoidal frequency over a time domain. On the other hand, discrete wavelet transformation captures both frequency domain information and location information over a time domain. So, wavelet based transformation provides better spatial localization than the other two.

   Feature extraction plays a crucial role in many machine learning and pattern recognition tasks. By extracting relevant and discriminative features, it enables algorithms to effectively learn patterns, make accurate predictions, and generalize well to new data.

V. IMPORTANCE SCORING

Importance scoring in video summarization is a technique used to identify and rank the importance of different segments or frames within a video. The goal is to select the most representative and informative content from the video and create a concise summary that captures the essential information.

There are various approaches to importance scoring in video summarization, and I'll outline a few commonly used methods:

1. **Motion-based Importance:** This method analyzes the motion patterns within a video. Typically, segments with significant motion or sudden changes are considered more important. Techniques like optical flow analysis or motion energy can be employed to estimate the motion information.

2. **Visual-based Importance:** This approach focuses on visual content analysis. It involves extracting visual features from frames or segments and using these features to determine importance. Features such as color histograms, texture descriptors, or object recognition can be utilized. Important frames might contain visually salient objects or regions.

3. **Audio-based Importance:** In addition to visual content, audio information can also play a role in determining importance. Segments with prominent or unusual audio events, such as speech, music, or sound effects, may be considered more important. Audio features like spectral characteristics or speech recognition can be employed for this purpose.

4. **Temporal Importance:** Temporal importance scoring takes into account the temporal structure of a video. Segments that represent transitions, key events, or summarize the overall content evolution might be considered more important. Techniques such as shot boundary detection or keyframe extraction can be used to identify these segments.

5. **User-based Importance:** This method involves incorporating user preferences or feedback to score the importance of video segments. Users may be allowed to rate or provide feedback on specific segments, and this information can be utilized to assign importance scores.

   It's worth noting that video summarization is a challenging task, and different approaches can be combined to achieve better results. Machine learning techniques, such as supervised or unsupervised learning, can also be employed to train models that can automatically learn the importance of different video segments.

   Overall, importance scoring in video summarization is an essential step to identify the most relevant content and create concise and informative video summaries. The specific method chosen depends on the characteristics of the video, the available features, and the desired objectives of the summarization system.

VI. SELECTION AND ORDERING

Selection and ordering are crucial aspects of video summarization that determine which segments or frames are included in the summary and how they are arranged to form a coherent and meaningful summary. Here's an overview of the selection and ordering techniques used in video summarization:

1. Selection Techniques

- **Importance Scoring:** As mentioned earlier, importance scoring methods can be employed to assign importance scores to different video segments or frames. These scores can then be used to select the most important content for the summary. Techniques such as motion analysis, visual analysis, audio analysis, or user feedback can be utilized to score the importance.

- **Diversity Maximization:** To ensure a diverse representation of the video content, selection techniques can focus on including segments that capture different aspects of the video. For example, diverse content can be achieved by selecting frames from various scenes, different camera angles, or different objects or people.

- **Keyframe Extraction:** Keyframes are frames that represent the most informative content within a segment or shot. Keyframe extraction techniques aim to identify these frames based on visual or temporal characteristics. Keyframes can serve as representative summaries of longer segments.

- **Content-based Clustering:** Clustering techniques can be employed to group similar video segments together. By selecting representative segments from each cluster, the summary can cover a wide range of content while avoiding redundancy.

2. Ordering Techniques

- **Temporal Ordering:** In many cases, maintaining the chronological order of video segments or frames can provide a coherent narrative. This is especially true for videos that capture a sequence of events or stories. The selected segments can be arranged in the summary based on their temporal positions in the original video.

- **Storytelling Structure:** Video summaries can be organized to follow a storytelling structure, where the selected segments are arranged in a logical and coherent sequence. This arrangement aims to convey the narrative or main points of the video effectively.

- **Transition Analysis:** Transitions between video segments, such as shot boundaries, can be analyzed to determine the smoothest ordering of the selected content. By considering the visual or audio continuity between segments, a more visually appealing and coherent summary can be achieved.

- **User Preference:** User preferences or predefined rules can also influence the ordering of video segments. For example, users may have specific requirements or guidelines for the summary structure, and the system can follow those preferences.

It's important to note that the selection and ordering techniques can be combined and customized based on the specific requirements and objectives of video summarization. Different approaches can be employed to strike a balance between the importance of content, diversity, coherence, and user preferences, ultimately resulting in a high-quality video summary.

Till now, in this chapter, we have discussed about different entities of the video summarization process. As we all know the video/image processing area is now highly influenced by different types of neural networks: Convolutional NN, Residual NN, Deep NN etc. and the high capability Graphics Processing Unit (GPU). So we can directly divide the video-processing approaches into two categories based on the arrival of neural networks into action, that is, prior to NN and After NN. Since video summarization was also practiced over a decade before neural networks came into action, we have divided it into following way (refer figure 2 for visualization) -

- Color Based Technique
- Motion Based Technique

- Local Features Based Technique
- Event Based Technique
- Time Based Technique
- Density Based Technique
- Hierarchical clustering based techniques
- Neural Network Based Technique

3. **Color Based Technique:** Color is one of the most expressive, simple, stable and effective feature of a frame. The color features are computed in terms of a histogram, which contains pixel level color information of an image. Color histograms are widely used due to its simplicity and robustness against small camera motion. Redundancy elimination can be done at low computational cost but it is sensitive to noise. Normally first frame is taken as a key-frame and the next key-frames are selected on basis of dissimilarity measures of color and texture features from the color histograms of consecutive frames. Video SUMMarization (VSUMM) is one example of color based technique that uses Hue components of color features in HVS color space to form a color histogram for performing a static video summarization.

4. **Motion Based Technique:** Motion based key-frame extraction techniques are gaining importance due to its expressiveness and informativeness. Motion consideration for a video can be two types: i) Object motion and ii) Camera motion. In case of video surveillance the camera is normally fixed mounted and hence camera motion is not considered. In this situation, only object motion is computed and hence computational complexity is in reduced form. But, in case of moving camera the computational complexity increases highly, which is a challenging task for using this technique. Motion estimation can be computed either by calculating pixel to pixel frame difference or by calculating the optical flow. Optical flow of each frame is calculated and results are stored in a simple motion metric, which is used for selecting the key-frames by finding the local minima of motion for a frame. Lucas- Kanade and Horn-Schunck are two popular optical flow algorithms which uses two different criteria of selecting key-frames.

   This approach is also known as domain specific approach due to its capability of catching high activity contents of sports domain videos. It is independent of skimming threshold but it fails to extract key-frames accurately when the video contains high level of motion or the video is motionless. This approach is suitable for surveillance videos with medium level of motion.

5. **Local Features Based Techniques:** As the name suggests itself this approach uses the local features of the key-points of a frame. Scale Invariant Feature Transform (SIFT) and Speeded-Up-Robust-Features are two prominent algorithms that uses local features. In implementation of SIFT, key-points(important locations) of an image frame are defined first by finding the maximum and minimum responses of features in scale space representation of Gaussian functions calculating the differences. Only the distinct and interesting key-points are considered thereafter and rests are discarded. All the local low level features from the selected key-points are then extracted to form a SIFT feature-set. In case of SURF a reduced feature-set is considered based on the dominance and robustness of features to secure less computational cost.

6. **Event Based Technique:** This approach deals with the highest semantic level of features, called events, detecting the interesting events and organizing them in essence of the original video. Normally events are detected by optical flow analysis and/or computing the energy difference of successive frames. Methods used for summarizing rare (important) events are [9]: i) RPCA-KFE, ii) Unified Framework, iii) Key-point-based Key-frame selection, iv) AJ Theft Prevention, v) Graph Modeling, vi) Two-Level Redundancy detection for personal video recorders and vii) CAFKF [14].

7. **Time Based Technique:** This approach uses a simple method of implementation by taking a constant time interval between two key-frames. Uniform Sampling [8] is one good example of time-based key-frame extraction. Here, every k^{th} frame of a video is selected as key-frame where k is evaluated from the length of video by which percentage the summary is required. For example, if we need 10% of summary of a video then every 10th frame is taken as the key frame. Similarly, for 5%, 15% or 20% of summary video the selected frames will be of 20th, 7th and 5th positions respectively. This approach doesn't require any feature to be extracted or analyzed and hence computing complexity is very less.

8. **Density Based Technique:** This approach uses density (number of frames) as criteria for clustering the frames. Normally clustering is done by feature-set similarity distances and a cluster grows when the neighborhood frame achieves the threshold of similarity distance. But here, a cluster grows when the density of its neighbors is greater than threshold and this approach is capable of discovering any arbitrary-shaped cluster and noise. Density-Based Clustering of applications with Noise (DBSCAN), DENsity-based CLUstEring(DENCLUE) and Ordering points to identify the clustering structure (OPTICS) are three well known algorithms that falls under this category. Like density, trajectory based techniques are also used in clustering of a feature extraction method [10].

9. **Hierarchical-clustering based techniques:** In this approach, hierarchy of clusters of frames is constructed based on distance, density or continuity with the independency of pre-defined clusters number [10]. Two different types of hierarchy clustering approaches are: i) Agglomerative and ii) Divisive. This approach poses a high complexity cost and the execution process gets slower when the video size increases.

10. **Neural Network Based Technique:** Neural Network uses regression method to train data and correlations between different features within the dataset are identified. Neural networks (NN) represent deep learning and use a number of hidden layers in between the input layer and output layer. Neural networks can be considered as the process of deep learning (DL) using artificial intelligence (AI). Different types of NN are available; common example includes: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long short-term memory (LSTM) , Residual Network (ResNet) etc. These deep learning networks use the training set to learn the procedure for getting predefined desired output and apply the learned procedure on new data. Reinforcement Learning (RL) is a dynamically learning network that uses continuous feedbacks for adjusting actions to maximize a reward.

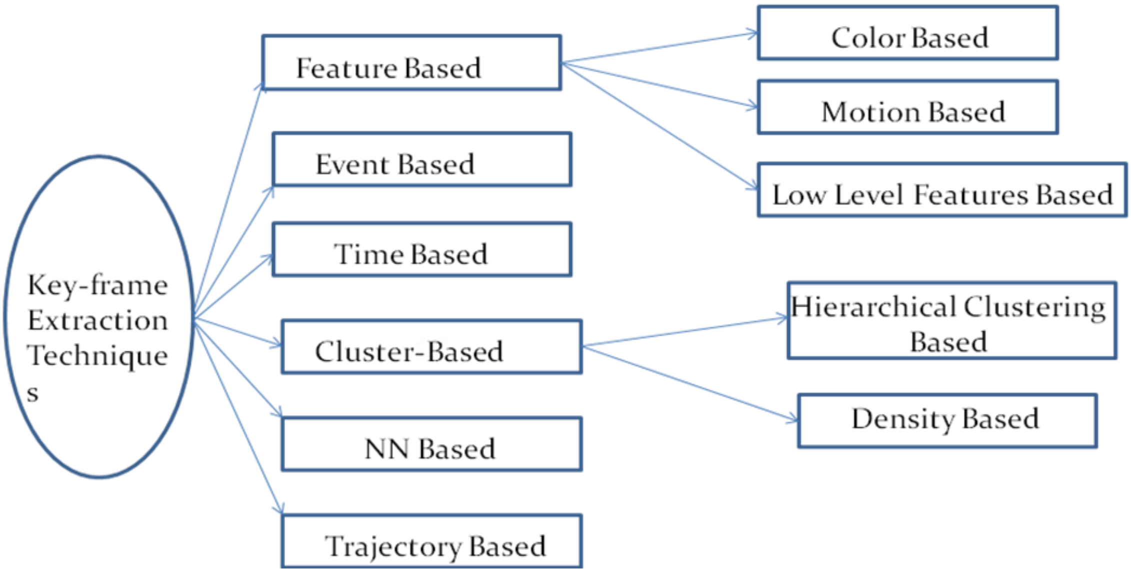

Figure 2: Types of Key-frame extraction techniques for video summarization

REFERENCES

[1] Shakya, Subarna, Suman Sharma, and Abinash Basnet. "Human behavior prediction using facial expression analysis." In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 399-404. IEEE, 2016.

[2] Bonchek-Dokow, Elisheva, and Gal A. Kaminka. "Towards computational models of intention detection and intention prediction." *Cognitive Systems Research* 28 (2014): 44-79.

[3] Kumar, Gaurav, and Pradeep Kumar Bhatia. "A detailed review of feature extraction in image processing systems." In *2014 Fourth international conference on advanced computing & communication technologies*, pp. 5-12. IEEE, 2014.

[4] Abouyahya, Anas, Sanaa El Fkihi, Rachid Oulad Haj Thami, and Driss Aboutajdine. "Features extraction for facial expressions recognition." In *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 46-49. IEEE, 2016.

[5] Stratou, Giota, Job Van Der Schalk, Rens Hoegen, and Jonathan Gratch. "Refactoring facial expressions: An automatic analysis of natural occurring facial expressions in iterative social dilemma." In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 427-433. IEEE, 2017.

[6] Suja, P., and Shikha Tripathi. "Real-time emotion recognition from facial images using Raspberry Pi II." In *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 666-670. IEEE, 2016.

[7] Gallese, Vittorio, and Alvin Goldman. "Mirror neurons and the simulation theory of mind-reading." *Trends in cognitive sciences* 2, no. 12 (1998): 493-501.

[8] Gordon, Robert M. "Folk psychology as simulation." *Mind & Language* 1, no. 2 (1986): 158-171.

[9] Janu, Neha, Pratistha Mathur, Sandeep Kumar Gupta, and Shubh Lakshmi Agrwal. "Performance analysis of frequency domain based feature extraction techniques for facial expression recognition." In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, pp. 591-594. IEEE, 2017.

[10] Bonchek-Dokow, Elisheva, and Gal A. Kaminka. "Towards computational models of intention detection and intention prediction." *Cognitive Systems Research* 28 (2014): 44-79.