CHAPTER 2
LITERATURE SURVEY

This Chapter introduces a brief survey of video summarization methods and technical achievements in this area that had motivated the research work described in the thesis. The survey includes numerous papers covering the research aspects of system design, applications of video summarization, video frame feature extraction, representation and video summary. Existing attempts on video summarization can be characterized by four main steps that includes i) shot boundary detection ii) key frame extraction iii) video representation and iv) video summarization. An open research issue is to achieve meaningful video summary for the ever-growing availability of multimedia. The video summarization scheme is categorized based on visual, textual, and audio features.

The system architecture of video summarization technique in general is as shown in Figure 2.1. It consists of input video stream, shot boundary detection, key frame extraction, video representation and video summary. The first step of video summarization is to detect the boundary of the shots in a video. Once the shot boundaries are detected, the next task is to determine the key frame(s) for each shot. In general, key-frames are selected according to certain key frame extraction schemes like Low level, medium level and high level feature extraction schemes. The extracted key frames mapped from 3D space to the 2D view screen for video representation. The representative frame(s) arranged chronologically can act as the access points to the user as a summary for browsing the video.

![Figure 2.1 General system architecture of video summarization.](image-url)
Organization of this Chapter is as follows. Section 2.1 presents input video stream as scripted and unscripted input video stream. Section 2.2 elaborates various shot boundary detection techniques and their comparison amongst each other in brief. Section 2.3 illustrates number of ways to extract key frames for video summarization. Section 2.4 intricate the representation schemes of extracted key frames. Section 2.5 describes the video summarization concepts followed by the concluding remarks.

2.1 Input video stream

The input video stream applied to the video summarization system may have scripted or unscripted contents. Scripted video contents are carefully produced according to a script or plan that is later edited, compiled and distributed for the end-users. Examples of scripted content are news videos, dramas & movies. Unscripted video content are not scripted according to a fixed plan. Examples of unscripted content are sports video, teaching video and surveillance video. It is not possible to process these video as a whole, so need to detect shot boundaries.

2.2 Shot boundary detection

Shot is a sequence of video frames which have similar characteristics. Shot boundary detection is the fundamental step in video summarization because it is not effective to process the entire video clip. Automatic shot boundary detection techniques can be classified into pixel based [1], statistics based [2], transform based [3, 4], feature based [5], and histogram based [6] detection. Costas et al [7] focused on the specific shot boundary detection algorithms and condensed video representation. The histogram-based shot boundary detection approach is most popular amongst all. The histogram based approach uses the histogram of the distribution of the intensity values in the displayed frames. Histogram difference between adjacent frames calculated and a shot change is detected when the histogram difference exceeds a threshold.

Hanjalic et al [8] discussed the shot-boundary detection problem in detail and the major issues that need to be considered. Shots extraction requires the computation of an appropriate metric (algorithm) to characterize the change of video content between two frames and a threshold to determine
whether the change is important enough to be defined as a shot boundary [9]. There are two types of shot changes, namely abrupt change and gradual change [10]. Abrupt change can be detected easily as it is always significant enough to be detected. The threshold value selected must be sensitive enough for the gradual change to be detected while not too sensitive that it would attract too many false cuts. Color features have been widely used for detecting shot boundaries due to its robustness to complex background (occlusion), scaling (image size), orientation and perspective [11, 12].

2.3 Key Frame Extraction

Once the shot boundaries are detected the next task is to select key frames based on audio-visual and text features. Video key frame extraction is one of the key problems in video summarization, video content indexing and retrieval. Extracted collection of salient images from a video sequence is used for visual content summarization. The audio-visual and textual information conveyed in a video stream is used to extract key frames. Cinematic features (shot types, replays), text information (from closed captions, text detection), audio features (whistle, excitement, applause, cheering, music, speech, speech with music, noise), object features that is spatial or spatio-temporal features (referee, players, object that use in games, color, object motion, intensity) can be used to extract key frames. Extracted collection of salient images from a video sequence is used for content summarization.

Video key frame(s) maps an entire video segment to a small collection of representative images. Most of earlier works in video key frames extraction choose to select key frames by randomly or uniformly sampling the video frames at a certain time intervals [14]. Video key frames provide a concise access to the video content [13].

Nagasaka et al [15] selected the first frame of the shot as the key frame. Rong et al [16] used a text retrieval method to extract key frames. Each frame is thought of as a word in a document (shot) and the most representative words of a given document are selected as key frames. An entropy-based method was introduced by Mentzelopoulos et al [17] where the entropy of a grays scale frame is computed and compared with that of the previous frame. If the entropy difference is higher than a user defined threshold, then the new frame is assumed to be a key frame. This method of key frame extraction is particularly helpful for browsing video contents as
users are provided with visual information about each video segment indexed. Thus, the selection of key frames is very important and there are many ways to automate the process. There are mainly two issues in selecting key frames, specifically, 1) the total of key frame(s) used and 2) the selection of good number representative frames contained by shot. For the first issue, the number of key frames for each shot can be dynamically determined based on the length of the shots such as allocating one temporary key-frame for each one second of the shot, especially, when the shot is very long [18]. Although this method considers only length of shots, the performance should be effective enough to save all the processing complexities and time needed to divide a shot into sub-shots and assign a key frame for them based on changes in contents such as motion information and optical flow [19]. For the second issue, it is generally difficult to select frames with the highest semantic usefulness. This can be tackled by reducing redundant frames using methods such as singular value decomposition [20, 21] or relevance ranking [22]. An ideal pictorial summary should present a video document in a limited 2D or flattened space while preserving the structure and sequential order. Xiang et al [23] explained some important terminologies used in the digital video research field. However, such kind of extraction method can not provide sufficient information about the video content of a video shot, especially for shots with much gradual change. For the extraction of key frames for video summarization, we can choose the scheme among discussed in the subsequent sections.

2.3.1 Visual processing
Feature extraction plays an important role in video information retrieval and summarization for key frames extraction from video database. Representation of these key frames can act as access points for video stream. Video stream involve multiple features like cinematic features, object based features, text information from closed captions, and audio features etc.

2.3.1.1 Color features
Color does not only add beauty to objects but also give more information, which is used as powerful tool in video summarization, indexing, and retrieval.
Ekin et al [24] proposed dominant color region detection, shot boundary detection and shot classification algorithms that are robust to variations in the dominant color. Also introduced new algorithms for automatic detection of goal events, referee, and penalty box in soccer videos. Color histogram, which represents the color distribution in an image, is one of the most widely used color features. It is invariant to image rotation, translation, and viewing axis. The effectiveness of the color histogram feature depends on the color coordinate used and the quantization method. Wan et al [25] studied the effect of different color quantization methods in different color spaces including RGB, YUV, HSV, and CIE L*u*v*. When it is not feasible to use the complete color histogram, one can also specify the first few dominant colors (the color values and their percentages) in an image. A problem with the color histogram is that it does not consider the spatial configuration of pixels with the same color. Therefore, images with similar histograms can have drastically different appearances. Several approaches have been proposed to circumvent this problem. Pass et al [26] proposed a histogram refinement algorithm. The algorithm is based on CCV (Color coherence vector), which partitions pixels based upon their spatial coherence. A pixel is considered coherent if it belongs to a sizable contiguous region with similar colors. A CCV is a collection of coherence pairs, which are numbers of coherent and incoherent pixels, for each quantized color. Similarly, Chen and Wong proposed an augmented image histogram [27], which includes, for each color, not only its probability, but also mean, variance, and entropy values of pair-wise distances among pixels with this color.

Peng et al [28] classify shots into different types based on playfield color. For the similarity in playfield color between medium shot and long shot, the accuracy rate of this method may be dissatisfied.

Because color histogram is robust to background noises and invariant to image orientations, most researchers proposed color-based key frame extraction methods [29, 30]. Ferman et al [30] constructed an alpha-trimmed average histogram describing the color distribution of a shot. Then compute the distance between the histogram of each frame in the shot and the alpha-trimmed average histogram. Key frame position is located based on the distribution of the distance curve. However, most of these color histogram based methods cannot well capture the underlying dynamics when there is lots of camera or object motion. Although HSI and CbCr space exploited [31,
Benjamas et al. [33] used color histogram comparison to detect shot boundaries. Flashlight detection is applied from the region histogram difference algorithm described by Benjamas et al. [34]. Detected flashlight and distance between players is utilized in efficient summarization of fighting sports videos. Proposed method composed of skin detection, enhancement image and calculation of distance between players. Sandra et al. [35] generated summaries based on color attributes and visual features. These attributes are necessary to identify the similarity among the video frames. They were extracted from color histogram adaptation. Cheng et al. [36] used shot detection process which adopts the color and edge information to make the shot boundaries more accurately. Shots are collected with similar motion type and color distribution together by adopting color and contour information and decide shot importance by employing the motion energy and color variation of shots.

A higher PV value means that this shot is more important of the cluster and the shot will be the highest of cluster. Panagiotakis et al. [37] Proposed a MINMAX key frame selection technique based on equipartition principle that is the frame cluster derived by the key frames boundaries are equal sized. The most representative frame (key frame) is selected according to the minimization of the maximum frame distortion of the sub-shot. Abd_Almageed [38] proposed algorithm based on sliding window singular value decomposition (SVD) approach. The algorithm extracts low-cost, multivariate color features from the video and constructs a 2D feature matrix. The matrix is then factorized using SVD. Sliding window SVD is then used to compute the rank of the current feature matrix. By analyzing the evaluation of the computed rank over time, shot boundaries and key frames are extracted. Gao et al. [39] analyzed the set distance that is shot distance proposed an advanced Hausdorff distance which can compare two shots from the global view. Here adopted the affinity propagation cluster method [40] to do video clip clustering. With affinity propagation cluster, the video shots are grouped into several clusters fast and without any predefined threshold, to remove redundant video content. Then the video summary is generated with important factor analysis.
Costas et al [41] presented a key frame selection algorithm based on three iso-content principles, iso-content distance, iso-content error and iso-content Distortion. In the framework of this work used the color layout Descriptor (CLD) of MPEG-7 standard to guarantee interoperability. Kamesh Namuduri [42] used features grass color; pitch color (sand color), audience texture, motion activity and number of edges to extract views or states from a cricket video. The states and their transitions in the cricket game are represented by Hidden Markov Model, based on which the game highlights are extracted. Konstantinos et al [43] propose detect salient region for classification. After taking cluster analysis for salient region, color and texture feature in each region are extracted to classify shot into different types. This method has a lower efficiency in the light of higher computational complexity.

Kenichi et al [44] used color histogram of a shot as color information and discovered important intervals having several color change patterns by using the probability model. Furthermore, extracted utterance information by using closed caption. Find the dialogue structure by analyzing the connectivity between utterances. Finally integrated them so as to skim meaningful portions in a video. Kokare et al [45] focused on visual features such as color, texture, shape, motion, spatial information and multidimensional indexing techniques.

Yihong et al [46] selected frames with a fixed interval for input video sequence and created feature frame matrix A and perform the SVD on A to obtain matrix $\gamma^T$ whose each column vector $\psi_i$ represents frame i in the refined feature space.

Visual content representation based on global field color distribution is used to create video summary.

### 2.3.1.2 Object features

Junyoung et al [47] integrated local motion model, contrast model, special scene model, and statistical Rhythm Model to form a perception curve based on linear and priority base fusion schemes. The frame that corresponds to the peak points in these individual models and curve are extracted as multilevel summarizations. Chekuri et al [48] extracted handwritten contents on the check board quantified the visual content in the video.
frames and then developed an algorithm to select the content-rich frames as key frames.

Wen-Nung et al [49] detected semantic features like zoom detection, caption-detection, explosion detection, human face detection, moving object detection and background classification, so that their relevance to user preference can be determined. A constrain optimization problem (subject to time constraint and video smoothness) is faced and solved to determine the non-uniform sampling rates for shots relevant to user preferences. Mohanta et al [50] used non-parametric hypothesis testing based on Wald-Wolfowitz runs test to identify the sub shots in the shot and then for each sub shot, the frame rendering the highest fidelity is extracted as the key frame.

Tomas et al [51] modeled the ability of ants to build live structures with their bodies in order to discover, in a distributed and unsupervised way, a tree-structured organization and summarization of the video data. The approach generates a set of representative shots and extracts the tree structure of a video sequence.

Yan et al [52] proposed a 3C-Diagnosis algorithm to diagnose the video summary from the perspective of coverage, conciseness and coherence.

Songhao et al [53] selected gray level variance histogram and wavelet texture variance histogram as the features to perform the shot detection and key frame extraction. These shots within the temporal constraint having visually similar activity are then grouped into scene. To retrieve interested video contents in terms of human habits, it is necessary to represent scene content from the semantic level e.g. a conversation scene, a suspense scene and an action scene.

Anan et al [54] proposed hierarchical framework consists of the low level part constructed the human attention model to depict the variation of human perception in multiple modalities and the high level part focuses on semantic understanding of different granularities of videos. Based on this hierarchical framework presented its applications on semantic retrieval, video summarization and content filter. Zhongna et al [55] proposed video-based eldercare monitoring. Developed a Silhouette extraction algorithm which handles shadows and dynamic background changes, followed by location and moving speed estimation algorithm using single-view camera videos. Then developed a manifold learning and dimension reduction scheme for
human action recognition to extract important activity statistics for functional assessment.

Pavan et al [56] proposed a completely unsupervised approach to learn the vocabulary from which a video segment can be decomposed into a collection of phrases for summarization. Luis et al [57] created a scalable representation of the set of summaries using an iterative ranking procedure, completely described as a ranked list of group of pictures (GOPs) for efficient analysis and bit stream generation.

Yueting et al [58] proposed a method for summarizing multi-view video contents such as semantic-face, video caption and visual features. Dirk et al [59] presented a clustering algorithm to solve the problem of frames selected out of transitions by two stage clustering process.

Zhong et al [60] have developed visual features including color clustering, object segmentation and line detection for classification of tennis and baseball shots.

2.3.1.3 Motion features

Brojeshwar et al [61] extracted motion based features by taking the statistical moments over an accumulator. Accumulator represents motion histogram over entire frame. A feature vector is generated using sliding window over time. The features are trained used a support vector machine (supervised learning) to classify the cut and normal scenes of the video. Vasileios et al [62] estimated video summary from edge motion profile which is calculated from edges.

Ramon et al [63] proposed a method based on imaginary straight line tracking to retrieve the projective transformations that describe the dominant motion of a sequence by estimating 2-D homographics.

To provide sufficient information about the video content of a video shot, especially for shots with much gradual change researchers proposed motion-based approaches, which compute motion activity-based image difference [64] or the motion energy distribution [65, 66].

2.3.2 Audio processing

Researchers have been using audio activities for video summarization of ball games, news video and in violence detection in movies. Qian et al [67] separated news and commercials based on nine acoustic features, text
An independent anchorperson was recognized using GMM and generated an abstraction of content for multimedia indexing and retrieval in the context of broadcast news.

Rui et al [68] used HMMs to model highlight sound effects (laughter, applause, and cheer) and log-likelihood scores based method to make final decision. Two new spectral features, sub band spectral flux and Harmonicity Prominence introduced to represent key effects and based on this proposal a Bayesian network based approach to discover the high-level semantics of an auditory context. Shao et al [69] created musical summary using features linear prediction coefficient, the short-time zero crossing rate and Mel-frequency cepstral coefficients (MFCCs) from audio track and shots are detected and clustered from the visual track. Finally created musical video summary by aligning the music summary and clustered video shots. Baoxin et al [70] analyzed visual and aural signals for events and non-events to form meaningful indexing points and video summarization is achieved by concatenating the event segment.

Leonardi et al [71] computed the loudness feature to find instances of goal scoring in soccer game, Wan et al [72, 73] extracted the frequency-domain audio feature to locate exciting segments in soccer and tennis videos, Huang et al [74] used six audio features, namely root mean square volume, zero crossing rate, pitch period, frequency centroid, frequency band-width and energy ratio to discriminate among news reports, commercials, weather forecasts, football videos and basketball video clips.

Dian et al [75] localized occurrences of highlights using detection of whistle sound, excitement, and text displays. Chen et al [76] extracted audio and video features and synchronized the input audio and video according to their content related features to form a musical video. Rui et al [77] proposed key audio effects with two new spectral features sub band flux and Harmonicity Prominence.

Audio content analysis plays an important role in video content parsing. Besides visual features, audio features are widely considered in many works, such as highlights extraction [78] and video summarization [79]. In order to extract the highlight shot more accurately, Rui et al [78] utilized announcer’s excited speech and baseball hit for TV baseball programs; Ma et al [79] proposed an audio attention model to measure the importance curve of an audio track. However, these works did not consider some general highlight
sound effects, such as laughter, applause and cheer. These sounds are usually semantically related with highlight events in general video, such as entertainments, spons, meeting, and home videos. The audience’s laughter often means a humor scene in TV shows and applause in meeting often imply wonderful presentations. Detection of these highlight sound effects in audio stream is very helpful for highlight extraction and video summarization. Most of previous works on audio content analysis focused on general audio segmentation and classification [80-82]. Where an audio track is segmented and then each segment is classified into one of predefined classes.

2.3.3 Textual processing

Shao et al [83] proposed method which detects and recognizes lyric captions to analyze music video structure and identify the most salient music part to create the summary of music video.

Wonjun et al [84] detected overlay text using transition map color-based thresholding method [85] to extract text strings. Jung et al [86] proposed key captions/text extraction method and provided a dual binarization method to segment texts easily with different color polarities from the background. Text in images and videos always contains useful information, which can help a machine to understand their content [87]. In the case of sports videos, text information plays a very important role because it explains valuable game information such as scores, players, and so on. We call the key texts as key captions, which can be used for video highlight or content search. Kim et al [88] developed an automatic summarization system for basketball videos. They detected caption box with score information by finding the region of high responses with vertical strokes within predefined time interval. They determined the score region among candidate regions within caption box using the characteristics of basketball game scores.

Some researchers have investigated for video summary using text information in broadcasting sports videos. Zhang et al [89, 90] developed a system for baseball videos. They detect caption box with score information using the compressed-domain features derived from DCT coefficient and motion vectors. Therefore, the system can work fast for caption box detection because of using compressed-domain features. They can achieve more accurate recognition performance using Zernike moments and domain
knowledge. However, they didn’t consider texts with the similar feature of caption box pattern with score information in the system. Therefore, there is a possibility that the system fails caption box detection on videos which have texts in advertisement broadcast. In the case of golf, the location of the key caption keeps changing during the game and key captions with a player name are located at the top-left and top-right positions, respectively [91]. Screen text processing has not received much interest in the sports video processing literature, except for logo detection from the text on the advertisement boards for revenue control purposes [92]. The availability of closed-captions and on-screen text is dependent on the geographical location and broadcaster, respectively.

Lazarescu et al [93] employed a natural language processing approach. The type of the action, the name and the type of the players involved, and the game statistics are extracted from closed-captions of football games. Babaguchi et al [94], on the other hand, propose an interesting combination of information retrieval and context-modeling for closed-caption processing of football games. Babaguchi et al [94] have been specified 90% recall, 55% precision rates for event detection by closed-caption analysis. The applicability of closed-caption processing is dependent on the geographical location. Zhang et al [95] extracted the text from the score-boxes of known broadcasters which are used to detect the occurrence of score and strike-out events as well as the time of the game.

2.3.4 Audio-visual and audio-video text features

Mark et al [96] outlined how BNE (Broadcast News Editor) and BNN (Broadcast News Navigator) systems analyze, select, condense and present news summaries.

Sudhir et al [97] extracted events from tennis using video analysis and the geometry of the court; Chang et al [98] have used HMMs for summarizing Baseball footage. Work is emerging that considers the audio signal for spotting important events [99,100]. In general for sports the audio signal is capable of characterizing much shorter duration events than the video signal. In sports like tennis, cricket, badminton it is the short and sharp noise of the ball hitting the racket or bats that defines the basic building block action of the game. Both the audio and video signals therefore contain
useful information and this work considers the use of both audio and video features for parsing tennis footage.

Leonardi et al [101] introduced an audio and visual model exploiting an hidden Markov models (HMM) to classify each pair of successive shots. Hanjalic [102] has proposed a deterministic excitement criterion based on the mean dominant motion magnitude per shot, the density of cuts, and the audio loudness to detect goals in soccer video.

Although many algorithms for scene segmentation using audio and video features have been presented [103, 104] make use of audio content to segment a sequence into scenes in order to keep the computational complexity reasonably low. For each scene, shots are segmented using the difference of luminance histogram between the adjacent frames. Subsequently, all the shots in one scene are clustered into different groups with different visual contents according to the similarity between the average histograms of each shot. After that, the semantic importance of segmented scenes, shots, and frames is modeled based on certain audiovisual features, such as motion, audio classification, loudness, shot change frequency, and attention region, which were selected based on the principles of human perception and cinematic techniques. It is proposed that the video summarization is generated according to the semantic importance of a scene, shot, or frame. Thus, the proposed approach is able to extract the most representative excerpts whilst excluding unimportant video segments from the summarization.

Xian et al [105] extracted certain editing rules from an input raw home video based on the content of the video and music. Also extracted a temporal structure, beats, and tempos from the incidental music. The final output video is rendered by connecting the selected video segments with specific transition effects and aligning them with the incidental music.

Katsutoshi et al [106] described indexing system for multimedia broadcast news content by integrating audio, speech and visual information. Alan Hanjalic [107] explored extraction of the affective content information from audiovisual signals and interesting research challenges.

Evangelopulos et al [108] integrated the various modality curves (i.e. audio, video and text saliency is extracted) in a single attention curve where presence of an event may be signified in one or multiple domains. Junyong et al [109] presented a semantic audiovisual approach for video
summarization. Based on selected audio characteristics the sequence is segmented into scenes which are further divided into shot groups with unrelated visual contents using luminance histograms. Selected audio and video features are used for the semantic importance.

2.3.5 Other features processing

Following other features are also used to extract key frames in video summarization.

2.3.5.1 Spatial-temporal features

Nuno et al [110] proposed the temporally localized motion model and the spatio-temporal model with a second order temporal constraint to the summarization of various sequences. The spatio-temporal model looks on to the body motion, leading to a map that summarizes the scene content in a much meaningful way. 

Ying et al [111] presented techniques for feature film skimming using audiovisual tempo analysis and specific cinematic rules. 

Zhonghua et al [112] proposed a Spatial-temporal color distribution based key frame extraction method. A temporally maximum occurrence frame used as reference frame is constructed [113] based on color distribution. Then compute the weighted distance between the color histogram of each frame in the shot and color histogram of the reference frame. The key frames are extracted at the peaks of the distance curve and can achieve high compression ratio and high fidelity.

2.3.5.2 Video redundancy

Ruo-gui et al [114] extracted video non-retrival repeating pattern and a video shot importance evaluation model (IEM) is used for constructing video summary. Victor et al [115] proposed generic on-line video summarization and retrieval system based on visual signatures, motion patterns or other characteristics. Emilie et al [116] initially removed uninteresting sequences and accelerated video according to motion activity. Agglomerative hierarchical event clustering used to cluster video segments of one-second and summaries produced with a strong positive correlation with the TRECVID campaign evaluation.
Jinchang et al [117] modeled hierarchically rush videos, employed adaptive clustering to determine retakes for redundancy removal, and each most representative shot selected from every cluster is ranked according to its length and sum of activity level for summarization.

More recently, Jiang et al [118] generated key frames using visual attention index (VAI) descriptor. Yi-xiong et al [119] used visual content and textual information for medical video summarization.

2.4 Video Representation

Video representation is the mapping from the 3-D space to the 2-D view screen. Different video representation techniques [23] for scripted content are discussed. Figure 2.2 shows a hierarchical video representation for scripted content. In sequential key frame representation the key frames of the video lay out sequentially, from top to bottom and from left to right. This technique works well when there are few key frames. The group-based representation is a tree structured video representation. In scene base representation the video sequence is first segmented into shots. Shots are then clustered by using time-constrained clustering i.e. based on the time flow of the clusters. These three representation techniques are suitable for video browsing and the video mosaic representation is especially suitable for video retrieval.

Figure 2.2 Hierarchical video representations for scripted content.
2.5 Video summarization

The methods of video summarization can be classified into two categories namely static video storyboard summary [120, 121], which involves a set of key-frames extracted from the original video, and dynamic video skimming [122-125], which collects a set of shots by computing the similarity or relationship of each shot. The purposed method clusters shots by considering scenes and texture information. Color and edge information are used to make shot boundary [126] more accurate. Key-frame extraction adopts motion attention model [122] from each shot to select the most salient frame. Still image abstract is the field that has attracted the most attention, since it is the simplest and most intuitive. The video abstraction systems can be categorized into shot-based [127], content-based [128], motion based [129] and cluster-based [130] key frame extraction systems.

2.6 Concluding remarks

Due to the unlimited production of digital video, video summarization has become an indispensable tool for video content management. Although there are many approaches developed by various researchers for video summarization, the results are not very promising. Several problems in video summarization systems remained untouched. It is required to develop a shot importance measure scheme for video summarization based on audio-visual and text features. Various videos like Medical videos, News videos, dramas, movies, sports and meeting videos need efficient and effective tools for video summarization, indexing, managing, cataloging, compressing and searching for online videos.

***