J. Vis. Commun. Image R. 62 (2019) 340–358

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Towards genre-specific frameworks for video summarisation: A survey ☆

Sreeja M.U., Binsu C. Kovoor *

Division of Information Technology, Cochin University of Science and Technology, Kochi, Kerala, India

ARTICLE INFO

Article history:
Received 2 August 2018
Revised 19 January 2019
Accepted 8 June 2019
Available online 11 June 2019

Keywords:
Video summarisation
Video summary
Genre-specific
Skim
Keyframe

ABSTRACT

Video summarisation is characterised as the process of extracting meaningful frames or segments from a video that best represents the content of the whole video. The proposed framework surveys and categorizes the existing video summarisation models in the recent research works on the basis of genre. The most important phase of video summarisation is the detection of key frames or segments in the video. The strategy for identifying key frames or segments vary for each genre. The various genre analysed are user generated videos, movies and documentary, sports, surveillance, egocentric and informational talk videos with a total of more than 25 varying parameters significant to the respective genre. Comprehensive evaluations of the results obtained from the models are also included based on quantitative and qualitative parameters. The framework will help the user in deciding the technology to be adopted for video summarisation in a particular domain. The framework also aids in deciding the type of summary suitable for each genre and the available datasets in each genre for experimental analysis.

1. Introduction

A vast majority of today's internet data are videos from various domains. In fact, with the invention of smart phones, wearable cameras, and other recording devices, anybody can record or create a video and upload. With the availability of high speed internet and cloud technology, millions of videos are getting uploaded day by day from a large variety of domains. Major collection of videos include entertainment (sports, movies), educational (lecture videos, informational talks), news (surveillance videos) and user videos (egocentric videos, consumer videos). But watching the whole video to extract useful information is time consuming and a tedious task. Video summarisation technology aids here by automatically providing a representative storyboard views of videos, making the process of information extraction easier and efficient. These summaries can also be used for annotation purposes which further eases the search and retrieval process. Instead of watching hours of video, video summaries provide a gist of the video by eliminating redundant frames and extracting keyframes or segments from the video. A generic framework for video summarisation suitable for all genres of videos is difficult to implement. This is because for each domain, keyframe selection depends on discrete factors. For example, in sports videos, the selection of key events is important in selecting the keyframes. This might require detection of court and logos which are generally done by dominant colour analysis. On the other hand, for movies, analysing frames for the presence of important characters plays a prime role in identifying keyframes. Hence, the need for discrete or genre specific video summarisation frameworks is of high significance. Video summarisation frameworks can be surveyed on several bases; summaries produced (static, dynamic, text, image), on the basis of keyframe identification technique or on the basis of source of information used for summarisation. Key frames or segment detection is the most important phase in video summarisation. Keyframe extraction method can be categorized on the basis of parameter identified for analysis (motion, object, event, colour, etc.), by evaluating if a keyframe draws visual attention or not (visual attention based models) or on the basis of factors analysed (internal, external or hybrid). But a detailed analysis of video summarisation frameworks in a genre specific categorization is not available. In the proposed survey, an overview of different frameworks for video summarisation in a genre-specific manner is detailed. The various domains considered are user generated videos that also spans general category including online and 3D videos, sports, movies and documentaries, surveillance videos, egocentric and informational talks. A conceptual framework of the proposed system is shown in Fig. 1.

The main contributions of the paper include:

1. A novel framework that not only identifies the need for genre specific models for video summarisation but also helps in determining the appropriate model for a domain.

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: binsu.kovoor@gmail.com (B.C. Kovoor).

https://doi.org/10.1016/j.jvcir.2019.06.004
1047-3203/

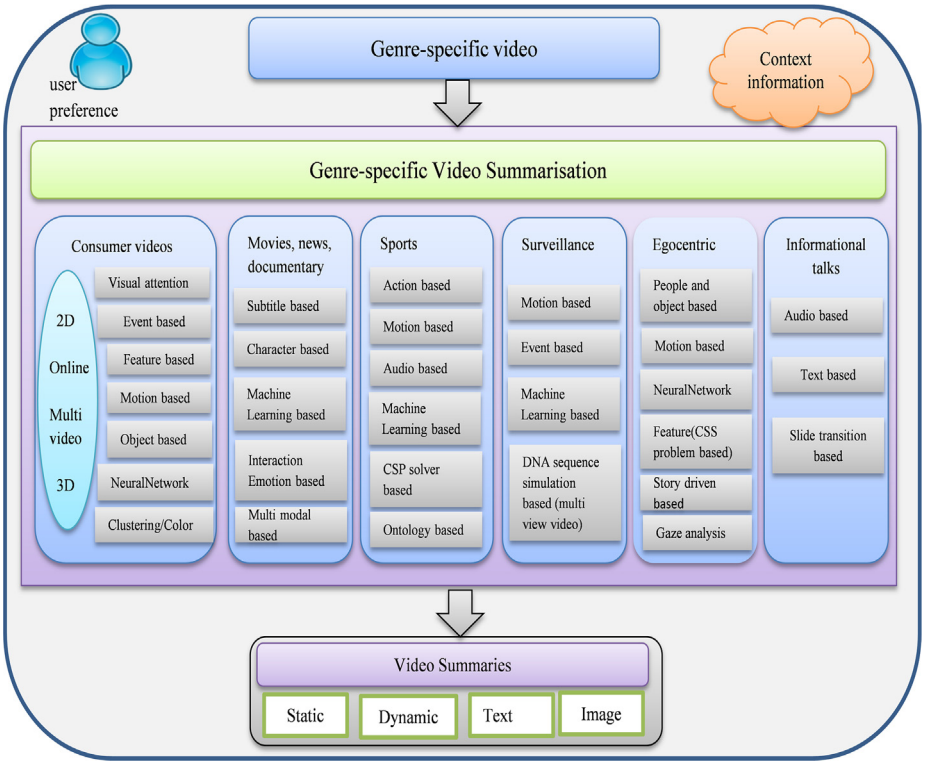


Fig. 1. Proposed framework.

2. A comprehensive survey that examines the significance of various parameters used for video summarisation in various domains.
3. A concise description on the available datasets used in the literature for experimental analysis.
4. A detailed analysis of the performance of various models based on quantitative and qualitative parameters.
5. The proposed framework also aids in identifying the type of summary to be produced that is appropriate for a particular domain.

The rest of the paper is organized as follows. Section 2 details the related surveys conducted in the video summarisation domain. Section 3 details the motivation behind surveying in a genre specific manner. Section 4 discusses the types of summaries produced followed by the genre specific video summarisation frameworks in Section 5. The datasets available for experimental analysis from the literature is detailed in Section 6 followed by evaluation of results obtained in Section 7, challenges and future recommendations in Section 8 and conclusion in Section 9.

2. Related surveys

Several surveys have been conducted on *general frameworks* for video summarisation. Money et al. [1] performs an extensive survey on the generic approaches for video summarisation. The survey is classified as (a) internal, external and hybrid on the basis of source of information used for summarisation, (b) on the basis of technology used as event based, object based, perception based and feature based, (c) on the basis of summary produced as interactive or personal. Kaur et al. [2] surveys the summarisation frameworks on the basis of general approaches used as feature based (high level, low level, single, multiple), technique based (shot boundary based, clustering and non-clustering based) and other recent methods. Ajmal et al. [3] performs a comprehensive survey on the basis of general techniques for summarisation such as feature based, cluster based, event based, shot selection based, trajectory based and mosaic based. The existing approaches for each genre of videos are also summarised in this framework. Moses et al. [4] surveys semantic based video summarisation frameworks. Various summarisation techniques analyzed are classified as object based, event based, egocentric, hierarchical, Index based, clustering based, content based semantic based and sparse dictionary based methods. Truong et al. [5] surveys the general video abstraction techniques. The work is initially classified on the basis of the type of output produced into key frames and video skims. The framework focuses on the various approaches to identify the key frames or important skims respectively. The key frame identification strategies are classified on the basis of (a) summary size as apriori (number of key frames known earlier), aposteriori (number of keyframes not known) and determined (number of keyframes identified before extraction), (b) units, (c) scope of key frame, (d) underlying technique used (content based, clustering, frame coverage, temporal variance, event based). Similarly, dynamic video summarisation frameworks are classified on the basis of size (apriori, aposteriori), target domain, skim generation process, preserved perspective coverage and the various features used. The datasets available for experiments and evaluation metrics are also analyzed in detail.

The major research issues and challenges in *sports videos* are examined by Wang et al. [6]. Identification of tactics for each sport is the most import challenge in sports video summarisation. Other research issues such as content insertion, feature extraction and tracking, highlight detection and landmark detection are also addressed. Tank et al. [7] surveys the approaches for sports video summarisation on the basis of outputs produced (key frames, skims) and the various techniques used (clustering, shot boundary detection, Delaunay triangulation).

Betancourt et al. [8] describes on how the *egocentric video* analysis evolved and the history behind it. The various existing methods on capturing the first person videos and upcoming technologies for analyzing the egocentric videos have been surveyed in detail. Bolanos et al. [9] reviews the methods for acquiring and summarising egocentric videos as a story telling approach. Available datasets for egocentric videos are analyzed in detail. The various parameters considered (motion, context, features) and the challenges involved in egocentric video summarisation have been detailed. The various approaches have been classified based on the various stages of egocentric video summarisation as video acquisition, segmentation, image detection, content based search and retrieval, object recognition, body movement based methods. Del Molino et al. [10] surveys the existing literature on egocentric videos on the basis of type of output preferred as skims, key frames or fast forward. Various approaches for key frame selection and segment selection are also detailed along with the quality of available datasets in detail.

Kalaivani et al. [11] reviews the existing *surveillance video* summarisation techniques on the basis of features used, datasets and approaches. The surveillance videos are classified and analyzed as crime prevention, traffic monitoring, home care or hospital assisting. The framework provides a clear concept on what are unusual events in surveillance videos.

The proposed framework is a novel approach for surveying the existing video summarisation models. The survey is conducted in a genre specific approach which was not seen in literature so far. The motivation behind conducting the survey is detailed in the following section.

3. Motivation

Genre specific videos are the videos that fall under a specific genre or category. The genres identified for review in the proposed framework are user generated videos, sports, movies, surveillance and news videos, egocentric videos and informational talk videos. The need for genre specific video summarisation framework stems from the fact that the key events in a video differ in each genre and hence, the approach to identify the keyframes that can well represent the key events, differ from genre to genre. A sample of possible key frame from different genre is shown in Fig. 2. The following facts demonstrate how the highlights differ in a genre specific manner.

- The keyframe or segment identification in *user generated videos* depends on many factors. The frames that the user intentionally covers will have high resolution and focus and hence identifying such frames may help in identifying key frames. These frames can be termed as the ones that have high visual attention score. Another strategy would be to identify frames having more motion since users might be interested in the frames that have more motion than those that are static. Features also play an important role in user generated videos. Summarisation frameworks that are based on important features in the video are also prevalent. Fig. 2a shows a key frame depicting an important scene from the video 'Jumps.mp4'. It shows the most important frame, a man jumping in a long shot.
- In *movies*, the category to which the movie belongs plays an important role. Identifying the key characters and relations may help in identifying the keyframes for summary. Identifying emotions is also a prime concern is movie summarisation frameworks. Movies fall under entertainment videos. Hence, unlike other genre, an additional goal of summarisation of entertainment videos is to ensure that the output summary exhibits enjoyability factor. Fig. 2b shows a keyframe of child's face in close up which is possibly an important scene from a movie.
- For *sports videos*, each kind of sport has different key events. Hence identifying the key events is a challenging task. For soccer, the key events may be goal, foul, free kick and so on. But for cricket it is the frames where the players acquire runs. Similar is the case with other sports like martial arts videos, wrestling, swimming, etc. Sports videos are also considered as entertainment ones. Hence, the importance of incorporating enjoyability factor is evident. Fig. 2c shows a keyframe representing a goal event from Soccer video.
- *Surveillance videos* are videos where camera remains static and videos are recorded throughout the day. The detection of important events is identical to detecting the unusual events based on motion parameters or optical flow analysis. Here, background details present in the video are less significant since the camera is mostly static. Fig. 2d depicts a keyframe from the surveillance footage of an ATM counter. A man engaged in cash withdrawal can be clearly seen from the frame.
- *News and documentary videos* generally consists of a narrator and a collection of important events. Summarising news and documentaries should ensure that important events are captured as with surveillance videos. The informative content from the summaries should be high compared to other factors. Fig. 2e shows two sample keyframes from news videos where a news reader is present and an accident scene is shown.

(a) User generated video, a man jumping

(b) an event from movie

(c) goal event from Soccer video from Sports genre

(d) event from ATM surveillance video footage

(e) highlights from news showing reader and accident victim

(f) egocentric video frames showing important places

(g) Frame from Information talk showing speaker with presentation

Fig. 2. Sample keyframes depicting major events from various genres.

- Summarisation of *egocentric videos* captured by first person using wearable cameras face many challenges the most important one being unallignment of the video. This is owing to the fact that the video is shot using a wearable camera. Here, the key segments can be identified either by analysing the wearer interactions or by the places visited by the wearer. Fig. 2f shows two sample keyframes from an egocentric video showing places of user interest.
- Summarisation of *informational talk videos* should aim at providing maximum informative content from the video. The information can either come from the speaker or from the power- point presentation accompanying the speech. Hence, identifying the important parts of the speech from the speaker gestures, from the slide transitions in power-point presentation, audio of the speaker or from the subtitles is an effective way of summarizing the video contents. Fig. 2g shows an important frame from an informational talk video where the speaker is present alongside the presentation.

From the above factors, the importance of surveying the existing literature in a genre specific manner is certainly obvious. Table 1 shows the comparison of various genres of videos on the basis of the differences in the properties that are utilized for summary creation. For instance, in surveillance genre, it can be depicted that the camera is stationary, the importance of shot boundaries is therefore nil, redundant frames are high, frames with focus and alignment present in the video is high, interaction factor is important for identifying important events, emotion is not important, event detection is of importance and the informativeness requirement is also high.

4. Video summaries

Video summaries are the output generated from the video summarisation frameworks which are the succinct representation of the original video. The types of summaries generally observed from the literature are static, dynamic, images and text. Example output summaries of the different types detailed surveyed are shown in Fig. 3. The characteristics of each summary type differ from each other as detailed below.

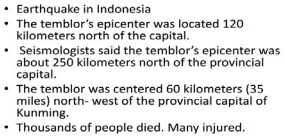
1. *Static* summaries are generated by combining a set of frames that are identified as key frames from the input video by the summarisation framework. Static summaries resemble stationary images. The extraction of key frames is performed on the basis of different summarisation approaches which are detailed in the subsequent sections. Static summaries are also referred to as storyboard layout. Static summaries lack continuity and audio cues and are generally less informative, but are considered efficient in terms of memory and computational time. Fig. 3a shows an example of static summary from the video 'Jumps.mp4' showing the important events as a static summary.
2. *Dynamic* summaries, also known as video skims, are summaries produced by combining segments or shots of videos together which can be played back. Apart from the fact that dynamic summaries offer continuity as a feature and additionally, audio elements can also be incorporated which makes dynamic summaries more preferable in certain genres like movies. Dynamic summaries can provide more information but are computationally complex in terms of memory and time. Additionally, dynamic summaries can be converted to static summaries anytime since it is possible to select a keyframe from the individual segments chosen for dynamic summary but the reverse is not possible. Fig. 3c shows the dynamic summary of the same video 'Jumps.mp4'. It can be seen that it is a combination of several segments or skims.
3. *Image* summaries: In certain frameworks, important objects or characters from different frames are combined to produce an image summary. Rather than frames or shots, images are presented as summary. It can be a single image or combination of image. Another variation for image summary is panoramic image where the flow of events can be represented in a single image and summary is represented as a combination of panoramic images. Panoramic summaries resemble dynamic summaries since they can represent the motion element in a video to a certain extent. Fig. 3 shows an example of image summaries where Fig. 3d is a combination of three single frames and Fig. 3e shows a panoramic image summary combining the first three frames. It is obvious from the figure that the motion element can be incorporated in the panoramic image summary [12].
4. *Text* summaries are produced by summarisation frameworks by employing Natural language Processing (NLP) techniques. Text summaries can either be produced by processing the sampled frames in the video for identifying the objects in the videos and later employing NLP techniques. Another option is by processing subtitles in a video if available and by summarising subtitles using NLP techniques. The drawback of text summaries is that it can't incorporate audio and motion elements and as a result the summaries produced can't provide a lot of information about the events in a video as compared to visual summaries. Though text summaries can involve lower computational costs and can incorporate continuity factor and storage efficiency, they are rarely used in video summarisation. Fig. 3b shows an example of a text summary from a news video.

Table 2 shows the properties of the various types of output summaries surveyed. For instance, it can be interpreted from the table that dynamic summaries have comparatively high information content and they can incorporate audio and motion elements. But the computational cost is high and storage efficiency is low. Similarly, storage efficiency and computational costs stand as an advantage for static summaries but they can't embed motion and audio elements which lowers the amount of information provided.

Table 1
Properties of videos in each genre used for summarisation.

Genre	Stationary camera	Mobile camera	Focus	Shot boundaries	Redundancy	Alignment	Interaction	Emotion	Events	Informative
User generated videos		✓		✓	✓					
Movies		✓	✓	✓		✓	✓	✓	✓	
News and documentary	✓	✓	✓	✓	✓	✓		✓	✓	✓
Sports		✓	✓			✓	✓		✓	
Surveillance	✓				✓	✓	✓		✓	✓
Egocentric		✓		✓	✓		✓	✓	✓	
Informational talks	✓	✓	✓			✓				✓

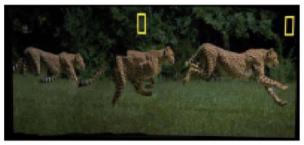
(a) Static summary



(b) Text summary

(c) Dynamic summary

(d) Single shot images



(e) Panoramic summary[12]

Fig. 3. Various types of output summaries.

Table 2
Characteristics of the various types of video summaries (H-High, L-Low).

Video Summaries	Properties						
	Output	Motion	Audio cues	Information content	Storage efficiency	Computatio- nal cost	Continuity
Static	Stationary frames	×	×	L	H	L	L
Dynamic	Video skims or segments	✓	✓	H	L	H	H
Images	Stationary image	×	×	L	H	L	L
Text	Text	×	×	H	H	L	H

5. Genre specific video summarisation framework

A video can be classified into any genre depending on the characteristics of the video. The video summarisation framework to be applied on a particular video can be decided by the genre it belongs to. The existing recent video summarisation frameworks analysed for this survey are classified in a genre specific manner along with the techniques associated with each genre.

5.1. User generated videos

User generated videos, also known as consumer videos are those that are created by users using normal video recording devices. Hence, the video lacks quality and resolution. In such cases, it is not easy to distinguish important shots in user videos. These videos are grouped into general category as these videos can be from any domain. Several models have been implemented for video summarisation in this domain which is classified mainly on the basis of visual attention, clustering based, neural network based, feature based, etc.

Visual attention or user attention models are the ones that aim at identifying the frames where viewers have shown more attention. Such frames can certainly be identified as key frames for the resultant summary. Visual attention scores are computed either based on features, time and motion or semantic tree encoding or a combination of these. For instance, while Fei et al. [13] and Gygli et al. [14] use features from the video, Kavitha et al. [15] and Thomas et al. [12] use motion characteristic for calculating the visual attention scores. Fei et al. [12] designs a shot segmentation approach based on Perceptual hashing using mutual information. It is proven to be more efficient than colour based shot segmentation. For extracting keyframes for summaries, an entropy value (density of an image) is combined with an image memorability score which is calculated from deep network (dataset of images with memorability scores) as proposed by Khosla et al. [16]. A pre-trained hybrid CNN is used for initializing the training (Zhou et al. [17]. Gygli et al. [14] performs summarisation based on super-frame segmentation. The framework, which produces dynamic summaries, divides the video into super-frames. Super-frames are sets of consecutive frames where the first and last frame of the set has the properties of shot boundaries. Segmentation is performed with colour histogram approach and human attention score for individual frames are considered. For summary generation, the superframe with the highest score which is the sum of scores of individual frames is selected. The model by Kavitha et al. [15] is based on the fact that, at times user might be interested in static objects as well as moving objects too. Hence, identifying both categories of frames will definitely improve the quality of the summary. User attention values are calculated from frames with static objects and Discrete Wavelet Transform is used to identify frames with more movements. Thomas et al. [12] requires image registration to from the panoramic image summary. Since, frames having more motion are of interest to the user, motion values are calculated using HSV colour model. Three factors used for motion estimation are motion contrast, motion energy and motion chromism. According to the estimated motion values, images are registered to the panoramic image using a transformation matrix which is calculated by setting the central frame as reference image rather than first frame.

On the other hand, Ejaz et al. [18], Srinivas et al. [19], and Fei et al. [20] calculate the visual importance score as a combination of features and motion. The four factors used for saliency computation by Ejaz et al. [18] are colour components (RG and BY values) corresponding to colour channels, motion value and intensity value. Then, an aggregated saliency value is computed on which Fourier transform is applied on the whole reducing computation cost significantly. Srinivas et al. [19] considers quality (brightness, colour, contrast), user attention (focus on individual frames), temporal coherence (which is time based and analyses maximum movements) and uniformity as factors for keyframe extraction. Keyframes are identified based on the computed score in the second stage and duplicate frames are eliminated in the third stage. Fei et al. [20] modifies the earlier version introducing two new phases to the framework; snap point detection and optical flow analysis. Snap points are the frames that the user has recorded intentionally which means that the respective frames have high memorability score. Since snap points are recorded intentionally, such frames have high resolution and better focus on objects. Optical flow analysis is performed in order to analyze the intensity of motion in frames because more user interest lies in the frames with more motion.

The model implemented by Yin et al. [21] considers user preferences which are encoded as a semantic tree apart from visual attention score calculation and segment importance computation. Semantic tree construction is performed based on pairwise similarity between images which is formerly calculated considering visual features, textual information and available user image connections from social networks. A normalized graph cut clustering is applied to generate semantic tree and image similarity is computed using Gaussian function. For summary generation, in addition to detection of shot boundaries using clustering, segment importance score is computed as the sum of personalised saliency (comparing video segments with user profile) and spatiotemporal saliency (visual attention score).

Clustering or colour based approaches: The models that implement summarisation framework using either colour or clustering or a combination of colour and clustering approaches fall under this category. For instance, while Jadhav et al. [22] and Sheena et al. [23] uses colour moments, Jeong et al. [24] and Mundur et al. [25] uses clustering for implementing the models. The model by De Avila et al. [26] uses a combination of both colour and clustering for implementing the summarisation framework. Jadhav et al. [22] bases selection of keyframes on the distribution of colours in the image computed by higher order colour moments. Frame with highest mean and variance is selected as keyframe. Image histograms, skewness and kurtosis are the colour moments considered for shot boundary detection. Sheena et al. [23] identifies keyframes by computing the absolute differences among histograms. Wherever the absolute difference crosses a threshold, shot boundary is detected and the frame in the shot with highest mean and variance is selected as key frame from the shot. Jeong et al. [24] uses spectral clustering with simple colour histograms for skimming the video. In the second step, sparse coding of Scale Invariant Feature Transform (SIFT) features is done. Keyframes are identified based on the sparse code. The clustering technique used by Mundur et al. [25] is Delaunay triangulation. The keyframes in the summary are not displayed in temporal order since keyframes are selected from clusters staring from the largest cluster. The difference with k means clustering is that in Delaunay clustering, it is not compulsory that the clusters formed will be of equal size. This ensures better quality summary. But, computational overhead is lesser in k-means clustering technique. The system by De Avila et al. [26] is efficient where only colour attributes are computed and similar frames are grouped together. Later, k-means clustering algorithm is performed and the most representative frame from each cluster is identified.

Feature based: The models where video segments or key frames for summary are identified based on features extracted from video frames are feature based video summarisation models. These features can be represented using DCT, vectors or matrices. For example, while Xiang-wei et al. [27] uses Discrete Cosine Transform for feature representation, Cong et al. [28] and Vivekraj et al. [29,30] represents features as vectors. Meng et al. [31] uses matrices for representing features of multiple views of a same shot. Xiang-wei et al. [27] implements video summarisation in compressed domain where keyframe extraction procedure is carried out by calculating the Discrete Cosine Transform coefficients and by applying Rough set theory to eliminate redundant frames. Video summarisation is viewed as a sparse dictionary problem by Cong et al. [28]. Each video frame is represented as a feature vector and a subset of keyframes is selected that best reconstructs the original video. Number of keyframes is based on user preference which provides scalability to the approach. In the framework proposed by Vivekraj et al. [29], a collective representation of features from segments in the form of vectors is implemented. And, relevance of the segments is found out by vector reduced ordering. Vector Reduced ordering or R-ordering is a distance metric based technique used to sort vectors. It selects a central vector from a set of vectors. Vector grouping is done on the basis of number of components in the feature vector and similar to clustering it is repeatedly performed until no new groups can be formed. Vivekraj et al. [30] have modified the above model to a multimodal framework. Here, r-ordering is performed on both audio and visual channels separately. A fused rank is generated either by low level method of clubbing together audio and visual features or by applying a round robin based fusion scheme on the separate audio and visual r-ordering based segment importance ranks audio and visual channels. The features analysed for audio and video are different; motion, intensity and average saturation values for the former and teager energy, amplitude and frequency for the latter. In the model proposed by Meng et al. [31], the goal is to find representative group of shots from multiple views. The framework differs from earlier works in the fact that multiple views are taken into consideration. The different views of features are concatenated as columns of matrix and the objective is a weighted multiview selection formula. A diversity regularizer parameter ensures that the resultant summary frames are as distinct as possible. The selected subshot from multiple views should agree with all views. Regularization is performed to ensure that the selected subsets are similar with respect to a selection. *Key objects based:* A novel approach where summary is presented as a set of key objects rather than frames is implemented by Meng et al. [32]. Here, video summarisation is viewed as a problem of identifying key object proposals rather than key frames. A set of objects are identified from the video frames and the candidate objects that can best represent the whole video are selected as the summary. Object proposals are extracted from video sequence and are represented as a feature vector. The framework involves sparse dictionary selection but in a modified way where weights and affinity are considered as factors.

Neural Network based: Summarisation frameworks where key segments or key frames are identified based on machine learning or deep learning models are neural network based models. Here, the models implemented can be supervised, unsupervised or weakly supervised. Results show that unsupervised approach fails to bring out the important segments whereas, supervised approach requires presence of enormous amount of videos with their respective summaries for training. The advantage of a weakly supervised system is that it requires only video level annotations for summarisation. Al Nahian et al. [33] proposed a framework that predicts the

importance of a shot with respect to the contents of the video using convolutional neural network. The predicted importance score is used to decide if a shot needs to be included in the summary or not. For training and optimization method by Glorot et al. [34] is utilized. Offline processing is performed for training weights and bias where bias is initially set to zeros and weights are assigned values from uniform distribution depending on the size of previous layer. Optimization used is Adam Stochastic optimization technique. A weakly supervised approach for summarisation is implemented by Panda et al. [35]. Datasets containing videos with annotations or tags are readily available nowadays. The approach makes use of three dimensional convolutional neural networks for training. Segment importance score is calculated depending on the category of the video. Importance map is calculated for segments which depend on video segments, predicted class labels, and a trained network. In the work done by Zhang K. et al. [36], keyframes and key shots are identified based on importance prediction using recurrent neural network approach. The type of neural network used is Long short term memory (LSTM) network. LSTM is a recurrent neural network which is built as a combination of cell, input gate, output gate and a forget gate where the cell block is responsible for remembering values over a certain period of time. Context memory is highly significant in video summarisation because identifying key frames or shots as significant purely depends on the context in which the frame appears. Two LSTM layers are used which helps in incorporating long term dependencies. The model is improved by adding determinantal point process (DPP) phase which aids in including a diversity parameter to the summary. Edited videos are those that are well processed and may not need a pre-processing stage and redundant frames will be lower. Raw videos are those that are not edited and are created by handheld devices by users. They require pre-processing and number of redundant frames will be high. A general framework for summarising both edited and raw videos is proposed by Li et al. [37] using neural networks. Four factors considered in the framework are importance, representativeness, diversity and storyness. Importance stands for identifying the key people and objects. Representativeness is considered for measuring how well the summary represents the important contents. Diversity measures the amount of distinctness in the frames selected. Storyness denotes how continuous or smooth the summary is. A score function is designed as a weighted combination of the above four factors. To generalise the framework the factor weights, also known as property weights are made to differ for edited and raw videos. Training set consists of both raw and edited videos.

Multi video based: Multi videos are the videos that cover the same scene or shot but are taken by different users. Each user captures similar scenes in a different way. To combine multiple videos into a single summary, video semantics has to be made use of. The works surveyed in this framework for multi video summarisation are neural network based. In the proposed framework by Nie et al. [38], this is achieved by manually assigning semantic tags. The methodology used is a weakly supervised algorithm. Training and testing videos should include video tags. Graph based segmentation is performed for each video where each frame is segmented into multiple regions. Each region is described as a feature vector which is a combination of histogram of gradient and colour values. The probabilistic output of an SVM classifier is used to compute the response of a region to a tag. An unsupervised framework for summarising a collection of videos is proposed by Panda et al. [39]. Videos in a collection are related to each other but are complementary or diverse in the information provided. The three major steps are video representation, diversity aware sparse representative selection and summary generation. Video representation includes shot boundary detection using colour changes and feature representation by using convolutional 3d neural networks. A diversity regularization function is also formulated that depends on pairs of segments from video pairs. For summary generation, representative segments are sorted by decreasing importance and according to user defined length summary is generated.

Online videos Single videos: Online videos are videos that are streamed directly from the Internet. Summarisation of online videos should ideally be computationally cost effective taking less time with lesser memory requirement. The frameworks for online video summarisation in the current literature are based on objects (Barhoumi et al. [40]), dictionary (Zhao et al. [41]), features (Almeida et al. [42]) or visual co-occurrence using neural networks (Chu et al. [43]). Barhoumi et al. [40] develops an object and event based approach where an incoming frame is divided into important objects and is compared with the previous key frames to identify if an important event has taken place. Here, first frame from each shot is considered as key frame. When further frames are scanned, evaluation for representativeness is done and selection of keyframes is changed accordingly for each shot. It is an on-the-fly process and memory requirement is less. The model implemented by Zhao et al. [41] known as LiveLight creates a sparse dictionary after scanning through first few segments. When a new video segment arrives, the system checks the sparse dictionary to check if the segment provides any new information and is added or deleted accordingly. The system utilizes histogram of Gradients and k-means clustering for dictionary creation. Almeida et al. [42] performs video summarisation on compressed domain by utilizing visual features extracted from the video stream to summarise the video contents efficiently. Compressed domain is created by converting the discrete cosine values of images to feature vectors by computing colour histograms. Key frames are selected as the frames with the features selected in the above step by computing the distance function using Zero mean normalized cross correlation metric. Video summarisation based on visual co-occurrence is proposed by Chu et al. [43]. The proposed approach does not require domain specific knowledge, which was a prerequisite for most of the frameworks discussed above. The system is an unsupervised model that aims to identify shots that co-occur frequently across videos of same topic. The framework also employs graph model for efficient keyframe selection.

Multi videos: The frameworks discussed above are suitable only for single video summarisation. These models cannot be utilized for multiple video summarisations. Hence, a multivideo summarisation framework for web videos using hyper graph dominant set is developed by Ji et al. [44]. Compact static or dynamic summaries are generated in this framework according to user preference. In this framework, large number of web videos are queried using the same search query and the results are summarised to a single video summary. The videos summarised are often of shorter length. The proposed system makes use of Hyper graph based Dominant Set(HDS), Query dependent Maximum Marginal Relevance(QD-MMR for query dependent frame extraction) and Graph based Topical closeness(for understanding the generated summaries easily). GTC is employed because the generated keyframes of summary are not temporally ordered.

3D videos Unlike 2 dimensional (2D) videos, 3Dimensional (3D) videos involve illusion of depth perception. Hence separate video summarisation techniques are necessary. 3D video summarisation techniques deploy shape and depth information rather than colour information in 2D videos. Huang et al. [45] uses 3D shape similarity whereas Valognes et al. [46] uses depth information for 3D video summarisation. Huang et al. [45] calculates 3D shape similarity on the basis of feature, graph or event. Volume sampling shape histogram is used for measuring 3D shape similarity. A self-similarity matrix based on volume sampling spherical shape histogram is applied to all frames of the 3D video sequence to extract keyframes. This method does not require shot boundary

detection. On the basis of selected keyframes graph is constructed and the shortest path gives the refined keyframes. Valognes et al. [46] uses RGB colour values and depth information in combination with image distance measures for summarisation. For incorporating depth information, the model relies on stereo videos. The main stages in the model include histogram comparisons for consecutive frames using cosine similarity and filtering using image distances that are Euclidean distance and Mean square error. A summary of the parameters used for summarisation and the summaries produced for user generated videos are represented in Table 3.

5.2. Movies, news and documentaries

Movies are story based moving pictures and fall under entertainment category. Movies take up major part of the video content. Unlike movie trailers which deliberately intend to hide the most important parts of the movie, movie summaries should include all the major events and the enjoyable portions too. The existing techniques for movie summarisation includes important character and scene identification (Tsai et al. [48], Qu et al. [49] and Kannan et al. [50]), subtitle based (Hesham et al. [51], Aparicio et al. [52]), multimodal (Evangelopoulos et al. [53] and (Mademlis et al. [54]) and machine learning algorithms (Ide et al. [55]).

Character and scene based: Some of the most effective works in movie summarisation focuses on identifying important characters and scenes. For instance, Tsai et al. [48] and Qu et al. [49] perform the task of character and scene recognition by using Role community models and Interaction Emotion Rolenets respectively. In the system by Tsai et al. [48], methods for identifying leading roles and role communities are developed. Role community is a combination of character and scene. Graphical representation is made use of for role community network creation and then clustering is applied. It is a two stage process which consists of analysis and summarisation stage where in the former social network is constructed for establishing relations between role communities in a movie and in the latter summarisation is seen as a social network pruning problem. An extension of this approach is implemented by Qu et al. [49]. Here role networks are replaced with interaction and emotion rolenet also known as IE-RoleNets. The system considers three key elements for semantic understanding of movie content-role, emotion and interaction. The system uses string of IE-RoleNets to model the summary and mine optimal substrings from the IE-RoleNets to generate summary at structure level and role level. Yet another framework using character recognition is implemented by Kannan et al. [50]. The system considers user's subjective preferences over visual and textual features in a movie. The system supports shot level and scene level real valued preferences. *Subtitle based:* A model for automatic movie trailer generation by only utilizing subtitles is proposed by Hesham et al. [51]. The system uses Natural language processing, Machine learning and deep learning to analyse the subtitle text. The three major phases in the model are movie genre detection, ranking of significant words in the subtitle based on the genre detected and applying PageRank algorithm to get time frame of subtitle. A training phase is developed for movie genre identification. A genre dictionary is created and in the processing phase, subtitles are processed for key phrase detection using K-nearest neighbour classifier.

Similarly, *news and documentary* summaries should include all the important events from the video. Hence, the techniques for summarisation for this genre are not restricted to important character and scene identification. Generally, in news videos, the speaker dictates the important news following which the visuals related to the news are displayed. In documentary videos too, narration is an inevitable part. It can be observed that most of these are accompanied by subtitles. The techniques for summarising news and documentaries therefore are based on processing *subtitles*. Even in the absence of subtitles, the audio present in the videos are unambiguous that can be audio processed and converted to captions. Apart from processing subtitles, the models also makes use of *Machine Learning and Deep learning techniques*. In the proposed framework by Ide et al. [55], the audio and video contents are made to sync with each other by utilizing closed captions or subtitles. It also makes use of ML models. The model involves text processing of captions and image processing of shots. A directed graph representing structural and temporal relations known as topic thread structure is built. The model uses temporal order and cosine distance between term frequency distributions in the captions. The representative sequence is estimated from the structure by Kato et al. [56]. Text processing is done using Term Frequency-Inverse Document Frequency and Image processing uses GoogleNet detector using deep neural network, SVM classifier for classification and ImageNet database. Aparicio et al. [52] develops a framework for movie and documentary video summarisation based on subtitles. The output of the proposed system as opposed to all other frameworks is text summary. By reading the text summary user can decide if that particular movie is of interest or not. Text for the summary is selected by analysing weights generated on the subtitles or scripts after applying a weighting algorithm.

Multi modal based: Evangelopoulos et al. [53] performs movie summarisation based on the detection of combination of audio and visual saliency features. It is a multi-modal framework where user attention values are made use of. Aural importance is calculated by dominant modulation identification and visual importance is calculated by feature based attention modelling after which a visual and audio attention saliency curves are generated and integrated. A multimodal framework that utilizes video, audio and depth (stereoscopic) data for movie summarisation is proposed by Mademlis et al. [54]. A modified parameter known as Position Invariant Frame moments descriptor is utilized in the model. Position Invariant feature removes the dependency on the spatial information as to where the frame lies in the video. This is done by converting the FMoD into a histogram. The produced key segments undergo a post-processing step that considers audio modality by detecting speech overlap using speech diarization algorithm. Later, a previously developed depth jump cut detection and characterization algorithm is applied on the produced video skim [57]. Kopf et al. [58] proposes a separate framework for *historical movies*. The challenges in processing historical movies are that the video lacks professional quality and are shaky, contains noise, lack of technology and are black and white. Filtering is performed initially and relevant features are calculated by assessing camera motion features, face detection and the various shot boundaries are identified and assembled.

5.3. Sports

Videos in sports genre comprises of a large variety of sports. For sports video summarisation, the major challenge is that key events differ for various sports. For example, in soccer the key events are goal, foul, replay, etc. whereas in cricket it is boundary, six, wide, etc. Hence the detection of major events in sports is a tedious task. This also demands a need for discrete video summarisation frameworks for different sports. Sports videos can be divided on the basis of the type of game played into continuous action sports (baseball, American football) and action-and-stop sports (Soccer) [59]. A generic framework on the basis of the former classification is proposed by Li et al. [59]. Another generic framework for user generated sports videos is developed by Tejere-de-Pablos et al. [60]. Research on video summarisation for sports videos have been conducted on the basis of audio, action, motion, ontology, machine learning and neural network based.

Table 3
Parameters analysed for video summarisation in user generated videos along with the summaries produced.

Parameters used for summarisation																					Summary produced			
Category	References	Transform	Feature	Motion/ Transition	Object	Colour	Clustering	CNN	Memorability score	Visual attention	Non visual attention	Depth information/ shape	Event	Audio	Graph/ decision tree	Depth	Histogram	User preference	External source	Static	Dynamic	Text	Image/ objects
Single video	Al Nahian et al. [33]		×					×			×									×			
	Cong et al. [28]		×								×							×		×			
	De Avila et al. [26]					×	×				×									×			
	Ejaz et al. [18]	×	×	×		×				×										×			
	Fei et al. [13]							×	×	×										×			
	Fei et al. [20]			×				×	×	×											×		
	Gygli et al. [14]		×			×				×							×				×		
	Jadhav et al. [22]					×					×									×			
	Jeong et al. [24]	×	×			×					×						×			×			
	Kavitha et al. [15]	×		×						×										×			
	Li et al. [37]		×		×					×									×				
	Meng et al. [32]		×		×					×									×			×	
	Meng et al. [31]		×								×									×			
	Mundur et al. [25]					×	×				×									×			
	Panda et al. [35]							×			×								×		×		
	Sheena et al. [23]					×											×			×			
	Srinivas et al. [19]					×				×										×			
	Thomas et al. [47]	×	×	×		×				×										×			×
	Vivekraj et al. [29]		×								×										×		
	Vivekraj et al. [30]		×								×			×							×		
	Xiang-wei et al. [27]	×									×									×			
	Yin et al. [21]						×			×								×			×		
	Zhang K. et al. [36]		×					×	×											×	×		
Multiple	Nie et al. [38]		×					×													×		
	Panda et al. [39]		×			×		×			×										×		
Online	Almeida et al. [42]		×			×											×			×			
	Barhoumi et al. [40]				×														×				
	Chu et al. [43]		×				×	×							×						×		
	Ji et al. [44]						×								×			×		×	×		
	Zhao et al. [41]		×			×	×										×				×		
3Dvideo	Huang et al. [45]		×									×	×		×					×			
	Valognes et al. [46]					×						×					×			×			

Audio based: A general framework for sports video summarisation is detailed by Li et al. [59] with its application to soccer videos. Here, sports are categorized as action-and-stop sports (baseball, American football) and continuous action sports (soccer). For the former category, keyframes are extracted by considering every pitch as an important event. Then the video is split into events and non-events. For the latter, audio of broadcaster is analysed for exciting portions and these portions are extracted as keyframes. The framework is demonstrated with application to soccer. Start of an exciting event is detected by a combination of replay with some close up segments and audio excitement. A hybrid video summarisation technique for cricket videos is implemented by Javed et al. [61]. Cricket video summarisation is challenging owing to the long duration of game. Here, audio signals are divided into short frames and the pitch in each frame is computed in order to evaluate the excitement in commentary and audience cheer. Then, the corresponding video frames of the excited audio clips are fed to the framework in order to detect the key events for cricket. A decision tree along with a set of rules for events is designed to categorize boundary, six, wicket and replay events. *Action recognition (Neural network):* Another generic framework for video summarisation for user generated sports video is implemented by Tejere-de-Pablos et al. [60] with focus on deep action recognition features. A sport is a combination of players' actions. This framework is implemented with the factor that in sports videos, one action leads to another and succession of actions leads to an interesting event. Based on this fact, highlights are modelled from videos. The player's action features are determined by deep neural network and by analysing body joint features.

Machine learning based: A framework for soccer video summarisation is implemented by Zawbaa et al. [62]. The framework is composed of six phases. Pre-processing phase (dominant colour detection, shot boundary detection, shot classification), shot processing(shots are classified on the basis of either shots themselves or play/break classifier), Machine learning based logo replay detection phase (presence of logo at start and end of an event), machine learning based scoreboard detection phase (scoreboard appearance at the bottom part of the screen for five seconds and appearance only after goal), excitement event detection phase (presence of goal post, goal net, or loudness in audio commentary or audience cheer), logo based event detection and summarisation phase (For example, a goal is detected as an event when a combination of long view, player close up, audience cheer, first replay, second replay and third replay occurs). Zawbaa et al. [63] again implements a modified version for soccer video summarisation. A combination of dominant colour region detection (court detection) and shot boundary detection is carried out for identification of goal mouth, shot classification, scoreboard detection and replay detection. Shots are classified into long shot, medium shot, close-up and audience shot. Each has its own relevance in event detection. One third portion on the bottom part of a frame is scanned for the presence of scoreboard. Slow motion replay from any camera and logo based replay detection are analysed. *Ontologies:* play a vital role in extracting semantic information from any data. Video summaries can be made meaningful if semantics can be incorporated in the form of ontologies. Hence, Ouyang et al. [64] implements an ontology based sports video summarisation framework. The most important factor is that an ontology reasoning and inference scheme is integrated in the framework. The advantage of using ontologies is that it is machine readable. Sports Video Description Language is used for constructing ontologies. The semantic information for the ontologies is obtained from image and video annotations. The reasoning engine used is extension of Tableau algorithm.

Motion based: Mendi et al. [65] proposes a video summarisation framework for Rugby sports video based on motion analysis. It is evident that in the earlier works, motion analysis was not given high priority. But for sports, motion analysis also plays a key part especially for sports like Rugby. In the proposed model, motion analysis is performed based on optical flow. To calculate optical flow algorithms by Lucas et al. [66] and Horn et al. [67] were used which makes use of the sum of squared error between two frames. *CSP on features:* Boukadida et al. [68] proposed a model where video summarisation is viewed as a constraint satisfaction problem. The conditions for producing the summary are represented as constraints. CSP Solver is used for deducing the summary from the stated problem. By varying the constraints different summaries can be produced for the same video. The general representation for CSP is (X, D, C) where X stands for the variables, D is the domain and C is the constraints. CSP solver finds values for X in a domain D that satisfies the constraints C. The features computed from audio and video signals are represented as segments. Examples for constraints include "last clip is audience", "scene previous to applause", etc.

5.4. Surveillance

Surveillance videos are videos that are recorded round the clock by cameras that monitor areas like roads, markets, shops, public areas, etc. for unusual activity. It is a tedious job to go through these hours of surveillance videos to extract useful information. The issue associated with surveillance videos are that the background remains constant most of the time that results in redundancies. Various frameworks have been developed which focus only on summarising surveillance videos. The techniques used are based on motion analysis, event detection, machine learning and human visual attention. Frameworks for multi view videos have also been surveyed in the proposed framework.

Motion based: Summarisation frameworks using motion analysis of objects for surveillance videos is implemented by Chen et al. [69], Lai et al. [70], Thomas et al. [71] and Zhang Y. et al. [72]. Chen et al. [69] employs a novel edge detection technique (multi scale morphological gradient edge detection) to improve accuracy of moving object detection. Later, video segmentation is performed using trajectory information from tracking. The framework by Lai et al. [70] produce compact summaries by extracting trajectory of object of interest and optimizing trajectory arrangement in the summary produced. Tublets, which are 3D representation of spatio temporal frames, are formed and multiple objects are brought into single trajectory by projecting the tublets over video cube planes. Thomas et al. [71] incorporates an additional phase in the surveillance video summarisation framework which is accident stage study with type of collision. Aggregated channel feature detection to extract moving objects and produce a bounding box around the object [73]. All frames are selected into a single frame. Decision on whether a frame has to be chosen for summary depends on cost function which is a sum of activity, saliency and collision costs. In the second phase, different collisions are studied and classified as head on, rear end, single vehicle and intersection collisions. Also, pre-accident, accident and post accident stages are extracted for analysis. Zhang Y. et al. [72] define keyframes as frames with motion state changes. Generally, optical flow method is used for motion analysis but it is extremely time consuming. Hence, another method is introduced in this framework known as STS-CS (spatio temporal slice) detection from which human attention value is computed from noise power spectrum and video power spectrum. A *machine learning* approach that uses external information is proposed by Zhu et al. [74]. The system analyses not only visual data from the video but also non visual data from external heterogeneous independent sources. A multi-source clustering forest is developed for sourcing external information that is

trained in every phases. Decision tree along with a set of rules are also deployed for decision making on selection of keyframes.

Event based: In majority surveillance videos, majority of the time no significant events are taking place. Hence detecting events from the video is crucial in surveillance genre. Damnjanovic et al. [75], Song et al. [76] and Zhang S. et al. [77] implements techniques for surveillance video summarisation that are event based. Damnjanovic et al. [75] identifies events by analysing energy difference between two frames. Later, clusters are extracted using spectral clustering for building dynamic skims. Song et al. [76] also implements event detection for large scale video summarisation without performing any shot boundary detection. Trajectories are created for the objects detected like vehicles and pedestrians. Abnormal events are detected from the trajectories and disjoint max coverage algorithm is applied for keyframe extraction. Zhang S. et al. [77] proposes the fact that spatiotemporal correlations between events are important as the events itself and video skims are generated on the basis of this fact. This makes the system context aware when compared to other existing systems. The proposed system learns a dictionary of these features during the training phase. A feature correlation graph is created that represents the spatiotemporal correlations between features. During scan of the video segments, if features and correlations can be well represented by the already learned features, the respective segment can be eliminated from the video summary. *Human attention* is considered as an important property for surveillance video summarisation by Thomas et al. [47]. In the proposed system, video summarisation problem is reduced to an efficient image retrieval problem. Human attention model is used for detecting key events. HVS (Human Visual System) is the system used for finding salient regions in the frames. For optimisation purposes, keyframes with same background are combined into a single frame in the resultant summary.

Multi view videos: Multi view videos are the videos of the same scene that are captured by different cameras or at different views. A potentially fast video summarisation framework for multi view videos using DNA sequence simulation is proposed by Kumar et al. [78]. Identical scenes captured by different cameras are compared with DNA sequence detection. DNA consists of Adenine, Cytosine, Guanine and Thymine (ACGT). Similarly, a frame can be divided as NE (No evidence meaning no moving objects), SH (Some hints or one moving object), SE (Significant evidence or two moving objects) and SV (frame has more than two moving objects). Pairwise local alignment FASTA algorithm used in DNA sequence is utilized to establish correlation between different views. Four phases are Visual feature extraction, nucleotide sequence formation, multi view video correlation with FASTA and event summarisation.

5.5. Egocentric

Egocentric videos are first person view videos. Due to the invention of wearable cameras, a large number of egocentric videos are being created daily. Egocentric videos exhibit a person's day to day activities. In certain wearable cameras like head mounted cameras, the video creator remains anonymous i.e. the person who shot the video might not appear in the footage. Egocentric videos are of long duration and summarising the videos to analyse the key events is crucially important. The different approaches to video summarisation of egocentric videos are based on objects and people detection (Lee et al. [79], Sun et al. [80], Yang et al. [81]), gaze analysis (Xu et al. [82]), object triggered query (Jain et al. [83]), story driven (Lu et al. [84] and (Kuncheva et al. [85]), neural networks (Varini et al. [86]) and visual features (Guo et al. [87] and Mademlis et al. [88]).

People and object based: Video summarisation models focusing on identifying important people and objects in the video are implemented by Lee et al. [79], Sun et al. [80] and Yang et al. [81]. Lee et al. [79] aims at creating a summary of a person's day by identifying important objects as those with which the user has significant interaction. The major egocentric features incorporated in the system are object interactions, user gaze, frequency of object detection, face, object motion, etc. On the basis of these features the frames are clustered and an affinity matrix is created which leads to selection of keyframes. In the framework proposed by Sun et al. [80], the objective is to identify important shots or montages. Montages are the shots that can be combined to form a summary. In other words, the goal is to identify important people and actions in a video particularly when multiple people exist. The framework makes use of poselet based robust human tracking algorithm to detect human poses and trajectories. A saliency detector using tracklets is trained to find important people. A tracklet is a fragment of the track followed by a moving object detected by the system. A seam term and montagability score is introduced for shot boundary detection and montage evaluation. A large montagability score indicates high montage. Yang et al. [81] proposes a video summarisation system where common features among all interactions known as interaction features are found out. IF is composed of head information, body language, and emotional expression. Here, Hidden Markov model with support vector machine is employed to model the interaction sequences and to extract meaningful frames from the video. This approach demonstrates good performance by covering all the important frames optimally.

Gaze analysis based: Xu et al. [82] proposes a video summarisation system that performs the task on the basis of gaze of the user. It depends on how a user views the world through a sequence of gaze measurements which conveys a person's interest. Relevance and diversity are two terminologies of interest where relevance considers frames of interest and diversity helps in eliminating redundant frames to ensure optimality. The optimization scheme used is non-monotone submodular maximisation with matroid constraints. *Object based Query retrieval:* Jain et al. [83] views video summarisation problem as an object triggered query retrieval model. Here, all the occurrences of a given query object is analysed and respective frames are extracted. Later, frame pruning is performed to sample only the significant frames which retain the sharp dissimilar adjacent frames. The scenes are modelled as an AND OR tree and a sparse dictionary is built over the objects present in different frames. *Story driven:* In this technique, summarisation is done by analysing relationship between events and by examining which event led to another. This model estimates influence of one visual event on another. While Lu et al. [84] analyses events by dividing video into sub shots, Kuncheva et al. [85] uses Greedy Tabu selector (GTS). In the former system, each subshot entities are identified and in the next phase, the individual importance of the entity as well as its importance on other sub shots is evaluated. Finally, an optimized energy function selects certain number of keyframes that well represents the relation between events and importance of entities. Kuncheva et al. [85] assumes that the video is already segmented into various classes of interest. Here, frames are represented as points in a n-dimensional space and keyframe selection is done using nearest neighbour classifier (1-nn).

Visual feature based: Guo et al. [87] and Mademlis et al. [88] develops video summarisation system for egocentric videos based on spatial and temporal importance. In the system by Guo et al. [87], weights for importance measure and diversity parameter are computed for each frame where the former is known as stable salience weight and the latter is discriminative weight. A weighted group sparse lasso model is deployed to embed weights of shots to extract relevant video shots. The framework proposed by Mademlis et al. [88] is a static summarisation problem that extracts key-

frames that acts as a dictionary. The frames that consist of elementary visual blocks are only considered. The methodology is formulated as a Column subset selection (CSS) problem where video is represented as a matrix D, summary is a matrix C that has lesser number of distinct frames but should exhibit the same amount of information. *Neural network based:* The framework proposed by Varini et al. [86] for summarising egocentric videos is by utilizing neural networks. Videos considered are taken in a cultural heritage scenario by considering user preferences also as input. By considering user preference the model will be able to consider user's intention for taking the video which otherwise will not be included in the summary. The framework makes use of a motion behaviour classifier and a three dimensional Convolutional Neural Network. Other factors considered equally important are semantic relatedness (user intention) and narrative importance (storyline). User preference is converted to lemmatized form. The lemmatized form is used for collecting images from DBpedia. DBpedia is a structured knowledge base that is machine readable and describes more than 4.5 million things including persons, places, creative works, etc. The various features for the 3d CNN are Optical flow, Motion boundary and blurriness. The classifiers are built on the basis of GPS tour information and DBpedia semantic information together. Optical flow analysis is performed by Farneback algorithm [89].

5.6. Informational talks

Informational talk videos are often accompanied by slides and speech. One way of summarising such videos is by analysing the presentation present in the video with focus on slide transitions. Some other techniques are based on combined analysis of the speech variations or audio content and gesture analysis of the speaker. He et al. [90] proposes a system for based on three vital information: a. audio signal, pitch and pause, b. slide transition points in presentation, 3. Information about access patterns of previous users. A combination of these three factors is used for video skim generation. This method has been proven to be efficient for educational domain as it incorporates all the necessary factors for summarisation. An easier approach to video summarisation is implemented by Taru et al. [91] which converts video content to audio content and the resulting audio content to text content using Application Programming Interface calls. Later the text is evaluated by weighting schemes and a text summary is generated which can be used by the user to evaluate if the video is of interest or not. A summary of the parameters used for various genre specific approaches mentioned above along with the summaries produced is represented in Table 4.

6. Datasets used

A large number of public video datasets are available for experiments. The generic video datasets are TVSUM, SumMe, YouTube, TRECVID, etc. Majority of the frameworks have used TVSUM, Open Video, YouTube and SumMe for experiments. These are large scale datasets that comprise of videos from all categories. Egocentric datasets are UTE, GTEA-gaze+, EgoSum + Gaze, and ADL. The datasets widely used for surveillance are Urbantracker, VIRAT, Tour20, TISI, and ERCe, OFFICE, LOBBY, BL-7F. The last three are multi view datasets. UGSV datasets are specifically designed for sports videos. Specific datasets for multivideo summarisation are also available such as MVS1K. Movie datasets used in the literature are MUSCLE and Kaggle. Three dimensional video dataset experimented in the literature is Princeton dataset. The news video dataset explored in the survey is NII TV-RECS which is live news recording from news programme NHK News7. A short description of the various datasets available for video analysis along with the references and the model where the dataset is used in the literature is summarised in Table 5.

7. Evaluation

Evaluating the summarisation frameworks is a challenge. There exists no unique method or metric for evaluating the video summarisation frameworks. Evaluation metrics used can be categorized as quantitative and qualitative evaluation metrics. The surveyed systems are evaluated quantitatively, qualitatively or as a combination of both.

7.1. Quantitative metrics

There are no standard metrics for quantitatively evaluating the video summarisation frameworks. The frequently used metrics for quantitative evaluation are the ones used in information retrieval systems; precision, recall, f-score and accuracy. F-score is the most commonly used metric since it is a trade-off between precision and recall and is computed as the harmonic mean of precision and recall. Other evaluation metrics include computational time, number of key frames, Compression ratio (CR) ([15,23,40]), Computational time ([26,12]), Area under Curve (AUC) [38] and Rogue score [52]. Error analysis is based on Mean square error, Mean absolute deviation, etc. In summarisation systems implemented using neural networks, the evaluation metric used is classification accuracy [86], training time, error rate, etc. Few of the important metrics is defined below.

- *Precision* is defined as the fraction of relevant frames among the retrieved frames and is evaluated from Eq. (1).

$$\text{Precision} = \frac{|\{\text{Relevant frames}\} \cap \{\text{retrieved frames}\}|}{|\{\text{retrieved frames}\}|} \quad (1)$$

- *Recall* is the fraction of relevant instances that have been retrieved over the total amount of relevant instances and is evaluated from Eq. (2).

$$\text{Recall} = \frac{|\{\text{Relevant frames}\} \cap \{\text{retrieved frames}\}|}{|\{\text{relevant frames}\}|} \quad (2)$$

- *F-score* is the harmonic mean of precision and recall and trades off the drawbacks of precision and recall and is evaluated from Eq. (3).

$$\text{F} - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

- *Accuracy* is the percentage of correctly classified frames and is evaluated from Eq. (4).

$$\text{Accuracy} = 100 * \frac{\text{Retrieved key frames}}{\text{Total number of frames}} \quad (4)$$

- *Number of key frames* is the total number of frames in the summary produced.
- *Computational time* is the time taken to generate the summary.
- *Compression ratio* denotes the extent to which the summary is compressed and is calculated from the Eq. (5).

$$\text{Compression ratio} = \frac{\text{Number of key frames}}{\text{Total number of frames}} \quad (5)$$

- *Rogue score*(*Recall-Oriented Understudy for Gisting Evaluation*) is a metric used for evaluating text summarisations in natural language processing.

Table 4
Parameters analysed in genre specific video summarisation frameworks along with the summaries produced.

		Parameters for summarisation																										Summary produced			
Genre	References	Feature	Motion/ Transition	Object	Colour	Clustering	Neural Network	user attention	Gaze	Shape	Event	Closed captions/text	Audio	Visual	NLP	Character	Graph/ decision tree	Depth	Interaction/ Emotion	CSP	Close up	DNA	Energy diff	Tublet	User preference	External source	Static	Dynamic	Text	Image/ object
Movies, news,documentary	Aparicio et al. [52]	×									×	×																	×	
	Evagelopoulos et al. [53]	×											×												×			×		
	Hesham et al. [51]											×			×													×		
	Ide et al. [55]						×					×	×				×											×		
	Kannan et al. [50]										×					×									×			×		
	Kopf et al. [58]	×	×								×					×												×		
	Mademlis et al. [54]	×	×			×							×															×		
	Qu et al. [49]																×		×									×		
	Tsai et al. [48]					×										×	×								×			×		
Sports	Boukadida et al. [68]												×							×								×		
	Javed et al. [61]				×						×		×				×											×		
	Li et al. [59]										×		×								×							×		
	Mendi et al. [65]		×																								×			
	Ouyang et al. [64]	×															×										×	×		
	Tejero-de-pablos et al. [60]		×				×																					×		
	Zawbaa et al. [63]		×		×		×				×										×							×		
	Zawbaa et al. [62]		×		×		×				×										×							×		
Surveillance	Chen et al. [69]	×	×	×																							×			
	Damnjanovic et al. [75]			×							×												×				×	×		
	Kumar et al. [78]	×									×											×					×	×		
	Lai et al. [70]	×		×																				×				×		
	Song et al. [76]		×	×																							×			
	Thomas et al. [12]							×			×																	×		
	Thomas et al. [71]	×	×		×																							×		
	Zhang S. et al. [77]	×									×																	×		
	Zhang Y. et al. [72]		×	×				×																			×			
	Zhu et al. [74]	×						×									×									×		×		
Egocentric	Guo et al. [87]	×												×														×		
	Jain et al. [83]	×		×													×										×			
	Kuncheva et al. [85]						×																				×			
	Lee et al. [79]	×		×		×			×							×			×								×			
	Lu et al. [84]			×							×			×													×			
	Mademlis et al. [88]	×																									×			
	Sun et al. [80]		×													×	×										×			×
	Varini et al. [86]		×				×	×																	×			×		
	Xu et al. [82]	×		×													×											×		
	Yang et al. [81]	×									×			×					×								×			
Infotalks	He et al. [90]		×					×				×																×		
	Taru et al. [91]											×	×		×														×	

Table 5
Datasets used in various frameworks with a short description and reference.

Datasets	Short description	Reference	Used in
TVSUM	Title-based Video Summarisation (TVSum) dataset; videos of various genre	[92]	[33,37,31,35,36,77]
Kodak Home Video	Consumer video dataset	[93]	[28]
Open Video	Shared digital video collection from various genre	[94]	[26,18,22,15,19,25,42]
YouTube/Youtube-8 M	Large scale video dataset with annotations	[95]	[13,12,36,42,43,41,71,83,80]
SumMe	Video dataset with 15 to 18 videos and summaries	[14]	[20,14,37,31,29,30,21,36,43,87]
UTE	First person daily activity videos in an uncontrolled setting	[96]	[24,87,79,84]
CoSum	Collection of videos from Summe covering 10 different categories	[43]	[35]
KTH	Human actions dataset	[97]	[23]
Vimeo	Generic video hosters and dataset	[98]	[12]
NUSEF	Eye fixation database for saliency detection	[99]	[38]
Tour20	140 videos of 20 tourist attraction places from YouTube	[100]	[39]
MVS1K	Largest annotated dataset for multi video summarisation	[101]	[44]
Princeton	General RGBD or 3d videos	[102]	[46]
MUSCLE	Multimodal movie database with rich annotation for dialogue and saliency detection	[103]	[53]
Kaggle	Multi genre movie database	[104]	[51]
NII TV-RECS	Video recordings from news program NHK News 7	[105]	[55]
UGSV Kendo	User generated sports videos of Japanese martial art Kendo	[60]	[60]
OFFICE	Multi view datasets	[106]	[78]
LOBBY	Multi view datasets	[107]	[78]
BL-7F	Multi view datasets	[108]	[78]
BBC motion gallery	Video collection of news, sport, natural history, wildlife, locations, celebrities, history, culture, science and stock	[109]	[47]
UCF 101	UCF101(action recognition dataset of realistic action videos from youtube)	[110]	[47]
Urban Tracker	Annotated video collection of urban traffic	[111]	[71]
VIRAT	Event recognition dataset for surveillance (Video Image Retrieval and Analysis tool	[112]	[77]
TISI	Time Square Intersection dataset collected from publicly accessible webcam	[113]	[74]
ERCe	Educational Research Center -indoor campus videos collected from publicly accessible webcam	[113]	[74]
ADL	First person daily activity videos at home in an uncontrolled setting	[114]	[79,84]
IMPART	Multimodal multiview datasets	[115]	[88]
GTEA- gaze+	Videos of cooking	[116]	[82]
EgoSum + Gaze	Daily activity dataset	[82]	[82]
TRECVID	TREC Video Retrieval evaluation- large collection of videos under all category	[117]	[47]
UCLA office	Dataset of three surveillance videos of single and two-person activities	[118]	[77]
CDVL	Consumer Digital Video Library-collection of videos for video processing intended for researchers and developers	[119]	[72]
CAVIAR	Context Aware Vision using Image-based Active Recognition-Collection of different video clips recorded from different scenarios.	[120]	[72]
EDUB 2015	Egocentric Dataset of University of Barcelona	[121]	[83]
FPO	First Person Outing dataset-10 unedited videos of 2.25 h long including parks, playgrounds, lakeside, etc. recorded using head mounted cameras	[80]	[80]
Art City Egocentric	Tourist videos using head mounted cameras of visits to six Italian Art cities	[86]	[86]

7.2. Qualitative metrics

As with quantitative metrics, there are no standard qualitative metrics to evaluate the performance of the summarisation frameworks. The evaluation is mostly done on the basis of user responses to a questionnaire that covers major subjective aspects such as informativeness, representativeness, diversity, conciseness, coverage, semantic essence, enjoyability, understandability, etc. The most common ones used to evaluate the models are described below.

- *Informativeness* is the measure of how well the generated summary provides information about the original video.
- *Enjoyability* measures the extent to which the generated summary covers the enjoyable moments of the video particularly in movie genre.
- *Coverage* evaluates how well the generated summary covers the important events in the original video.
- *Rank* is a user preference value given by the user by comparing the performance of the model with other summarisation models.

7.3. Evaluation of user generated video summarisation frameworks based on F-score

A vast majority of the user generated or general movie summarisation frameworks are performance evaluated based on quantitative metrics. Very few models are evaluated qualitatively and as a combination of both. Comparison of the different models based on f-score is plotted in Fig. 4. The datasets used for experiments in the models are generic datasets like YouTube, SumMe, TVSUM50, OpenVideo, MVS1K, etc. The characteristics of the datasets are detailed in Table 5. It is observed that a high f-score is achieved in the model implemented by Chu et al. [43]. The model is implemented for online videos and detects visually co-occurring frames in a collection of videos online. The model implemented by Fei et al. [20] achieves an f-score of 90.4 which is high when compared to other models. The model is based on memorability scores where frames with more memorability are detected as key frames. Fei et al. [20] implements a visual attention based summarisation model that makes use of memorability score. Here, user's intention to record a video is identified by analysing snap points in a video. Snap points can be easily identified since these frames will have better focus and quality. Since user interested frames are separately analysed, information loss in the extracted summary is minimum. Keyframes are also identified on the basis of motion by generating optical flow diagrams. Since the method utilizes image memorability, user interest and motion analysis, it successfully generates high quality dynamic video summaries and can be successfully employed for user generated videos in any domain. The model implemented by Vivekraj et al. [30] is a multimodal one and achieves a relative f-score (actual f-score not available) of 20 which is considerably high particularly in multimodal frameworks.

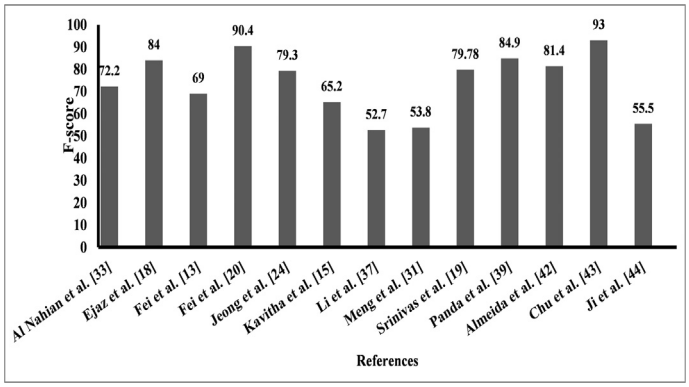


Fig. 4. Comparison based on f-score for user generated videos.

For 3D videos, rate distortion curves are analysed for performance evaluation. Rate distortion curve determines the minimum number of frames (rate-R) that should be present in the resulting summary so that the original video can be reconstructed without exceeding a given information loss (distortion-D). R-D curves should ideally be lower for optimal systems. The model implemented by Huang et al. [45] is evaluated against another 3D video summarisation model developed by [122]. The former gives improved key frame selection for a variety of 3D videos and the rate distortion curve is lower compared to the latter model. Hence, it is an acceptable model for 3D video summarisation.

7.4. Evaluation of summarisation framework in the movie and informational talks genre based on qualitative parameters

Majority of movie summarisation frameworks are evaluated qualitatively based on subjective parameters. For movies and informational talks, subjective elements such as informativeness, enjoyability, etc. are more important than evaluating accuracy of the system. Comparison of the different models based on these qualitative parameters is plotted in Fig. 5. The datasets used for experiments in the referred models in movies include MUSCLE, ECHO datasets and random movies. For informational talks, the dataset used for experiment is MSTE dataset from Microsoft. The characteristics of the datasets used for experiments are summarized in Table 5. The overall score shows that the model implemented by He et al. [90] for informational talks performs well in the genre. No recent works have been observed in the informational talk video summarisation category. Apart from that, the model proposed by Evangelopoulos et al. [53] and Kannan et al. [50] achieves an overall score of 84 and 80.6 respectively. The model by Evangelopoulos et al. [53] is based on user preference which outperforms the other models. The model proposed by Kannan et al. [50] is a multimodal framework that considers both audio and visual elements and has achieved an equally high overall preference score in the movie genre. It demonstrates that the system generates better summaries by considering both audio and visual saliencies.

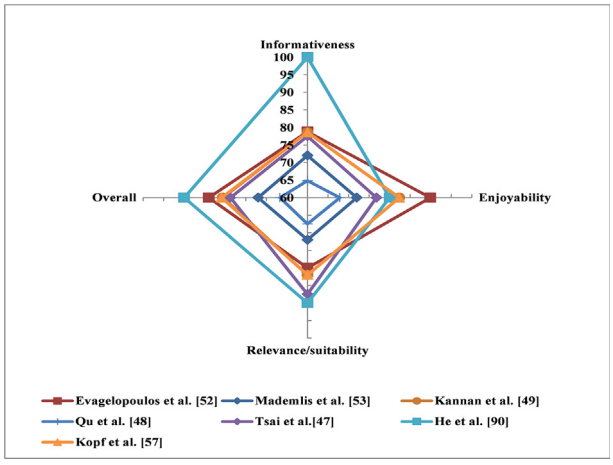


Fig. 5. Comparison based on qualitative parameters for movie summarisation frameworks.

7.5. Evaluation of sports videos based on quantitative (f-score) and qualitative (user rating) metrics

For sports videos, it is observed that the models surveyed are evaluated quantitatively and qualitatively on an equal ratio. Comparison of the different models based on f-scores and user scores are plotted in Fig. 6 (a) and (b) respectively. There is no general datasets for sports videos. All the datasets taken for experiments in the reviewed models are from championship videos or user generated sports videos detailed in Table 5. The different sports videos considered are Soccer, Cricket, Kendo, Diving, Rugby and Tennis. It is observed that the model proposed by Zawbaa et a. [62] achieves highest f-score (93.97) in sports domain. Zawbaa et al. [62] implements a soccer video summarisation framework with machine learning and support vector machine. This demonstrates that SVM classifier works efficiently for soccer videos. But in subjective evaluation, the model implemented by Boukadida et al. [68] outperforms the other models. The model is a novel approach for general videos based on Constraint Satisfaction programming. It has more application to sports where the conditions for each sport are specified as constraints. The model outperforms other state of the art models based on user rating.

7.6. Evaluation of surveillance and egocentric videos based on precision

Majority of surveillance and egocentric videos are performance evaluated on the basis of Precision metric. Subjective or qualitative evaluations are rare. Quantitative evaluations can be performed easily since the major events in surveillance videos in a particular domain are always fixed and are easily identifiable. Comparison based on the same is plotted in Fig. 7 (a) and (b) respectively. The datasets used for experimentation in the models for surveillance videos are OFFICE, Lobby, BL-7F, UrbanTracker, UCLA, VIRAT, SumMe, CDVL and CAVIAR which are detailed in Table 5. Similarly, the datasets used for evaluation on egocentric videos are UTE, OpenVideo, SumMe, EDUB 2015, FPO and Art City Egocentric data-

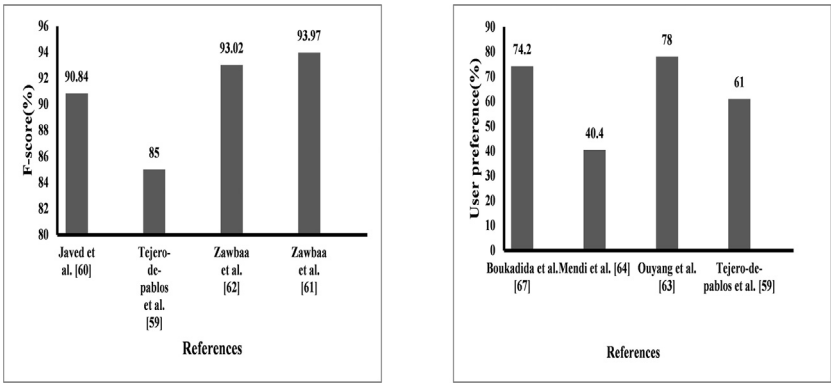


Fig. 6. Comparison of sports smmarisation frameworks based on (a) f-score(b) user ratings.

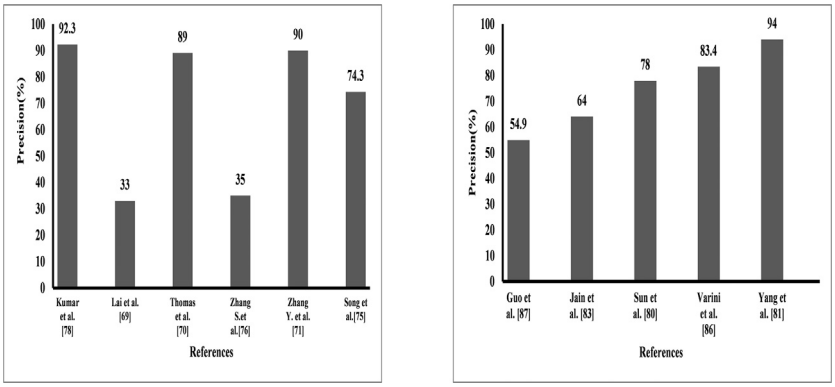


Fig. 7. Comparison based on precision of (a) surveillance video summarisation frameworks (b) egocentric video summarisation frameworks.

sets detailed in Table 5. The model implemented by Kumar et al. [78] achieves high precision in the surveillance video domain. The framework extracts unusual events in surveillance videos from multiple videos of the same location by comparing the process to DNA nucleotide detection. It is a novel technique experimented and has demonstrated high performance which makes it suitable for surveillance videos especially which are taken from multiple cameras. In egocentric videos, the model proposed by Yang et al. [81] has achieved a precision of 94%. The model utilises Interaction Features (third person interaction to identify important objects) in combination with Hidden Markov Model and Support vector machine. Although, design of the model is complex, the results achieved are ideal and optimum which shows that the proposed model can be successfully utilised for egocentric video summarisation.

8. Challenges and future recommendations

Identifying the challenges and future research scope is one of the prime outcomes of a survey. Following are some of the challenges identified while surveying the works undertaken in this framework. It is worth pointing out that addressing these factors could possibly stem for future research concerns.

- It can be observed that a large number of works using different techniques have been done in user generated videos, sports videos and egocentric videos. But, the number of works cited in news, documentaries and informational talk genre are comparatively less. Informational talk videos include online educational videos that are gaining high popularity these days. Efforts to produce reliable summaries in this genre will obviously be a great advancement in the area. It is noteworthy that summarisation frameworks for news and documentary videos are also comparatively less. More efforts utilizing novel technologies to produce fruitful summarisation frameworks in these genres is a potential scope for research.
- Although, various techniques have been addressed for video summarisation problems, it is noteworthy that only very few works addresses the use of semantics in video summarisation problems. Extracting semantics from the video could aid in identifying the key segments for yielding reliable summaries that are semantically related. Hence, developing novel techniques for expoliting semantics from the video will certainly amount to a prospective research concern. Additionally, semantics can be put to use for automatic video genre identification which could be a pre-phase in summarisation. Another prospective concern identified is the need for incorporating user preferences in the frameworks which will also result in meaningful summaries.
- Utilizing external sources of information could aid in generating more expressive summaries. But, exploiting external information in a standardized form is not an easy task. Though, many frameworks surveyed have utilized context information for summary generation, there is still need for significant research in this direction on how to model the context information in a standardised form.
- A standard benchmark evaluation metric for video summaries is not realized until now. The typical metrics used for evaluation are precision, recall, fscore, compression ratio, computation time, etc. These are common for all information retrieval systems. Comparison of the performance of a system with another has become difficult with the absence of standard evaluation

metrics. Research efforts in this direction is a notable research concern.

- Standard benchmark datasets in each genre also needs to be realized. Absence of standard datasets makes efforts for system experimentation and analysis futile. So, the available datasets and new ones could be combined under standard benchmark categories for easier experimental analysis under different genre.
- It has been observed that deep learning and machine learning models is prevailing in almost all genres. But, the training time taken and the availiability of training data for genre specific videos is a potential challenge. Research endeavours to build optimized deep learning frameworks addressing these issues will be worthwhile.

9. Conclusion

Videos belonging to different genre exhibit dissimilar characteristics. Extraction of significant information from videos of varying characteristic involves the implementation of genre-specific models. The proposed framework identifies the need for having genre-specific models for video summarisation. A survey of the recent research works in the different domains and detailed analysis on the results are discussed in this framework. The proposed framework encompasses user generated videos (consumer videos or general videos), movies, sports, surveillance, egocentric, and educational. The framework could successfully identify the variables that needs to be analysed in a framework for a video summarisation problem from a particular genre. Additionally, the type of summary suitable for a particular genre and the available datasets for experimental analysis have been analyzed in detail. The techniques implemented for particular genre can be successfully applied to videos of identical genre.The challenges and future research recommendations in this area have also been identified. Although, the results are promising in every models, there still needs a lot of research in this area for extracting semantically meaningful frames or segments with more accuracy, that does not result in information loss and can best represent a video in optimum frames or time.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

References

[1] A.G. Money, H. Agius, Video summarisation: A conceptual framework and survey of the state of the art, J. Vis. Commun. Image Represent. 19 (2) (2008) 121–143.

[2] P. Kaur, R. Kumar, Analysis of video summarisation techniques, Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) 6 (1) (2018) 1157–1162.

[3] M. Ajmal, M.H. Ashraf, M. Shakir, Y. Abbas, F.A. Shah, Video summarization: techniques and classification, in: International Conference on Computer Vision and Graphics, Springer, 2012, pp. 1–13.

[4] T.M. Moses, K. Balachandran, A classified study on semantic analysis of video summarization, in: 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), IEEE, 2017, pp. 1–6.

[5] B.T. Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM Trans. Multimedia Comput., Commun. Appl. (TOMM) 3 (1) (2007) 3.

[6] J.R. Wang, N. Parameswaran, Survey of sports video analysis: research issues and applications, in: Proceedings of the Pan-Sydney Area Workshop on Visual information processing, Australian Computer Society, Inc., 2004, pp. 87–90.

[7] D. Tank, A survey on sport video summarisation, Int. J. Sci. Adv. Res. Technol 2 (10) (2016) 435–439.

[8] A. Betancourt, P. Morerio, C.S. Regazzoni, M. Rauterberg, The evolution of first person vision methods: A survey, IEEE Trans. Circuits Syst. Video Technol. 25 (5) (2015) 744–760.

[9] M. Bolanos, M. Dimiccoli, P. Radeva, Toward storytelling from visual lifelogging: An overview, IEEE Trans. Human-Mach. Syst. 47 (1) (2017) 77–90.

[10] A.G. del Molino, C. Tan, J.-H. Lim, A.-H. Tan, Summarization of egocentric videos: A comprehensive survey, IEEE Trans. Human-Mach. Syst. 47 (1) (2017) 65–76.

[11] P. Kalaivani, S.M.M. Roomi, Towards comprehensive understanding of event detection and video summarization approaches, in: 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), IEEE, 2017, pp. 61–66.

[12] S.S. Thomas, S. Gupta, V.K. Subramanian, Perceptual video summarization–a new framework for video summarization, IEEE Trans. Circuits Syst. Video Technol. 27 (8) (2017) 1790–1802.

[13] M. Fei, W. Jiang, W. Mao, Memorable and rich video summarization, J. Vis. Commun. Image Represent. 42 (2017) 207–217.

[14] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: European Conference on Computer Vision, Springer, 2014, pp. 505–520.

[15] J. Kavitha, P.A.J. Rani, Static and multiresolution feature extraction for video summarization, Procedia Comput. Sci. 47 (2015) 292–300.

[16] A. Khosla, A.S. Raju, A. Torralba, A. Oliva, Understanding and predicting image memorability at a large scale, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2390–2398.

[17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, Adv. Neural Informat. Process. Syst. (2014) 487–495.

[18] N. Ejaz, I. Mehmood, S.W. Baik, Feature aggregation based visual attention model for video summarization, Comput. Electr. Eng. 40 (3) (2014) 993–1005.

[19] M. Srinivas, M.M. Pai, R.M. Pai, An improved algorithm for video summarization–a rank based approach, Procedia Comput. Sci. 89 (2016) 812–819.

[20] M. Fei, W. Jiang, W. Mao, Creating memorable video summaries that satisfy the user's intention for taking the videos, Neurocomputing 275 (2018) 1911–1920.

[21] Y. Yin, R. Thapliya, R. Zimmermann, Encoded semantic tree for automatic user profiling applied to personalized video summarization, IEEE Trans. Circuits Syst. Video Technol. 28 (1) (2018) 181–192.

[22] M.P.S. Jadhav, D.S. Jadhav, Video summarization using higher order color moments (vsuhcm), Procedia Comput. Sci. 45 (2015) 275–281.

[23] C.V. Sheena, N. Narayanan, Key-frame extraction by analysis of histograms of video frames using statistical methods, Procedia Comput. Sci. 70 (2015) 36–40.

[24] D.-J. Jeong, H.J. Yoo, N.I. Cho, A static video summarization method based on the sparse coding of features and representativeness of frames, EURASIP J. Image Video Process. 2017 (1) (2016) 1.

[25] P. Mundur, Y. Rao, Y. Yesha, Keyframe-based video summarization using delaunay clustering, Int. J. Digit. Libr. 6 (2) (2006) 219–232.

[26] S.E.F. De Avila, A.P.B. Lopes, A. da Luz Jr, A. de Albuquerque Araújo, Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recogn. Lett. 32 (1) (2011) 56–68.

[27] L. Xiang-wei, Z. Li-dong, Z. Kai, Hierarchical video summarization extraction algorithm in compressed domain, Phys. Procedia 24 (2012) 2360–2366.

[28] Y. Cong, J. Yuan, J. Luo, Towards scalable summarization of consumer videos via sparse dictionary selection, IEEE Trans. Multimedia 14 (1) (2012) 66–75.

[29] V. Vivekraj, R. Balasubramanian, D. Sen, Vector r-ordering based selection of segments for video skimming, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 871–876.

[30] V. Vivekraj, D. Sen, R. Balasubramanian, Vector ordering based multimodal video skimming for user videos, in: Region 10 Conference, TENCON 2017-2017 IEEE, IEEE, 2017, pp. 775–780.

[31] J. Meng, S. Wang, H. Wang, J. Yuan, Y.-P. Tan, Video summarization via multi-view representative selection, IEEE Trans. Image Process. (2018) 2134–2145.

[32] J. Meng, H. Wang, J. Yuan, Y.-P. Tan, From keyframes to key objects: Video summarization by representative object proposal selection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1039–1048.

[33] M. Al Nahian, A. Iftekhar, M.T. Islam, S.M. Rahman, D. Hatzinakos, Cnn-based prediction of frame-level shot importance for video summarization, in: 2017 International Conference on New Trends in Computing Sciences (ICTCS), IEEE, 2017, pp. 24–29.

[34] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[35] R. Panda, A. Das, Z. Wu, J. Ernst, A.K. Roy-Chowdhury, Weakly supervised summarization of web videos, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 3677–3686.

[36] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: European Conference on Computer Vision, Springer, 2016, pp. 766–782.

[37] X. Li, B. Zhao, X. Lu, A general framework for edited video and raw video summarization, IEEE Trans. Image Process. 26 (8) (2017) 3652–3664.

[38] L. Nie, R. Hong, L. Zhang, Y. Xia, D. Tao, N. Sebe, Perceptual attributes optimization for multivideo summarization, IEEE Trans. Cybernet. 46 (12) (2016) 2991–3003.

[39] R. Panda, N.C. Mithun, A.K. Roy-Chowdhury, Diversity-aware multi-video summarization, IEEE Trans. Image Process. 26 (10) (2017) 4712–4724.

[40] W. Barhoumi, E. Zagrouba, On-the-fly extraction of key frames for efficient video summarization, AASRI Procedia 4 (2013) 78–84.
[41] B. Zhao, E.P. Xing, Quasi real-time summarization for consumer videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2513–2520.
[42] J. Almeida, N.J. Leite, R.d. S. Torres, Vison: Video summarization for online applications, Pattern Recogn. Lett. 33 (4) (2012) 397–409.
[43] W.-S. Chu, Y. Song, A. Jaimes, Video co-summarization: Video summarization by visual co-occurrence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3584–3592.
[44] Z. Ji, Y. Zhang, Y. Pang, X. Li, Hypergraph dominant set based multi-video summarization, Signal Process. 148 (2018) 114–123.
[45] P. Huang, A. Hilton, J. Starck, Automatic 3d video summarization: Key frame extraction from self-similarity, in: International Symposium on 3D Data Processing, Visualization and Transmission, 2008.
[46] J. Valognes, M.A. Amer, N.S. Dastjerdi, Effective keyframe extraction from rgb and rgb-d video sequences, in: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2017, pp. 1–5.
[47] S.S. Thomas, S. Gupta, V.K. Subramanian, Smart surveillance based on video summarization, in: IEEE Region 10 Symposium (TENSYMP), 2017, IEEE, 2017, pp. 1–5.
[48] C.-M. Tsai, L.-W. Kang, C.-W. Lin, W. Lin, Scene-based movie summarization via role-community networks, IEEE Trans. Circuits Syst. Video Techn. 23 (11) (2013) 1927–1940.
[49] W. Qu, Y. Zhang, D. Wang, S. Feng, G. Yu, Semantic movie summarization based on string of ie-rolenets, Computat. Visual Media 1 (2) (2015) 129–141.
[50] R. Kannan, G. Ghinea, S. Swaminathan, What do you wish to see? a summarization system for movies based on user preferences, Informat. Process. Manage. 51 (3) (2015) 286–305.
[51] M. Hesham, B. Hani, N. Fouad, E. Amer, Smart trailer: Automatic generation of movie trailer using only subtitles, in: 2018 First International Workshop on Deep and Representation Learning (IWDRL), IEEE, 2018, pp. 26–30.
[52] M. Aparício, P. Figueiredo, F. Raposo, D.M. de Matos, R. Ribeiro, L. Marujo, Summarization of films and documentaries based on subtitles and scripts, Pattern Recogn. Lett. 73 (2016) 7–12.
[53] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, Y. Avrithis, Movie summarization based on audiovisual saliency detection, in: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, IEEE, 2008, pp. 2528–2531.
[54] I. Mademlis, A. Tefas, N. Nikolaidis, I. Pitas, Multimodal stereoscopic movie summarization conforming to narrative characteristics, IEEE Trans. Image Process. 25 (12) (2016) 5828–5840.
[55] I. Ide, Y. Zhang, R. Tanishige, K. Doman, Y. Kawanishi, D. Deguchi, H. Murase, Summarization of news videos considering the consistency of auditory and visual contents, in: 2017 IEEE International Symposium on Multimedia (ISM), IEEE, 2017, pp. 193–199.
[56] K. Kato, I. Ide, D. Deguchi, H. Murase, Estimation of the representative story transition in a chronological semantic structure of news topics, in: Proceedings of International Conference on Multimedia Retrieval, ACM, 2014, p. 487.
[57] S. Delis, I. Mademlis, N. Nikolaidis, I. Pitas, Automatic detection of 3d quality defects in stereoscopic videos using binocular disparity, IEEE Trans. Circuits Syst. Video Techn. 27 (5) (2017) 977–991.
[58] S. Kopf, T. Haenselmann, D. Farin, W. Effelsberg, Automatic generation of video summaries for historical films, 2004 IEEE International Conference on Multimedia and Expo, 2004. ICME'04, vol. 3, IEEE, 2004, pp. 2067–2070.
[59] B. Li, H. Pan, I. Sezan, A general framework for sports video summarization with its application to soccer, 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03), vol. 3, IEEE, 2003, pp. III–169.
[60] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, E. Rahtu, Summarization of user-generated sports video by using deep action recognition features, IEEE Trans. Multimedia 20 (8) (2018) 2000–2011.
[61] A. Javed, K.B. Bajwa, H. Malik, A. Irtaza, M.T. Mahmood, A hybrid approach for summarization of cricket videos, in: IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), IEEE, 2016, pp. 1–4.
[62] H.M. Zawbaa, N. El-Bendary, A.E. Hassanien, T.-H. Kim, Machine learning-based soccer video summarization system, in: Multimedia, Computer Graphics and Broadcasting, Springer, 2011, pp. 19–28.
[63] H.M. Zawbaa, N. El-Bendary, A.E. Hassanien, Automatic soccer video summarisation, 2012.
[64] J.-Q. Ouyang, R. Liu, Ontology reasoning scheme for constructing meaningful sports video summarisation, IET Image Proc. 7 (4) (2013) 324–334.
[65] E. Mendi, H.B. Clemente, C. Bayrak, Sports video summarization based on motion analysis, Comput. Electr. Eng. 39 (3) (2013) 790–796.
[66] B.D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision, 1981.
[67] B.K. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1–3) (1981) 185–203.
[68] H. Boukadida, S.-A. Berrani, P. Gros, Automatically creating adaptive video summaries using constraint satisfaction programming: Application to sport content, IEEE Trans. Circuits Syst. Video Technol. 27 (4) (2017) 920–934.
[69] Y. Chen, B. Zhang, Surveillance video summarisation by jointly applying moving object detection and tracking, Int. J. Comput. Vision Robot. 4 (3) (2014) 212–234.
[70] P.K. Lai, M. Décombas, K. Moutet, R. Laganière, Video summarization of surveillance cameras, in: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2016, pp. 286–294.
[71] S.S. Thomas, S. Gupta, V.K. Subramanian, Event detection on roads using perceptual video summarization, IEEE Trans. Intell. Transport. Syst. (2017).
[72] Y. Zhang, R. Tao, Y. Wang, Motion-state-adaptive video summarization via spatiotemporal analysis, IEEE Trans. Circuits Syst. Video Technol. 27 (6) (2017) 1340–1352.
[73] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.
[74] X. Zhu, C.C. Loy, S. Gong, Learning from multiple sources for video summarisation, Int. J. Comput. Vision 117 (3) (2016) 247–268.
[75] U. Damnjanovic, V. Fernandez, E. Izquierdo, J.M. Martinez, Event detection and clustering for surveillance video summarization, in: Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08, IEEE, 2008, pp. 63–66.
[76] X. Song, L. Sun, J. Lei, D. Tao, G. Yuan, M. Song, Event-based large scale surveillance video summarization, Neurocomputing 187 (2016) 66–74.
[77] S. Zhang, Y. Zhu, A.K. Roy-Chowdhury, Context-aware surveillance video summarization, IEEE Trans. Image Process. 25 (11) (2016) 5469–5478.
[78] K. Kumar, D.D. Shrimankar, F-des: Fast and deep event summarization, IEEE Trans. Multimedia 20 (2) (2018) 323–334.
[79] Y.J. Lee, K. Grauman, Predicting important objects for egocentric video summarization, Int. J. Comput. Vision 114 (1) (2015) 38–55.
[80] M. Sun, A. Farhadi, B. Taskar, S. Seitz, Summarizing unconstrained videos using salient montages, IEEE Trans. Pattern Anal. Machine Intell. 39 (11) (2017) 2256–2269.
[81] J.-A. Yang, C.-H. Lee, S.-W. Yang, V.S. Somayazulu, Y.-K. Chen, S.-Y. Chien, Wearable social camera: Egocentric video summarization for social interaction, in: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2016, pp. 1–6.
[82] J. Xu, L. Mukherjee, Y. Li, J. Warner, J.M. Rehg, V. Singh, Gaze-enabled egocentric video summarization via constrained submodular maximization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2235–2244.
[83] S. Jain, R.M. Rameshan, A. Nigam, Object triggered egocentric video summarization, in: International Conference on Computer Analysis of Images and Patterns, Springer, 2017, pp. 428–439.
[84] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2714–2721.
[85] L.I. Kuncheva, P. Yousefi, J. Almeida, Edited nearest neighbour for selecting keyframe summaries of egocentric videos, J. Vis. Commun. Image Represent. 52 (2018) 118–130.
[86] P. Varini, G. Serra, R. Cucchiara, Personalized egocentric video summarization of cultural tour on user preferences input, IEEE Trans. Multimedia 19 (12) (2017) 2832–2845.
[87] Z. Guo, L. Gao, X. Zhen, F. Zou, F. Shen, K. Zheng, Spatial and temporal scoring for egocentric video summarization, Neurocomputing 208 (2016) 299–308.
[88] I. Mademlis, A. Tefas, I. Pitas, Summarization of human activity videos using a salient dictionary, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 625–629.
[89] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian Conference on Image Analysis, Springer, 2003, pp. 363–370.
[90] L. He, E. Sanocki, A. Gupta, J. Grudin, Auto-summarization of audio-video presentations, in: Proceedings of the seventh ACM International Conference on Multimedia (Part 1), ACM, 1999, pp. 489–498.
[91] P. Taru, S. Hiray, S. Gurnalkar, A. Gokhale, Video summarization, International Research Journal of Engineering and Technology 4 (3) (2017) 111–112.
[92] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5179–5187.
[93] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, A. Yanagawa, Kodak's consumer video benchmark data set: concept definition and annotation, in: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, ACM, 2007, pp. 245–254.
[94] G. Marchionini, G. Geisler, The open video digital library, D-Lib Magazine 8 (12) (2002) 1082–9873.
[95] Youtube-8m, https://research.google.com/youtube8m/, accessed: 17-01-2019.
[96] http://vision.cs.utexas.edu/projects/egocentric/, accessed: 17-01-2019.
[97] http://www.nada.kth.sc/cvap/actions/, accessed: 17-01-2019.
[98] https://vimeo.com/, accessed: 17-01-2019.
[99] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, T.-S. Chua, An eye fixation database for saliency detection in images, in: European Conference on Computer Vision, Springer, 2010, pp. 30–43.
[100] http://www.ee.ucr.edu/amitrc/datasets.php, accessed: 17-01-2019.
[101] Z. Ji, Y. Ma, Y. Pang, X. Li, Query-aware sparse coding for multi-video summarization, arXiv preprint arXiv:1707.04021.
[102] S. Song, J. Xiao, Tracking revisited using rgbd camera: Unified benchmark and baselines, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 233–240.
[103] D. Spachos, A. Zlantintsi, V. Moschou, P. Antonopoulos, E. Benetos, M. Kotti, K. Tzimouli, C. Kotropoulos, N. Nikolaidis, P. Maragos, et al., Muscle movie-database: A multimodal corpus with rich annotation for dialogue and

saliency detection, in: 6th Language Resources and Evaluation Conference, 2008, pp. 16–19.
[104] Kaggle, https://www.kaggle.com/, accessed: 17-01-2019.
[105] N. Katayama, H. Mo, I. Ide, S. Satoh, Mining large-scale broadcast video archives towards inter-video structuring, in: Pacific-Rim Conference on Multimedia, Springer, 2004, pp. 489–496.
[106] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, Z.-H. Zhou, Multi-view video summarization, IEEE Trans. Multimedia 12 (7) (2010) 717–729.
[107] S.K. Kuanar, K.B. Ranga, A.S. Chowdhury, Multi-view video summarization using bipartite matching constrained optimum-path forest clustering, IEEE Trans. Multimedia 17 (8) (2015) 1166–1173.
[108] S.-H. Ou, C.-H. Lee, V.S. Somayazulu, Y.-K. Chen, S.-Y. Chien, On-line multi-view video summarization for wireless video sensor network, IEEE J. Sel. Top. Signal Process. 9 (1) (2015) 165–179.
[109] Bbc motion gallery, https://www.gettyimages.in/bbcmotiongallery, accessed: 17-01-2019.
[110] K. Soomro, A.R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv: 1212.0402.
[111] J.-P. Jodoin, G.-A. Bilodeau, N. Saunier, Urban tracker: Multiple object tracking in urban mixed traffic, in: 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2014, pp. 885–892.
[112] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J.T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al., A large-scale benchmark dataset for event recognition in surveillance video, in: Computer vision and pattern recognition, in: 2011 IEEE conference on Computer vision and pattern recognition (CVPR), IEEE, 2011, pp. 3153–3160.
[113] X. Zhu, C. Change Loy, S. Gong, Video synopsis by heterogeneous multi-source correlation, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 81–88.
[114] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2847–2854.
[115] H. Kim, A. Hilton, Influence of colour and feature geometry on multi-modal 3d point clouds data registration, 2014 2nd International Conference on 3D Vision (3DV), vol. 1, IEEE, 2014, pp. 202–209.
[116] A. Fathi, Y. Li, J.M. Rehg, Learning to recognize daily actions using gaze, in: European Conference on Computer Vision, Springer, 2012, pp. 314–327.
[117] Trecvid, https://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html, accessed: 17-01-2019.
[118] M. Pei, Y. Jia, S.-C. Zhu, Parsing video events with goal inference and intent prediction, in: 2011 IEEE International Conference on Computer Vision (ICCV) , IEEE, 2011, pp. 487–494.
[119] W. Yodel, The consumer digital video library, 2011.
[120] R. Fisher, J. Santos-Victor, J. Crowley, Caviar test case scenarios, http://homepages.inf.ed.ac.uk/rbf/CAVIAR.
[121] M. Bolanos, P. Radeva, Ego-object discovery, arXiv preprint arXiv: 1504.01639.
[122] J. Xu, T. Yamasaki, K. Aizawa, Summarization of 3d video by rate-distortion trade-off, IEICE Trans. Informat. Syst. 90 (9) (2007) 1430–1438.