



# A Review on Key Features and Novel Methods for Video Summarization

Vinsent Paramanatham, S. Suresh Kumar

**Abstract:** This paper discusses techniques, algorithms, and evaluation methods used in online, offline, supervised, unsupervised, multi-video, and clustering methods for video summarisation and multi-view video summarisation, drawing on various references. We have studied multiple techniques in the literature and described the features used for generating video summaries, along with evaluation methods, including supervised and unsupervised algorithms, as well as the datasets employed. We have covered the survey on the new frontier of research in computational intelligence techniques, such as Artificial Neural Networks (ANN) and other evolutionary algorithms for VS, using both supervised and unsupervised methods. We highlight single and multi-video summarisation, considering features such as video, audio, and semantic embeddings for VS in the literature. A careful presentation is attempted to bring the performance comparison with Precision, Recall, F-Score, and manual methods to evaluate the VS.

**Keywords:** Video Summarization, Multi-View Video Summarization, Online Offline Video Highlighting, Key Frames, Sparse Coding, Feature Extraction, Sparse Land, CNN, RNN, LSTM.

## I. INTRODUCTION

Video summarisation (VS) is a mechanism that produces a summary, synopsis, or montage of a given video. The surveyed literature broadly classifies VS as supervised, unsupervised, static or dynamic. A more recent review of text-based NLP methods has yielded better evaluation scores. A few references with Reinforcement learning for VS have also gained attention in the research [58]. However, a careful selection of video summarisation techniques is necessary, based on your domain and the specific challenge to address. This survey will aid researchers in selecting the proper methods and approaches to tackle their challenges in VS, although the selection of VS approaches and techniques is mostly subjective. Some of the prominent use cases for VS are to provide efficient indexing, browsing, search, storage reduction, synopsis/montage generation, anomaly detection [7].

Manuscript received on 22 July 2022 | Revised Manuscript received on 16 February 2023 | Manuscript Accepted on 15 February 2023 | Manuscript published on 28 February 2023.

\*Correspondence Author(s)

Vinsent Paramanatham\*, Faculty of Computing, Sathyabama University, Chennai (Tamil Nadu), India. E-mail: [vinsent.storage@gmail.com](mailto:vinsent.storage@gmail.com), ORCID ID: <https://orcid.org/0000-0002-8824-1911>

Dr. S. Suresh Kumar, Principal, Swarnandhra College of Engineering and Technology, Narasapur (A.P), India. E-mail: [nice.ssk@gmail.com](mailto:nice.ssk@gmail.com), ORCID ID: <https://orcid.org/0000-0001-9912-5927>

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

### 1.1. VS Approach

The VS and MVS approaches have been carefully reviewed from the literature, as they differ from cluster-based approaches, unsupervised and supervised methods, ANN, SVD, sparse methods, Graph Methods, DPP, and Evolutionary methods, such as genetic algorithms, MVS, and egocentric methods. The traditional approach to VS involves feature extraction, wherein an ANN-based approach utilises the inherent Video features to perform VS. In approaches like clustering, the temporal sequences are lost, which is not required in certain domains but is required in others. Certain VS approaches are targeted towards specific applications like summarizing scripted content as used in surveillance videos, anomaly detection [55]. The choice of supervised or unsupervised approach also varies depending on the application domain for VS. In the supervised approach, there are additional annotations of video, which solves the VS challenge only to a particular domain like surveillance, anomaly detection [17]

### 1.2. Features, Methods and Evaluation Techniques

The feature extraction techniques plays a vital role in producing the desired VS. Spatial features like the RGB colors, SIFT, iDT, HOG, HOF, color histogram, GMM, Trajectory-pooled Deep Convolutional Descriptors are good feature candidates for VS. Traditional methods like clustering and more recent methods like supervised and unsupervised ANN methods, Sparseland, SVD, Graph-based methods for VS are prominent in the literature [16, 33, 44, 57]. The most commonly used methods for evaluating VS results are individual human ratings, precision, recall, and F-score. The datasets widely studied for video summarization are SumMe [28], TVSum [28], VSUMM [4], OVP (Open Video Project) [1], STIMO [24], Of-office [11]. Combined spatio-temporal, audio, and text features are yet to take a route with matured proof of techniques.

### 1.3. Problem Domain

VS solves challenges like summary, synopsis, montage generation, anomaly detection, storage reduction, information retrieval, transmission [44, 55, 57]. In information retrieval, VS can be used for faster indexing and retrieval [12, 33, 43, 53]. A common framework for summarisation, whether supervised or unsupervised, remains a challenge. As most of the VS is a domain-specific challenge, a standard annotation of frames captured from one domain may not apply to other domains.

### 1.4. Classification of VS Methods

There are various methods for VS in the literature.



In this paper, we have broadly classified the approaches into various categories, as depicted in Figure 1. The main classification of VS is either supervised or unsupervised, involving frames and shots, which are either extractive or spatiotemporal clips, known as compositional. The other approach to classification is based on either keyframes or clustering methods. The advent of ANN in VS has far exceeded other methods, with various nuances that achieve higher evaluation scores. The ANN-based video summarization can be subdivided based on the techniques, such as CNN, RNN, transformer variation [8], LSTM, RL [58]. For a full citation of ANN techniques, refer to Table 5. Other prominent VS methods based on the sparse-land approach are found in the following references [6, 27, 44, 57]. Reference to ANN techniques-based VS can be seen in Table 5. MVS with ANN and DPP-based have gained attention [5, 11, 17, 18, 20], a complete reference list is found in table 5. Graph-based methods also bring an inherent queryable video summarization [35],[20]. Clustering methods which generate the keyframe based on the video features are still used as final summarization techniques in advanced VS, references of clustering and keyframe based VS can be found in table 2.

## 1.5. Tables and Figures in the Table

In [Table 1](#), we have covered the type of system, application, Key features, and processing techniques. This highlights the key features used for the VS work. [Table 2](#) covers about the key process in performing the Video summarization as can be seen that the works based on a keyframe, clustering, and sparse-land based approach are more common in VS. [Table 3](#) refers to the main feature like Optical flow, HOF, HOG, Color Histogram, and other features in the literature. [Table 4](#) presents the empirical results and evaluation methods, as well as the scores used in the VS evaluation. [Table 5](#) refers to the ANN techniques used for VS. [Table 6](#) refers to the multi-view and egocentric approach to do MVS. [Figure 1](#) provides a broad classification of VS work, primarily categorised as static (offline) and dynamic (online), which internally comprises either supervised or unsupervised methods, followed by other classifications based on the techniques employed.

## II. RELATED WORKS

### 2.1. VS Based on Unsupervised Approach

The authors propose an algorithm for VS with human activities that utilises a linear combination of visual words, exploiting semantic scene content properties. The recommended method extracts features to form a salient dictionary given an input video. The dictionary component identifies video frames with simple visual building blocks. The saliency components modulate the construction of the Dictionary, meant to ensure outlier inclusion and broad content coverage, which operates in the traditional inter-frame distance-based measures. Thus, the process strikes a balance between video frame representativeness and saliency. The Column Subset Selection Problem (CSSP) with an optimisation framework helps to achieve VS. The paper also compares its results with those of the baseline clustering approach and sparse land dictionary learning methods [25]. The authors present a context-aware VS (CAVS) framework that uses two dictionaries: one with sparse coding and

generalised sparse group lasso, and a dictionary of spatiotemporal features. The features are modelled in a correlation graph. Sparsity within the groups captures the important inter-relationships and features in Video. The spatiotemporal feature correlation graph provides information on object motion and region correlation, maintaining global information. The CAVS updates both dictionaries online as new sparse videos are received. The objects in VS are to identify the inter-relation between frames given context information as an event occurs [52]. The authors discuss an incremental subset selection framework that, at each time instant, uses a comparison technique to select a subset of representatives from a previously selected set and a new batch of representatives with minimal overlap, thereby forming the current set. VS is modelled as an integer binary optimisation problem that minimises the encoding cost via representatives regularised by the number of selected items. Real-time experiments are conducted using a randomised greedy approach for sub-modular optimisation. Tests are conducted to check the effectiveness of the subset selection framework in online VS. A proper cluster selection method is proposed [6]. The authors present an advanced silhouette extraction, human detection and tracking algorithm for indoor environments. In silhouette extraction, shadows and background changes are removed to facilitate further classification of the activity. Hue Movement Invariants (HMIs) are used to classify actions based on body shape and movements. Human action recognition is detected using adaptive features learned from the video, which is guided by a hierarchical decision tree and dimension reduction techniques. VS is done by detecting video segments and shot boundary detections via a high-level fuzzy Petri net (HLFPN). Shot Boundary Detection and grayscale histogram difference, with a zero-mean difference for three or five consecutive frames, have modelled three features that are part of the membership function for feature extraction. The proposed model aims to reduce the detection of improper shots with camera and object motion, as compared to human labour performing the same shot correction. Delaunay triangulation (DT) is used for VS, where the video contents are represented as colour histograms. Clustering is performed using DT, and the size of the cluster depends on the video content. Significant compression, overlap, and other factors are considered for VS. Automatic VS with batch processing is possible in DT compared to K-means clustering. The number of clusters is determined automatically. Multidimensional point data is given to DT clustering, where DT clustering is formed for each frame. The inter-frame similarity is measured among the data points in DT. Inter- and intra-clusters are formed to select the keyframe from the cluster, which is then used to form VS [32]. Presents a VS technique based on a motion analysis metric using optical flow algorithms, by selecting keyframes. In sports video, motions are a critical measure, and optical flow captures them well. The brightness features between two frames are computed as a threshold measure. The low and high motion threshold functions help to obtain the keyframes. Optical flow video segment features are obtained by the Lukas and Kanade method.

The keyframes formed are part of a motion analysis pattern. Presents a VS method using unsupervised cluster validation for a given arbitrary video sequence. The cluster formation is based on similar visual content. The colour histogram in the YUV colour space is used to represent the frame's features. An adaptive visual content selects the keyframes in the video and a discontinuity threshold value. A content component of cluster 'A' and 'B' forms the representative frame from the cluster obtained, followed by concatenating the representative frame to form a sequence for VS [14]. Presents an approach for VS by clustering all the frames of similar concepts and discarding the repeated sequence. The approach is well-suited for travel guides, documentaries, and dramas that frequently repeat concepts. A frame similarity threshold is used to determine whether a frame should be considered part of the frame set. Similar frames with various features are clustered by taking a Euclidean distance measure to form the VS. The frame similarity determination is generic, and the approach can be applied to multi-video summarisation. Proposes an automatic VS method by using constraint satisfaction programming (CSP) by taking various user inputs like desired length for summarisation and other video features as the solver's input.

The summarised video output is a solver's output function that incorporates constraints such as audio-visual features, user parameters, and other hard optimisation constraints, which are to be solved by the dynamic programming solver. Constraints such as the length of the video and neighbourhood constraints for sports segments are desirable in real-time sports video. Shot boundary segments are a key feature for the solver.

The authors discuss a structured learning approach, which involves taking semantic user-created inputs and a raw video as paired inputs to form a VS output. A combined objective function, formed from the input video feature and text, is given to a structure learning formulation with a greedy optimisation that maximises a submodular function. Then, a new summary is presented that is both exciting and representative of the input video. The subset selection is used to arrive at a summarisation that maximises the objective function, making it exciting and representative for a video in a weighted approach.

The objective function is solved by projected sub-gradient descent, and the segments form the atomic entities. Local segment features are considered to create a complete objective function for interestingness, for representativeness of how well a summary represents the initial video is considered, for uniformity, the objective function to remove similar adjacent frames is used [13]. Analyzed different optimization methods by comparing the time and storage reduction in VS. The online large videos can be a summarized into smaller parts in real time. To achieve this, the authors utilise feature extraction algorithms, including the gradient (HOG and HOF), and optical flow features. The effectiveness of VS is measured across various algorithms, including PSO, GA, ADMM, and ABC, with the parameters being compared in terms of time and storage reduction. The paper also determines the most effective methods for reducing storage and improving time complexity in computation. The use of sparse land [44] approach to VS is a highlight in this paper. Unedited, unstructured video is used to learn the

dictionary via sparse coding, where the atoms in the dictionary are updated online. The summary video is generated using the video that cannot be reconstructed using the learned dictionary in sparse-land. High reconstruction error causes the frame to be included in the summarisation. As an online framework, the dictionary begins to build after the initial frames are displayed. The paper proposes VS as a sparse-land problem with ADMM as the optimization choice [57]. The authors discuss a Minimum Sparse Reconstruction (MSR) that differs from the convex relaxation sparse dictionary, as L2,1 norms do not directly consider the sparsity. MSR uses a valid sparse constraint L0 norm; they also proposed an additional percentage of reconstruction (POR) criterion to intuitively obtain a summary with a desired length [27]. Localisation in surveillance applications is a vital feature extraction technique for faster anomaly detection. The authors propose a feature map selection algorithm which can intelligently choose appropriate feature maps from the convolution layers of the trained CNN, and the localization segmentation helps to identify anomalies efficiently rather than by manual localization creation, also improving the anomaly detection accuracy [30]. In the MSKVS framework to produce VS, the authors introduce a novel key frame extraction method. A hybrid descriptor, termed GFFV (Global Frame Feature Vector), is used to represent a frame. The global and local visual features from the frame are used to eliminate redundant frames by employing a linear algorithm called the adaptive mean shift algorithm (MSKVS) to select the keyframe. The key points are samples using the Difference-of-Gaussian (DoG) filter, which yields entropy-based singular values. Representativeness and Compactness Ratio (RCR) is used to perform objective comparisons among different approaches for Keyframe detection [15].

## 2.2. VS Based on Supervised Approach

The authors present a VS based on feature modalities, selecting the most multi-view representative features in the video. A multi-view sparse dictionary selection with centroid co-regularisation (MSDS-CC) is used to select features in each view, forming a representative collection for VS summarisation. The centroid co-regularisation is solved using FISTA by incorporating the view features into a consensus selection matrix, thereby creating an objective function for the sparse reconstruction summarisation model. The MSDS-CC helps to scale the solution for VS of any input video size [28].

The paper discusses a temporal collaborative representation (TCR) model by considering the visual similarity between frames for VS. TCR regards each frame as a linear combination of keyframes formed by a representative selection, ignoring the adjacent frames by defining a reconstruction error and an average mean score. Transitional frames are discarded by considering adjacent frames rather than individual frames. The adjacent frames are represented collaboratively, and the influence of transitional frames is reduced. In TCR-based VS, Keyframes are selected for the frames with low reconstruction errors [24]. The authors present a motion-state-adaptive VS based on a spatiotemporal analysis. spatio-temporal

slices are utilized to track object state changes, and keyframes are formed from the object state changes between the frames by forming an attention curve (Spatio-temporal-slices-curve-based) STS-CS. Keyframes are refined using a visually distinguishable attention curve to track motion state changes in the STS-CS model. The attention curve measured by state changes outperforms other computationally intensive saliency maps [54]. Proposes a VS system with shot boundary detection, shot view classification, and mid-level visual features, followed by the construction of a Bayesian network. HMM is used to convert the videos into semantic units; the semantic units form the features of VS. In Sports video, the basic unit of information is composed of play-break sequences, as there is an action and a stop, which covers most of the meaningful features. The Farlie-Gumbel-Morgenstern family of Copulas is used to form the Bayes net, providing a joint distribution. The structures for the network are estimated using the Chow–Liu tree.

A multi-feature-based framework approach utilising multimedia, with sentence highlights as a text feature, segment-level features from the video, and acoustic highlights. The authors have tested their framework on lecture notes (subtitles) and acoustic features from MOOCs. The audio signals are processed as voiced and unvoiced, which gives an acoustic emphasis to the features. Other characteristics of sound, such as pitch, syllable duration, and Pause rate, are also assigned an importance score to extract the essential parts for VS. The factors for the type of slide include T-SU (text slide), NT-SU (no Text, just illustration), and HT-SU (a mixture of both prior types, T-SU and NT-SU). A score to identify the critical relationship between the sentences and the video segments is proposed [3]. A Video summarisation method called VSUKFE (video summarisation using keyframe extraction). VSUKFE utilises the correlation of RGB colour channels, colour histograms, and moment of inertia for the inter-frame difference. Keyframe extraction using the aggregation method is the primary approach in this paper. An adaptive framework is also used to compensate for varying lighting conditions. Additionally, the computational complexity has been measured and compared against the known dataset VSUMM. The results have been compared against CUA (comparison of user summary) and their method, VSUKFE.

The paper proposes an interactive visual analytics system based on action shots and the movement of objects with trajectory visualisation. The system (sViSIT) can perform spatio-temporal queries and object tracking as part of summarisation. The motion of the object is detected by taking the difference in pixel values between the foreground and background. The generalised Mahalanobis distance function is used to track objects by deploying a graph-based approach. Layers of frames  $L_1, \dots, L_n$  are modelled as nodes, with edges pointing to the relationships between the layers. The Mahalanobis distance weight connects the nodes. The graph is mined with a Depth-first search to identify the object tracking path. The space-time cube offers an option to view the summarisation in a cube with time as the height dimension. The authors discuss the static vs. clustering challenge based on the work of High-Density Peaks Search (HDPS). They utilise a video representation based on High-

Density Peaks Search (VRHDPS) as a clustering algorithm. Keyframe extraction is done via the SVD technique [33]. The removal of useless frames is done by considering black and shot boundaries. Video frame representation is performed by identifying keyframes, along with SIFT descriptors, and then clustering the descriptors. Identifying representative frames by cluster similarity measures from the VS [45].

Features like shot detection, static features, and DWT (Discrete Wavelet Transform) are utilised. These features behave differently in fast and slow-moving videos. Combining these two results in a trade-off in VS. A priority fusion method is used to determine which of the above techniques to use, taking into account whether they are too slow or too fast. The static methods utilise LMS colour space values derived from RGB values to obtain the keyframes. The DWT method utilises high-resolution values derived from the shots. The frames are formed into blocks, and RGB data is converted to LMS colour space values for calculating colour opponency (RG, BY), contrast, and intensity. A centre-surround difference from neighbouring blocks and Visual attention is used to calculate the visual attention index for each shot. The authors have also compared the work against static and fast-moving videos [21]. To tackle users subjectivity and interest the authors frame a query focused VS approach, for user preferences the text queries are used in VS process, rather than visual features or temporal overlaps, the work focuses on semantic information only humans can perceive, Dense tag with caption generation is used to obtain the semantic text features. A binary mapping for the presence and absence of concepts in a dictionary forms a semantic network. The semantic information is marked as 1's for the correct semantic concepts, which can be queried upon. Vine search options covering various concepts are also discussed. The memory network incorporates the attention-based VS technique with question and answer options. Evaluation and benchmarking are conducted using ROUGE-SU4 compared to other methods, primarily for text semantic processing. [38]. Presents an unsupervised Adaptive key frame extraction using a clustering approach. A visual content feature in HSV space is utilised to construct the colour histogram. A similarity measure for two adjacent frames in HSV space and a threshold function to control the cluster density. The centroid of the cluster serves as the keyframe for summarisation, as it contains the highest visual content.

Exploits a subspace selection method to achieve a rapid and accurate video event classification method. The approach can identify the intra-modal geometry of samples inside a matching class and within classes for event detection. Modality Mixture Projections (MMP) is a dimensionality reduction technique that utilises linear discriminant-preserving geometry projections. Kernel MMP (KMPP) is proposed to have a unified subspace for identifying events. With visual and acoustic feature vectors, a different type of multi-data is used, which can efficiently predict events modelled in the MMP and KMPP subspace, which are formed to be in a unified subspace. The event detection happens on the unified subspace formed. The basic idea of novel subspace selection is a crucial area to explore for generating video descriptors via a learning process that provides

more discriminative abilities than the global inter-class points when compared to other classifiers, such as SVM.

Presents a static video summarisation framework by extracting the keyframe using the fuzzy c-means clustering (FCM) technique. Presents a static video summarization framework by extracting the keyframe using the fuzzy c-means cluster technique. The proposed clustering method differs from other summarisation approaches, which utilise colour histograms, luminance, and motion vectors using k-means or hierarchical clustering. The membership selection technique from FCM clustering gives a better representation of the keyframe in the cluster, and then the redundant frames are eliminated by Euclidean distance measures, also showing a low error rate when compared to other cluster-based approaches [2].

Introduces the VS technique for intelligent surveillance, utilising a human attention model based on human perception. A Human Visual system (HVS) with a salient region helps to identify the keyframes in the video. The identified keyframes with similar static contents are formed into bucket bins. Aggregated Channel Feature (ACF) detection is used to extract features of moving objects. An object tube with the various cost functions is measured; the cost function (saliency measures) is used as a criterion to identify the keyframes for summarization. The authors discuss a retrieval framework by searching the summarized keyframes [43].

The author presents a framework with Time-Sync Comments (TSC) that online users provide. TSC is used to identify events in the video in a time-series approach. The events in the video form a set, 'e', which contains 'p' events that interest the user. The event is a combination of (topic, tc(time slice), density), the topic is from the topic modeling, tc is the time range of the exciting video, density is the number of users that have given TSC, then arriving at the commonality between these scores (topic, tc, density) to form a relevancy matrix  $M \times N$  users and comments in time slice [23]. Presents a multi-modal feature saliency VS method, where both audio-visual signals and visual semantic text are used in the VS framework. Features such as dominant modulation energy and amplitude, as well as the use of frequency filters, are well-tested against the movie data sets. Visual features, energy formation, Energy minimization, visual saliency, and text analysis technique like Audio Segmentation using Forced Alignment, Syntactic Text Tagging, Text saliency, are the methods used to form the model for VS. Different fusion techniques like low, high-level fusion feature vectors, saliency curves are also tested for VS effectiveness. The authors compare the effectiveness of VS with various normalisation features, including global, scene-level, and shot-level normalisation, among others.

The paper presents a framework with a recommendation engine built using Bayesian and LDA techniques, bridging the intention gap and the semantic gap. The framework's recommendation is a model-based approach. Videos based on low-level and high-level semantic features are displayed in the browser history approach for VS. A hypergraph construction-based approach for visual similarity and user preference is modelled in the graph. Weights are included based on the number of clip similarities. The paper presents a framework for VS with real-time video summarization based

on segments of videos and a tree-based rank method. Unwanted frames are eliminated by measuring luminance, sharpness, and uniformity in the frames using a threshold measure. Frame-level features, such as aesthetic, number of faces, and interestingness, are captured for each frame. FHOG is used for salient face detection, and the Aesthetic score is calculated using an XGBoost classifier. Segment features are ranked using the knapsack method, assigning a rank of 0 or 1 to each feature. The feature importance is computed using a Decision tree using XGBoost methods, and a maximization function for all the features to arrive at the final VS. The paper compares the author's process and LSTM methods for a benchmark against the known datasets like SumMe [42, 50]. The authors present a VS for mobile devices based on the Visual Attention Model (VAM) and FAST Directional Motion Intensity Estimation (FDMIE). VAM considers the action of the Human Vision System (HVS) in forming features. VAM aims to capture the features that are a combination of video's low-level and semantic features from the HVS. FDMIE captures the intensity differences in frames. The frame sampling is performed to remove visually redundant frames, utilising the static attention module and motion attention module. The features used in static attention are contrast and luminance, and then a saliency map is generated for both colour and grayscale videos. The Static attention curve and the Motion attention curve obtained by FDMIE are merged to form a proper attention curve. The peak attention curve guidance forms the final VS [9]. The authors propose a keyframe selection method based on features such as video quality, colourfulness, brightness, edge distortion, hue, contrast, and others. A weighted approach is used to select the keyframe, measuring the mean and standard deviation scores of the weighted features. Duplicate frame elimination uses the Euclidean distance measure with a threshold [39].

### 2.3. VS Based on Genetic Algorithm Approach

Discusses a genetic segmentation algorithm for video segmentation and summarisation. The video is sub-sampled at half a second per frame, and the similarity is measured using colour histograms. Standard deviation and mean are used to select the keyframe. Genetic string encoding is performed on the input video, with segment boundaries corresponding to the chromosomes. Similarity adjacency is used as a fitness function; crossover is performed only at segment boundaries. In online VS techniques, GA variation helps to achieve better convergence, as it can capture features across the entire video. The paper proposes a Genetic algorithm (GA) approach for VS of soccer videos having audio signals and relative frame distance. Typical GA operations like crossover and mutation are performed to identify the right generation of VS. The features extracted from audio signals include average sound energy, average sound peak, and response time, which indicate the occurrence of a critical event in sports video. The subsampling process utilises both audio and video colour histogram features. The sub-sampled video forms the input to GA. Binary encoding is well-suited for frame-equidistant, extended scenes and short shots. The proposed approach yields better

keyframes compared to other methods for identifying keyframes [47].

## 2.4. VS Based on Determinantal Point Process

In this paper, the authors firmly believe that supervised learning is required, as summarisation and event detection occur based on user needs. They have also treated the summarisation problem as a supervised subset selection problem and used a sequential determinantal point process (seqDPP) to achieve better results compared to DPP. seqDPP-derived metrics are far closer to the quality of the human-perceived metric for VS. Their approach to summarisation consists of three steps: humanly created ground truth summaries, a new probabilistic model, seqDPP, and a novel way for subset selection from standard visual and contextual features. Proposes a new DPP-LSTM model that combines vsLSTM and DPP (determinantal point process) to capture long-range dependencies and pair-wise frame-level repulsiveness. Due to memory constraints in LSTM models, capturing long-term relationships in frames is impossible, resulting in a high recall and low precision. DPPs, on the other hand, produce high precision and low recall. Therefore, combining LSTM and DPP complements each other. VsLSTM predicts the frame-level important likelihood scores for inclusion in the VS. The proposed supervised learning approach uses square loss and F-score as the stopping criteria for training the network [50].

## 2.5. VS Based on SVD Approach

The author proposes a technique for VS by the use of SVD, PCA, and a binary blob detection algorithm to summarise the input video, to do a content-based summary retrieval. The background keyframe and object keyframes are stored separately for easy retrieval and access. For VS, the interesting points are captured using PCA techniques. The frames with the most significant eigenvalues are used for VS [33]. The paper presents a VS method utilising Singular Value Decomposition (SVD). From the given input video, a set of frames is selected, and an RGB colour histogram is derived. Each video frame is constructed into 3x3 blocks, forming a 3D histogram that contains all the spatial information. Nine histograms form a feature vector. From the input video, a feature frame matrix 'A' (usually sparse) is obtained, and then SVD is performed to obtain the matrix 'V', where the column vector represents the frame in feature space. Clustering the shots in a top 'n' frames gives a measure of the content for VS. The paper utilises SVD properties from the temporal and spatial characteristics of the input video, along with a desired length and granularity, to achieve VS. The colour histogram and colour distribution are mapped into the SVD features matrix. The right singular matrix of the SVD is formed with the matching colour histogram. Finally, a singularity threshold value is used to form VS. The number of frames is reduced by shot boundary detection, and redundant frames are removed using a Euclidean distance measure. The shot boundary detection gives static clusters with the frames in which the shortest shot boundary frames are ignored in the VS [12].

## 2.6. VS Based on Graph Theory Approach

The authors model the VS challenge as a graph colouring problem by introducing L(q)-colouring in the graph. The

length of the VS is tunable. A tube arrangement is discussed to facilitate understanding of the collision in the frames. A global energy function is introduced to measure the cost of the maximum activity, similar to the background, chronological, and tube collision. A rearrangement of tubes to avoid collision with the previous tubes and inter-tube collision. Graph colouring is used to solve the tube collision challenges. M-nodes are used to track the tube index, and s-nodes record the collision times of other M-nodes. The S nodes and M-nodes are connected if a collision exists. The S-nodes are depicted with colour differences and timestamps in the graph for VS. Threefold contributions are tube collision identification, tube rearrangement, and a greedy algorithm solution to the graph L(q)-colouring problem [16]. The authors introduce a framework based on the Hypergraph Dominant Set (HDS) framework. The proposed Multi-Video Summarisation (MVS) task identifies a dominant set within a hypergraph. Query-Dependent Maximum Marginal Relevance (QD-MMR) determines the keyframe by adjusting conciseness and adapting to the query. Conciseness helps to avoid redundant frames, and query adaptation helps to maintain the relevance of the MVS. A Graph-based Topical Closeness (GTC) process brings more meaning and relevance to the generated summary. The paper also highlights the use of HDS to address the MVS by tackling the common visual discovery problem, as well as the application of QD-MMR to find keyframes by minimising variation between query adaptive criteria and achieving minimal redundancy [20]. The paper proposes a new framework to tackle camera movement and illumination changes, as the human eye can easily follow object movements for VS. As smooth pursuit provides the location of the object in the video frame, a motion saliency score is used to identify smooth pursuit by a distance score. Spatial and motion saliency maps are used to remove unwanted frames and help form better keyframe selection. The VS is modelled as smooth pursuit identification based on Gaze Data, Spatial Saliency Prediction, motion saliency estimation, and saliency score generation, followed by the VS. In Spatial Saliency mapping, the overlapping of frames is considered redundant by a saliency score, and the movement of the smooth pursuit frame is in descending order of saliency with a summarisation ratio [35]. The paper discusses the formation of a storyline or summarisation from the given video by modelling the causal relationship between actions as edges and objects as nodes in an AND-OR graph. Unlike graphical models like HMM and DBN, the proposed AND/OR graph can model dynamic changes in node structure. In contrast, graphical models are fixed in terms of node and relationship structure and do not grow dynamically. A video with weak captions and an Expectation Maximisation (EM) recursive search for a query, incorporating actions and relationships with conditional probability, forms the storyline. The AND-OR graph also forms an alternate storyline as an OR condition for the query match. Different matches to the queries are compared using recursive matching patterns to create a storyline, with a threshold for appearance based on the entropy of distribution, structural complexity, and spatiotemporal relationships formed in the AND-OR graph.

## 2.7. VS Based on ANN Unsupervised Approach

proposes a fully unsupervised deep summarisation network (DSN) trained using Reinforcement learning and a novel reward function that takes into account both diversity and representativeness. Diversity takes into account different frames, and representativeness measures the similarity between adjacent frames.

The reward functions in reinforcement learning facilitate the learning process by introducing novelty to the solution. The paper also compares the results with supervised, unsupervised, augmented, and transfer learning setups for VS [59]. The authors propose an approach to VS using an ANN with an attention module instead of an LSTM/GRU combined with an Encoder and Decoder. A regressor module outputs the importance score for the frame by replacing the LSTM encoder and decoder. The multiplicative attention helps in parallelising the operation, similar to matrix multiplication. The regression score originates from the final single layer, which includes dropouts for regularisation. The regression scores are used to form the key shots, and then these key shots are constrained to 15% of the actual video length by applying a knapsack method. Local attention can be used in the case of longer videos as compared to global attention, as the sequential operation will involve more computation in the adversarial network [8]. The authors discuss a multi-task spatio-temporal ANN and attention framework called See, Understand, and Summarise it Network (SUSiNet) for VS. As an extension to the ResNet, the paper proposes a Deeply Supervised Attention Module (DSAM) for saliency and a summarisation module. The activation function at the last layer gives the probability of including the video segment in the summarisation. Asynchronous Stochastic Gradient Descent (SGD) is discussed as having task-specific training (Kokkinos, 2017). A weighted variant of Binary cross-entropy (BCE) loss function is used to identify the video segments for VS, as there are only a few annotated video frames used in the training [22]. Uses an LSTM encoder-decoder framework to learn the video representation. The Encoder is used to form the video representation, and the Decoder is used to reconstruct the target. This is the natural extension of the LSTM decoder, where one decodes the input representation and the other predicts the feature. Images, patches, and high-level percepts are the critical features used [40]. Proposes an unsupervised setting to solve VS trained on the videos that are available for a specific domain. A recurrent neural network trained with an auto-encoder on edited videos for identifying sub-events in the video, having a shrinking exponential loss function to mitigate noisy inputs. The autoencoder receives feedback from a training set of videos to train on. A temporal segmentation performed on the collected videos in a particular domain falls in the number of frames ranging between [48]. C3D features are considered from the CNN network, followed by a bidirectional LSTM autoencoder network for frame reconstruction. For video highlight and outlier detection, a threshold of small or large reconstruction error is considered [46]. The paper proposes an iterative quantisation (ITQ) method to remove duplicate frames using a distance measure. Keyframe selection is formed by SNIP and CNN features, which are generated through rank pooling to capture the temporal evolution of the frame-to-frame appearance. SNIP is then applied to select the key frames from a time series of frames. Temporal data is again ranked to get the final summarization, which involves finding peaks as well as the local minima [36]. Uses a

generative adversarial framework with a summarizer and a discriminator using LSTM, where the LSTM discriminator's purpose is to identify the gap between the input video and the reconstructed video for VS. Mean threshold importance score criteria in the sLSTM give a subset of frames. Then, the encoder encodes to form the deep features for eLSTM, sLSTM outputs the importance score, and the decoder dLSTM takes the encoder input and reconstructs the frame. For VS, the reconstruction error threshold forms a criterion. A new reconstruction loss function with a hidden representation in cLSTM (classifier LSTM) is used, rather than the Euclidean. Diversity Regularization selects frames with high visual diversity, and Effective regularization is allowed for the number of keyframe selection in the VS, maximizing the visual diversity in keyframe selection [26].

## 2.8. VS based on ANN Supervised Approach

The authors have examined the use of global temporal features in the video clip, augmenting the action features from local temporal features, and then described the action features for the subset of frames. Authors have compared the results with those of YouTube2Text also on DVS tracks. They also proposed a novel 3-D CNN-RNN encoder-decoder to capture local spatiotemporal features and state the importance of local temporal features for video description through experiments. This video description can be combined with other VS work to have a text summary of the video [48]. A framework for VS, called Deep Side Semantic Embedding (DSSE) model, is implemented using the available semantic information about the input video. Two unimodal autoencoders are used to input the video and side semantic features, and a latent subspace learning process occurs between the two features. The loss functions for semantic relevance and video reconstruction are measured to find the effectiveness of VS. A latent subspace distance is measured to see the significance of the VS, a small distance score gives a high relevance for VS. A graphical representation of both semantic and video features are constructed into a bipartite graph for the user query and response. The number of clicks weights gives a high score for the query match in the bipartite graph. [49]. The paper discusses a framework with Spatiotemporal and high-level features of shot segmentation derived from motion magnitude. Local phase quantisation features, such as the Local Phase Quantisation features of the Three Orthogonal Planes (LPQ-TOP) in the XY, XT, and YT spaces, as well as space-time volume features, sparse auto-encoder (SAE) high-dimensional input vectors, a Chebyshev distance measure between frames, and a mean threshold of the distance scores, are used to arrive at the keyshots. SAE takes the 768-D LPQ-TOP features for key shot threshold scores. Shot segmentation is considered to remove most of the redundant frames by motion vectors that are formed as grouped shots. SAE is used for feature reconstruction between shots to compare frames for VS [29]. The authors have implemented work on the Semantic Attribute-assisted VS framework (SASUM). Visual and semantic features are built for the given video. NLP is performed on the text corpus, which forms semantic attributes of the input visual content. A joint approach of visual and semantic features is used to create the essential parts of the video, which are taken into consideration in the proposed framework. Deep learning

techniques are used to learn the Semantic features. The clustering approach is used to form a group of clusters, and then a temporal order is established to obtain VS [41]. A fine-grained, unsupervised VS with an online motion autoencoder vsLSTM is explored. Extracting key motions from the participating objects and learning and summarising them online is the key difference in this paper.

The framework works by super-segmenting into multiple objects, then following the motion clips by involving an Auto Encoder (online motion AE). Involving semantic information in this framework facilitates downstream processing, such as object retrieval. The main contribution in this framework involves key object-motion based VS, unsupervised online dictionary learning, Orange Ville-benchmark, object and frame-level VS [53]. The paper exploits the hierarchical structure present in the video, as shots are composed of frames, and frames contain objects. Presents an adaptive video summarization technique that looks into shot segmentation and VS into a Hierarchical Structure-Adaptive RNN(HSA-RNN). A sliding bidirectional LSTM detects shots, the detected shot boundaries are then passed to the top layers, followed by a second layer capturing temporal dependencies and then a shot level probability are assigned for VS. LSTM captures the bidirectional in formation of the frames by having a sliding window operation; this helps to evade from having irrelevant global information [56].

The authors discuss a discriminative loss function for VS, a loss function which measures the predicted summary to the original video in the form of a Retrospective Encoder. The Retrospective Encoders serve as a metric learner, enabling learning without human intervention. The seq2seq models are used to generate the summary. Also, re-seq2seq (retrospective sequence-to-sequence) gives a similarity measure between the machine-generated summary and the input video in an abstract semantic space. The output of the encoders is a vector embedding containing the semantic meaning of the original video. The output of the decoder forms the frames for the video summary; then the retrospective encoder infers a vector embedding of the summary. The model also uses similar and far-away embeddings. The standard loss function (regression) is used to measure the summary at the local frame or shot level. In their approach, human annotation can be excluded, making this a semi-supervised learning approach. The primary goal here is to perform embedding matching between the video summary and the input video, utilising the retrospective encoder to capture the input video embedding. The regression loss function to measure embedding outputs in the decoder will result in the VS as shots [51].

Proposes a framework to learn VS from unpaired data, a set of raw video (V), and a summary video set (S). No correspondence is required between these (V) and (S) videos. To generate a summary with (V) and (S), a mapping function is learned,  $F: V \rightarrow S$ , and a summary video  $F(V)$  is generated, similar to the distribution of S. In a traditional supervised setup, the summary and the raw video are paired. It's easier to have unpaired videos. The paper proposes a model for keyframe selection as a mapping function. For the summary, a summary discriminator network is used. The summary discriminator provides discrimination between the real summary in learning and the summary generated with adversarial, reconstruction, and diversity loss functions [37].

### 2.9. VS Based on ANN Supervised and Unsupervised Approach

The paper presents a VS technique for energy-efficient and resource-constrained devices by utilising the CNN method. The approach consists of three steps: Initially, a shot segmentation is performed using CNN deep features, followed by the computation of the entropy of each frame, and then a keyframe selection for summarisation. Shot segmentation is a primary feature extraction technique that utilises memorability prediction and Entropy score calculation to identify Keyframe selection. Object motion identified via the attention curve method also helps to identify keyframes for VS. A deep feature comparison using an Euclidean distance measure is then employed to form the essential features. The approach is tested in video surveillance [31]. In this paper, the authors have taken an approach of combining video features segmented using a deep neural network (DNN) and deep semantic features extracted using RNNs. They have demonstrated how features of semantic space can add value to VS. The segmentation and cluster approach using DNN is used for VS as a new path to VS. The results compared with baseline and SumMe show the value added to the summarization technique [19].

The framework includes an external augmented memory to record visual information and predicts the importance of shot-level scores in a video shot based on global frame details. Shot-level feature representation is significantly more memory-efficient, enabling storage on external memory. The two main features incorporated are the pooled information of objects, scenes, and shot feature representation. The ANN used here identifies the shot-level importance score to add the shots to VS. The paper also demonstrates the global attention modelling having a good understanding among datasets and is well suited for noisy videos [10].

### 2.10. VS Based on Multi-View and Egocentric Approach

Discusses on multi-view video summarization (MVS) based on query aware Sparse coding (QUASC) approach, which gives keyframes as the summary from multiple retrieved videos, In QUASC both the candidate query frame and the search returned multiple videos are provided to a reconstruction framework, and the frames with high importance score are selected. The least-squares reconstruction error (LSRE) is used as the objective function and is solved by the Coordinate Descent Method (CDM). The similarity score is then calculated between the candidate and the returned web images. The Event-Keyframe Presentation (EKP) structure, formed from both video and text features, has an edge weight based on text similarity, thus creating a Multi-Graph Fusion (MGF). A graph cut algorithm is used to separate videos into multiple categories for VS [18]. The authors present a VS approach for video data from egocentric or “wearable” devices. The authors propose a storyboard summary approach to the data, utilising a date-based summary that tracks people and objects. Their methods predict essential events in the egocentric video by identifying the candidate regions. Egocentric, Object, and Region features are used to indicate the importance of areas in the frame. These temporal features are then used to

identify the key people and objects, ultimately generating a storyboard that encompasses all these key individuals and events. The authors have collected a multi-view, egocentric dataset by carefully annotating videos for summarisation. Consensus analysis is performed on the dataset collection using the F1-measure and selection ratio. For the supervised training,

The oracle summary is followed by a greedy approach using a Determinantal Point Process (DPP) for shot selection. In an egocentric, multi-view approach, a multi-DPP passes each view's data to a CNN and Bi-LSTM for spatial and temporal feature learning, respectively. The MLP layer acts as a classification of opinions; the classified views are processed further with the Multi-DPP measure to obtain a VS [5]. The author proposes MVS as a framework for interview and intra-view correlations, which is achieved through a two-tier approach of online and cloud tiers. The online tier utilises shot segments to reduce redundant frames, which are then transmitted to the cloud for summary generation. The MVS involves three main steps: pre-processing, feature selection, and post-processing to remove redundant frames. The authors introduce a target object-based shot segmentation method primarily designed to include humans and vehicles. A lookup table comprising segmented shots from various views is constructed promptly, and the interview correlation is computed using this table. Unlike other methods that use low-level features to generate a summary, in this work, a CNN, LSTM-based DB-LSTM is proposed to bring out the informativeness from the sequence of frames with a probability score; higher entropy score frames are selected for the summary [17].

### 2.11. VS Based on Text/NLP

The authors propose new methods to extract text summaries from long videos by annotating them. The paper also presents strategies for splitting the video into super-frame segments, ranking each segment based on image quality, cinematography rules, and end-user preferences. Proposes techniques to produce text summaries of the video by having control over variable-length video summaries. The authors also proposed techniques to identify super frame segmentation using key elements in the frames, such as boundary, Attention, Contrast, sharpness, colourfulness, facial Impact, and Fusion. For keyframe identification after the super-frame cuts, optical flow estimation is performed to capture significant changes between frames. ROUGE scores.

### III. CONCLUSIONS

In this paper, we have extensively surveyed the literature related to VS and MVS. We delve into the latest techniques and process pipelines for VS. This literature survey will benefit researchers in accurately choosing between techniques in VS, such as single, multi-view, egocentric, and application-specific methods. The current research on VS focuses on deep learning-based supervised and unsupervised techniques. The VS Framework's applicability to various challenges in information retrieval, summarisation, anomaly detection, and others is highlighted. VS is a subjective approach to various domain-specific needs. The Techniques that are surveyed in this paper help to identify the key features and methods for VS. Some of the primary focuses on

VS/MVS are as follows: ANN network to perform multiple tasks for VS [22], real-time application support for anomaly detection and storage reduction, and query-based retrieval along with VS in a multiview framework [5, 17]. VS with subdivided task for object identification and then summarizing the input video frames by identifying the high entropy frames [17, 22] is also gaining at tension. Recent advancement in the sparse-land CSC approach also gives new frontiers for research in VS by enhancing the approaches taken in [44, 57].

### DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article.

### REFERENCES

- Open video project.
- E. ASADI AND N. M. CHARKARI, Video summarization using fuzzy c-means clustering, in Electrical Engineering (ICEE), 2012 20th Iranian Conference on, IEEE, 2012, pp. 690–694. [CrossRef]
- X. CHE, H. YANG, AND C. MEINEL, Automatic online lecture highlighting based on multimedia analysis, IEEE Transactions on Learning Technologies, 11 (2018), pp. 27–40. [CrossRef]
- S. E. F. DE AVILA, A. P. B. LOPES, A. DA LUZ JR, AND A. DE ALBU-QUERQUE ARAÚJO, Vsum: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters, 32 (2011), pp. 56–68. [CrossRef]
- M. ELFEKI, A. SHARGHI, S. KARANAM, Z. WU, AND A. BORJI, Multi-view egocentric video summarization, arXiv preprint arXiv:1812.00108, (2018).
- E. ELHAMIFAR AND M. C. D. P. KALUZA, Online summarization via submodular and convex optimization., in CVPR, 2017, pp. 1818–1826. [CrossRef]
- Z. ELKHATTABI, Y. TABII, AND A. BENKADDOUR, Video summarization: techniques and applications, World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9 (2015), pp. 928–933.
- J. FAJTL, H. S. SOKEH, V. ARGYRIOU, D. MONEKOSSO, AND P. REMAGNINO, Summarizing videos with attention, arXiv preprint arXiv:1812.01969, (2018).
- H. FAROUK, K. ELDAHSHAN, AND A. A. E. ABOZEID, Effective and efficient video summarization approach for mobile devices, International Journal of Interactive Mobile Technologies (iJIM), 10 (2016), pp. 19–26. [CrossRef]
- L. FENG, Z. LI, Z. KUANG, AND W. ZHANG, Extractive video summarizer with memory augmented neural networks, in 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 976–983. [CrossRef]
- Y. FU, Y. GUO, Y. ZHU, F. LIU, C. SONG, AND Z.-H. ZHOU, Multi-view video summarization, IEEE Transactions on Multimedia, 12 (2010), pp. 717–729. [CrossRef]
- Y. GONG AND X. LIU, Video summarization and retrieval using singular value decomposition, Multimedia Systems, 9 (2003), pp. 157–168. [CrossRef]
- M. GYGLI, H. GRABNER, AND L. VAN GOOL, Video summarization by learning

Published By:

Blue Eyes Intelligence Engineering

and Sciences Publication (BEIESP)

© Copyright: All rights reserved.

# A Review on Key Features and Novel Methods for Video Summarization

- submodular mixtures of objectives, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3090–3098. [CrossRef]
- 14 A. HANJALIC AND H. ZHANG, An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis, IEEE Transactions on Circuits and Systems for Video Technology, 9 (1999), pp. 1280–1289. [CrossRef]
- 15 R. Hannane, A. Elboushaki, and K. Afdel, "Mskvs: Adaptive mean shift-based keyframe extraction for video summarisation and a new objective verification approach," Journal of Visual Communication and Image Representation (2018). [CrossRef]
- 16 Y. HE, C. GAO, N. SANG, Z. QU, AND J. HAN, Graph colouring-based surveillance video synopsis, Neurocomputing, 225 (2017), pp. 64–79. [CrossRef]
- 17 T. HUSSAIN, K. MUHAMMAD, A. ULLAH, Z. CAO, S. W. BAIK, AND V. H. C. DE ALBUQUERQUE, Cloud-assisted multiview video summarisation using CNN and bidirectional LSTM, IEEE Transactions on Industrial Informatics, 16 (2019), pp. 77–86. [CrossRef]
- 18 Z. JI, Y. MA, Y. PANG, AND X. LI, Query-aware sparse coding for web multi-video summarization, Information Sciences, (2018).
- 19 Z. JI, K. XIONG, Y. PANG, AND X. LI, Video summarization with attention-based encoder-decoder networks, arXiv preprint arXiv:1708.09545, (2017).
- 20 Z. JI, Y. ZHANG, Y. PANG, AND X. LI, Hypergraph dominant set based multi-video summarisation, Signal Processing, 148 (2018), pp. 114–123. [CrossRef]
- 21 J. KAVITHA AND P. A. J. RANI, Static and multiresolution feature extraction for video summarization, Procedia Computer Science, 47 (2015), pp. 292–300. [CrossRef]
- 22 P. KOUTRAS AND P. MARAGOS, Susinet: See, understand and summarize it, arXiv preprint arXiv:1812.00722, (2018). [CrossRef]
- 23 J. LI, Z. LIAO, C. ZHANG, AND J. WANG, Event detection on online videos using crowdsourced time-sync comment, in Cloud Computing and Big Data (CCBD), 2016 7th International Conference on, IEEE, 2016, pp. 52–57.
- 24 M. MA, S. MET, J. HOU, S. WAN, AND Z. WANG, Video summarization via temporal collaborative representation of adjacent frames, in Intelligent Signal Processing and Communication Systems (ISPACS), 2017 International Symposium on, IEEE, 2017, pp. 164–169.
- 25 I. MADEMLIS, A. TEFAS, AND I. PITAS, Summarization of human activity videos using a salient dictionary, in Image Processing (ICIP), 2017 IEEE International Conference on, IEEE, 2017, pp. 625–629. [CrossRef]
- 26 B. MAHASSENI, M. LAM, AND S. TODOROVIC, Unsupervised video summarization with adversarial lstm networks, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2017. [CrossRef]
- 27 S. MEI, G. GUAN, Z. WANG, S. WAN, M. HE, AND D. D. FENG, Video summarization via minimum sparse reconstruction, Pattern Recognition, 48 (2015), pp. 522–533. [CrossRef]
- 28 J. MENG, S. WANG, H. WANG, J. YUAN, AND Y.-P. TAN, Video summarization via multi-view representative selection, IEEE Trans. on Image Processing, (2018), pp. 2134–2145. [CrossRef]
- 29 J. MOHAN AND M. S. NAIR, Dynamic summarization of videos based on descriptors in space-time video volumes and sparse autoencoder, IEEE Access, 6 (2018), pp. 59768–59778. [CrossRef]
- 30 K. MUHAMMAD, J. AHMAD, Z. LV, P. BELLAVISTA, P. YANG, AND S. W. BAIK, Efficient deep CNN-based fire detection and localisation in video surveillance applications, IEEE Transactions on Systems, Man, and Cybernetics: Systems, (2018), pp. 1–16.
- 31 K. MUHAMMAD, T. HUSSAIN, AND S. W. BAIK, Efficient cnn based summarization of surveillance videos for resource-constrained devices, Pattern Recognition Letters, (2018).
- 32 P. MUNDUR, Y. RAO, AND Y. YESHA, Keyframe-based video summarization using delaunay clustering, International Journal on Digital Libraries, 6 (2006), pp. 219–232. [CrossRef]
- 33 A. PACKIALATHA AND A. CHANDRASEKAR, Effective video summarization using eigen based classification, Transylvanian Review, (2016).
- 34 V. PARAMANANTHAM AND D. S. SURESHKUMAR, Multi view video summarization using rnn and surf based high level moving object feature frames, International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 9 (2022). [CrossRef]
- 35 M. PAUL AND M. M. SALEHIN, Spatial and motion saliency prediction method using eye tracker data for video summarization, IEEE Transactions on Circuits and Systems for Video Technology, (2018). [CrossRef]
- 36 D. PURWANTO, Y.-T. CHEN, W.-H. FANG, AND W.-C. WU, Video summarization: How to use deep-learned features without a large-scale dataset, in 2018 9th International Conference on Awareness Science and Technology (iCAST), IEEE, 2018, pp. 220–225. [CrossRef]
- 37 M. ROCHAN AND Y. WANG, Learning video summarization using unpaired data, arXiv preprint arXiv:1805.12174, (2018). [CrossRef]
- 38 A. SHARGHI, J. S. LAUREL, AND B. GONG, Query-focused video summarization: Dataset, evaluation, and a memory network based approach, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2127–2136. [CrossRef]
- 39 M. SRINIVAS, M. M. PAI, AND R. M. PAI, An improved algorithm for video summarization—a rank based approach, Procedia Computer Science, 89 (2016), pp. 812–819. [CrossRef]
- 40 N. SRIVASTAVA, E. MANSIMOV, AND R. SALAKHUDINOV, Unsupervised learning of video representations using lstms, in International conference on machine learning, 2015, pp. 843–852.
- 41 K. SUN, J. ZHU, Z. LEI, X. HOU, Q. ZHANG, J. DUAN, AND G. QIU, Learning deep semantic attributes for user video summarization, in Multimedia and Expo (ICME), 2017 IEEE International Conference on, IEEE, 2017, pp. 643–648. [CrossRef]
- 42 W. TAYLOR AND F. Z. QURESHI, Real-time video summarization on commodity hardware, in Proceedings of the 12th International Conference on Distributed Smart Cameras, ACM, 2018, p. 16. [CrossRef]
- 43 S. S. THOMAS, S. GUPTA, AND V. K. SUBRAMANIAN, Smart surveillance based on video summarization, in IEEE Region 10 Symposium (TENSYP), 2017, IEEE, 2017, pp. 1–5. [CrossRef]
- 44 D. S. K. VINSENT PARAMANANTHAM, A real time video summarization for youtube videos and evaluation of computational algorithms for their time and storage reduction, International Journal on Recent and Innovation Trends in Computing and Communication, 6 (2018), pp. 176–186.
- 45 J. WU, S.-H. ZHONG, J. JIANG, AND Y. YANG, A novel clustering method for static video summarization, Multimedia Tools and Applications, 76 (2017), pp. 9625–9641. [CrossRef]
- 46 H. YANG, B. WANG, S. LIN, D. WIPF, M. GUO, AND B. GUO, Unsupervised extraction of video highlights via robust recurrent autoencoders, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4633–4641. [CrossRef]
- 47 X. YANG AND Z. WEI, Genetic keyframe extraction for soccer video, Procedia Engineering, 23 (2011), pp. 713–717. [CrossRef]
- 48 L. YAO, A. TORABI, K. CHO, N. BALLAS, C. PAL, H. LAROCHELLE, AND COURVILLE, Describing videos by exploiting temporal structure, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4507–4515. [CrossRef]
- 49 Y. YUAN, T. MEI, P. CUI, AND W. ZHU, Video summarization by learning deep side semantic embedding, IEEE Transactions on Circuits and Systems for Video Technology, 29 (2017), pp. 226–237. [CrossRef]
- 50 K. ZHANG, W.-L. CHAO, F. SHA, AND K. GRAUMAN, Video summarization with long short-term memory, in European conference on computer vision, Springer, 2016, pp. 766–782. [CrossRef]
- 51 K. ZHANG, K. GRAUMAN, AND F. SHA, Retrospective encoders for video summarization, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 383–399. [CrossRef]
- 52 S. ZHANG, Y. ZHU, AND A. K. ROY-CHOWDHURY, Context-aware surveillance video summarization, IEEE Trans. Image Processing, 25 (2016), pp. 5469–5478. [CrossRef]
- 53 Y. ZHANG, X. LIANG, D. ZHANG, M. TAN, AND E. P. XING, Unsupervised object-level video summarization with online motion auto-encoder, arXiv preprint arXiv:1801.00543, (2018).
- 54 Y. ZHANG, R. TAO, AND Y. WANG, Motion-state-adaptive video summarization via spatiotemporal analysis, IEEE Transactions on Circuits and Systems for Video Technology, 27 (2017), pp. 1340–1352. [CrossRef]
- 55 B. ZHAO, L. FEI-FEI, AND E. P. XING, Online detection of unusual events in videos via dynamic sparse coding, in CVPR 2011, IEEE, 2011, pp. 3313–3320. [CrossRef]
- 56 B. ZHAO, X. LI, AND X. LU, Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7405–7414. [CrossRef]
- 57 B. ZHAO AND 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. [CrossRef]
- 58 K. ZHOU, Y. QIAO, AND T. XIANG, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, arXiv preprint arXiv:1801.00054, (2017). [CrossRef]

59 Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

**AUTHOR PROFILE**



Reinforcement Learning. An amalgamation of both Industry and academia keeps my research curiosity consistently high.

Vincent Paramanatham, B.E. from Government College of Engineering, Tirunelveli, M.S. from BITS Pilani, and M.Tech by research degree from Sathyabama Deemed University, has over 18 years of experience in industry and research. His core areas of research include image processing and video summarisation. Other areas of interest include computer vision, natural language processing, and



**Dr. S. Suresh Kumar**, Principal of Swarnandha College of Engineering, Narasapur, Andhra Pradesh, India. His area of research is Smart Energy, Image Processing, Big Data, and Network Security. He has exemplary academic records. He completed his Doctoral Degree (Ph.D.) in the Faculty of Information Communication Engineering from Anna University in 2009. And he obtained two Master's Degrees in India from Premier institutions. M.Tech from the Indian Institute of Technology, Kharagpur, in 1999, M.S. from Birla Institute of Technology and Science, Pilani, in 1993, and his bachelor's degree (B.E., Computer Engineering) from Madurai Kamaraj University in 1988. He is a recognised supervisor for the Ph.D. programme at Anna University and Manonmaniam Sundaranar University. Dr. Suresh Kumar has published 92 papers in various International and National Journals, as well as 61 papers in National and International Conference proceedings. He also published six books with co-authors. He has thirty-one years of experience in the field of Engineering Education. He plays a vital role in many professional organizations. He is an active member of the IET (UK) Institution of Engineering Technology, UK. He served as Chairman of the IET Chennai Local Networks from 2014 to 2016 in India. Additionally, he was an elected council member of IET(UK) for the period 2013-2016. He is a fellow member of IE(I) and IETE. Additionally, he is a life member of CSI and ISTE, a Senior member of IACSIT, and a member of IAENG, ICSES, and ACEEE.

**Table 1: Captures the Feature, Technique, Domain of Application, Main Summarization Methods Used in the References.**

First author, year	Type of System / Application	Features Used	Processing Techniques
[25] Mademlis, I. 2017	Clustering	CSSP (Column Subset Selection), Bag-of-Features (BoF), LMoD, Trajectories, human activities detection, exploiting semantic scene content properties	Baseline clustering approach and sparse dictionary learning
[28] Meng, J. 2018	VS with multi-view representation-presentation, multi-view representative selection (feature modalities)	consensus selection, GIST CNN features	sparse dictionary selection (MSDS-CC), FISTA (Fast Iterative Shrinkage Thresholding)
[24] Ma, M. 2017	TCR (temporal collaborative representation), Static camera	visual similarity of adjacent frames	A greedy iterative algorithm
[52] Zhang, S. 2016b	CAVS (context-aware VS), Static camera	Sparsity, spatio-temporal interest point (STIP), correlation feature	sparse coding, sparse group lasso
[54] Zhang, Y. 2017	STS-CS (motion-state-adaptive), video motiondynamic maintenance	motion state changes by collinear trajectory	spatio-temporal analysis
[6] Elhamifar, E. 2017	Incremental subset selection, Traffic monitoring	Convolutional 3D,	integer binary optimization, unconstrained sub-modular optimization, online subset selection algorithms, greedy approach
[3] Che, X. 2018	Scene detection, Lecture Highlighting, Acoustic Analysis, Statistical Analysis, e-learning	acoustic features, pitch and energy	Voice and unvoiced sound classification, Acoustic Emphasis Analysis
[45] Wu, J. 2017	Static VS, Static video summary	Video representation: High density peaks search, SIFT descriptors,	peak search clustering algorithm, classical clustering
[48] Yao, L. 2015	Video description, summary-Introduction, Text summary of video	HOG, HOF, MBH	3-D CNN, RNN
[32] Mundur, P. 2006	VS, Static camera, OV storyboard	Delaunay Triangulation, Colour features,	
[21] Kavitha, J. 2015	VS, Static camera	LMS, colour contrast and intensity, Colour Opponency (RG, BY)	Static, wavelet features dominate

## A Review on Key Features and Novel Methods for Video Summarization

[38] Sharghi, A. 2017	Query-based VS, Video retrieval	semantic(dense tag, captions)	sequential determinantal point process (DPP)
[14] Hanjalic, A. 1999	Video content Analysis, human detection and tracking	color histograms	Key frame extraction, a partitional clustering
[59] Zhou, K. 2018	VS	CNN	
[2] Asadi, E. 2012	c-means clustering, games	colour histogram, Hue histogram from the HSV colour space	FCM (fuzzy c-means clustering) clustering, k-means, hierarchical games
[43] Sinnu S. 2017	Smart surveillance, summarization epitomize	Aggregated Channel Feature (ACF)	visual saliency
[8] Fajtl, J. 2018	VS	CNN	Regressor and Attention network
[22] Koutras, P. 2018	VS, Object-detection, saliency estimation	CNN	
[46] Yang, H. 2015	Video highlight, surveillance	Temporal structure of video highlight, C3D, HoG, HoF and MBH, Fisher vector	parametric highlight detection model
[49] Yuan, Y. 2017	VS	Semantic, CNN	latent subspace
[23] Li, J. 2016	event detection, movies	Time-Sync comments,	TSC Matrix, weighted word frequency
[13] Gygli, M. 2015	Video light, surveillance high	Deep features (DeCAF)	Sub-modular function, segments local feature, k-medoids,
Hypergraph framework, Bayesian technique,			
<b>First author, year</b>	<b>Type of System /Application</b>	<b>Features Used</b>	<b>Processing Techniques</b>
[20] Ji, Z. 2018b	Multi-Video, Video retrieval	textual graph, word2vec	Query-Aware Sparse Coding, MVS, Multi-Graph Fusion (MGF)
[29] Mohan, J. 2018	VS	shot segments, histogram of spatio-temporal volume, orthogonal planes (LPQ-TOP)	sparse autoencoders, Chebyshev
[42] Taylor, W. 2018		luminance, sharpness, uniformity, FHOG,	Resnet-34, FaceScrub, VGG-Face, Chinese whisper graph, knapsack, XGboost
[26] Mahasseni, B. 2017	VS	LSTM encoder deep features,	Adversarial LSTM
[9] Farouk, H. 2016	Action recognition and Human tracking	HOGs, a boosted particle filter,	Multi-class sparse classifiers, HMM
[41] Sun, K. 2017	SASUM Semantic Attribute assisted VS, surveillance	Text captions, visual, and semantic features- tures	CNN, ResNet, ImageNet, affinity matrix, cosine distance
[47] Yang, X. 2011	hierarchical VS soccer videos	audio features, colour histogram motion activity	genetic algorithm
[44] Vinsent 2018 P.	VS, Anomaly detection, surveillance, Human tracking	HOG, HOF, optical flow	PSO, GA, ADMM and ABC
[19] Ji, Z. 2017	VS	video segments	Multiplicative attention mechanism
[57] Zhao, B. 2014	Online VS	HOG, HOF	ADMM
[27] Mei, S. 2015	VS, online shopping, recommendation	CENTRIST, HSV colour space	MSR, OffMSR
[31] Muhammad, K. 2018b	VS, Video Analysis, Surveillance	Deep features of CNN, shot-selection with memorability, salient objects	image memorability and entropy, Euclidean distance
[30] Muhammad, K. 2018a	Video Surveillance	CNN deep features	SqueezeNet, AlexNet
[39] Srinivas, M. 2016	VS	Quality score, temporal attention score, HSV (component, count), Edge distribution, Contrast, Static attention (saliency value), image signature, temporal attention,	Histogram normalize,

[50] Zhang, K. 2016		Visual, penultimate layer outputs, shallow and deep features	
[53] Zhang, Y. 2018	VS	Faster-RCNN region feature, masked images,	motion-clip-based, online motion Auto-Encoder (online motion-AE)
[40] Srivastava, N. 2015	VS	Image patches, RGB percepts,	
[36] Purwanto, D. 2018	VS	CNN	peak-searching algorithm, iterative quantization, peak-searching algorithm, SNIP
[10] Feng, L. 2018	VS	CNN penultimate layer,	Shot boundary detection, Shot Segmentation
[33] Packialatha, A. 2016	VS/ Content retrieval, Video indexing	RGB features	SVD, PCA
[16] He, Y. 2017	Video Synopsis	Tubes with frame sequences model into graph	tube rearrangement, graph colouring, Potential collision graph
[56] Zhao, B. 2018	VS	Shot-detection,	Hierarchical Structure-Adaptive RNN (HSA-RNN)
[51] Zhang, K. 2018a	VS	visual feature vectors (of shots)	sequence-to-sequence learning
[37] Rochan, M. 2018	VS	Image features, pooling features	Function mapping and summary discriminator network

**Table 2: Key Processes for Video Summarization**

Key process	Reference papers
Keyframe	[2, 15, 20, 20, 21, 24–26, 35–37, 39, 43, 47, 54]
Clustering	[2, 9, 14, 20, 28, 32, 41, 45]
Sparseland	[6, 27, 44, 57]
Audio features	[3]
Text features	[48]
SVD	[12]
Shot boundary/sub-shot	[10, 31, 50]
subset selection	[13]
key objects	[33]
attentive-curve Key Frame	[31]

**Table 3: Captures the Features for VS used in the references.**

First Author, Year	Optical Flow	HOF	HOG	Colour His-to Gram	Others
[25] Mademlis, I. 2017		X			
[28] Meng, J. 2018					GIST CNN
[52] Zhang, S. 2016b		X	X		
[54] Zhang, Y. 2017					STS-CS collinear Trajectory
[6] Elhamifar, E. 2017					
[3] Che, X. 2018					
[45] Wu, J. 2017					SIFT
[48] Yao, L. 2015		X	X		MBH
[32] Mundur, P. 2006					Delaunay Triangulation
[21] Kavitha, J. 2015					LMS colour space Intensity
[38] Sharghi, A. 2017					Semantic Features dense tag captions
[14] Hanjalic, A. 1999					D-dimensional feature vector D3 bin colour histogram
[59] Zhou, K. 2018					DPP-LSTM
[2] Asadi, E. 2012					HSV colour space
[43] Sinnu S. 2017					HOOF
[8] Fajtl, J. 2018					CNN

## A Review on Key Features and Novel Methods for Video Summarization

[22] Koutras, P. 2018			X		CNN
[46] Yang, H. 2015	X	X	X		Fisher vector CNN LSTM
[49] Yuan, Y. 2017					Deep Side Semantic, CNN, Encoders and decoders
[23] Li, J. 2016					Time-Sync comments
[13] Gygli, M. 2015					deep features (DeCAF)
[18] Ji, Z. 2018					Visual features, textual features (tags)
[12] Gong, Y. 2003				X	RGB features Trajectory
[29] Mohan, J. 2018					shot segments
[42] Taylor, W. 2018			X		
[26] Mahasseni, B. 2017					Encoder Deep feature from CNN RGB network
[9] Farouk, H. 2016		X	X		HSV SMLR 5-NN DIG DOF descriptor PF BPF
[41] Sun, K. 2017			X		
[47] [44] Vinsent P. 2018	X	X	X		
[19] Ji, Z. 2017					Bi-LSTM
[57] Zhao, B. 2014		X	X		
[27] Mei, S. 2015					CENTRIST HSV colour space
[31] Muhammad, K. 2018b				X	
[30] Muhammad, K. 2018a				X	CNN deep features
[39] Srinivas, M. 2016					HSV Colour space
[50] Zhang, K. 2016			X	X	GIST Dense SIFT
[53] Zhang, Y. 2018					RCNN features
[40] Srivastava, N. 2015					Image patches, High-Level percepts.
[36] Purwanto, D. 2018					ITQ SNIP
[10] Feng, L. 2018					CNN features penultimate layer (pool5)
[33] Packialatha, A. 2016					RGB colour features
[16] He, Y. 2017					Tubes in 3D spatio-temporal
[56] Zhao, B. 2018					Shot detection features to the LSTM.
[51] Zhang, K. 2018a					visual feature vectors (of shots)
[37] Rochan, M. 2018					Visual pooling features
[20] Ji, Z. 2018					Video features
[15] Hannane, R. 2018					SIFT GFFV (Global Frame Feature Vector) Difference-of-Gaussian(DoG)
[35] Paul, M. 2018					Single saliency feature
[5] Elfeki, M. 2018					Multi-DPP

**Table 4: Empirical Results, Comparison Methods and Evaluation Techniques**

First Author, Year	Empirical results	Evaluation methods	Precision, Recall, F-Score
--------------------	-------------------	--------------------	----------------------------

[25] Mademlis, I. 2017	Column Subset Selection Problem (CSSP), clustering approach, sparsedictionary learning		
[28] Meng, J. 2018	MSDSCC	Clustering subspace learning	X
[24] Ma, M. 2017	TCR, Delaunay Clustering (DT), SOMP, MSRa, MSRm, adaptive greedy dictionary selection(AGDS)		X
[52] Zhang, S. 2016b	AC DSVS LL CAVS		X
[54] Zhang, Y. 2017		PSNR techniques to identify keyframe, Computational Efficiency Comparison, SRD Comparison, Mean Opinion Score	
[6] Elhamifar, E. 2017		Matthews correlation coefficient (MCC), segment rankings	
[3] Che, X. 2018		Precision analysis, Expected Explanation Rate, Speaking Rate and Matching Rate for text	
[45] Wu, J. 2017		KNN, VRHDPS (video representation-based high-density peaks) and HDPS (high density peaks) clustering, k-means (KVS), SC spectral clustering (SCVS), AP (APVS)	
[48] Yao, L. 2015		Local 3D CNN, Global temporal attention, BLEU METEOR CIDEr	
[32] Mundur, P. 2006	K-means clustering	Significance Factor for the frame in cluster overlap factor compression factor	
[21] Kavitha, J. 2015	DWT-VS, Static-VS, HIST-VS, DCT-VS		X
[38] Sharghi, A. 2017		SH-DPP Conditional DPP, SeqDPP, ROUGE-SU4	X
[14] Hanjalic, A. 1999	unsupervised procedure for cluster-validity analysis	Manual abstraction, Automatic abstraction, Cluster validity analysis, reliability measure	
<b>First Author, Year</b>	<b>Empirical results</b>	<b>Evaluation methods</b>	<b>Precision, Recall, F-Score</b>
[59] Zhou, K. 2018		DSN sup, D-DSN, D-DSN, R-DSN, DR-DSN, DR-DSN sup	X
[2] Asadi, E. 2012	(Fuzzy Video Summarization MethodFVSM), Delaunay Triangulation (DT)	Mean accuracy, Mean error rate	
[43] Sinnu S. 2017	NN (Nearest neighbour), KLSH (Kernel Locality Sensitivity Hashing), SH(Spectral HashingKSH(Kernel Spectral Hashing), RSH(Riemannian Spectral Hashing), DKSH(Distributed Kernel Spectral Hashing)	Information rate IR, reduction ratio RR, Top retrievals	
[8] Fajtl, J. 2018			X
[22] Koutras, P. 2018	DIEM, DFK1K, ETMD	SUSiNet (1-task) (multi), saliency score	
[46] Yang, H. 2015	robust recurrent auto-encoder (RRAE), auto-encoder (AE), One-class Support Vector Machines(OCSVM), latent ranking SVM, SVM with C3D features, LRSVM	PCA OCSVM CNN LSTM	
[49] Yuan, Y. 2017			X

## A Review on Key Features and Novel Methods for Video Summarization

[23] Li, J. 2016	Density of TSCs(time sync comments)	TF IDTF, LDA	
[13] Gygli, M. 2015		Gygli et.al Video MMR Uniformity Interestingness Representative	X
[18] Ji, Z. 2018		MSR Clustering, DSC, Video-MMR, CAA, QUASC Subjective experiments	X
[29] Mohan, J. 2018			X
[42] Taylor, W. 2018			X
[26] [9] Farouk, H. 2016	Switching Probabilistic Principal Component Analysis (SPPCA), Sparse Multinomial Logistic Regression (SMLR), boosted particle filter (BPF)	HMM SPPCA 3, SPPCA 5, SPPCA 7, SPPCA 9, SPPCA 10, SPPCA 20, SPPCA 30	
[41] Sun, K. 2017	Affinity matrix temporal constraints, cross-entropy loss, Cosine distance, Bundling Centre Clustering (BCC)	(VF 2048-d), (SF 186-d), (VSF 2234-d), (PCA-VS 442-d), (PCA-V+S 256-d + 186-d), Interestingness Submodular DPP, dppLSTM, Video MMR, Uniform sampling, WebPrior quasi	X
[47] Yang, X. 2011	average sound energy, average sound peak length, commonality precedence	GA with mutation possibility of 0.2 crossover possibility of 0.8	
[44] Vinsent P. 2018	Online dictionary update	PSO, GA, ADMM, and ABC	
[19] Ji, Z. 2017			X
[57] Zhao, B. 2014	K-means, DSVS Livelight	computational time and video Length is compared	
[27] Mei, S. 2015	Keypoint-Based Keyframe Selection (KBKS)	Automatic summaries (AS), User summaries	X
[31] Muhammad, K. 2018b			X
[30] Muhammad, K. 2018a			X
[39] Srinivas, M. 2016			X
<b>First Author, Year</b>	<b>Empirical results</b>	<b>Evaluation methods</b>	<b>Precision, Recall, F-Score</b>
[50] Zhang, K. 2016		MLP-Shot, MLP-Frame, vsLSTM, dppLSTM, Canonical Augmented Transfer	X
[53] Zhang, Y. 2018	stacked GRU, Online Motion AE, Webimage prior		X
[40] Srivastava, N. 2015		Single frame LSTM, composite LSTM	
[36] Purwanto, D. 2018	SumTransfer, SUM-GAN, SeqDPP, LSTM, MSDS-CC, LLR-SDS, and Online Motion AE	iterative quantization (ITQ), sen-positive non-linear iterative peak-clipping (SNIP)	
[10] Feng, L. 2018		MAVS MLP LSTM	X
[33] Packialatha, A. 2016			X
[16] He, Y. 2017		Frame condensation ratio (FR), Frame compact rate (CR), Overlap ratio (OR)	
[56] Zhao, B. 2018	CoSum and VTW	Sliding single LSTM, Sliding bidirectional LSTM	X
[51] Zhang, K. 2018a	VTW	re-seq2seq	
[37] Rochan, M. 2018			X

[20] Ji, Z. 2018	MVS1K CoSum GeoVid YSL		X
[15] Hannane, R. 2018	DS[1 2 3 4] Checkpoint	Mean Opinion Score (MOS) RCR	X
[5] Elfeki, M. 2018	Multi-DPP + CE		X

**Table 5: VS using ANN Techniques**

First Author, Year	CNN	LSTM, RNN	Graph	Others
[28] Meng, J. 2018	X			
[48] Yao, L. 2015	X	X		
[59] Zhou, K. 2018	X	X		DPP-LSTM, Reinforcement learning (RL)
[8] Fajtl, J. 2018	X	X		LSTM/GRU, Attention and Regressor Network
[22] Koutras, P. 2018	X			Intensely Supervised Attention Module (DSAM)
[49] Yuan, Y. 2017	X			Deep Side Semantic Embedding (DSSE)
[29] Mohan, J. 2018				sparse autoencoder (SAE)
[41] Sun, K. 2017	X	X		Google Net GAN VAE
[30] Muhammad, K. 2018a	X			Squeeze Net Alex Net
[53] Zhang, Y. 2018				RCNN
[40] Srivastava, N. 2015		X		
[36] Purwanto, D. 2018	X			
[10] Feng, L. 2018	X			
[56] Zhao, B. 2018				HSA-RNN
[51] Zhang, K. 2018a		X		
[37] Rochan, M. 2018		X		Hierarchical Structure-Adaptive RNN (HSA-RNN)
[20] Ji, Z. 2018b			X	
[35] Paul, M. 2018			X	
[5] Elfeki, M. 2018	X	X		Bi-LSTM
[34] Vinsent, P. 2022		X		Bi-LSTM, SVM

**Table 6: VS using Multiview, Egocentric Approach**

First Author, Year	multi-view	Egocentric
[18] Ji, Z. 2018a	X	
[20] Ji, Z. 2018b	X	
[5] Elfeki, M. 2018	X	X
[17] Hussain, T. 2019	X	
[11] Fu, Yanwei 2010	X	

# A Review on Key Features and Novel Methods for Video Summarization

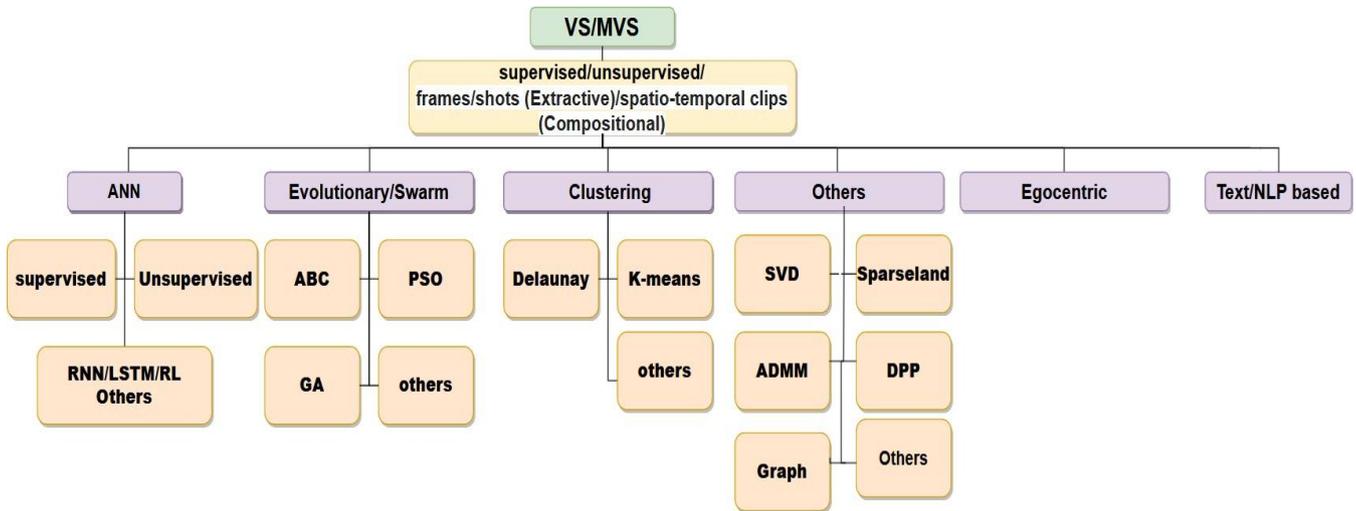


Fig. 1: Broad Classification VS/MVS approaches

- VS..... Video Summarization
- MVS ..... Multi-View Video Summarization
- ANN ..... Artificial Neural Network
- CNN ..... Convolutional Neural Network
- RNN ..... Recurrent Neural Network
- LSTM..... Long Short-Term Memory
- RL ..... Reinforcement Learning
- NLP ..... Natural Language Processing
- Vlog ..... Video Logging
- OVP..... Open Video Project [1]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.