International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-12 Issue-3, February 2023

A Review on Key Features and Novel Methods for Video Summarization

Vinsent Paramanantham, S. Suresh Kumar

Abstract: ***In this paper, we discuss techniques, algorithms, evaluation methods used in online, offline, supervised, unsupervised, multi-video and clustering methods used for Video Summarization/Multi-view Video Summarization from various references. We have studied different techniques in the literature and described the features used for generating video summaries with evaluation methods, supervised, unsupervised, algorithms and the datasets used. We have covered the survey towards the new frontier of research in computational intelligence technique like ANN (Artificial Neural Network) and other evolutionary algorithms for VS using both supervised and unsupervised methods. We highlight on single, multi-video summarization with features like video, audio, and semantic embeddings considered 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 summarization (VS) is a mechanism to produce a summary/synopsis/montage of a given video. The surveyed literature broadly classifies VS as supervised, unsupervised, static or dynamic. A more recent review on Text-based NLP methods has yielded VS better evaluation scores. Few references with Reinforcement learning for VS has also gained attention in the research [58]. However, there has to be a careful selection of Video summarization techniques based on your domain and challenge to address. This survey will aid the researchers in selecting the right techniques and approaches to tackle their challenges in VS, although the selection of VS approach and techniques are mostly subjective. Some of the prominent use cases for VS is 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 Paramanantham***,** Faculty of Computing, Sathyabama University, Chennai (Tamil Nadu), India. E-mail: 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, ORCID ID: https://orcid.org/0000-0001-9912-5927

1.1. VS Approach

VS and MVS approach have been carefully reviewed from the literature, as it varies from cluster-based approach, unsupervised, supervised, ANN, SVD, Sparse land, Graph, DPP, Evolutionary methods like genetic algorithm, MVS egocentric. The traditional approach to VS comes up with the feature extraction wherein ANN based approach uses the inherent Video features to perform VS. In approaches like clustering the temporal sequences are lost which is not required in certain domains and 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 supervised approach there are additional annotation 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 evaluation of VS results are using 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- fifice [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 and 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 summarization with either supervised or unsupervised is still a challenge. As most of the VS is a domain-specific challenge, hence a common annotation of frames captured from one domain may not be applicable for other domains.

1.4. Classification of VS Methods

There are various methods for VS in the literature.

In this paper, we have broadly classified into various approaches depicted in figure 1. The main classification of VS is either supervised or unsupervised, involving frames and shots called extractive or Spatio-temporal clips known as compositional. The other way to look at the classification is based either on keyframes or on clustering methods. The advent of ANN into VS has far exceeded other methods with various nuances to achieve higher evaluation scores. The ANN based video summarization can be subdivided based on the techniques like 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 full reference list is found in table 5. Graph-based methods also bring an inherent queriable 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 Table

In Table 1 we have covered the type of system, application, Key features, and processing techniques, this gives a highlight on 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 talks about the empirical results and evaluation methods and scores used in 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 gives a broad classification of VS work mainly as static(offline) and dynamic (online) which internally contain either supervised or unsupervised methods, followed by other classifications based on the techniques used.

II. RELATED WORKS

2.1. VS Based on Unsupervised Approach

The authors propose an algorithm for VS with human activities with a linear combination of visual words, exploiting semantic scene con tent properties. The recommended method extracts feature 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 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. Column Subset Selection Problem (CSSP) with optimization framework helps to achieve VS. The paper also com pares their results with baseline clustering approach and sparse-land dictionary learning methods [25]. The authors present a context-aware VS (CAVS) framework which uses two dictionaries, one with sparse coding with generalized sparse group lasso and a dictionary of spatiotemporal feature. The features are modelled in a correlation graph. Sparsity within the groups captures the import inter-relationships and features in Video. The Spatio-temporal feature correlation graph gives the object motion and regions correlation; the correlation graph maintains the global information. The CAVS updates both the dictionaries in an online fashion as the new sparse video comes in. The objects in VS is to identify the inter-relation between frames given a context information as event occurs [52]. The authors discuss an incremental subset selection framework that, at each time instant, uses a comparison technique between a previously selected set of representatives and the new batch of the representative with minimal overlap to form a current set. VS is modeled as an integer binary optimization problem minimizing the encoding cost via representatives regularized by the number of selected items. Real time experiments are conducted with a randomized greedy approach for sub-modular optimization. Tests are conducted to check the effectiveness of the subset selection frame work 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 shadow and background-changes are removed for further classification of the activity. Hue Movement Invariants (HMI) are used to classify the action from the body shape and movements. Human action recognition is detected by the adaptive features learned from the video given in the 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 Gray scale histogram difference, the zero-mean difference for three, five consecutive frames, three features have modeled a part of the membership function for feature extraction. The model proposed is to avoid lots of improper shots detection with camera and object motion as when compared to human labor doing the same shot correction. Delaunay triangulation (DT) is used for VS representing the video contents as color histogram, clustering is done by DT, and the size of the cluster depends on the content of the video. Significant, compression, overlap are the factors considered for VS. Automatic VS with batch processing is possible in DT as compared with K-means clustering. The number of clusters is determined automatically. A Multi-dimensional point data is given to DT clustering, in DT clustering, DT is formed for individual frames. The inter-frame similarity is measured among the data points in DT. Inter clusters and intra-clusters are formed to have the keyframe that is selected from the cluster to form VS [32]. Presents a VS technique based on motion analysis metric using optical flow algorithms, by selecting the keyframes, as in sports video motions are a critical measure, optical flow captures the motions well, and the brightness features between two frames are computed as a threshold measure. Low and high motion threshold function helps to obtain the keyframes. Optical flow video segment features are obtained by Lukas and Kanade method.

The keyframes formed are part of 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 color histogram in YUV color space is used to represent the frames 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 a travel guide, documentaries, dramas where the concepts keep repeating. A frame similarity threshold is used to consider the frame to 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 followed for multi-video summarization. Proposes an automatic VS method by using constraint satisfaction programming (CSP) by taking various user inputs like desired length for summarization and other video features as solvers input.

The summarized video output is a solvers output function having constraints like Audio-visual features, user parameters and other hard optimization constraints which are to be solved by the dynamic programming solver. Constraints like the length of the video, neighborhood constraints for sports segments are desired constraints in real time sports video. Shot boundary segments are one of the key features for the solver.

The authors discuss a structured learning approach, by taking a semantic user created inputs and a raw video as paired input to form a VS output. Combined objective function formed from the input video feature and text is given to a structure learning formulation with a greedy optimization maximizing a submodular function. Then a new summary is presented with both exciting and representative of the input video. The subset selection is used to arrive at the summarization which falls under maximizing the objective function for the 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 form 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 the similar adjacent frame 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. In order to achieve this, the authors use feature extracting algorithms called the gradient (HOG and HOF) and optical flow features. The effectiveness of VS is measured across various algorithms like PSO, GA, ADMM and ABC, the parameters compared are time and storage reduction. The paper also determines the best methods against storage reduction and 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 updating the atom online in the dictionary. The summary video is generated using the video that cannot be reconstructed us ing the learned dictionary in sparse-land, high reconstruction error makes the frame included in the summarization. As an online frame work, the dictionary starts building after seeing the initial frames. The paper proposes VS as a sparse-land problem with ADMM as the optimization choice [57]. The authors discuss a Minimum sparse reconstruction (MSR) which differs from convex relaxation sparse dictionary as the sparsity is not considered directly by L2,1 norms, MSR uses true sparse constraint L0 norm, they also proposed an additional percentage of reconstruction (POR) criterion to intuitively obtaining a summary with a desired length [27]. Localization in Surveillance application 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 anomaly efficiently rather than by manual localization creation, also improves the anomaly detection accuracy [30]. In MSKVS framework to produce VS the authors introduce a novel key frame extraction, 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 the redundant frames by having a linear algorithm called adaptive mean shift algorithm (MSKVS) to select the keyframe. The key points are sample using Difference-of-Gaussian (DoG) which gives 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-regularization (MSDS-CC) for choosing the features in the view to form a representative collection for VS summarization is used. The centroid co-regularization is solved using FISTA by taking the view features into a consensus selection matrix, to create an objective function for the sparse reconstruction summarization 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 spatio temporal 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 in the motion state changes in STS-CS model. The attention curve measured by state changes outperforms other computational intensive saliency maps [54]. Proposes VS with shot boundary detection, shot view classification, mid-level visual feature 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 ba sic unit of information is formed with the play-break sequences for there is an action and stop which covers most of the meaning full features. Farlie-Gumbel-Morgenstern family of Copulas is used to form the Bayes net giving a joint distribution. The structures for the network are estimated using Chow–Liu tree.

A multi-feature based framework approach based on the multimedia having the 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 processes as voiced, and un-voiced which gives an acoustic emphasis to the features, the other characteristic of sound like a pitch, syllable duration, Pause rate is also given to an importance score to extract the essential parts for VS. The factor for the type slide includes T-SU (text slide), NT-SU (no Text just illustration), HT-SU 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 summarization method called VSUKFE (video summarization using keyframe extraction). VSUKFE uses the correlation of RGB color channels, color histogram, and moment of inertia for the inter-frame difference. Keyframe extraction using the aggregation method is the primary approach in this paper, Also used an adaptive framework used to negate lighting conditions, they have also measured the computational complexity and compared against the known dataset VSUMM, compared the results against CUA (comparison of user summary) and their method VSUKFE.

The pa per proposes an interactive visual analytic system based on action shot and movement of the object with trajectory visualization. The system (sViSIT) can perform a spatio-temporal query and object tracking as part of summarization. The motion of the object is detected taking the difference in the foreground and background pixel values. Generalized Mahalanobis distance function is used to track the objects, by deploying a graph-based approach. Layers of frames L1..., Ln are modeled as nodes, 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 provides an option to view the summarization in a cube with the height as the time domain. Discusses static VS as clustering challenge based on the works of High-density peaks search (HDPS), the authors use video representation based on high-density peaks search (VRHDPS) as a clustering algorithm. Keyframe extraction is done via the SVD technique [33]. The useless frames removal is done by considering black and shot boundaries. Video frame representation is performed by identifying the keyframes, along with SIFT descriptors, then the descriptors are clustered. Identifying representative frames by cluster similarity measures form the VS [45].

Uses features like shot detection, static features, and DWT (Discrete wavelet transform) features, these features behave differently in the fast, slow-moving video, combining these two gives a trade off in VS. A priority fusion method is used to identify which of the above too slow, fast-moving methods to use. The static methods use LMS color space values derived from RGB values to get the keyframes. DWT method uses the high-resolution values formed from the shots. The frames are formed into blocks, and RGB data is converted to LMS color space values for calculating color Opponency (RG, BY), contrast and intensity. A center-surround difference from neighboring blocks and Visual Attentions 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 semantic network is formed by binary mapping for the presence and absence of concepts in a dictionary. The semantic information is marked as 1's for the correct semantic concepts, which can be queried upon. Vine search options covering various concept is also discussed. The memory network brings the attention-based VS technique with a question and answers options. Evaluation and benchmarking are done with ROUGE-SU4 vs other methods mainly to do text semantic processing. [38].

Presents an unsupervised Adaptive key frame extraction using a clustering approach. A visual content feature in HSV space is used to build the color 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 gives the keyframe for summarization having the highest visual content.

Exploits a subspace selection method to achieve a rapid and accu rate 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) as a dimensionality reduction technique uses linear dis criminate preserving geometry projections. Kernel MMP (KMPP) is proposed to have a unified subspace to identify the events, With the visual and acoustic feature vectors a different sort of multi-data which can efficiently predict the events modeled in MPP and KMPP subspace, which are formed to be in unified subspace. The event de taction 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 gives more discriminate abilities than the global inter-class points when compared to other classifiers like SVM.

Presents a static video summarization framework by extracting the keyframe using the fuzzy c-means cluster (FCM) technique. Presents a static video summarization framework by extracting the keyframe using the fuzzy c-means cluster technique. The proposed cluster ing method is different from other summarization approaches with color histogram, luminance, and other motion vector 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 shown low error rate when compared to other cluster-based approaches [2].

Introduces VS technique for smart surveillance with the human attention model, a human perception based. 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 the moving object features. An object tube with the various cost function is measured; the cost function (saliency measures) is used as a criterion to identify the keyframes for sum marization. The authors discuss a retrieval framework by searching the summarized keyframes [43].

The author presents a framework with Time-Sync Comments (TSC) that are provided by online users. TSC is used to identify events in the video in a time-series approach. The events in the video are as 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 X N users and comments in time slice [23]. Presents a multi-modal feature saliency VS method, Both audio visual signal, visual semantic text are used is in the VS framework. Features like dominant modulation energy and amplitude and 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 normalization features like global, scene-level, shot-level, and others.

The paper presents a framework with a recommendation engine built by Bayesian and LDA techniques while bridging between 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 shown in the browser history approach for VS. A hyper-graph construction-based approach for visual similarity and user preference is modelled in the graph and weights are included from the number of clip similarity/. 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 the luminance, sharpness, and uniformity in the frames by having a threshold measure. Frame level features like aesthetic, number of faces, and interestingness for each frame are captured, and FHOG is used to do salient face detection, Aesthetic score is calculated using XGBoost classifier. Segment features are ranked using the knapsack method as 0 or 1 as a rank. The feature importance is calculated using a Decision tree using XGBoost methods, and maximization function for all the features to arrive at the final VS. The paper compares the author's method 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), FAST Directional Motion Intensity Estimation (FDMIE). VAM considers the action of the Human Vi sion System (HVS) to form the features. VAM aims to capture the features that are a combination of video's low level and semantic features from HVS. FDMIE captures the intensity differences in frames. The frame sampling is done to remove the visually redundant frames, the static attention module, motion attention module. The feature used in static attention are contrast, luminance and then a saliency map is generated for the color and grayscale video. 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 the features like the quality of the videos colorfulness, brightness, edge distortion, hue, contrast, and others. A weighted approach is followed to choose 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 summarization, the video is sub-sampled at half a second for a frame, the similarity is measured using color histograms, standard deviation and mean are used to select the keyframe. Genetic string encoding is done for the input video, and the segment boundaries form the chromosomes. Similarity adjacency is used as a fitness function; cross over is done only in segment boundaries. In online VS techniques, GA variation helps to arrive at a better convergence as this can capture features on the whole 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 from audio signals are average sound energy, average sound peak, and response time which indicates the occurrence of a critical event in sports video. The sub-sampling process uses both the audio and video color histogram features. The sub-sampled video forms the input to GA. Binary en coding is adapted for frame equidistant, long scenes, short shots feature. The proposed approach obtains better keyframes compared to other approaches to identify keyframes [47].

2.4. VS Based on Determinantal Point Process

In this paper the authors firmly believe that supervised learning is required as the summarization and event detection happens on the user need, they have also taken the summarization problem as a supervised subset selection problem and used a sequential determinantal point process (seqDPP) to prove better results compared to DPP. seqDPP derived metrics are far close to the quality of the human perceived metric for VS. Their approach to summarization is in three steps namely humanly created ground truth summaries, a new probabilistic model seqDPP, a novel way for subset selection problem from standard visual and contextual features. Proposes a new dppLSTM 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 relationships in frames is impossible yielding a high recall and low precision, DPPs on the other hand yields high precision and low recall, So combining LSTM and DPP complements each other. VsLSTM predicts the frame-level important likelihood scores for including into the VS. The proposed supervised learning approach uses square loss and F-score 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 blog detection algorithm technique to summarize the input video, to do a content-based summary retrieval. The back ground keyframe and object keyframes are stored separately for easy retrieval. For VS the interest points are captured using PCA techniques. The frames with the largest eigen values are used for VS [33]. The paper presents a VS method using Singular value decomposition (SVD). From the given input video a set of frames are selected, an RGB color histogram is derived, each video frame is constructed into 3*3 blocks forming a 3D-histogram having all the spatial information, nine histogram forms a feature vector. From the input video, a feature frame matrix 'A' (usually sparse) is obtained, then SVD is performed to obtain the matrix 'V' where the column vector rep resents the frame in feature space. Clustering the shots in a top 'n' frames gives the measure of the content for VS. The paper uses SVD properties from temporal and spatial characteristics in the input video along with a desired length and granularity for achieving VS. Color histogram, color distribution is mapped into the SVD features matrix. The right singular matrix of the SVD is formed with the matching color 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 re moved by having 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 coloring problem by introducing L(q) - coloring in the graph. The length of the VS is tunable. A tube arrangement is discussed to know the collision in the frames; A global energy function is introduced to measure the cost of the max activity, similar background, chronological, and the tube collision. A rearrangement of tubes to avoid collision with the previous tubes and inter-tube collision. Graph coloring is used to solve the tube collision challenges. M-nodes are used to track the tube index, s-nodes records the collision times of other M-nodes. The S nodes and M-nodes connected if the collision exists. The S-nodesis depicted with colors differences and timestamp in the graph for VS. Threefold contributions are tubes collision identification, tube re-arrangement, greedy algorithm solution to the graph L(q)-coloringproblem [16]. The authors introduce a framework based on a framework with Hypergraph Dominant Set (HDS). The proposed Multi-Video Summarization (MVS) task finds a dominant set in a hypergraph. Query Dependent Maximum Marginal Relevance (QD-MMR) finds the keyframe by adjusting the conciseness and query adaptation. 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 summary generated. The other highlights in the paper are HDS used to tackle the MVS by common visual discovery problem, QD-MMRis used to find the keyframe by having less variation between query adaptive criteria and minimal redundancy [20]. The paper proposes a new framework to tackle camera movement, illumination changes as the human's eye can easily follow the 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 the smooth pursuit by a distance score. Spatial and motion saliency maps are used to remove the unwanted frames and helps to form better keyframe selection. The VS modeled 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 overlapping of the frame is considered as redundant by a saliency score, movement of smooth pursuit frame is in the descending order of the saliency with a summarization ratio [35]. The paper discusses the formation of a storyline or summarization from the given video by modeling the casual relationship of actions as edges and objects as nodes in an AND-OR graph relationship. Unlike the graphical models like HMM, DBN, the proposed AND OR graph can model the structure changes in nodes dynamically, whereas the Graphical models are fixed in node and relationship structure and do not grow dynamically. A video with weak captions and Expectation maximization (EM) recursive search for a query with the actions and relationships with a conditional probability forms the storyline. The AND-OR graph also forms an alternate story-line as an OR condition for the query match. Different matches to the queries are compared with recursive matching patterns to forma story-line with a threshold for appearance. The entropy of distribution, complexity in structure, and spatiotemporal relationship formed in the AND-OR graph.

2.7. VS Based on ANN Unsupervised Approach

proposes a fully unsupervised deep summarization network (DSN) trained using Reinforcement learning and a novel reward function taking into account diversity and representativeness. Diversity takes into account of different frames and representative takes similarity measures between adjacent frames.

The reward functions negate the learning process in reinforcement learning which brings a 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 ANN by having an attention module instead of LSTM/GRU combined with 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 parallelizing the operation as matrix multiplication. The regression score comes from the final single layer having dropouts for regularization. The regression scores are used to form the key shots, and then the key shots are constrained to15% of the actual video length by passing to 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 Summarize it Network(SUSiNet) for VS. As an extension to the ResNet an attention module the paper proposes, a Deeply Supervised Attention Module(DSAM) for saliency and a summarization module. The activation function at the last layers gives the probability to include the video segment in summarization. Asynchronous Stochastic Gradient De scent (SGD) is discussed to have 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 rep resentation, The Encoder is used to form the video representation, and the Decoder is used to find the target reconstruction, the natural extension of LSTM decoder, one to decode the input representation and the other to predict 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 auto-encoder gets the input from a training set of videos to train. A temporal segmentation performed on the collected videos in a particular domain fall in the number of frames ranging between [48]. C3D features are considered from the CNN network, followed by a bi-directional LSTM auto-encoder network to do a frame reconstruction. For video high light and outlier detection a threshold of small or large reconstruction error is considered [46]. The paper proposes an iterative quantization (ITQ) method to re move duplicate frames with a distance measure. Keyframe selectionis formed by SNIP and CNN features formed by rank pooling to capture the temporal evolution of the frame-to-frame appearance; SNIPis applied again to pick the frames from a time series of frames. Temporal data is again ranked to get the final summarization, which involves finding peaks as well to find the local minima [36]. Uses a generative adversarial framework with summarizer and dis criminator using LSTM, where the LSTM discriminators 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 forms the deep features for eLSTM, sLSTM outputs the importance score, the decoder dLSTM takes encoder input and reconstructs the frame. For VS the reconstruction error threshold forms a criterion. A new reconstruction loss function with hidden representation in cLSTM (classifier LSTM) is used other 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 and augmenting the action features from local temporal features and then describing the action features for the subset of frames. Authors have compared the results with Youtube2Textalso on DVS tracks, also proposed a novel 3-D CNN-RNN encoder decoder is used to capture the local spatiotemporal features, also states the importance of local temporal features for video description by 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 with the available semantic information about the input video. Two Uni-modal auto encoders are used to input the video and side semantic features, and a latent subspace learning happens 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 find the relevance 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 click's weights gives a high score for the query match in the bipartite graph. [49]. The paper discusses a framework with Spatiotemporal and high-levelfeatures of shot segmentation coming from motion magnitude. Lo cal phase quantization features such as Local Phase Quantization features of the Three Orthogonal Planes (LPQ-TOP) XY, XT, YTspace-time volume features, sparse-auto-encoder (SAE) high dimensional input vectors, Chebyshev distance measure between frames, and 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 frame by motion vectors that are formed as grouped shots. SAEis used for feature reconstruction between shots to compare frames for VS [29]. The authors have implemented work on Semantic Attribute assisted VS framework (SASUM). Visual and semantic features are built for the given video, NLP performed on the text corpus forms semantic attributes of input visual content. A joint approach of visual and semantic features to form the essential parts of the video taken into consideration in the proposed framework. Deep learning techniques are used to learn the Semantic features. The clustering approach is taken to form a group of clusters; then a temporal order is formed to obtain VS [41]. A fine-grained unsupervised VS with online motion auto-encoder vsLSTM is explored. Extracting key motions from the participating objects and an online manner to learn and summarize is the key difference in this paper.

The framework works on super segmented into multiple objects, then following the motion clips by involving an Auto Encoder (online motion- AE). Involving semantic information in this framework helps downstream processing like 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 contains 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 acts as a metric learner that allows the learning without the human-created VS. 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 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 mea sure the summary to the local frame/shot level summary. In their approach, human annotation can be excluded making this a semi supervised learning approach. The primary goal here is to do an embedding matching for the video summary and the input video, adding the retrospective encoder to captures the input video embed ding. 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, set of raw video (V), and summary video set (S), No correspondence is required between these (V) and (S) video. To generate a summary with(V) and (S) a mapping function is learned F: V → S, 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 it uses a summary discriminator network. The summary discriminator pro vides discrimination between the real summary in learning and the summary generated with adversarial, reconstruction, diversity loss functions [37].

2.9. VS Based on ANN Supervised and Unsupervised Approach

The paper presents a VS technique in an energy-efficient and resource-constrained devices by using CNN method. The approach is in three steps: Initially, a shot segmentation using CNN deep features, computing the entropy of each frame, followed by a keyframe selection for summarization. Shot segmentation is the primary feature extraction technique done via memorability prediction and Entropy score calculation to identify Keyframe selection. Object motion identified via the attention curve method also helps to identify the keyframes for VS. Deep feature comparison using Euclidean distance measure is followed to form the essential features, the approach is tested in video surveillance [31]. In this paper, authors have taken an approach of combining video features segmented using the deep neural network (DNN) and deep semantic features extracted using RNN's. 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 add to summarization technique [19].

The framework includes an external augmented memory to record the visual information and predicts the importance of shot-level scores of a video shot based on the global details of the frames, shot level feature representation is much more memory efficient which enables to store on external memory. The two main features incorporated are pooled information of objects, scenes, and shot feature representation. The ANN used here identifies the shot level importance score to add the shorts to VS. The paper also demonstrates the global attention modeling 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 given to a reconstruction framework, and the frames with high importance score are selected. Least-square reconstruction error (LSRE) is used as the objective function and is solved by Co-ordinate Descent Method (CDM), the similarity score is calculated between the candidate and returned web images. Event-Keyframe Presentation (EKP) structure formed from both video and text features has edge weight as text similarity thus forming a Multi-Graph Fusion (MGF). A graph cut algorithm is used to separate videos in 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 on a date-based summary tracking the people and object. Their methods predict essential events in the egocentric video by identifying the candidate regions. Egocentric, Object, Region features are used to predicts the importance of the areas in the frame; these temporal features are then used to determine the key people and objects, finally generating a storyboard of all these key people and events. The authors have collected a multi-view-egocentric dataset by carefully annotating the video for summarization. Consensus analysis is performed on the dataset collection using the F1-measure and selection ratio. For the supervised training,

The oracle-summary followed from is referenced with a greedy approach of Determinantal Point Process (DPP) for shot selection is followed. In an egocentric, multi-view approach a multi-DPP, each view's data is passed to CNN and Bi-LSTM for spatial and temporal feature learning respectively, the MLP layer acts as a classification of views, the classified views are processed further with Multi-DPP measure to obtain a VS [5]. The author proposes MVS as an interview and intra-view correlations the framework is achieved with a two-tier approach of online and cloud tier. The online tier uses the shot segments for reducing the redundant frames, the reduced frames are transmitted to the cloud for a summary generation. The MVS carries three main steps involving pre-processing, feature selection followed by post-processing to re move the redundant frames. The authors introduce a target object-based shot segmentation for mainly including humans and vehicles. A lookup table consisting of the segmented shots of different views in a timely order is built, the interview correlation is computed using the lookup 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 summary [17].

2.11. VS Based on Text/NLP

The authors propose new methods to bring out text summaries by annotating long videos. The paper also presents methods to split the video into super-frame segments, ranking each segment depending on image quality, cinematography rules, and end-user preferences. Proposes techniques to produce text summaries of the video by having control for variable-length video summaries. The authors also proposed techniques to identify super frame segmentation using some of the key elements in the frames like Boundary, Attention, Contrast, sharpness, colourfulness, facial Impact, and Fusing. For the Keyframe identification after the super-frame cuts optical flow estimation is done to capture major 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 the researchers to accurately choose between techniques in VS like single, multi-view egocentric, and application-specific methods. The current research on VS focuses on deep learning-based supervised and unsupervised techniques. The VS Frameworks applicability on various challenges for information retrieval, summarization, anomaly detection, and others are highlighted. VS is a subjective approach to various domain-specific needs, techniques that are surveyed in this paper helps to identify the key features and methods for VS. Some of the main focus on VS/MVS is 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 mult-view framework [5, 17]. VS with sub-divided 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 and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1 Open video project.
2 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]
3 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]
4 S. E. F. DE AVILA, A. P. B. LOPES, A. DA LUZ JR, AND A. DE ALBU- QUERQUE ARAÚJO, Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters, 32 (2011), pp. 56–68. [CrossRef]
5 M. ELFEKI, A. SHARGHI, S. KARANAM, Z. WU, AND A. BORJI, Multi-view egocentric video summarization, arXiv preprint arXiv:1812.00108, (2018).
6 E. ELHAMIFAR AND M. C. D. P. KALUZA, Online summarization via submodular and convex optimization., in CVPR, 2017, pp. 1818–1826. [CrossRef]
7 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.
8 J. FAJTL, H. S. SOKEH, V. ARGYRIOU, D. MONEKOSSO, AND P. REMAGNINO, Summarizing videos with attention, arXiv preprint arXiv:1812.01969, (2018).
9 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]
10 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]
11 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]
12 Y. GONG AND X. LIU, Video summarization and retrieval using singular value decomposition, Multimedia Systems, 9 (2003), pp. 157–168. [CrossRef]
13 M. GYGLI, H. GRABNER, AND L. VAN GOOL, Video summarization by learning 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 summarization and a new objective verifi- cation approach, Journal of Visual Communication and Image Representation, (2018). [CrossRef]
16 Y. HE, C. GAO, N. SANG, Z. QU, AND J. HAN, Graph coloring-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 summarization 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 summarization, 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 descrip- tors 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 localization 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, Inter- national 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 summa- rization: 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 learn- ing 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 (TENSYMP), 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 auto-encoders, 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 Proceed- ings 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 un- supervised 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

Vinsent Paramanantham, B.E., from Government College of Engineering; Tirunelveli, M.S. from BITS Pilani, M. Tech by research degree from Sathyabama Deemed University, he has got around 18 years of industry and research experience. His core area of research includes Image processing and Video summarization, other areas of interest include Computer Vision and Natural Language processing and Reinforcement Learning. An amalgamation of both Industry and academic keeps my research curiosity always high.

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 Doctoral Degree (Ph.D) in the faculty of Infor- mation Communication Engineering from Anna University in 2009. And he obtained two Master Degrees in India from Premier institutions. M. Tech from Indian Institute of Technology, Kharagpur in 1999, M.S from Birla Institute of Technology and Science, Pilani, in 1993, and his bachelor de- gree(B.E.,) Computer Engineering from Madurai Kamaraj University in 1988. He is a recognized Supervisor for Ph.D Programme in Anna University and Manonmaniam Sundara- nar University. Dr. Suresh Kumar has published 92 papers in various International and National Journals 61 papers in National and International Conference proceedings. He also published 6 books with co- authors. He has Thirthy one years of experience in the field of Engineering Education. He plays a vital role in many professional organizations. He is an active member in IET (UK) Insti- tution of Engineering Technology, UK. He was a Chairman for IET Chennai Local Networks during 2014-2016, India. Also, He was a elected council member of IET(UK) for the period 2013-2016. He is a fellow member in IE(I) and IETE. Also He is life member in CSI and ISTE, Senior member in IACSIT and member in 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 Selec- tion), Bag-of-Features (BoF), LMoD, Trajectories, human activi- ties detection, exploiting semantic scene content properties	Baseline clustering approach and sparse dictionary learning
[28] Meng, J. 2018	VS with multi view repre- sentation, multi-view repre-sentative selection (feature modalities)	consensus selection, GIST CNN features	sparse dictionary selection (MSDS-CC), FISTA (Fast Iterative Shrink-age Thresholding)
[24] Ma, M. 2017	TCR (temporal collabora- tive 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 select- tion, Traffic monitoring	Convolutional 3D,	integer binary optimization, uncon- strained sub modular optimization,online subset selection algorithms greedy approach
[3] Che, X. 2018	Scene detection, Lecture Highlighting, Acous-tic Analysis, Statistical Analysis,e-learning	acoustic features, pitch and energy	Voice and unvoiced sound classifi- cation, Acoustic Emphasis Analysis
[45] Wu, J. 2017	Static VS, Static vi de o summ	video representation High density peaks search, SIFT descriptors,	peaks search clustering algorithm, classical clustering
[48] Yao, L. 2015	Video description, summa- rization, Text summary of video	HOG, HOF, MBH	3-D CNN, RNN
[32] Mundur, P. 2006	VS, Static camera, OV sto- ryboard	Delaunay Triangulation, Color fea- tures,	
[21] Kavitha, J. 2015	VS, Static camera	LMS, color contrast and intensity, Color Opponency (RG, BY)	Static,wavelet features dominate
[38] Sharghi, A. 2017	Query based VS, Video re- trieval	semantic(dense tag, captions)	sequential determinantal point pro- cess (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	color histogram, Hue histogram from HSV color space	FCM (fuzzy c-means clustering) clustering, kmeans, 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 model highlight detection
[49] Yuan, Y. 2017	VS	Semantic, CNN	latent subspace
[23] Li, J. 2016	event detection, movies	Time-Sync comments,	TSC Matrix, weighted word free-Quincy
[13] Gygli, M. 2015	Video light, surveillance high	Deep features (DeCAF)	Sub modular function, segments local feature, k-medoids,
Hypergraph framework, Bayesian technique,			

First author, year	Type of System /Application	Features Used	Processing Techniques
[20] Ji, Z. 2018b	Multi-VS, 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		luminence, sharpeness, 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	Actions recognition and, Human tracking	HOGs, an boosted particle filter,	Multi-class sparse classifiers, HMM
[41] Sun, K. 2017	SASUM Semantic Attribute assisted VS, surveillance	Text captions, visual, semantic features	CNN, ResNet, ImageNet, affinity matrix, cosine distance
[47] Yang, X. 2011	hierarchical VS, soccer videos	audio features, color histogram motion activity ,	genetic algorithm
[44] Vinsent P. 2018	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,on-line shopping, recommendation	CENTRIST, HSV color 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 (compoent, 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 coloring, 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 Process 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	Color 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 color space Intensity
[38] Sharghi, A. 2017					Semantic Features dense tag captions
[14] Hanjalic, A. 1999					D-dimensional feature vector+ D3 bin color histogram
[59] Zhou, K. 2018					DPP-LSTM
[2] Asadi, E. 2012					HSV color space
[43] Sinnu S. 2017					HOOF
[8] Fajtl, J. 2018					CNN
[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, En-coders 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 de-scriptor 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 color space
[31] Muhammad, K. 2018b				X	
[30] Muhammad, K. 2018a				X	CNN deep features
[39] Srinivas, M. 2016					HSV Color 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 color 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		

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

[28] Meng, J. 2018	MSDSCC	Clustering subspace learning	X
[24] Ma, M. 2017	TCR, Delaunay Clustering (DT), SOMP, MSRa, MSRm, adap- tive greedy dictionary selection(AGDS)		X
[52] Zhang, S. 2016b	AC DSVS LL CAVS		X
[54] Zhang, Y. 2017		PSNR techniques to identify keyframe, Computational Ef- ficiency Comparison, SRD Comparison Mean opinion score	
[6] Elhamifar, E. 2017		Matthews correlation coefficient (MCC), segment rankings	
[3] Che, X. 2018		Precision analysis, Expected Ex- planation Rate, Speaking Rate and Matching Rate for text	
[45] Wu, J. 2017		KNN, VRHDPS (video representa- tion based high density peaks) and HDPS (high density peaks) cluster- ing, 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 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 ab- straction, Cluster validity analysis, reliability measure	
First Author, Year	**Empirical results**	**Evaluation methods**	**Precision, Recall, F-Score**
[59] Zhou, K. 2018		DSN sup, D-DSN, D-DSN, R-DSN, DR-DSN, DR-DSN sup	X
[2] Asadi, E. 2012	(Fuzzy Video Summariza- tion MethodFVSM), Delaunay Triangulation (DT)	Mean accuracy, Mean error rate	
[43] Sinnu S. 2017	NN (Nearest neighbor), KLSH (Kernel Locality Sensi-tive Hashing) SH(Spectral HashingKSH(Kernel Spectral Hashing), RSH(Riemannian Spectral Hash- ing), 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
[23] Li, J. 2016	Density of TSCs(time sync com- ments)	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 Re- gression (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 contrain- sts, cross-entropy loss, Cosine dis- tance, Bundling Center Clustering (BCC)	(VF 2048-d), (SF 186-d), (VSF 2234-d), (PCA-VS 442-d), (PCA-V+S 256-d + 186-d), Interesting- ness Submodular DPP, dppLSTM, Video MMR Uniform sampling, WebPrior quasi	X
[47] Yang, X. 2011	average sound energy, average sound peak length, commonalityprecedence	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 Selec- tion (KBKS)	Automatic summaries (AS), User summaries	X
[31] Muhammad, K. 2018b			X
[30] Muhammad, K. 2018a			X
[39] Srinivas, M. 2016			X
First Author, Year	**Empirical results**	**Evaluation methods**	**Precision, Recall, F-Score**
[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- sitive 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 bidi- rectional 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			Deeply 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	

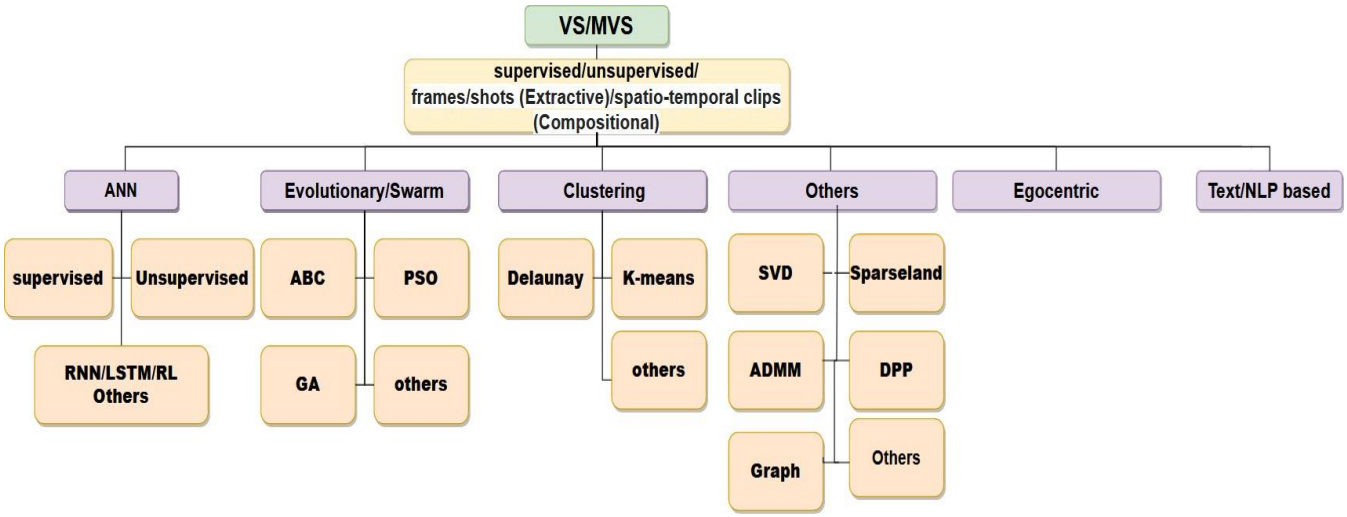

Fig. 1: Broad Classification VS/MVS approaches

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]