IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 2, April, 2023, pp. 555-565
Journal homepage: http://iieta.org/journals/ts

Statistical Evaluation of Video Summarization Models from an Empirical Perspective

Ankit Kumar[1], Saroj Kumar Pandey[1], Chetan Swarup[2*], Kamred Udham Singh[3], Teekam Singh[4], Manoj Kumar Ojha[5], Pankaj Kumar Mishra[6]

[1] Department of Computer Engineering & Applications, GLA University, Mathura 281406, India
[2] Department of Basic Science, College of Science and Theoretical Studies, Saudi Electronic University, Riyadh-Male Campus, Riyadh 13316, Saudi Arabia
[3] School of Computing, Graphic Era Hill University, Dehradun 248002, India
[4] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India
[5] Department of Computer Science and Engineering, K.R. Mangalam University, Gurugram 122103, India
[6] Department of Electronics and Communication Engineering, Rungta College of Engineering and Technology, Bhilai 490024, India

Corresponding Author Email: c.swarup@seu.edu.sa

https://doi.org/10.18280/ts.400214

Received: 27 December 2022
Accepted: 8 March 2023

Keywords:
video, summarization, machine learning, neural network, multimedia

ABSTRACT

Video summarization is the process of creating a shortened version of a longer video while retaining its essential content and meaning. It entails automatically identifying the most important parts of the video and selecting the relevant frames, shots, or scenes that best represent the original video's content. Video summarization entails complex image signal analysis and processing to extract the most important frames or shots from a video while discarding redundant or less informative ones. Several stages of analysis and processing are typically involved in the process, which may include video segmentation, feature extraction, frame selection, classification, and quality assessment. A variety of algorithms and system models are available for this task. Classification architectures such as convolutional neural networks, recurrent neural networks, and others are used to categorize video frames as redundant or non-redundant. This article provides a categorization and analysis of video summarization methodologies, with a focus on methods from the real-time video summarizing (RVS) domain. The current study will aid in laying the groundwork for future research and investigating potential research avenues by combining key research findings and data for quick reference. Video summarization has been shown to be useful in a variety of real-world contexts in smart cities, such as detecting anomalies in a video surveillance system. To address this issue, research studies can be conducted to evaluate and compare different video summarization algorithms in terms of their effectiveness, efficiency, and suitability for various applications. These studies can use benchmark datasets and standardized evaluation metrics to provide objective and quantitative comparisons of different algorithms. Based on the findings of these studies, researchers and multimedia system designers can make informed decisions about which algorithmic combination will work best for their application.

1. INTRODUCTION

Video processing is a complex task that involves processing a series of images or frames to extract useful information. Compared to image processing, video processing is more computationally complex due to the higher dimensionality of the input data. There are numerous applications of video processing, including surveillance, object tracking, user tracing, path monitoring, and more. In all of these applications, the event of interest occurs only for a short period, so it is crucial to capture and analyze the data effectively to identify the event accurately. To achieve this, video-capturing systems gather a large amount of data and apply application-specific classification algorithms to identify the relevant events. These algorithms may involve various techniques such as object detection, motion analysis, and machine learning-based classification. The output of video processing systems can be used in a variety of ways, such as identifying security threats in surveillance applications or tracking user behavior for marketing purposes. It is an important field with numerous applications, and continued research and development in this area can lead to significant advancements in fields such as artificial intelligence, computer vision, and robotics. This system can generate large amounts of data, which can make it computationally impossible to extract specific events of interest in real time. Video summarization architectures have been introduced to address this challenge. It involves reducing the size of a video while still retaining the important information. This can be achieved by identifying redundancies and outliers in the video frames and removing them. The result is a summarized version of the video that contains only the most relevant information. These architectures use various mathematical operations to identify redundancies and outliers. These operations can include feature extraction, clustering, classification, and filtering. The choice of architecture and operations depends on the specific application and the desired

output. It has numerous applications, including in surveillance systems where it can help security personnel quickly identify potential threats, or in video-sharing platforms where it can help users quickly find the most relevant videos. The computational limitations of video processing systems make it easier to extract relevant information from large video datasets.

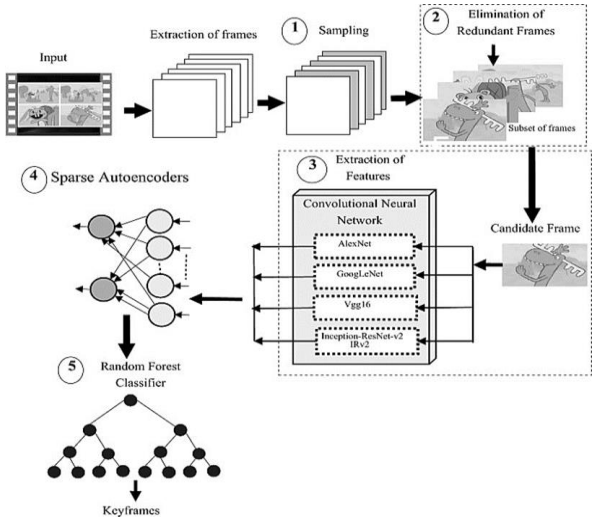

Figure 1. General architecture of video summarization

When these outliers are filtered away, the remaining video is reduced to only the relevant frames, which make up less than 10% of the original file size. Many deep learning models, as seen in Figure 1, can be employed for this purpose. To identify and eliminate anomalies in a video feed, various computer vision algorithms can be used. Object detection algorithms are frequently used to locate and follow moving targets in a video feed. These algorithms can be taught to identify particular objects or motion patterns, and then applied to the task of discarding frames that don't contain those objects or patterns.

Another strategy involves the use of motion detection algorithms to spot shifts in the video feed that could point to the presence of significant events or actions. These algorithms help to filter out unnecessary frames while highlighting those that contain significant motion or activity. These algorithms can learn from massive labeled video frame datasets to spot regularities and outliers. It is a crucial task in many applications to eliminate anomalous elements from video streams, which necessitates the use of cutting-edge computer vision and deep learning methods. The effectiveness and speed of video analysis and processing can be enhanced by using these methods to eliminate unnecessary frames and zero in on relevant ones.

Various deep-learning models can be used for video summarization, including CNNs, RNNs, and their combinations. These models are effective in identifying important frames and generating high-quality summaries. It is an important technique for processing and analyzing large video datasets, and the use of deep learning models has greatly improved the efficiency and accuracy of this process. Deep learning-based approaches use neural networks to learn features that are representative of the video and then generate a summary based on those features. To evaluate the performance of these algorithms, researchers typically use metrics such as precision, recall, F1-score, and diversity. These metrics measure how well the summary captures the important information in the video, and how diverse the selected frames.

2. LITERATURE REVIEW

Concurrently applying a complicated set of signal processing procedures on the provided video sequence is necessary for the summarising process. Among these tasks, feature extraction, selection, and analysis stand out as particularly important. A very efficient approach for these activities is advised for the development of a video summarising model. Using convolutional neural networks (CNNs) to analyze histogram of gradients (HoG) and Temporal Difference Map (TDM) feature sets is proposed in Elharrouss et al. [1] as a paradigm for video summarization. The model first employs a convolutional neural network (CNN) based on VGGNet for action identification, then uses the results of this process to locate relevant frames in the video. Silhouette detection is used to isolate the regions of interest, and then temporal difference maps and HoG features are used to characterize them. Identification of summarised frames is achieved by measuring the gap between TDM and HoG values. Figure 2 depicts the system's overall architecture and features visualizations of the CNN training and testing processes.

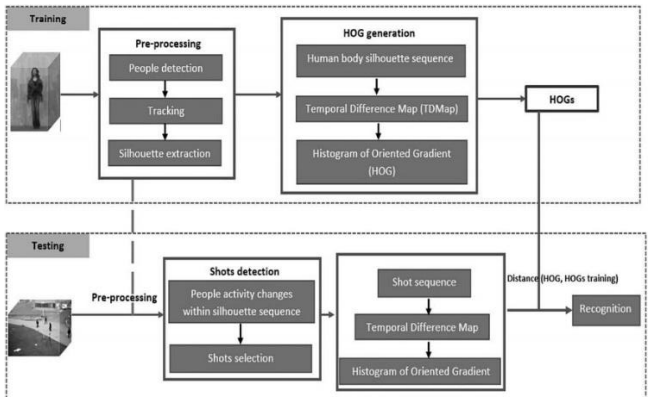

Figure 2. CNN with HoG and TDM for summarization [1]

The following formula is used to compute the sum of similarity (SS) between two frames to rank HoG features:

$$SS(i,j) = \sum_{t=1}^{T-1} \sum_{i=1}^{w} \sum_{j=1}^{h} \frac{(\sum I_t(i,j) * I_{t+1}(i,j))}{(\sum I_t(i,j))^2 * (\sum I_{t+1}(i,j))^2} \quad (1)$$

where, *I* is the image, *T* is the video's frames, *w* is the width of a single frame, and *h* is the height. Using these feature descriptors, CNN with HoG and TDM achieves an accuracy of 98% on the Kungliga Tekniska högskolan (KTH) dataset, 97% on the UCF-ARG dataset, 87% on the UT-interaction dataset, and 98% on the IXMAS dataset.

For the KTH dataset, CNN with TDM features achieves an accuracy of 99.82 percent, for the Weizmann dataset, 99.85 percent, for the UCF-ARG dataset, 98.9 percent, for the UT-interaction dataset, and 99.6 percent, for the IXMAS dataset. when comparing various models, CNN using only TDM characteristics is the superior option. It is claimed in Pan et al. [2] that energy entropy-based summarization can be used to further refine this complicated model. Assigning an entropy block to each paired frame, we may calculate their energy using the following formulas:

$$E_d(i) = \frac{\sum f(I(j,k,i) - I(j,k,i+1) > \partial)}{j * k} \quad (2)$$

where, I is it input image, (j, k) are the pixel numbers, f is the activation function, and ∂ is the difference threshold, which is decided based on input video type.

$$x_t = \frac{1}{L_W - 1} * \sum x \tag{3}$$

$$E_r = \frac{1}{|x_t - x`_t|} \tag{4}$$

$$R = \frac{\sum E_d}{\sum E_r} \tag{5}$$

$$E(t) = \propto * E_d(t) + \partial * R * E_r(t) \tag{6}$$

where, L_W is the length of the sliding window, x is the number of pixels in the frame, it is the average number of pixels in the sliding window, E_d and E_r are the ratio and division values for energy, and $E(t)$ is the energy for the frames in question. A frameset is retrieved and utilized for summarization if and only if it has a high energy level. Because of this oversimplified assessment, the approach achieves a middling 75% accuracy compared to deep learning models.

Textual descriptions can be used by summarization models to locate crucial scenes. This is a possibility for videos that include annotated texts or have a person speaking directly into the camera. Nonetheless, Otani et al. [3] provides a methodology to extract video summaries from such movies, which might be valuable for certain applications like teaching, product evaluations video analysis, etc. despite their relative scarcity in comparison to the vast array of multimedia data available online. To locate 'entity mentions,' the proposed model employs noun extraction from the text and NLP. To pinpoint shifts in time, we compare these noun entities to what appears in the film. Summary frames, which are frames from a video that have been extracted because of their significant changes, are then organized for entity-based analysis. Figure 3 displays the process's overall flow diagram, with entity analysis and video analysis presented in their respective panels. Because of this association, the model achieves an accuracy of 83%; replacing the video segmentation process with deep learning models might further increase this accuracy. Recurrent neural networks (RNNs), memory networks (MNs), convolutional long short-term memory (LSTM) networks (CNNs), gated recurrent units (GRUs), capsule networks (CNs), generative adversarial networks (GANs), etc. are just some of the methods described in Apostolidis et al. [4], which provides a summary of these models.

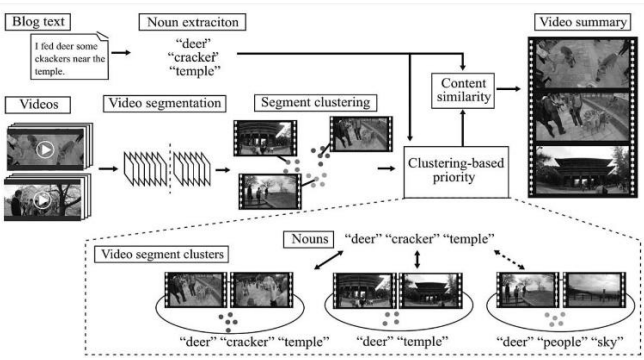

Figure 3. Text-based analysis of videos for summarization [3]

Comparing the performance of these networks with that of other summarization models reveals that CNN-based architectures and Memory Augmented Neural Networks (MANN) get superior results (97% and 94%, respectively). Using a rank-based method, as suggested in Srinivas et al. [5], can further boost these models' efficiency. Here, the ultimate score for each video frame is calculated by combining scores from many processes, such as quality, static attention, temporal attention, and representativeness. The frames are given a score, and those with the highest scores are flagged as "key-frames" and taken out for summary. To get a final video summary, this model employs a post-processing phase in which duplicate key-frames are deleted from the output. This rank-based method only achieves an accuracy of around 79%, which is unacceptable and has to be enhanced with deep learning models. Atencio et al. [6] is an example of a deep learning model that makes advantage of both visual and category variation. Using a deep convolutional network and a pre-trained word-embedding matrix, the authors of this study offer a method for extracting images that are highly distinct from one another. Redundant frames are filtered out by running them via a thresholding mechanism. Figure 4 shows the model's architecture, which relies on the combination of deep features and means activation functions to generate the final video summary. Achieving accuracies between 80% and 90%, this approach is extremely reliant on the existence of word content in the video. In Zhou et al. [7], we see another approach for character-based summarization that relies on the video's textual contents, this one employing entity identification to assess the temporal locations of these entities. These timestamps, along with the user's query, allow the video summarising process to produce only the frames that pertain to the specified entities. For entity detection in input movies, it combines a character-level LSTM model, a neural topic model, and a skip-gram model. Accurate video summarization, measured across several datasets, is possible thanks to the mapping of these items to their associated frames.

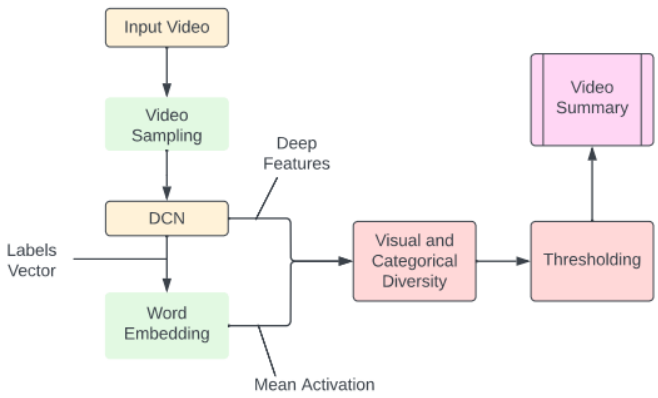

Figure 4. Word embedding extraction from video sequences to obtain final summarization [6]

Algorithms are assessed based on their performance on several datasets, and the correctness of each dataset is measured. For example, the accuracy of LSTM + Random Forest is 70%, the accuracy of the Kronecker-Product Matching Model is 81%, the accuracy of the Multi-level Factorisation Net is 97%, the accuracy of Mancs is 61%, the accuracy of Textual-based Methods is between 70% and 80%, and the accuracy of Deep-Semantic Models is between 87% and 98% for both animation and real-time datasets. Also, deep

learning models may be used to enhance the classification performance of application-specific video summarization algorithms. Coarse-grained and fine-grained refinement models based on versions of CNN are proposed for high accuracy [8]. Figure 5 depicts the model's architecture, which shows how the final summary generation process is formed from the coarse refining and fine refining models. Based on the specifics of the dataset, the suggested model may reach an accuracy of between 79% and 92%. Because of its excellent accuracy, the method may be used for any type of video summary. Using the same dataset, we find that AlexNet has 35% accuracy, GoogLeNet has 39% accuracy, SqueezeNet has 79% accuracy, and MobileNet V2 has 45% accuracy, all of which improve the proposed model's real-time deploy-ability. Hybridizing the models' designs to decrease feedback error increases their effectiveness. To improve summarization precision, Ji et al. [9] propose a hybrid model that blends GoogLeNet architecture with several BiLSTM models. To minimize the effects of regression and distribution error in the input video, this model employs a self-attention architecture that gives both additive and multiplicative attention. The following formula represents this mistake (E):

$$E_{kl} = \frac{1}{n} * \sum_{i=1}^{N} \mathrm{W} \underline{\mathrm{softmax}}(Y_i) * \log[\mathrm{softmax}(Y_i)] - \mathrm{softmax}(Y_{i+1}) * \log[\mathrm{softmax}(Y_{i+1})] \quad (7)$$

The accuracy level achieved by the proposed model, which uses a combination of deep learning models and a rank-based approach, is impressive, with a maximum accuracy level of 64.5% achieved on multiple datasets. This level of accuracy is higher than that achieved by individual models such as AlexNet, GoogLeNet with LSTM, GoogLeNet with Generative Adversarial Network, and GoogLeNet with attention-based encoder model, which achieve accuracies ranging from 52% to 60% when evaluated on similar datasets. The proposed model has the potential to outperform other existing models for video summarization tasks. However, it is important to note that the choice of dataset and evaluation metrics can significantly affect the performance of the model. Therefore, it is important to evaluate the proposed model on a variety of datasets and metrics to gain a comprehensive understanding of its effectiveness and limitations.

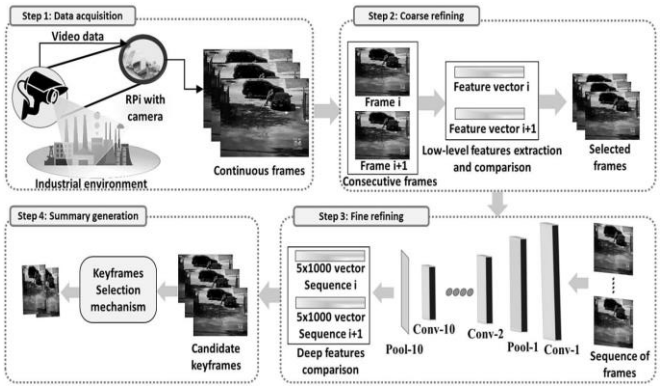

Figure 5. Coarse-grained and fine-grained deep learning models for video summarization [8]

The use of the BiLSTM model with attention-based encoder and decoder models proposed in Ji et al. [10] has shown significant improvements in accuracy for video summarization tasks, achieving an impressive accuracy of 68%. This improvement can be attributed to the high-complexity feature extraction method used in the model. Other models, such as RNN with VGGNet, have a lower accuracy of 52% when evaluated on the same dataset. Dynamic sequence deep neural network with GoogLeNet has a higher accuracy of 65% but still falls short of the accuracy achieved by the proposed BiLSTM model with attention-based encoder and decoder models. The performance of these models can vary depending on the dataset and evaluation metrics used. Therefore, it is important to evaluate these models on a variety of datasets and metrics to gain a comprehensive understanding of their effectiveness and limitations for video summarization tasks.

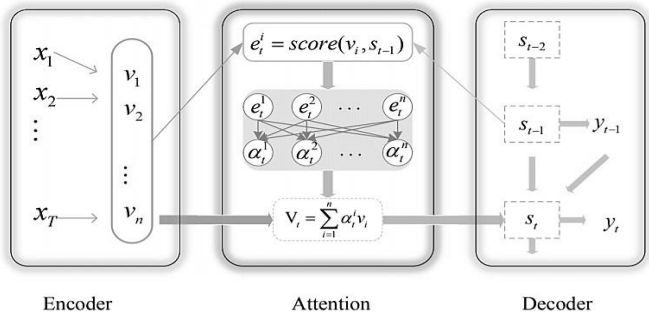

Figure 6. Encoder, attention, and decoder blocks for improved video feature extraction [10]

Figure 6 shows the interconnections between the many building elements that contribute to the system's high accuracy and reveals the role played by the auto-encoder module in performing high-level data augmentation. Patch-level video summarization using block sparse representation, as proposed in Mei et al. [11], can further enhance this accuracy, which is already high compared to previous methods. In this study, researchers extracted patches from input video frames and evaluated the feature values of each patch independently. With these feature values in hand, an Orthogonal Matching Pursuit (OMP) is used to find an exact fit between frames. Keyframes are taken from the video and matched based on the criteria. This methodology allows for real-time deployments with accuracies between 65% and 80%. In Yuan et al. [12], we design a new architecture that employs lengthy short-term memory models within a cycle-consistent Adversarial network. In this model, we combine two GANs in a way that allows them to learn from one another to improve their accuracy. This function for reducing errors may be derived from equation 8 using the frame difference metric (Dfi) as the final error reducer. A4 (8.27×11.69 inches) is the recommended page size for a printed book. While using the template, keep the existing page settings.

$$E_r = \frac{1}{2} * \left[\sum_{i=1}^{\frac{N}{2}} p_i * D_{f_i} + \sum_{i=1+\frac{N}{2}}^{N} p_i * D_{f_i} \right] \quad (8)$$

If we say that the video is summarised across N frames, then we mean that the video is compressed over N frames. The primary goal is to minimize this mistake, which would boost video summarising accuracy from 52% to 65%, surpassing the results of approaches like Sum-based GAN (52% accuracy) and deep reinforcement learning (57% accuracy). By including the TED model proposed in Huang and Wang [13],

in which a capsule net and a self-attention network are fused to pinpoint the exact moments in time when frames change, we can increase the precision of our results. Figure 7 displays capsule networks and their link with sequence self-attention networks, which are used to assess these time points via inter-frames motion curve.

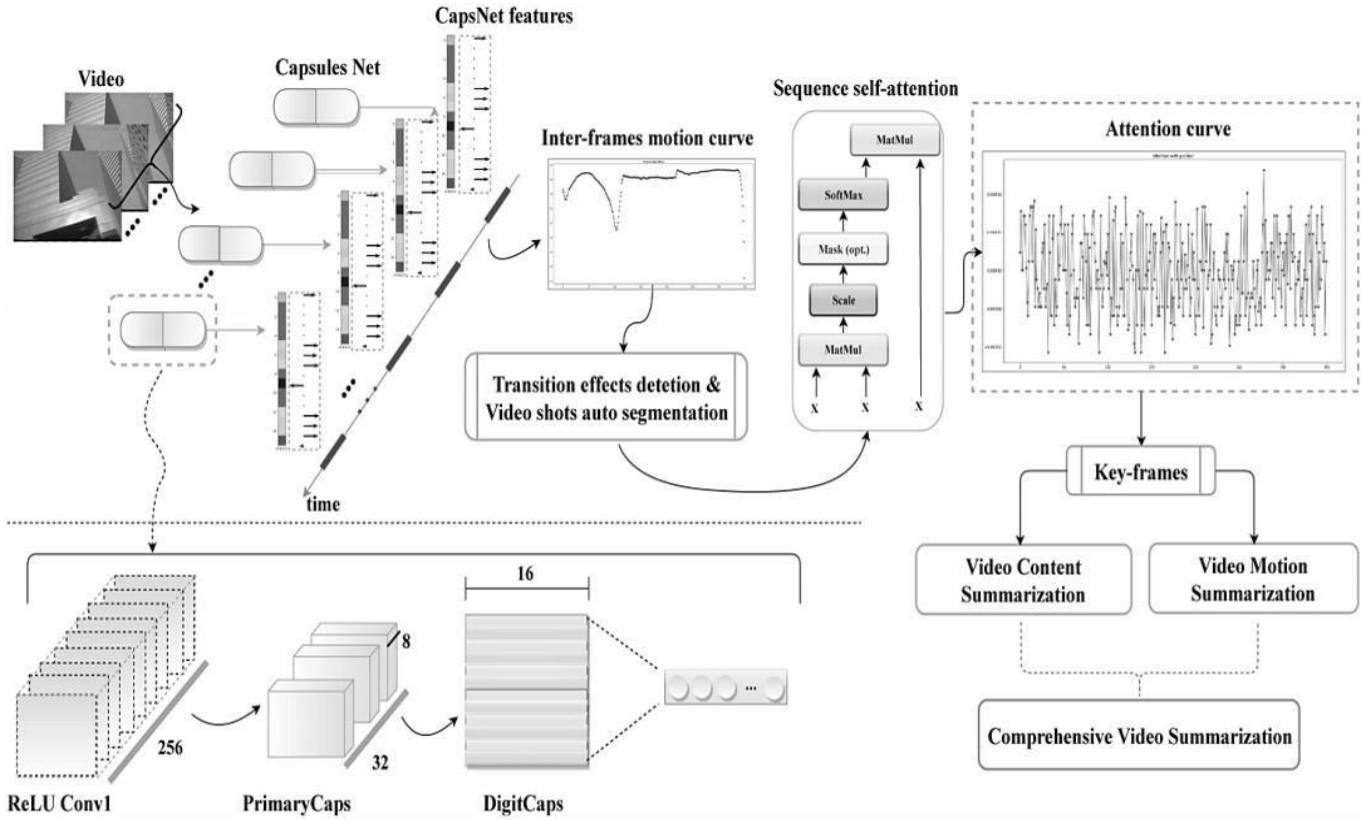

Figure 7. Capsule net with sequential attention network for video summarization [13]

Architecturally, the system is designed to collect an attention curve and assign it to a key-frame block, where the video's content and motion are examined to provide a final, all-encompassing summary. The model's 91% accuracy may be attributed to the attention model, which is defined as follows in Eq. (9):

$$\begin{aligned} A_m &= \left[\frac{(\partial - t) - \sin(\partial - t)}{\partial - \sin(\partial)}\right] * h_0 + \left[\frac{t - \sin(t)}{\partial - \sin(\partial)}\right] * h_3 + M \\ &* \left[\frac{\partial * \cos(t) - t * \cos(t) - \sin(\partial) + \sin(t)}{\partial * \cos(\partial) - \sin(\partial)}\right. \\ &- \left[\frac{(\partial - t) - sin(\partial - t)}{\partial - sin(\partial)}\right] * h_0 \Bigg] * h_1 + M \\ &* \left[\frac{sin(\partial - t) - t * cos(\partial) - sin(\partial)}{\partial * cos(\partial) - sin(\partial)} - \left[\frac{t - sin(t)}{\partial - sin(\partial)}\right] * h_3\right] * h_2 \end{aligned} \tag{9}$$

where, h represents the H-Bézier curve's coefficients, t represents the time interval, " represents the difference between frames, and M represents the dynamic change variable specified by the following Eq. (10):

$$M = \frac{(\cos\ \partial + 1) * (\sin\ \partial + 1)}{(\partial * \cos(\partial)) + (\partial - \sin(\partial))} \tag{10}$$

The proposed model described earlier is highly efficient for a wide variety of datasets, outperforming other models such as shot clustering (44% accuracy), scene transition graph (54% accuracy), hierarchical clustering (70% accuracy), fast CNN (88% accuracy), and spatiotemporal CNN (81% accuracy) according to comparisons presented in the study of Huang and Wang [13]. Another approach using the Markov Decision Process (MDP) based on the optimization of the reward function described in Eq. (11) has also been proposed for video summarization in Lei et al. [14]. This model achieves a comparable high accuracy of 84.7% across multiple datasets.

$$E_{\text{reward}} = n^1|_{i=1}^{N} * \sum_{\emptyset} \gamma^k * I_{j-i,k-i} * R_t\left(I_{j,k}\right) \tag{11}$$

where, N represents the total number of frames, p starting error probability, I represents the frame, and R represents the reward function for the frame with dimensions j by k. Large computational delays are incurred by these models; however, by processing these films on the cloud, using the elastic video summarising technique, as described in Wang et al. [15], we may eliminate these delays and achieve a more effective summary. This technique, which is very similar to CNN models, runs the full video summarising process on high-performance cloud machines, resulting in a 25% increase in total system speed and approximately 85% accuracy of summary.

The context of the video can also be taken into account while summarising it; for example, if the video is based on a sporting event, aspects such as the playing field, the players and movement, sports-specific activities, etc. can be identified and included in the summary. Time series analysis of these actions can be utilized to enhance video summary results. Using a perceptual retrieval model based on feature optimization, Thomas et al. [16] provides a model for such context-based video retrieval, which achieves an accuracy

level of 71% on a wide set of movies. For even higher precision, it is recommended to apply powerful feature selection techniques, such as principle component analysis (PCA) and Linear Discriminant Analysis (LDA), as discussed in Raksha et al. [17]. With an accuracy of 93.7%, sufficient for real-time video summarization systems, these techniques attempt to reduce input size by extracting maximal variance features, which are subsequently used for classification using CNN. Further evaluation of system performance and deploy ability requires testing this accuracy on numerous datasets. In Jiang et al. [18], we find a discussion of a different deep learning model that employs multi-task learning with CNN to get an accuracy of 82%. In Xiao et al. [19], for example, the authors combine local attention models with a global attention model to boost classification precision. Figure 8 depicts the model's architecture, which involves splitting an input video into several frames and then clustering each of those frames into related segments.

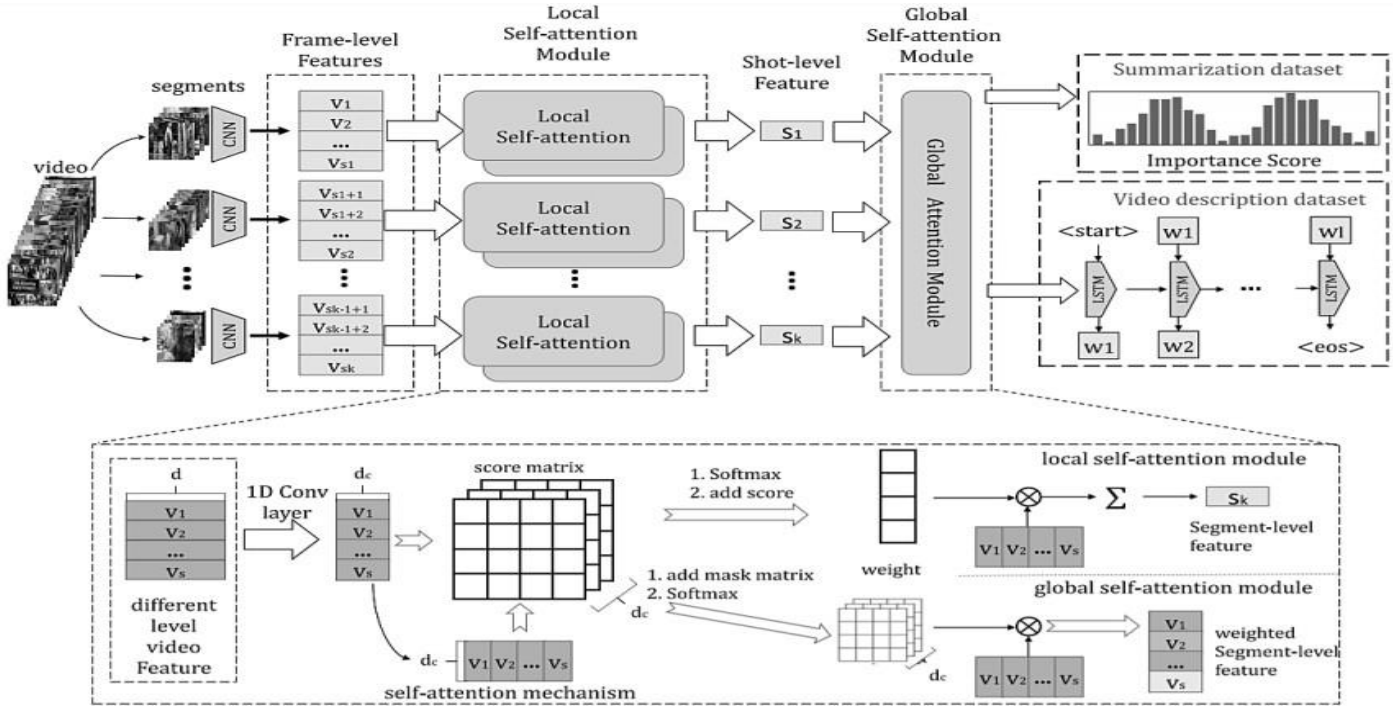

Figure 8. Multiple attention models for video summarization [19]

Table 1. Accuracy and area of application-based comparison for different models

Method	Accuracy (%)	Application
CNN with HoG [1]	98	Natural scenes
CNN with HoG [1]	97.9	Sports events
CNN with HoG [1]	97	Medical sets
CNN with TDM [1]	99.82	Natural scenes
CNN with TDM [1]	99.85	Sports events
CNN with TDM [1]	99.73	Medical sets
Entropy-based linear classifier [2]	75	General purpose
Semantic feature clustering [3]	83	Blog & Product Reviews
VGGNet CNN [4]	97	General purpose
MANN [4]	94	General purpose
Rank-based approach [5]	79	General purpose
Visual and categorical CNN [6]	85	Blog & Product Reviews
Entity identification with LSTM [7]	91	Person & entity-based videos
LSTM + Random Forest [7]	70	Person & entity-based videos
Kronecker-Product Matching Model [7]	81	Person & entity-based videos
Multi-level Factorisation Net [7]	97	Person & entity-based videos
Mancs [7]	61	Person & entity-based videos
Textual model [7]	75	Person & entity-based videos
Deep semantics [7]	94	Person & entity-based videos
Coarse- and fine-grained CNN [8]	92	Real-time videos
AlexNet [8]	38	Real-time videos
GoogLeNet [8]	39	Real-time videos
SqueezeNet [8]	79	Real-time videos
MobileNet V2 [8]	45	Real time videos
GoogLeNet with BiLSTM [9]	64.5	Real time videos
AlexNet [9]	52	Real-time videos
GoogLeNet with LSTM [9]	53	Real time videos
GoogLeNet with GAN [9]	57	Real time videos

GoogLeNet with Attention model [9]	60	Real-time videos
BiLSTM with Attention model [10]	68	Real-time videos
RNN with VGGNet [10]	52	Real-time videos
Deep neural network with GoogLeNet [10]	65	Real-time videos
Orthogonal Matching Pursuit [11]	79	General purpose videos
Cycle-consistent Adversarial network with LSTM [12]	65	General purpose videos
Sum-based GAN [12]	52	General purpose videos
Transition effect detection (TED) [13]	91	General purpose videos
Shot clustering [13]	44	General purpose videos
Scene transition graph [13]	54	General purpose videos
Hierarchical clustering [13]	70	General purpose videos
Fast CNN [13]	88	General purpose videos
Spatio-temporal CNN [13]	81	General purpose videos
Markov Decision Process [14]	84.7	General purpose videos
Cloud-based elastic CNN [15]	85	General purpose videos
Context-based video with perceptual retrieval model [16]	71	General purpose videos
PCA & LDA with CNN [17]	93.7	General purpose videos
Multi-task CNN [18]	82	General purpose videos
Local and global attention models [19]	97	Entity based summarization
Uniform sampling method [19]	75	Entity based summarization
k-Medoids [19]	64	Entity based summarization
Dictionary selection [19]	79	Entity based summarization
GAN [19]	91	Entity based summarization
DSSE [20]	81	Real-time videos
Hybrid 2D CNN, 1D CNN, and LSTM [21]	83	Entity based recognition
Visual LSTM [21]	53	Entity based recognition
Deep LSTM [21]	54	Entity based recognition
GAN [21]	59	Entity based recognition
Deep regression [21]	61	Entity based recognition
Audio-video alignment [22]	65	Audio-based systems
Nonlinear sparse dictionary selection [23]	69	Real-time videos
Graph-Based sentence summarization [24]	91.3	Blog & Product Reviews
EMD [25]	65	Real-time videos
Dual CNNs for abstractive summarization [26]	91	Abstractive summarization
Visual cues [27]	75	General purpose videos
Eye tracking with CNN [28]	92.8	General purpose videos
Spatial-temporal modeling [29]	76	General purpose videos
Pairwise deep ranking [30]	84	General purpose videos
Cutting-merging adjusting [31]	76	Real-time videos
Weighted neighborhood-based representation [32]	74	Real-time videos
CNN with BiLSTM [33]	94	Cloud-based general-purpose videos
Multicriteria Decision-Making [34]	86	General purpose videos
Correlation of modality with linear classifier [35]	53	General purpose videos
CNN with Bi-LSTM [36]	60	General purpose videos
SIFT with CNN [37]	75	General purpose videos
ABC [38]	70	Real-time videos
Keyframe with thresholding [39]	67	General purpose videos

Accuracy levels of 65, 69, and 91.3% are attained by using distinct CNN implementations in the three models. Trajectory, sparse dictionary learning, empirical mode decomposition (EMD), and other approaches are all detailed in Bora and Sharma [25]. The results of this study indicate that, for many different tasks, CNNs built on LSTMs provide the highest levels of retrieval accuracy. The method is based on trajectory analysis has achieved a high accuracy rate of 91.3% for object detection. This high accuracy can be attributed to the use of a more advanced and complex approach for analyzing video data. This method utilizes a combination of feature extraction and trajectory analysis to detect and track objects in video streams. The approach involves extracting features from video frames using a CNN, and then using these features to track the movement of objects over time. The trajectory analysis component of this approach involves analyzing the movement patterns of objects in the video stream to identify potential targets. By analyzing the trajectory of an object over time, it is possible to distinguish it from other objects and track its movement even in complex or cluttered environments. The success of this approach can also be attributed to the use of a large dataset for training the model. By training on a large and diverse dataset, the model is able to learn to recognize a wide range of object types and adapt to different environmental conditions. The combination of feature extraction and trajectory analysis, along with the use of a large and diverse training dataset, has enabled this method to achieve a high level of accuracy for object detection in video streams.

In Dilawari and Khan [26], a framework for abstractive video summarising is discussed to collect textual information from the video for summation and then utilize this textual information to pull frames from other video sources to construct an abstractive summary. This method employs a two-network setup, with the first network extracting frames with an eye toward video semantics and the second network using these same semantics to pull together condensed versions of video sequences from the video dataset. As compared to comparable state-of-the-art approaches, our extractive video summarising accuracy of 91% and our abstractive video summarization accuracy of 40% are both quite respectable results. Other video characteristics including visual cues [27], eye tracking [28], spatial-temporal modeling

[29], and highlight identification with pairwise deep ranking [30] might help increase this accuracy. Using comparable datasets, these models obtain accuracies of around 75%, 92.8%, 76%, and 84%, respectively. All of these models, which are versions of the convolutional neural network (CNN), seek to optimize feature extraction for efficient summarization.

The simpler models like cutting-merging adjusting scheme and weighted neighborhood-based representation can also be effective for video summarization, albeit with limited accuracy. However, the use of multi-view videos processed with CNNs and bi-directional LSTMs over the cloud can significantly improve accuracy by up to 94%. Furthermore, the application of Multicriteria Decision-Making can enhance performance by 10% through the use of mapping and feature selection operations. This suggests that there are various approaches to video summarization, and choosing the appropriate model depends on the specific use case and the available resources.

You can see how linear models like the ones in studies [35-37] leverage efficient feature extraction for video summarization by looking at the techniques used, such as the correlation of modality, the bidirectional LSTM, and the scale-invariant feature transform (SIFT). These algorithms only employ deep learning and bioinspired models sparingly, yet they nonetheless manage to obtain respectable results (53–60–75%). In Bhattacharjee et al. [38], the use of Artificial Bee Colony (ABC) Optimization in conjunction with a bio-inspired model for video summarization is proposed. To attain its target accuracy of 70% across datasets, this model maximizes the difference in frames between successive images. The model may be made more accurate by including color characteristics with key frame extraction, as proposed in Asim et al. [39], which uses a thresholding approach to simplify the system while increasing accuracy to 67%. This variety of models shows a broad range of accuracy, performance, and utility. So, the next part compares these algorithms in terms of field of application and approximate accuracy of summarization, simplifying the process of algorithmic selection.

3. STATISTICAL ANALYSIS

CNNs, RNNs, LSTMs, GRUs, and Bi-LSTMs are only a few of the deep learning architectures that are crucial to the reviewed models. To achieve a satisfactory level of retrieval performance, it is also common to resort to more elementary models like support vector machines (SVMs), random forests (RF), etc. Table 1 provides a statistical comparison of the models' accuracy performances; for more informed algorithm selection, the models' areas of application are also included.

Table 1 compares various video summarization models based on their accuracy and application areas. According to the table, different models are better suited to different applications and have varying levels of accuracy. Some models are intended for real-time video summarization applications, while others are better suited for offline video data analysis. Models designed for real-time summarization have lower accuracy rates, whereas models designed for offline analysis have higher accuracy rates. Aside from the differences in accuracy and application, the table emphasizes the importance of considering other factors such as computational complexity and scalability when choosing a video summarization model. Models that are highly accurate but computationally expensive may be unsuitable for large-scale applications, whereas models that are less accurate but more computationally efficient may be better. Overall, the table emphasizes the importance of selecting an appropriate model after carefully considering the specific requirements and constraints of a video summarization application. Accuracy, computational complexity, and scalability should all be considered to ensure that the chosen model is both effective and practical for the intended use case.

As can be seen from the outcomes, the studied algorithms have several potential uses. Figures 9, 10, 11, 12, 13, 14, and 15 illustrate the accuracy results of these algorithms as they were sorted into groups according to their respective applications before the algorithmic selection.

Blog & product reviews videos can be best summarized using Graph-based sentence summarization algorithm as mentioned in Zhang et al. [24].

Entity-based summarization videos can be best identified using local and global attention models asdesigned in Xiao et al. [19].

General purpose videos can be summarized using MANN-based models as mentioned in Apostolidis et al. [4], while CNN & LSTM models also perform with good accuracy.

While person & entity-based videos can be summarized using Multi-level Factorisation Net as suggested in [7], where an accuracy of 97% is achieved making it suitable for real-time deployments.

Real-time videos taken from CCTV cameras can be identified with 92% accuracy using coarse and fine-grained models as designed in Muhammad et al. [8].

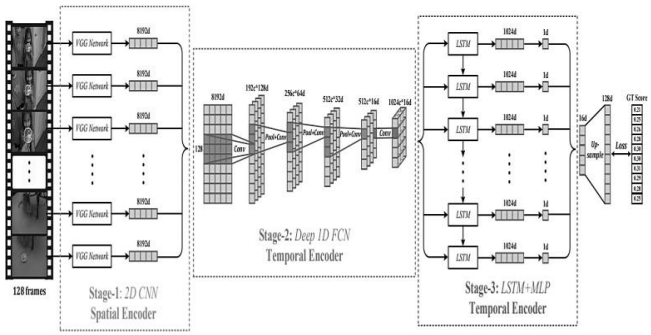

Figure 9. Combination of different CNNs for effective summarization [21]

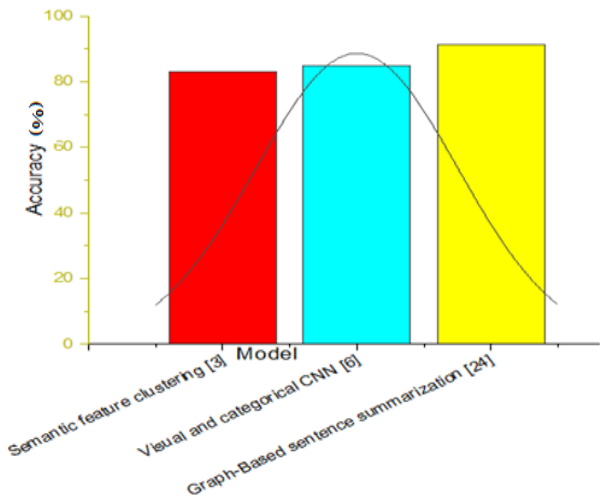

Figure 10. Algorithms used for Blog & product reviews

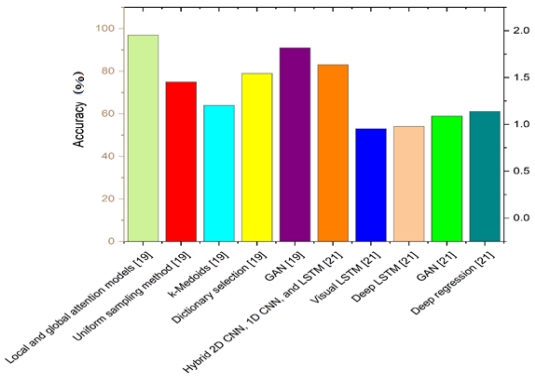

Figure 11. Performance of entity-based summarization models

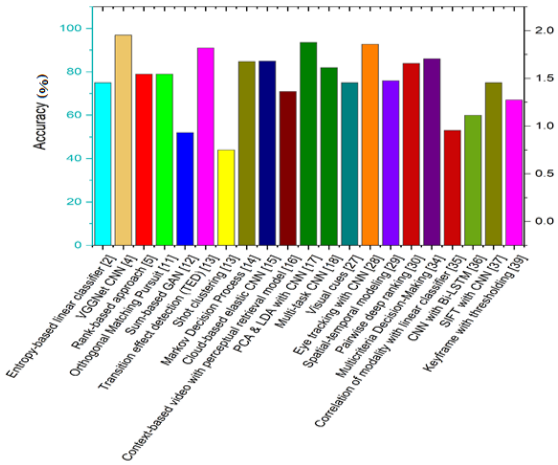

Figure 12. Performance for general-purpose videos

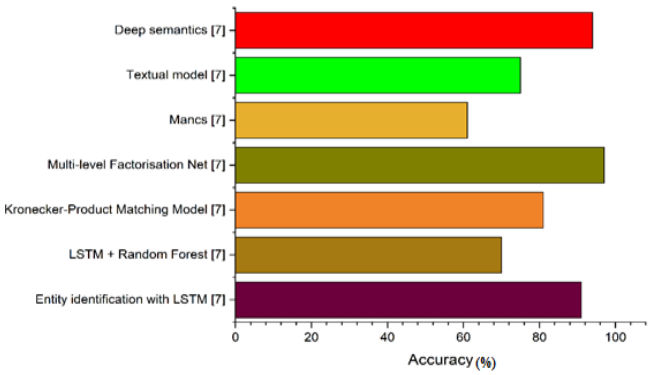

Figure 13. Performance comparison for person & entity-based videos

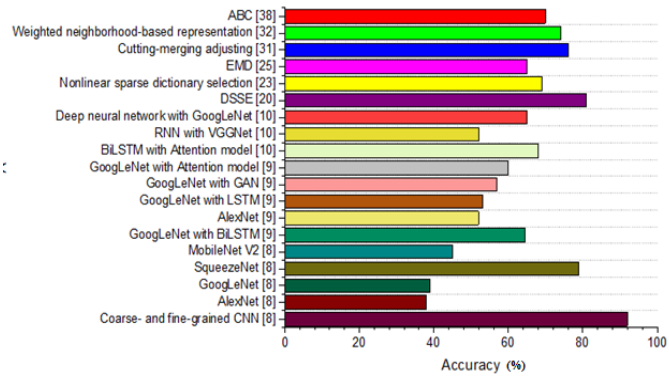

Figure 14. Performance comparison for real-time videos

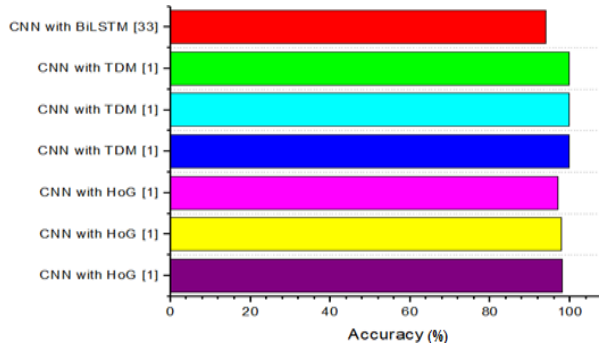

Figure 15. Performance comparison for miscellaneous applications

Then, all other random applications CNN models all have similar performance, and CNN with a TDM model, as proposed in Elharrouss et al. [1] may be utilized to classify data accurately. Scientists can use this comparison to determine which algorithms are ideal for their project. The specific use case, resources, and performance requirements all play a role in deciding which algorithm to use for video summarization. By comparing each algorithm's performance metrics and characteristics, researchers can find the best algorithms for their specific use case using the available statistical data. The reliability of the algorithm is an essential aspect to consider. Researchers can determine which algorithms are most accurate for their application by analyzing the accuracy rates reported in existing studies or datasets. Nonetheless, there are other factors that should be thought about, such as computational complexity, scalability, and real-time performance. The context in which it will be used is also crucial. Surveillance, sports analysis, and entertainment videos could all benefit from the use of distinct algorithms. Researchers can determine which algorithms are best suited for their unique use case by analyzing the domain of application information provided in prior studies. Additional considerations include computational complexity and scalability. To determine which algorithms are feasible with their current resources, researchers can compare their computational requirements, such as the number of parameters and processing time. Some algorithms may not be suited for use on a large scale, so scalability is also crucial. In general, when looking for a video summarization algorithm, it's important to consider the algorithm's performance metrics, field of use, and computational requirements. Researchers can effectively identify the best algorithms for their given application by analyzing existing statistical data and thinking about these factors.

4. CONCLUSION AND FUTURE WORK

According to the research done in the past, it is possible to conclude that CNN and its variations outperform other algorithms in terms of core accuracy when compared to different applications. Real-time video captured by CCTV cameras can be recognized using coarse and fine-grained CNN models. For all other applications, CNN with the TDM model can be used for high-accuracy classification. For example, blogs and product review videos may be best summarized using a graph-based sentence summarization algorithm. General-purpose videos can be summarized using MANN-based models. Person and entity-based videos can be

summarized using multi-level factorization nets. It is advised that researchers can employ algorithms from one application with a high degree of accuracy and test them on their application to evaluating the performance of their application. In addition, transfer learning and its variations can be used to produce more efficient video summaries.

REFERENCES

[1] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., Beghdadi, A. (2021). A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence, 51: 690-712. https://doi.org/10.1007/s10489-020-01823-z

[2] Pan, G., Zheng, Y., Zhang, R., Han, Z., Sun, D., Qu, X. (2019). A bottom-up summarization algorithm for videos in the wild. EURASIP Journal on Advances in Signal Processing, 2019: 1-11. https://doi.org/10.1186/s13634-019-0611-y

[3] Otani, M., Nakashima, Y., Sato, T., Yokoya, N. (2017). Video summarization using textual descriptions for authoring video blogs. Multimedia Tools and Applications, 76: 12097-12115. https://doi.org/10.1007/s11042-016-4061-3

[4] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I. (2021). Video summarization using deep neural networks: A survey. Proceedings of the IEEE, 109(11): 1838-1863. https://doi.org/10.1109/JPROC.2021.3117472

[5] Srinivas, M., Pai, M.M., Pai, R.M. (2016). An improved algorithm for video summarization–A rank based approach. Procedia Computer Science, 89: 812-819. https://doi.org/10.1016/j.procs.2016.06.065

[6] Atencio, P., German, S.T., Branch, J.W., Delrieux, C. (2019). Video summarisation by deep visual and categorical diversity. IET Computer Vision, 13(6): 569-577. https://doi.org/10.1049/iet- cvi.2018.5436

[7] Zhou, P., Xu, T., Yin, Z., Liu, D., Chen, E., Lv, G., Li, C. (2019). Character-oriented video summarization with visual and textual cues. IEEE Transactions on Multimedia, 22(10): 2684-2697. https://doi.org/10.1109/TMM.2019.2960594

[8] Muhammad, K., Hussain, T., Del Ser, J., Palade, V., De Albuquerque, V.H.C. (2019). DeepReS: A deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. IEEE Transactions on Industrial Informatics, 16(9): 5938-5947. https://doi.org/10.1109/TII.2019.2960536

[9] Ji, Z., Zhao, Y., Pang, Y., Li, X., Han, J. (2020). Deep attentive video summarization with distribution consistency learning. IEEE Transactions on Neural Networks and Learning Systems, 32(4): 1765-1775. https://doi.org/10.1109/TNNLS.2020.2991083

[10] Ji, Z., Xiong, K., Pang, Y., Li, X. (2019). Video summarization with attention-based encoder–decoder networks. IEEE Transactions on Circuits and Systems for Video Technology, 30(6): 1709-1717. https://doi.org/10.1109/TCSVT.2019.2904996

[11] Mei, S., Ma, M., Wan, S., Hou, J., Wang, Z., Feng, D.D. (2020). Patch based video summarization with block sparse representation. IEEE Transactions on Multimedia, 23: 732-747. https://doi.org/10.1109/TMM.2020.2987683

[12] Yuan, L., Tay, F.E.H., Li, P., Feng, J. (2019). Unsupervised video summarization with cycle-consistent adversarial LSTM networks. IEEE Transactions on Multimedia, 22(10): 2711-2722. https://doi.org/10.1109/TMM.2019.2959451

[13] Huang, C., Wang, H. (2019). A novel key-frames selection framework for comprehensive video summarization. IEEE Transactions on Circuits and Systems for Video Technology, 30(2): 577-589. https://doi.org/10.1109/TCSVT.2019.2890899

[14] Lei, J., Luan, Q., Song, X., Liu, X., Tao, D., Song, M. (2018). Action parsing-driven video summarization based on reinforcement learning. IEEE Transactions on Circuits and Systems for Video Technology, 29(7): 2126-2137. https://doi.org/10.1109/TCSVT.2018.2860797

[15] Wang, Y., Dong, Y., Guo, S., Yang, Y., Liao, X. (2019). Latency-aware adaptive video summarization for mobile edge clouds. IEEE Transactions on Multimedia, 22(5): 1193-1207. https://doi.org/10.1109/TMM.2019.2939753

[16] Thomas, S.S., Gupta, S., Subramanian, V.K. (2018). Context driven optimized perceptual video summarization and retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 29(10): 3132-3145. https://doi.org/10.1109/TCSVT.2018.2873185

[17] Raksha, H., Namitha, G., Sejal, N. (2019). Action based video summarization. TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 457-462. https://doi.org/10.1109/TENCON.2019.8929597

[18] Jiang, Y., Cui, K., Peng, B., Xu, C. (2019). Comprehensive video understanding: Video summarization with content-based video recommender design. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 1562-1569. https://doi.org/10.1109/ICCVW.2019.00195

[19] Xiao, S., Zhao, Z., Zhang, Z., Guan, Z., Cai, D. (2020). Query-biased self-attentive network for query-focused video summarization. IEEE Transactions on Image Processing, 29: 5889-5899. https://doi.org/10.1109/TIP.2020.2985868

[20] Yuan, Y., Mei, T., Cui, P., Zhu, W. (2017). Video summarization by learning deep side semantic embedding. IEEE Transactions on Circuits and Systems for Video Technology, 29(1): 226-237. https://doi.org/10.1109/TCSVT.2017.2771247

[21] Huang, S., Li, X., Zhang, Z., Wu, F., Han, J. (2018). User-ranking video summarization with multi-stage spatio–temporal representation. IEEE Transactions on Image Processing, 28(6): 2654-2664. https://doi.org/10.1109/TIP.2018.2889265

[22] Rajpoot, V., Girase, S. (2018). A study on application scenario of video summarization. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 936-943. https://doi.org/10.1109/ICECA.2018.8474699

[23] Ma, M., Mei, S., Wan, S., Wang, Z., Feng, D. (2019). Video summarization via nonlinear sparse dictionary selection. IEEE Access, 7: 11763-11774. https://doi.org/10.1109/ACCESS.2019.2891834

[24] Zhang, Z., Xu, D., Ouyang, W., Zhou, L. (2020). Dense video captioning using graph-based sentence summarization. IEEE Transactions on Multimedia, 23:

1799-1810. https://doi.org/10.1109/TMM.2020.3003592
[25] Bora, A., Sharma, S. (2018). A review on video summarization approcahes: recent advances and directions. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 601-606. https://doi.org/10.1109/ICACCCN.2018.8748574
[26] Dilawari, A., Khan, M.U.G. (2019). ASoVS: Abstractive summarization of video sequences. IEEE Access, 7: 29253-29263. https://doi.org/10.1109/ACCESS.2019.2902507
[27] Zhang, Z., Xu, D., Ouyang, W., Tan, C. (2019). Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. IEEE Transactions on Circuits and Systems for Video Technology, 30(9): 3130-3139. https://doi.org/10.1109/TCSVT.2019.2936526
[28] Paul, M., Salehin, M.M. (2018). Spatial and motion saliency prediction method using eye tracker data for video summarization. IEEE Transactions on Circuits and Systems for Video Technology, 29(6): 1856-1867. https://doi.org/10.1109/TCSVT.2018.2844780
[29] Yuan, Y., Li, H., Wang, Q. (2019). Spatiotemporal modeling for video summarization using convolutional recurrent neural network. IEEE Access, 7: 64676-64685. https://doi.org/10.1109/ACCESS.2019.2916989
[30] Sridevi, M., Kharde, M. (2020). Video summarization using highlight detection and pairwise deep ranking model. Procedia Computer Science, 167: 1839-1848. https://doi.org/10.1016/j.procs.2020.03.203
[31] Ai, X., Song, Y., Li, Z. (2018). Unsupervised video summarization based on consistent clip generation. In 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1-7. IEEE. https://doi.org/10.1109/BigMM.2018.8499188
[32] Ma, M., Mei, S., Wan, S., Wang, Z., Tsoi, A.C., Feng, D. (2018). Video summarization via weighted neighborhood based representation. In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1273-1277. https://doi.org/10.1109/ICIP.2018.8451722
[33] Hussain, T., Muhammad, K., Ullah, A., Cao, Z., Baik, S.W., de Albuquerque, V.H.C. (2019). Cloud-assisted multiview video summarization using CNN and bidirectional LSTM. IEEE Transactions on Industrial Informatics, 16(1): 77-86. https://doi.org/10.1109/TII.2019.2929228
[34] Yang, M., Nazir, S., Xu, Q., Ali, S. (2020). Deep learning algorithms and multicriteria decision-making used in big data: A systematic literature review. Complexity, 2020: 2836064. https://doi.org/10.1155/2020/2836064
[35] Wang, X., Nie, X., Liu, X., Wang, B., Yin, Y. (2020). Modality correlation-based video summarization. Multimedia Tools and Applications, 79: 33875-33890. https://doi.org/10.1007/s11042-020-08690-3
[36] Datt, M., Mukhopadhyay, J. (2018). Content based video summarization: Finding interesting temporal sequences of frames. In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1268-1272. https://doi.org/10.1109/ICIP.2018.8451282
[37] Majumdar, J., Awale, M., Santhosh, K.K. (2018). Video shot detection based on sift features and video summarization using expectation-maximization. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1033-1037. https://doi.org/10.1109/ICACCI.2018.8554662
[38] Bhattacharjee, T., Saha, S., Konar, A., Nagar, A.K. (2018). Static video summarization using artificial bee colony optimization. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 777-784. https://doi.org/10.1109/SSCI.2018.8628784
[39] Asim, M., Almaadeed, N., Al-Máadeed, S., Bouridane, A., Beghdadi, A. (2018). A key frame based video summarization using color features. In 2018 Colour and Visual Computing Symposium (CVCS), pp. 1-6. https://doi.org/10.1109/CVCS.2018.8496473