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

Traitement du Signal
Vol. 40, No. 3, June, 2023, pp. 1227-1234
Journal homepage: http://iieta.org/journals/ts

Gated Recurrent Units and Recurrent Neural Network Based Multimodal Approach for Automatic Video Summarization

Lakhwinder Kaur[1], Turki Aljrees[2], Ankit Kumar[3], Saroj Kumar Pandey[3], Kamred Udham Singh[4,5*], Pankaj Kumar Mishra[1], Teekam Singh[6]

[1] Department of Electronics & Telecommunication Engineering, Rungta College of Engineering and Technology, Bhilai 490024, India
[2] College of Computer Science and Engineering, University of Hafr Al Batin, Hafar Al Batin 39524, Saudi Arabia
[3] Department of Computer Engineering & Applications, GLA University, Mathura 281406, India
[4] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan
[5] School of Computing, Graphic Era Hill University, Dehradun 248002, India
[6] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India

Corresponding Author Email: 11004033@gs.ncku.edu.tw

https://doi.org/10.18280/ts.400340

Received: 29 September 2022
Accepted: 10 March 2023

Keywords:
video segments, key frames, gated recurrent units (GRU), recurrent neural network (RNN), feature extraction

ABSTRACT

A typical video record aggregation system requires the concurrent performance of a large number of image processing tasks, including but not limited to image acquisition, pre-processing, segmentation, feature extraction, verification, and description. These tasks must be executed with utmost precision to ensure smooth system performance. Among these tasks, feature extraction and selection are the most critical. Feature extraction involves converting the large-scale image data into smaller mathematical vectors, and this process requires great skill. Various feature extraction models are available, including wavelet, cosine, Fourier, histogram-based, and edge-based models. The key objective of any feature extraction model is to represent the image data with minimal attributes and no loss of information. In this study, we propose a novel feature-variance model that detects differences in video features and generates feature-reduced video frames. These frames are then fed into a GRU-based RNN model, which classifies them as either keyframes or non-keyframes. Keyframes are then extracted to create a summarized video, while non-keyframes are reduced. Various key-frame extraction models are also discussed in this section, followed by a detailed analysis of the proposed summarization model and its results. Finally, we present some interesting observations about the proposed model and suggest ways to improve it.

1. INTRODUCTION

Video summarization is a critical task in efficiently processing vast video sequences. It involves designing various video processing models, including shot boundary detection, clustering of distinct shot boundaries, event-based classification, feature extraction from shot boundaries and events, and ultimately generating a concise summary. The goal of video summarization is to reduce large-sized videos into smaller segments or strips through event time-stamp estimation. For an effective summarization approach, each of these segments is analyzed using multiple feature extraction and classification methods. Numerous researchers have developed a wide range of event-based key-frame extraction (KFE) or video summarization algorithms to accomplish this task.

In the domain of signal processing, video summarization aims to eliminate redundant information from a series of frames. To create an effective video summarization model, it is necessary to analyze frame features and assess their similarity. By employing this approach, frames with similar feature sets can be grouped, which in turn can be used to reduce frame redundancy. Moreover, the summarized output should reflect all key events present in the input video while maintaining a minimal output video size. High-speed algorithms that generate large-sized summarized videos limit their usability for event-based summarization applications.

Video processing operations are among the most computationally complex operations due to the high dimensionality of the input data. Applications of video processing include surveillance, object tracking, user tracing, path monitoring, and more. Each of these applications requires a substantial amount of video data to produce the desired output. In all these applications, the event of interest occurs for a short duration, which must be effectively captured for efficient identification. Video capturing systems collect vast amounts of data and apply application-specific classification algorithms to identify the underlying events of interest.

However, the immense data size makes it computationally challenging to extract the event of interest within a given time frame. This limitation is overcome by introducing video summarization architectures, which accept input videos and apply a series of complex mathematical operations on the frame sets to identify redundancies and outliers. By removing these outliers, only frames of interest remain, typically constituting less than 10% of the entire video's size. Various deep learning models can be employed to perform this task, as shown in Figure 1. The architecture operates through the following steps:

- The input video is divided into frames, which are provided to a sampling unit.
- The sampling unit is based on a shot boundary detection algorithm that identifies video effects such as fade in, fade out, and cuts.
- Shot boundaries allow for the initial removal of redundancy from the image, generating candidate frames.
- These candidate frames are processed using deep learning models, e.g., AlexNet and GoogLeNet, to obtain relevant features.
- The extracted features are fed to classification models, such as Auto Encoders and Random Forest classifiers, to determine if the current frame is useful or redundant.

This method enables the generation of a concise summary that conveys the essential parts of the complete video. Considering the vast amount of video content available on the web or systems, effective video summarization facilitates users' browsing and navigation through extensive video collections, thus increasing user engagement and content consumption. This motivation underlies the current study.

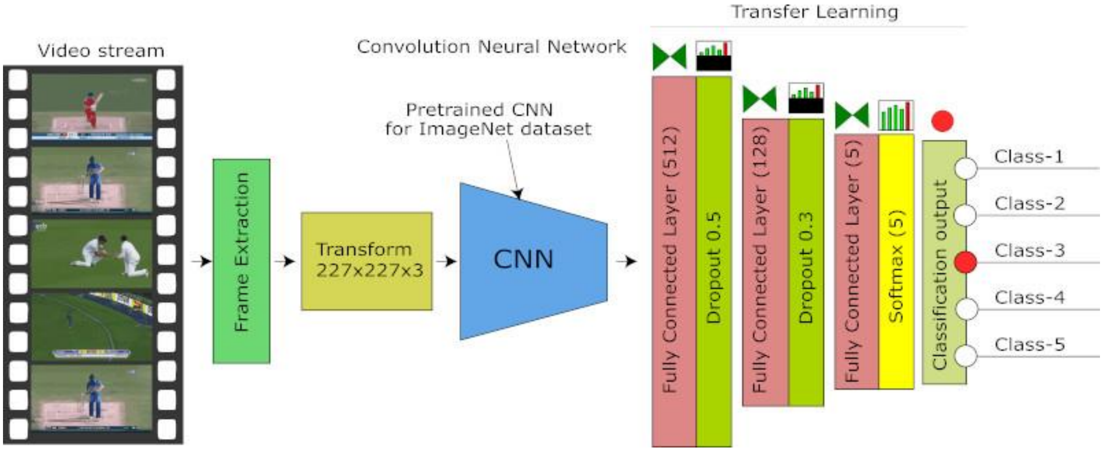

Figure 1. Proposed feature variance-based GRU RNN video summarization model

2. LITERATURE REVIEW DEEP LEARNING-BASED VIDEO SUMMARIZATION

Video summation necessitates the appearance of predetermined picture administration squares with the goal of these squares performing consistent subordinate picture to-picture comparability evaluations for discovering plenitude lodgings. Such a design is represented in the study [1] where approach affiliation is examined. Such a design is represented in the study performed by author of the study [1] where attitude affiliation is examined. The association combines data from visual and text-based highlights to assess outline significance; and after a brief timeframe, removes the essential lodgings from the given educational rundown. It makes use of highlights like the number of worked up area, the relationship from text to design, the disaster from text to format, the number of text sections in a substance portrayal, the scene from bundling to message, the weaving relationship between edge and text, and the significance score of video shot. The organization additionally employs a nearby idea model of long-passing memory (LSTM) to discover visual substance, and thus interfaces these substances with text data to discover layout significance. Considering this multimodal framework, exactness of plan is around 53%, which is higher than fearless taking a gander at (33%), Dictionary Selection based Video Summarization (DSVS) (49%) and Co-Model Analysis (CA) (half); in any case is low for tireless use. This precision can be other than outstanding utilizing the work in the study [2], wherein an essential idea model with LSTM is proposed. The model uses a mix of multi-facet perceptron with twofold LSTMs to work on the value of once-wrapped up. By morals of this blend, an exactness of 63% is refined utilizing visual saliency LSTM with thought model (vsLSTM+Attn.), which is higher when separated and visual saliency LSTM (vsLSTM) (57%), and fundamental LSTM dppLSTM (52.5%), and accordingly, has astounding obvious use. This show can be comparatively improved by utilizing different set up CNN models (irrefutable CNN) as proposed in the study [3], wherein key-frame extraction is finished utilizing CNN consolidate vectors. This work utilizes lacking auto-encoder (AE) nearby odd woods (RF) classifier to pass on an accuracy of 83.3%, which makes it reasonable for shut circuit TV (CCTV) applications. The assorted CNN model has high accuracy when confined and AlexNet (66%), GoogLeNet (61%), VGG16 (67%), and Inception Net (71%); appropriately, showing that obvious CNN structures are unparalleled than single CNN ones. This show can consider other summation models as depicted in the study [4], wherein approaches like Motion based way of thinking, Event based procedure, Color based point of view, Audio-visual based framework, Trajectory assessment, Gesture based, Clustering, shot confirmation approach and Mosaic based construction is portrayed. These designs can be applied for different video news format as displayed in the study [5], wherein human methods for unprecedented of video are portrayed. These methodologies consider partition data; and oblige it with PC vision to review best models for news design. It is seen that colossal learning models that usage CNN and relative enhancements beat different models the degree that news and occasion-based video rundown. This can be seen from the study [6], wherein foreboding neural affiliations (RNNs) with multi-edge revived LSTM (MOLRVS) is utilized for once-wrapped up. The model adjusts the LSTM plan by adding different heaps of LSTM layers for better execution. It is seen that the proposed MOLRVS organizing has an accuracy of 85.3%; which is higher than Memorable Rich Video Summarization (MRVS) (69.3%), Video Summarization with Attention-based Encoder–Decoder Networks (VSABEDN) (71.6%), Unsupervised Object-level Video Summarization with Online Motion Auto-encoder (UOVSOMAE) (76.9%), Framework towards Domain Specific Video Summarization (FDSVS) (82.2%), and Video Summarization through Nonlinear Sparse Dictionary Selection (VSNSDS) (75%); in this way making it fundamentally basic determinedly video outline applications. A quick overview of other such monstrous taking in constructions can be seen from the study [7], wherein assessments like Action Ranking (56.3%), cross plan RNN (57.7%), MAVS (66.8%), and SMN (64.5%) are looked at. These models produce moderate level precision; and thusly, can be utilized for coarse approximated video once-over applications. These outcomes can be other than forefront if once-over is done by client question. The work described in the study [8] proposes such a masterminding that utilizes LSTM model with Generative Adversarial Network (GAN) to get accuracy of 74.5%; by integrity of relationship of client obligation during chart evaluation. This work is invigorated by the models proposed in the study [9], wherein explicit fundamental learning and AI methods are portrayed for outline.

A close to display is proposed in the study [10], wherein content-based video thought is utilized for extraction of accurate edges from input video. The model joins Frame-based highlights, Segment-based highlights, Visual Semantic highlights, Visual Raw highlights and Audio highlights to make a wide segment vector. This vector is given to a CNN-based association named as HighlightNet; which gives a section significance score. The pieces that have high significance score are withdrawn, and utilized for verifiable outline. This model can be broadened utilizing the work in the study [11], wherein a base up video stand-out framework is portrayed. This structure can consider yield video improvement choices subject to various part designs. A precision of 36.6% is obtained utilizing this approach, which can be utilized for changing the yield of different procedures. A calculation that uses super-pixels for video once-over can be seen from the study [12], here, every pixel is changed over into super-pixel through similarity evaluation. This closeness respect permits the pixels to be tended to regarding each other; and starting there on by utilizing a district blending calculation; the pixels are cemented to diagram a super-pixel. These super-pixels are then given to an indistinguishable quality assessment motor to play out the last synopsis. This outcome into a precision of 62.6% which is higher when segregated and VSUMM (54.4%), lacking word reference methodologies (SD) (48.3%), and Key Point-Based Key-frame Selection (KBKS) (46%). Comparative models are depicted in the studies [13-16], wherein Event-Related Potential Responses, network video overview utilizing super pixels, Joint Integer Linear Programming (JILP), and cautious video summary improvements are portrayed. Out of these, the JILP approach beats different procedures, by giving an exactness of 59.9% across different datasets. Strategies like JILP are depicted in the studies [16-18], wherein Sequential Determinant Point Process, Feature Discrimination, and Synthetic Coordinate based Recommendation are inspected.

A novel Deep Side Semantic Embedding (DSSE) model is proposed, wherein side information about videos is used in order to generate video summaries. The model uses information like click rates, semantic relevance, textual-space to visual-space mapping and feature classification using CNNs in order to achieve classification accuracy of 81%, which makes the system deployable for real-time use cases. Another novel model that uses user-ranking for video segments in order to summarize the video is proposed in the studies [19-21]. This model uses a combination of 2D CNNs, 1D CNNs and LSTMs in order to generate a refined importance score for each video segment. This score is combined with user rankings in order to obtain the final score about each segment, and based on this score the final summarization is done. Architecture of this system can be observed from Figure 1, wherein the fused 2D CNN [22], 1D CNN and LSTM model [23] can be seen. This model is able to achieve an accuracy of 83% on SumMe and TVSum datasets, which is high enough considering that standard CNN architectures like visual LSTM is 53% accurate, deep LSTM is 54% accurate, GAN [24] is 59% accurate while deep regression learning is only 61% accurate on the same datasets. Similar deep learning models are proposed, wherein audio-video alignment, Nonlinear sparse dictionary selection, and Graph Based sentence summarization algorithms respectively are used.

Linear models that use effective feature extraction for video summarization, wherein algorithms like correlation of modality, bidirectional LSTM, and scale invariant feature transform (SIFT) are used respectively. These algorithms achieve a moderate performance of 53%, 60% and 75% are achieved due to minimal use deep learning or bioinspired models. A high accuracy bio-inspired model-based video summarization architecture that uses Artificial Bee Colony (ABC) Optimization is also proposed. This model aims at maximizing frame difference between consecutive frames in order to achieve an accuracy of 70% on different datasets. This model can be improved by addition of colour features with key frame extraction as suggested, wherein an accuracy of 67% is achieved with help of thresholding algorithm, which reduces overall system complexity. It can be observed that these models vary widely in terms of accuracy performance and area of application. Thus, in order to reduce complexity of algorithmic selection, the next section compares these algorithms in terms of area of application and approximate accuracy of summarization.

3. GRU BASED RNN MODEL

The proposed feature variance-based GRU RNN video summarization model uses a combination of efficient feature extraction, feature selection and RNN based classification. Overall work flow of the system is shown in Figure 1, can be described as follows,

- The input training set video is given to feature extraction, wherein grey level co-occurrence matrix (GLCM) [25] and Wavelet features are extracted.
- These features are given to a variance-based feature selection model, wherein feature reduction is performed.
- The selected features are converted into feature frames, and given to GRU based RNN model for classification.

Output of classification indicates whether the given frame is key-frame or non-key-frame as shown in Figure 2 [26].

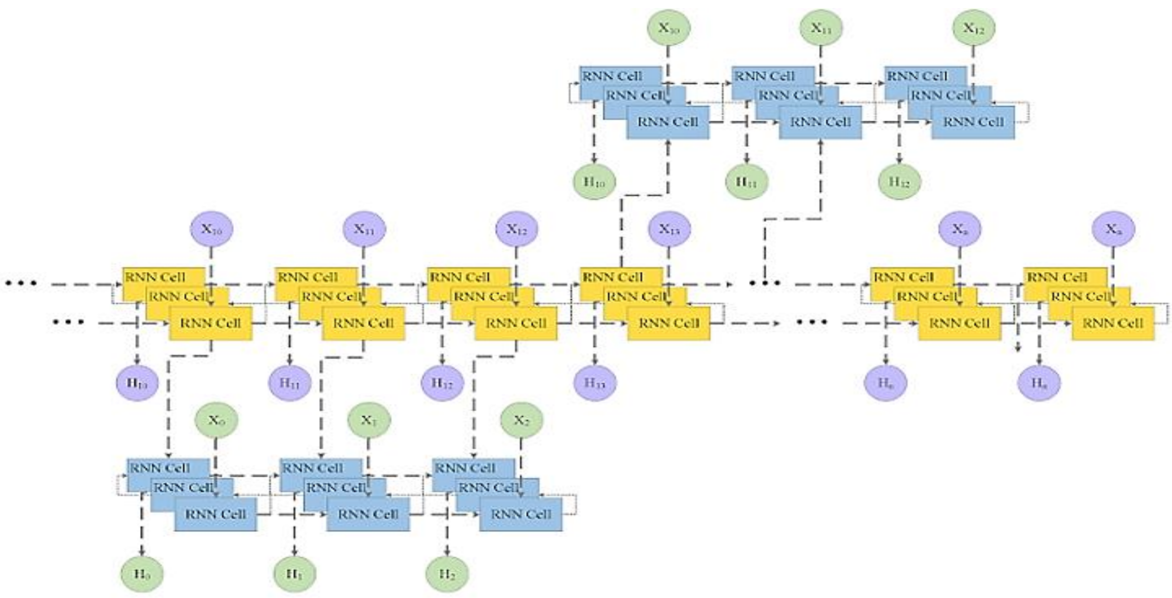

Figure 2. GRU's RNN model to be classified into key frames and non-key frames

There is no standard or universal way to classify GRU and RNN models into key frames and non-key frames. The classification of frames into key frames and non-key frames typically depends on the specific problem and the desired outcome. In video processing, for example, key frames are often selected based on their visual importance or their ability to represent the content of the video, while non-key frames are frames that can be discarded without losing much information. The choice of which frames to classify as key or non-key frames depends on the specific application and the requirements of the user.

Based on these steps, it can be observed that the input training video is initially given to a feature extraction block, wherein GLCM features are evaluated. These features include angular 2nd moment (A2M), contrast (C), and correlation (Cr) and are estimated using Eq. (1), Eq. (2), Eq. (3) respectively.

$$A2M = \sum_{i=1}^{R}\sum_{j=1}^{C} p(i,j)^2 \tag{1}$$

$$C = \sum_{i=1}^{R}\sum_{j=1}^{C} R * C * p(i,j) \tag{2}$$

$$Cr = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} R * C * p(i,j) - \mu_i * \mu_j}{\sigma_i * \sigma_j} \tag{3}$$

where, R, C represent video dimensions, and, 'p' is co-occurrence probability for the video pixels, while σ *and* μ are standard deviation and mean of the video pixels. After GLCM, Haar wavelet transform is applied to the video sequence, and approximate component (AC), horizontal component (HC), vertical component (VC) and diagonal component (DC) [27] are estimated. Evaluation of these components is done using Eq. (4), Eq. (5), Eq. (6), Eq. (7) as follows:

$$AC_i = \frac{\left(Pix_{1_i} + Pix_{2_i}\right)}{2} \tag{4}$$

$$HC_i = \frac{\left(Pix_{1_i} - Pix_{2_i}\right)}{2} \tag{5}$$

$$VC_i = \frac{\left(-Pix_{1_i} + Pix_{2_i}\right)}{2} \tag{6}$$

$$DC_i = \frac{\left(-Pix_{1_i} - Pix_{2_i}\right)}{2} \tag{7}$$

where, Pix_1 *and* Pix_2 current and next video pixels, and $'i'$ represents number of pixels present in each video frame [28]. Both these feature sets are combined in order to estimate feature variance, which is evaluated using Eq. (8) as follows:

$$V_{avg} = \sqrt{\frac{\sum_{a=1}^{m}\left(x_a - \frac{\sum_{i=1}^{m}\sqrt{\frac{\sum_{j=1}^{n}\left(x_j - \frac{\sum_{k=1}^{n} x_k}{n}\right)^2}{n-1}}}{m}\right)^2}{m-1}} \tag{8}$$

where, 'm' is the number of video frames present in current class (key-frame and non-key-frame), 'n' is number of samples in other class, while 'x' represents value of the feature. Due to selection of most variant features dimensionality of the features is reduced, which allows for faster and more accurate classification. The goal of dimensionality reduction is to remove redundant or less important features while retaining as much information as possible. There are many methods for dimensionality reduction, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-SNE, and each method will result in a different reduction in feature dimension. Typically, the reduced feature dimension is a parameter that can be adjusted to trade off between computation time and accuracy. A lower dimensionality leads to faster computation times but potentially lower accuracy, while a higher dimensionality results in more accurate predictions but longer computation times. The optimal feature dimension will depend on the specific problem and the desired level of accuracy. The actual features are around 50% higher than selected features.

These features are given to a GRU RNN [29, 30] model as observed from Figure 3, wherein video features are classified into key-frame and non-key-frame classes. Here, the input image is given to 3 different convolutional layers, each layer consists of the following sub-layers,

- Convolutional layer for feature augmentation.
- Rectilinear unit with max-pooling for feature selection
- Dropout layer for feature reduction

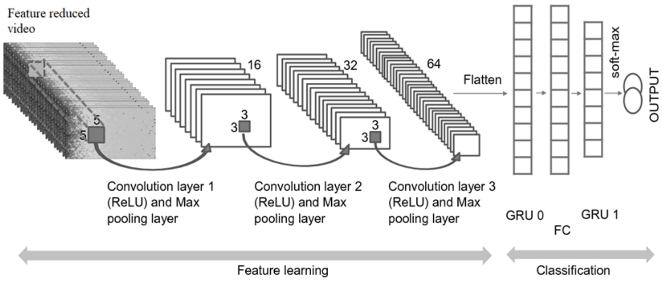
Feature reduced video
5
5
16
3
3
32
3
3
64
Flatten
soft-max
OUTPUT
Convolution layer 1 (ReLU) and Max pooling layer
Convolution layer 2 (ReLU) and Max pooling layer
Convolution layer 3 (ReLU) and Max pooling layer
GRU 0
FC
GRU 1
Feature learning
Classification

Figure 3. Proposed GRU based RNN classifier design for video summarization

The network structure of each convolutional layer typically consists of the following sub-layers:

Convolutional layer: This layer performs a convolution operation on the input, which involves element-wise multiplication of the input with a set of filters or kernels, followed by a summation over all elements. The result of the convolution operation is a feature map.

Activation layer: This layer applies an activation function to the output of the convolutional layer. Common activation functions used in convolutional neural networks (CNNs) include ReLU (rectified linear unit), sigmoid, and tanh. The activation function adds non-linearity to the network, allowing it to model more complex functions.

Pooling layer: This layer performs down-sampling by summarizing the information in a subset of the feature map. Common pooling methods include max pooling and average pooling.

All these layers are connected in cascade such that the final layer has maximum number of relevant features. These features are given to a GRU model, wherein fully connected neural network is used. Outputs of this GRU layer are connected to a soft max activation model, wherein the final key-frame and non-key-frame information is obtained. All the key-frames are stored at the output, while all the non-key-frames are removed from the sequence. This model is tested on a large number of videos, and compared with existing state-of-the-art models. Result analysis of the model can be observed from the next section, wherein accuracy and delay values are evaluated and compared with reviewed models.

4. ANALYSIS OF RESULT

We compared the performance of the proposed GRU RNN-based feature variant methods in terms of average accuracy (A), and delay (R) values. The proposed model was run in Python, which is a general purpose environment that can be used for analysis of different video data samples. According to the evaluation of these values for the various tested models shown in Tables 1 and 2, the proposed model outperforms the other models in terms of accuracy and delay values. These results come from testing the system using the TRECVID dataset, which includes more than 5000 movies grouped into different categories.

The results showcase that the proposed model is over 4% more accurate than the models described in the studies [3, 14], thereby making it useful for real time deployments. This can also be observed from Figure 4, wherein accuracy values are visualized. In Figure 4 higher ordinate values indicate greater values or quantities, while lower ordinate values indicate smaller values or quantities.

Similar analysis is done in terms of delay of execution, and can be observed from Table 2.

It is clear from the delay values that the suggested GRU RNN model performs better than other models and increases speed by 25%. This is also evident in Figure 5, which shows these numbers in visual form.

Table 1. Comparative analysis of different video summarization models

Number of videos	Avg. A [3]	Avg. A [14]	Avg. A [Proposed]
10	00.80	00.90	00.90
20	00.85	00.90	00.95
30	00.67	00.95	00.95
40	00.70	00.94	00.94
50	00.79	00.85	00.96
60	00.83	00.86	00.95
70	00.86	00.85	00.95
80	00.78	00.85	00.94
90	00.81	00.85	00.93
100	00.83	00.86	00.92
110	00.85	00.86	00.92
120	00.86	00.86	00.92
130	00.87	00.86	00.94
140	00.88	00.87	00.92
150	00.86	00.88	00.93
160	00.87	00.89	00.93
170	00.87	00.89	00.93
180	00.88	00.90	00.94
190	00.89	00.90	00.94
200	00.89	00.91	00.95
210	00.90	00.91	00.95

Table 2. Delay of different video summarization models

Number of videos	Avg. D (s) [3]	Avg. D (s) [14]	Avg. D (s) [Proposed]
10	31.60	15.80	15.80
20	23.70	15.80	7.90
30	52.67	7.90	7.90
40	47.40	9.48	9.48
50	33.71	23.70	6.32
60	26.33	22.80	8.26
70	21.67	22.97	8.21
80	34.56	23.50	8.98
90	30.10	23.24	10.69
100	26.66	22.88	12.22
110	23.94	22.07	12.74
120	21.73	21.75	13.13
130	19.89	21.63	9.71
140	18.34	19.94	11.89
150	22.69	18.39	11.68
160	21.16	17.53	11.15
170	19.83	16.77	10.52
180	18.65	16.07	9.73
190	17.61	15.37	9.16
200	16.68	14.83	8.40
210	15.84	13.85	7.69

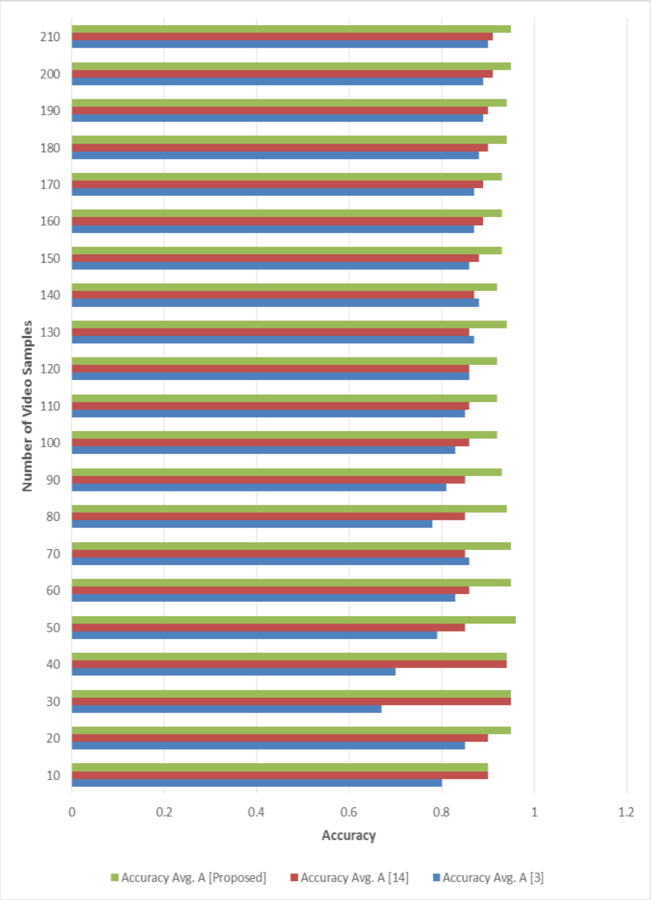

Figure 4. Comparative analysis of average accuracy of different model used for video summarization

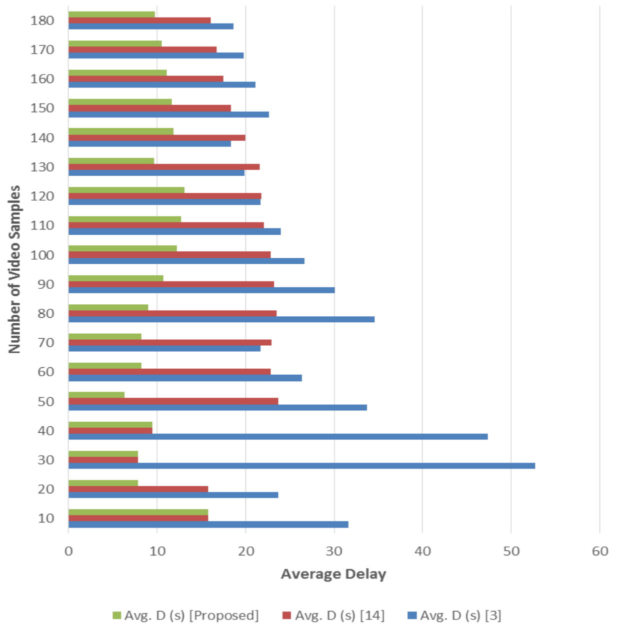

Figure 5. Average delay for different models

Thus, it can be observed that RNN and its subtypes are most suited for video summarization. In order to perform statistical analysis of the reviewed models; their general-purpose average delay is compared. Comparison of this delay can be observed from Table 2, wherein it is seen that deep learning RNN model have better performance. These findings show that the suggested methodology is quite useful for real-time video summarising applications. This method uses bio inspired computing to extract multiple feature sets, Bioinspired optimization algorithms are based on principles from nature, such as evolution, genetics, and swarm behaviour. The advantages of these algorithms over traditional optimization methods include:

1. Global Optimization: Bioinspired algorithms are capable of finding global optima, unlike local optimization algorithms, which only find the optimal solution in a local neighborhood.

2. Robustness: Bioinspired algorithms are robust to changes in the environment and can handle problems with multiple objectives and constraints.

3. Diversity: Bioinspired algorithms promote diversity in the search space, which can help avoid getting stuck in a local optimum and increase the chances of finding a global optimum.

4. Simplicity: Many bioinspired algorithms have a simple mathematical structure and are easy to implement.

5. Flexibility: Bioinspired algorithms can be easily adapted to a wide range of optimization problems and can handle problems with complex, non-linear objective functions.

In summary, bioinspired optimization algorithms offer several advantages over traditional optimization models, such as robustness, diversity, simplicity, and flexibility, which can lead to improved optimization results and solutions.

5. CONCLUSION AND FUTURE SCOPE

After testing the proposed system on over 500 videos, it is evident that the GRU classification engine can optimize video summarization performance. It is observed that the proposed model is over 4% more effective than state-of-the-art techniques, and has 25% lower delay as well. These metrics make the system model perform in a better and sophisticated manner when applied to large number of datasets. In future, researchers can apply LSTM and other models and check their performance on the proposed summarization application. In order to increase the effectiveness of video condensation, there is a need to build an effective feature extraction and analysis model. Models that try to increase this effectiveness typically increase the computational complexity of the summarization, which restricts their usefulness. This essay suggests a high-speed recurrent neural network (RNN) model based on gated recurrent units (GRU) that uses condensed feature sets for video condensation to overcome this limitation. The proposed model initially scans a series of different test video sequences, and estimates their frame-level features. Each of these features are compared with the features of resultant ground truth videos, and feature variance is evaluated. The features with maximum variance are then used in order to represent the underlying frames, thereby reducing frame dimensions. The reduced dimension frames are then given to a GRU-based RNN model for high-speed summarization. The results show that the suggested model can achieve 28% speed improvement and 95% accuracy for video summarization compared to the original RNN model. Large-scale video summarization requires video pre-processing, feature extraction, feature analysis, and post-processing operations. Each of these tasks requires large computational delays, which limits their performance for long length video sequences.

REFERENCES

[1] 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

[2] Lebron Casas, L., Koblents, E. (2019). Video summarization with LSTM and deep attention models. In MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II 25, Springer International Publishing, pp. 67-79. https://doi.org/10.1007/978-3-030-05716-9_6

[3] Nair, M.S., Mohan, J. (2021). Static video summarization using multi-CNN with sparse autoencoder and random forest classifier. Signal, Image and Video Processing, 15: 735-742. https://doi.org/10.1007/s11760-020-01791-4

[4] Ajmal, M., Ashraf, M.H., Shakir, M., Abbas, Y., Shah, F. A. (2012). Video summarization: techniques and classification. In International Conference on Computer Vision and Graphics. (ICCVG), pp. 1-13. https://doi.org/10.1007/978-3-642-33564-8_1

[5] Barbieri, T.T.D.S., Goularte, R. (2021). Content selection criteria for news multi-video summarization based on human strategies. International Journal on Digital Libraries, 22: 1-14. https://doi.org/10.1007/s00799-020-00281-9

[6] Archana, N., Malmurugan, N. (2021). RETRACTED ARTICLE: Multi-edge optimized LSTM RNN for video summarization. Journal of Ambient Intelligence and Humanized Computing, 12(5): 5381-5395. https://doi.org/10.1007/s12652-020-02025-8

[7] 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

[8] Nalla, S., Agrawal, M., Kaushal, V., Ramakrishnan, G., Iyer, R. (2020). Watch hours in minutes: Summarizing videos with user intent. In Computer Vision-ECCV 2020 Workshops: Glasgow, UK, August 23-28, 2020, Proceedings, Part 16, Springer International Publishing, 714-730. https://doi.org/10.1007/978-3-030-68238-5_47

[9] Workie, A., Sharma, R., Chung, Y.K. (2020). Digital video summarization techniques: A survey. International Journal of Engineering Research and Technology, 9: 81-85. https://doi.org/10.17577/IJERTV9IS010026

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

[11] 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

[12] Jin, H., Yu, Y., Li, Y., Xiao, Z. (2022). Network video summarization based on key frame extraction via superpixel segmentation. Transactions on Emerging Telecommunications Technologies, 33(6): e3940. https://doi.org/10.1002/ett.3940

[13] Kim, H.H., Kim, Y.H. (2019). Video summarization using event‐related potential responses to shot boundaries in real‐time video watching. Journal of the Association for Information Science and Technology, 70(2): 164-175. https://doi.org/10.1002/asi.24103

[14] Lei, S., Xie, G., Yan, G. (2014). A novel key-frame extraction approach for both video summary and video index. The Scientific World Journal. https://doi.org/10.1155/2014/695168

[15] Jangra, A., Jatowt, A., Hasanuzzaman, M., Saha, S. (2020). Text-image-video summary generation using joint integer linear programming. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II, Springer International Publishing, 42: 190-198. https://doi.org/10.1007/978-3-030-45442-5_24

[16] Zheng, J., Lu, G. (2021). k-SDPP: fixed-size video summarization via sequential determinantal point processes. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 774-781. http://dx.doi.org/10.24963/ijcai.2020/108

[17] Ismail, M.M.B., Bchir, O., Emam, A.Z. (2013). Endoscopy video summarization based on unsupervised learning and feature discrimination. In 2013 Visual Communications and Image Processing (VCIP), IEEE, pp. 1-6. https://doi.org/10.1109/VCIP.2013.6706410

[18] Panagiotakis, C., Papadakis, H., Fragopoulou, P. (2020). Personalized video summarization based exclusively on user preferences. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II, Cham: Springer International Publishing, pp. 305-311. https://doi.org/10.1007/978-3-030-45442-5_38

[19] Challapalle, N., Chandran, M., Rampalli, S., Narayanan, V. (2020). X-VS: Crossbar-based processing-in-memory architecture for video summarization. In 2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), IEEE, pp. 592-597. https://doi.org/10.1109/ISVLSI49217.2020.00091

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

[21] Archana, N., Malmurugan, N. (2021). RETRACTED ARTICLE: Multi-edge optimized LSTM RNN for video summarization. Journal of Ambient Intelligence and Humanized Computing, 12(5): 5381-5395. https://doi.org/10.1007/s12652-020-02025-8

[22] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I. (2020). AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. In IEEE Transactions on Circuits and Systems for Video Technology, 31(8): 3278-3292. https://doi.org/10.1109/TCSVT.2020.3037883

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

[24] Zhang, C., Hu, B., Suo, Y., Zou, Z., Ji, Y. (2020). Large-scale video retrieval via deep local convolutional features. Advances in Multimedia, 2020: 1-8. https://doi.org/10.1155/2020/7862894

[25] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G., de Albuquerque, V.H.C. (2019). Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. In IEEE Internet of Things Journal, 7(5): 4455-4463. https://doi.org/10.1109/JIOT.2019.2950469

[26] 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. In IEEE Transactions on Industrial Informatics, 16(1): 77-86. https://doi.org/10.1109/TII.2019.2929228

[27] Hussain, T., Muhammad, K., Del Ser, J., Baik, S.W., de Albuquerque, V.H.C. (2019). Intelligent embedded vision for summarization of multiview videos in IIoT. In IEEE Transactions on Industrial Informatics, 16(4): 2592-2602. https://doi.org/10.1109/TII.2019.2937905

[28] Wu, G., Lin, J., Silva, C.T. (2022). Intentvizor: Towards generic query guided interactive video summarization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 10503-10512. https://doi.org/10.48550/arXiv.2109.14834

[29] Workie, A., Sharma, R., Chung, Y.K. (2020). Digital video summarization techniques: A survey. International Journal of Engineering and Technology, 9: 81-85. https://doi.org/10.17577/IJERTV9IS010026

[30] 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