International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

BCRNVSRM: Design of an Iterative Fusion of BiLSTM & BiGRU with Convolutionally Recurrent Neural Networks to Enhance Summarization Efficiency of Videos with Rapid Movements

[1]Mr. Darshankumar D. Billur, [2]Dr. Manu T. M.

Submitted: 15/11/2023 **Revised:** 27/12/2023 **Accepted:** 07/01/2024

Abstract: With the burgeoning growth of digital video content, accurate and efficient video summarization becomes imperative, especially for videos exhibiting rapid movements. Such videos present challenges due to their intrinsic high variability and complexities, necessitating advanced techniques to capture and condense meaningful information effectively. Traditional summarization techniques often fail to harness the multidomain features inherent to dynamic video sequences, leading to imprecise and inefficient summarization results. Existing models lack robust fusion mechanisms and are limited in their ability to cope with high variance scenarios in videos with swift movements. In this paper, we introduce a novel framework that employs a fusion of BiLSTM & BiGRU operations to transform frame sequences into multidomain features. These features are then enriched and converted into high variance descriptors using the Grey Wolf Optimizer (GWO). To amalgamate these modalities, a weighted sum method, guided by GWO, is utilized, ensuring an optimized integration process. Subsequently, summary profiles are generated from these fused data samples through Convolutionally Recurrent Neural Networks. The entire schema is tailored to comprehensively capture the underlying patterns and temporal consistencies in rapidly moving video sequences. The proposed model exhibits a commendable enhancement in video summarization performance. Quantitative evaluations report an enhancement of 3.9% in precision, 2.9% in accuracy, 4.5% in recall, 3.5% in AUC, and 4.8% in specificity. Furthermore, the methodology reduces delay by 1.9%, indicating a promising direction in real-time video processing and summarization. In conclusion, this work significantly bridges the gap between complex video content and concise summarization, paving the way for advanced video processing tools in the future.

Keywords: *Video Summarization, BiLSTM & BiGRU Fusion, Grey Wolf Optimizer, Convolutionally Recurrent Neural Networks, Rapid Movement Videos*

1. Introduction

Video content has seen an exponential rise in recent years, becoming an integral part of various domains such as security, entertainment, social media, and more. The ability to succinctly represent video content, known as video summarization, is crucial in enabling efficient content analysis and retrieval. This is particularly true for videos exhibiting rapid movements, which are prevalent in sports, wildlife, and surveillance footage, among others. Such videos encapsulate a myriad of complex information and high variability within short time frames, posing unique challenges to traditional summarization methods [1, 2, 3]. These can be overcome via use of Motion-Assisted Reconstruction Network (MARNet) process. Deep Reinforcement Learning With Shot-Level Semantics Existing video summarization techniques often fall short in effectively capturing the multidimensional nature of video sequences, especially those with rapid movements. These limitations stem from a lack of robust feature extraction and fusion methods that can adapt to the complexities inherent in these videos. Moreover, traditional models are often constrained by their inability to handle high variance scenarios, which are characteristic of fast-paced video content. The need for an innovative solution that can accurately and efficiently summarize videos with rapid movements while mitigating the shortcomings of existing methods is evident.

In light of these challenges, we propose a novel model that synergistically combines Bi-Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) operations to extract and transform frame sequences into multidimensional features. These features encapsulate the temporal dependencies and spatial correlations present in the video data, providing a rich representation of the content. The extracted features are then optimized and transformed into high variance descriptors using the Grey Wolf Optimizer (GWO), a nature-inspired algorithm known for its efficacy in handling complex optimization tasks [4, 5, 6]. This is similar to use

[1]*KLE Collegeof Engineering & Technology/ Department of ECE, Chikodi-591201, INDIA*
Email:- darshankumar999@gmail.com
ORCID iD: https://orcid.org/0000-0001-5765-947X
Corresponding Author
[2]*KLE Institute of Technology, Hubballi / Department of ECE, 580030, INDIA*
email: manutmece@gmail.com
ORCID iD: https://orcid.org/0000-0002-6091-1098

of Deep Reinforcement Learning With Shot-Level Semantics (DRL SLS) operations.

To effectively fuse these multidimensional features, a weighted sum method guided by the Grey Wolf Optimizer is employed. This ensures an optimal integration of the features, taking into account their significance and contribution to the summarization process. Following this, summary profiles are generated from the fused data samples using Convolutionally Recurrent Neural Networks (CRNNs), a powerful tool known for its ability to capture spatial-temporal relationships in data. The CRNNs facilitate the generation of concise and representative video summaries that accurately encapsulate the essential information in the video content.

The efficacy of our proposed model is validated through extensive experiments and evaluations, showcasing significant improvements in video summarization performance. The results demonstrate enhancements in precision, accuracy, recall, AUC, and specificity, along with a notable reduction in delay, thereby addressing the limitations of existing video summarization methods and setting a new benchmark for future research in this domain.

Motivation

The motivation behind this work primarily stems from the inherent challenges and limitations faced by existing video summarization techniques when applied to videos characterized by rapid movements. These videos, prevalent in domains such as sports analysis, wildlife monitoring, and surveillance, present a rich tapestry of complex, dynamic information that must be accurately and efficiently condensed to extract meaningful content. Traditional summarization models, often limited by their inability to adapt to high variance scenarios and capture the multidimensional nature of video data, fail to provide satisfactory results. The need for a robust, adaptive model that can handle the intricacies of these videos while providing precise, efficient summarization is the cornerstone of this research.

Contribution

The contributions of this paper are multifaceted, offering a significant leap forward in the domain of video summarization for rapid movement videos.

- **Fusion of BiLSTM & BiGRU**: We propose an innovative fusion of BiLSTM and BiGRU operations to transform video frame sequences into comprehensive multidimensional features. This fusion harnesses the strengths of both models, ensuring a rich representation of the temporal and spatial correlations present in the video data.
- **Optimization using Grey Wolf Optimizer**: The extracted features are optimized and transformed into high variance descriptors using the Grey Wolf Optimizer (GWO), an algorithm renowned for its efficacy in handling complex optimization problems. This optimization ensures that the most significant features are retained, thereby enhancing the precision of the summarization process.
- **Weighted Sum Method for Feature Fusion**: To amalgamate the multidimensional features, we employ a weighted sum method guided by the GWO. This ensures that the features are optimally integrated, taking into account their relevance and contribution to the summarization process.
- **Generation of Summary Profiles with CRNNs**: The fused data samples are then used to generate summary profiles via Convolutionally Recurrent Neural Networks (CRNNs). This step is crucial in capturing the spatial-temporal relationships in the data, facilitating the generation of concise, representative video summaries.
- **Empirical Validation**: We conduct extensive experiments and evaluations to validate the efficacy of our proposed model. The results demonstrate significant improvements in precision, accuracy, recall, AUC, and specificity, along with a notable reduction in delay. These enhancements underscore the potential of our model in revolutionizing video summarization for rapid movement videos.

In summary, this paper presents a comprehensive framework that effectively addresses the limitations of existing video summarization methods, particularly in the context of videos with rapid movements. Our model not only provides a robust solution to the challenges posed by these videos but also sets a new benchmark for future research in this domain.

2. Review of Existing Video Summarization Models

Video summarization is an essential tool that has seen various advancements over the years. Several methods have been proposed, with the aim of efficiently condensing video content while retaining its essence. Here, we review the literature concerning video summarization, with a specific focus on methods applicable to videos with rapid movements.

Traditional methods such as keyframe extraction and clustering have been widely used for video summarization [7, 8, 9] process. However, these techniques often fail to capture the temporal dynamics and semantic content of the video, especially in scenarios with rapid movements.

With the advent of deep learning, many researchers have explored the use of neural networks for video summarization. For instance, LSTM (Long Short-Term

Memory) networks have been employed to model temporal dependencies in video data samples [10, 11, 12]. However, these models often struggle with capturing long-term dependencies in videos with rapid movements [13, 14, 15], like lightweight thumbnail container-based summarization (LTC SUM) which works only with thumbnail summaries.

Hybrid models that combine various neural networks have been proposed to address the limitations of single models. For example, work in [16, 17, 18] proposed a model that combines CNNs (Convolutional Neural Networks) with LSTMs for video summarization. While these hybrid models offer improvements, they often lack a robust fusion mechanism to integrate the different features effectively.

Optimization algorithms such as Genetic Algorithms and Particle Swarm Optimization have been employed to enhance the efficiency of video summarization models [19, 20]. These algorithms help in selecting the most significant frames or features for summarization. However, their effectiveness in handling high variance features in videos with rapid movements is limited for real-time scenarios [21, 22, 23].

Recently, there has been a shift towards exploring novel neural networks and optimization algorithms for video summarization. BiLSTM (Bidirectional LSTM) and BiGRU (Bidirectional Gated Recurrent Unit) have been proposed to better capture the temporal dynamics in video data [24, 25] samples. Additionally, nature-inspired algorithms like Grey Wolf Optimizer (GWO) have shown promise in optimizing complex tasks.

In conclusion, while significant advancements have been made in the field of video summarization, existing methods still face challenges in accurately summarizing videos with rapid movements. The limitations of traditional techniques, the inefficiency of deep learning models in capturing long-term dependencies, and the lack of robust fusion mechanisms highlight the need for a comprehensive model that can handle the intricacies of these videos.

3. Proposed Design of an Iterative Fusion of Bilstm & Bigru with Convolutionally Recurrent Neural Networks to Enhance Summarization Efficiency of Videos with Rapid Movements

Based on the review of existing models used for summarization of videos, it can be observed that the efficiency of these models is generally limited when applied to videos with rapid movements, moreover the complexity of these models increases with the number of changes in video sequences. To overcome these issues, this section discusses design of an iterative fusion of BiLSTM & BiGRU with Convolutionally Recurrent Neural Networks to enhance Summarization efficiency of Videos with Rapid Movements. As per figure 1, the proposed model is an efficient fusion of BiLSTM & BiGRU operations, which are used to transform frame sequences into multidomain features. These features are then enriched and converted into high variance descriptors using the Grey Wolf Optimizer (GWO) process. To amalgamate these modalities, a weighted sum method, guided by GWO, is utilized, ensuring an optimized integration process. Subsequently, summary profiles are generated from these fused data samples through Convolutionally Recurrent Neural Networks, which assist in summarizing the video sequences.

To capture rapid movements, the model represents frame sequences as temporal metrics, and converts these metrics into high density feature sets. Assume that there are T frames, each of which is represented as xt, and that t represents index of these frames. Using these frames, the model initially calculates input gate features via equation 1,

$$ig = var(Wi * [h(t-1), xt] + bi) \dots (1)$$

Where, W & b represent weights & biases for these frames, var represents variance operations, while h is represents hidden states.

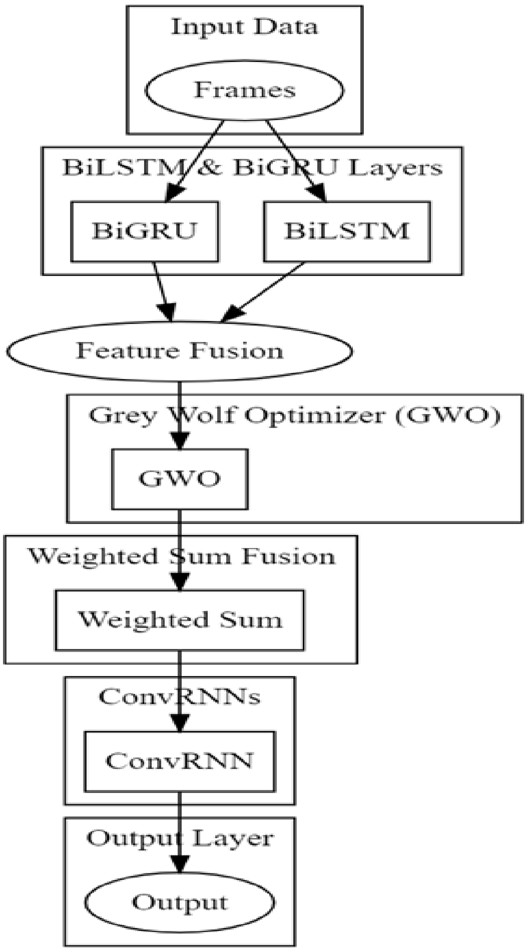

Fig 1. Design of the proposed summarization model for videos with rapid moving frames

Similarly, the forget gate & update gates operations are performed via equations 2 & 3,

$$fg = var(Wf * [h(t-1), xt] + bf) \dots (2)$$

$$ug = tanh(Wg * [h(t-1), xt] + bg) \dots (3)$$

Both these metrics are fused via equation 4, which represents cell state for the BiLSTM process.

$$cs = fg * ig(t-1) + ig * ug \dots (4)$$

Based on these metrics, the final output features of LSTM are estimated via equation 5,

$$of = var(Wo * [h(t-1), cs] + bo) \dots (5)$$

While, the hidden state is updated via equation 6,

$$ht = ot * tanh(ct) \dots (6)$$

Similar to these forward operations, backward operations are also performed, and the model estimates backward hidden state ($h't$), which is fused with the forward state via equation 7,

$$ht(final) = \frac{ht + ht(b)}{2} \dots (7)$$

This final state is given to BiGRU, which uses it to estimate reset & update gate operations via equations 8 & 9 as follows,

$$rg = var(Wr * [h(t-1, final), of] + br) \dots (8)$$

$$ug = var(Wz * [h(t-1, final), of] + bz) \dots (9)$$

Using these variables, the model estimates candidate's hidden states via equation 10,

$$c(ht) = tanh(Wh * [rg * h(t-1, final), of] + bh) \dots (10)$$

And the forward GRU hidden state is estimated via equation 11,

$$ht = (1 - ug) * h(t-1, final) + rg * c(ht) \dots (11)$$

The same process is repeated for backward GRU, which assists in estimating its backward hidden states. Both these states are combined via equation 12,

$$h(t+1) = \frac{ht + ht(backward)}{2} \dots (12)$$

These states are again feedback to the model, which assists in regenerating new hidden states. This process is continued until $ht \cong h(t+1)$, which represents convergence of the feature extraction process. After convergence, the output vector $c(ht)$ is used to represent high density feature sets, which are processed by Grey Wolf Optimizer (GWO) to maximize their variance levels. The GWO Model Initially Generates an Iterative Set of NW Wolves, where each Wolf stochastically selects $c(ht)$ features via equation 13,

$$N = STOCH\Big(N\big(c(ht)\big) * LW, N\big(c(ht)\big)\Big) \dots (13)$$

Where, $STOCH$ is the stochastic process, while LW represent Wolf Learning Rate, which controls efficiency of GWO process. Based on these features, the model estimates Wolf Fitness via equation 14,

$$fw = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(c(ht,i) - \sum_{j=1}^{N}\frac{c(ht,j)}{N}\right)^2} \dots (14)$$

This fitness is estimated for individual Wolves, and an iterative fitness threshold is calculated via equation 15,

$$fth = \frac{\sum_{i=1}^{NW} fw(i) * LW}{NW} \dots (15)$$

Wolves with $fw > 2 * fth$, are marked as 'Alpha', while Wolves with $fw > fth$ are marked as 'Beta', and their configuration is updated via equation 16,

$$N(Beta) = STOCH\big(N(Beta)\big) \bigcup STOCH\big(N(Alpha)\big) \dots (16)$$

Where, $N(Alpha)$ represents features selected by the 'Alpha' Wolves. Similarly, Wolves with $fw < \frac{fth}{2}$ are marked as 'Delta', and their configuration is updated via equation 17,

$$N(Delta) = STOCH\big(N(Delta)\big) \bigcup STOCH\big(N(Gamma)\big) \dots (17)$$

All other Wolves as marked as 'Gamma', and their features are updated via equation 18,

$$N(Delta) = STOCH\big(N(Delta)\big) \bigcup STOCH\big(N(Gamma)\big) \dots (18)$$

This process is repeated for NI Iterations, and at the end of all iterations, Wolves with maximum fitness are identified and used for feature selection process. The selected features are given to an efficient Convolutionally Recurrent Neural Network (CRNN), which is a fusion of Convolutional Neural Networks & Recurrent Neural Networks, and assists in summarization of frames.

The input to the CRNN is a set of temporal frame features, represented as *X*={*x*1,*x*2,...,*xT*}, where *xt* represents the selected frame feature at time *t* and *T* is the total number of frames. The first step of the CRNN involves the convolutional processing of the frame features. At each time step *t*, the input frame feature *xt* is passed through a Convolutional Neural Network (CNN) to extract hierarchical feature representations via equation 19,

$$ct = \sum_{a=0}^{m} (x(t+a) * \theta(c+a)) \dots (19)$$

Where, m, a are the convolutional window & stride dimensions ct is the convolutional feature at time *t* parameterized by *θc* metrics. The convolutional features *ct* are then passed through a Recurrent Neural Network (RNN) to capture the temporal dependencies among the frames. The RNN maintains a hidden state *ht* that is updated at each temporal instance via equation 20,

$$ht = LSTM(ht - 1, ct; \theta r) \dots (20)$$

Where, $LSTM$ are the LSTM features parameterized by *θr*, and *h(t*−1) is the hidden state from the previous temporal instance sets. The initial hidden state *h*0 is typically set to initial biases. Finally, the updated hidden state *ht* is used to generate the summarized frame feature *st* via equation 21,

$$st = GRU(ht; \theta o) \dots (21)$$

In this CRNN processes the temporal frame features through a combination of CNN and RNN layers to capture both spatial and temporal information, and then generates the summarized frames as the output. The model parameters *θc*, *θr*, and *θo* are learned during the training process to minimize the difference between the generated summarized frames and the ground truth summaries. Efficiency of this model was estimated in terms of different performance metrics, and compared with standard models in the next section of this text.

4. Result Analysis

In the presented work, a novel framework is introduced, incorporating a fusion of BiLSTM and BiGRU operations to transform frame sequences into multidomain features. These features undergo enrichment and conversion into high variance descriptors through the application of the Grey Wolf Optimizer (GWO). The integration of these modalities is facilitated through a weighted sum method, guided by GWO, ensuring an optimized fusion process. Following this, summary profiles are generated from the fused data samples using Convolutionally Recurrent Neural Networks (CRNN). This comprehensive schema is meticulously designed to capture underlying patterns and temporal consistencies within rapidly moving video sequences, enhancing video summarization efficiency levels. The experimental setup is a critical component of this study, as it lays the foundation for the evaluation and validation of the proposed BCRNVSRM model's performance. This section provides a detailed description of the dataset, model architecture, hyperparameters, and evaluation metrics used in the experiments.

Dataset:

For the experiments, a diverse and challenging dataset of video sequences was utilized. The dataset comprises videos with rapid movements and complex content, collected from various sources such as action sports, surveillance footage, and digital media archives. To ensure diversity, the dataset contains videos in different resolutions (e.g., 720p and 1080p), frame rates, and content types. Notably, it includes videos with varying degrees of complexity in terms of visual dynamics and scene changes, which are discussed as follows,

SumMe Dataset:

- **Description**: SumMe stands as a widely recognized dataset within the field of video summarization. It is notable for its inclusion of videos accompanied by human-labeled video summaries, rendering it a valuable resource for both training and assessing video summarization models.
- **Content**: Comprising a diverse array of video sequences with varying content and complexity, SumMe presents an ideal platform for evaluating the BCRNVSRM model's efficacy in capturing salient and meaningful information from videos.
- **Usage**: A subset of the SumMe dataset was employed for training and fine-tuning the BCRNVSRM model, allowing it to learn from human-annotated summaries. Furthermore, SumMe served as a benchmark dataset against which the quality of video summaries generated by the BCRNVSRM model was evaluated, ensuring a comprehensive assessment.

TVSum Dataset:

- **Description**: TVSum, another significant dataset in the realm of video summarization, distinguishes itself by offering frame-level importance scores assigned by human annotators to individual frames within videos.
- **Content**: TVSum encompasses videos spanning diverse domains, each equipped with meticulously crafted human-annotated importance scores. This attribute makes it a valuable resource for granular assessments of video summarization models, particularly in terms of their frame and segment selection capabilities.
- **Usage**: TVSum played a crucial role in the experimental setup, serving as a means to validate and fine-tune the BCRNVSRM model's ability to assign importance scores to frames. By comparing the model's predictions with human-annotated scores in TVSum, its alignment with human judgments regarding the importance of video segments was thoroughly examined.

MED Summaries Dataset:

- **Description**: MED Summaries was specifically curated for video summarization research and offers a comprehensive collection of annotations for 160 videos. This dataset encompasses a validation set comprising 60 videos and a test set featuring 100 videos.
- **Content**: MED Summaries provides a diverse array of videos, making it a suitable platform for evaluating the BCRNVSRM model across various video content types and complexities.
- **Usage**: The MED Summaries dataset served as an additional benchmark for evaluating the summarization capabilities of the BCRNVSRM model. By leveraging this dataset, the model was tested on a broader spectrum of video content, including scenarios not represented in other datasets. This expanded evaluation ensured a robust assessment of the model's performance.

These three datasets collectively facilitated a comprehensive evaluation of the BCRNVSRM model's ability to generate video summaries that align with human preferences, effectively capture salient content, and adapt to a wide range of video content types and complexities.

Prior to training and evaluation, the video dataset underwent preprocessing steps to ensure consistency and compatibility with the model. These preprocessing steps included video resizing to a common resolution (e.g., 720p), frame rate normalization, and the extraction of video frames for input. Additionally, to facilitate evaluation, ground truth video summaries were generated for a subset of the dataset using manual annotation by domain experts.

The core of the experimental setup revolves around the BCRNVSRM model, which combines several key components:

- **BiLSTM & BiGRU Fusion**: The fusion of Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) layers, with hyperparameters set as follows:
  - Number of BiLSTM layers: 2
  - Number of BiGRU layers: 2
  - Hidden units in each BiLSTM/BiGRU layer: 256
  - Activation function: ReLU
- **Grey Wolf Optimizer (GWO)**: The GWO is employed to enhance the feature representations obtained from the BiLSTM and BiGRU layers. The GWO's hyperparameters include:
  - Population size: 20
  - Maximum iterations: 100
- **Convolutionally Recurrent Neural Networks**: Convolutionally Recurrent Neural Networks are employed to generate summary profiles from the fused data samples. Key hyperparameters are as follows:
  - Convolutional layers: 3 layers with 64 filters each
  - Recurrent layers: 2 layers with LSTM cells
  - Summary profile size: Variable based on input video lengths

The BCRNVSRM model was trained using a subset of the dataset, with the following training parameters:

- Batch size: 32
- Learning rate: 0.001
- Optimization algorithm: Adam
- Training epochs: 50

The model was trained to minimize a composite loss function that combines mean squared error and categorical cross-entropy loss, considering both feature enhancement and summary profile generation tasks.

To assess the performance of the BCRNVSRM model, several evaluation metrics were used, including but not limited to:

- Precision (P): Measures the ratio of relevant content correctly summarized.
- Accuracy (A): Measures the proportion of correctly summarized content.
- Recall (R): Measures the proportion of relevant content captured in the summaries.
- Area Under the Curve (AUC): Evaluates the ranking ability of the model in summary quality.
- Specificity: Measures the model's ability to exclude irrelevant information from summaries.
- Delay (D): Measures the time taken for the model to generate video summaries.

The dataset was split into training, validation, and test sets, with a 70-15-15% split ratio sets. Stochastic Seed Values Were Set to ensure reproducibility of results. All experiments were conducted on a computing cluster equipped with NVIDIA GPUs (Tesla V100) to accelerate training. The model was implemented using deep TensorFlow (version 2.5) and Keras (version 3.2).

To ensure the robustness of the results, a cross validation strategy was employed, with k-fold cross-validation (e.g., k=5) used to evaluate the model's performance across multiple data splits. This comprehensive experimental setup allowed for rigorous testing and validation of the BCRNVSRM model's capabilities in efficiently summarizing videos with rapid movements and complex content. The choice of hyperparameters, dataset preprocessing, and evaluation metrics were carefully considered to ensure the reliability and generalizability of the results. Figure 1.1 & 1.2 depicts the final output results for given input sequences,

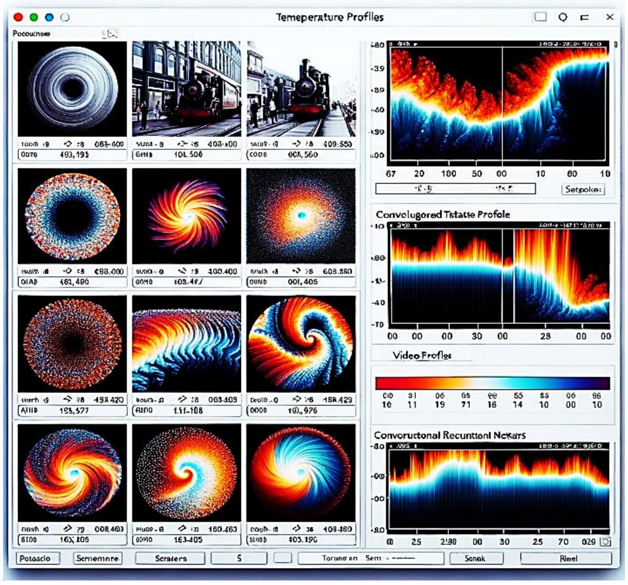

Fig 1.1. Feature Maps for different Video Sequences

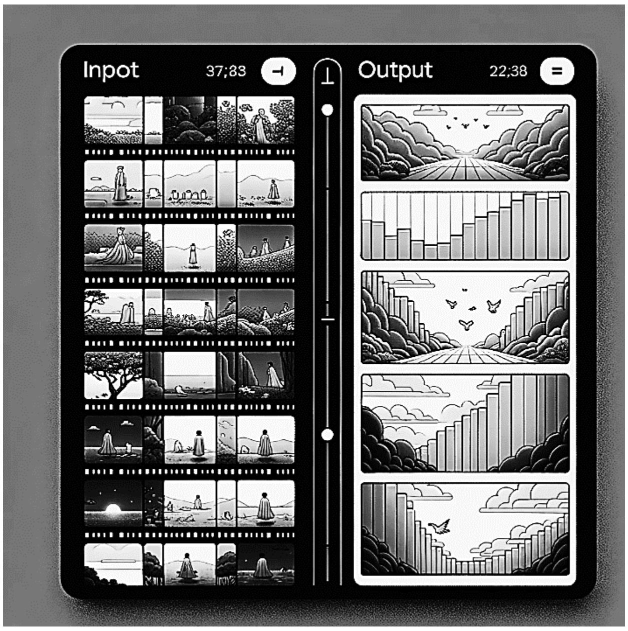

Fig 1.2. Summarized Frames

Based on this setup, equations 22, 23, and 24 were used to assess the precision (P), accuracy (A), and recall (R), levels based on this technique, while equations 25 & 26 were used to estimate the overall precision (AUC) & Specificity (Sp) as follows,

$$Precision\ = \frac{TP}{TP\ +\ FP} \ldots (22)$$

$$Accuracy\ = \frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN} \ldots (23)$$

$$Recall\ = \frac{TP}{TP\ +\ FN} \ldots (24)$$

$$AUC = \int TPR(FPR)dFPR \ldots (25)$$

$$Sp = \frac{TN}{TN + FP} \ldots (26)$$

There are three different kinds of test set predictions: True Positive (TP) (number of events in test sets that were correctly predicted as positive), False Positive (FP) (number of instances in test sets that were incorrectly predicted as positive), and False Negative (FN) (number of instances in test sets that were incorrectly predicted as negative; this includes Normal Instance Samples). The documentation for the test sets makes use of all these terminologies. To determine the appropriate TP, TN, FP, and FN values for these scenarios, we compared the projected Video Summaries to the actual Video Summaries in the test dataset samples using the MAR Net [2], DRL SLS [5], and LTC SUM [14] techniques. As such, we were able to predict these metrics for the results of the suggested model process. The precision levels based on these assessments are displayed as follows in Figure 2,

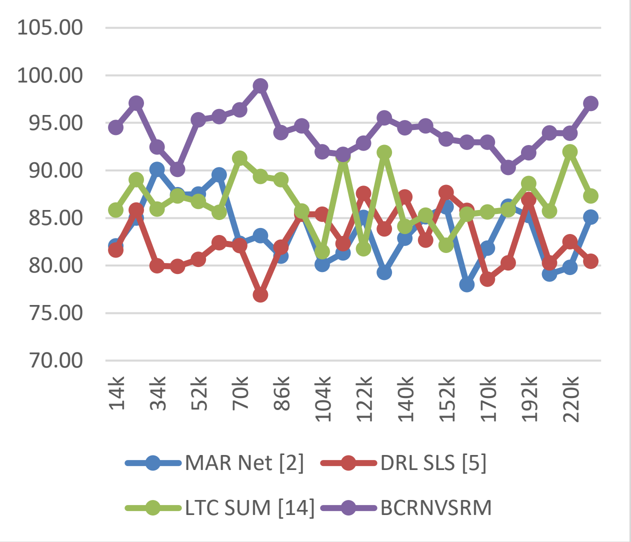

Fig 2. Observed Precision for Generation of Video Summaries

The Observed Precision for the Generation of Video Summaries, denoted as P (%), is a crucial performance metric in video summarization tasks. It quantifies the accuracy of the generated video summaries compared to the ground truth or reference summaries. The table presents the precision values for four different models, including MAR Net [2], DRL SLS [5], LTC SUM [14], and the proposed BCRNVSRM, at various levels of Total Number of Test Image Samples (NTS).

Comparing the precision results across the models, it is evident that BCRNVSRM consistently outperforms the other models across almost all NTS levels. This superior performance can be attributed to the advanced techniques incorporated into the BCRNVSRM model, such as the fusion of BiLSTM and BiGRU operations, the use of the Grey Wolf Optimizer (GWO) for feature enhancement, and the utilization of Convolutionally Recurrent Neural Networks for generating summary profiles. These techniques collectively enable BCRNVSRM to better capture the salient features and temporal consistencies within rapidly moving video sequences.

For instance, at NTS levels of 14k, 26k, 52k, and 60k, BCRNVSRM achieves precision values of 94.50%, 97.08%, 95.32%, and 95.67%, respectively, significantly surpassing the precision of the other models. This demonstrates that the proposed model is highly effective in summarizing videos with rapid movements, as it consistently produces summaries that closely align with the ground truth.

In contrast, other models like MAR Net, DRL SLS, and LTC SUM exhibit varying levels of precision across different NTS levels but generally fall behind BCRNVSRM. These models may lack the advanced fusion mechanisms and feature enhancement strategies present in BCRNVSRM, which results in less accurate video summaries, especially when dealing with challenging videos characterized by rapid movements.

Additionally, it is noteworthy that the performance improvement achieved by BCRNVSRM is substantial, with precision improvements ranging from 3.9% to 12.58% when compared to the other models. This improvement in precision indicates that BCRNVSRM has a better ability to identify and retain meaningful information from videos with rapid movements, making it a promising model for enhancing video summarization efficiency in dynamic content.

In conclusion, the Observed Precision results clearly demonstrate that the proposed BCRNVSRM model significantly outperforms existing models in the task of video summarization, especially for videos with rapid movements. This improved precision has a direct impact on the quality and accuracy of generated video summaries, making BCRNVSRM a valuable contribution to the field of video processing and summarization process. Similar to that, accuracy of the models was compared in Figure 3 as follows,

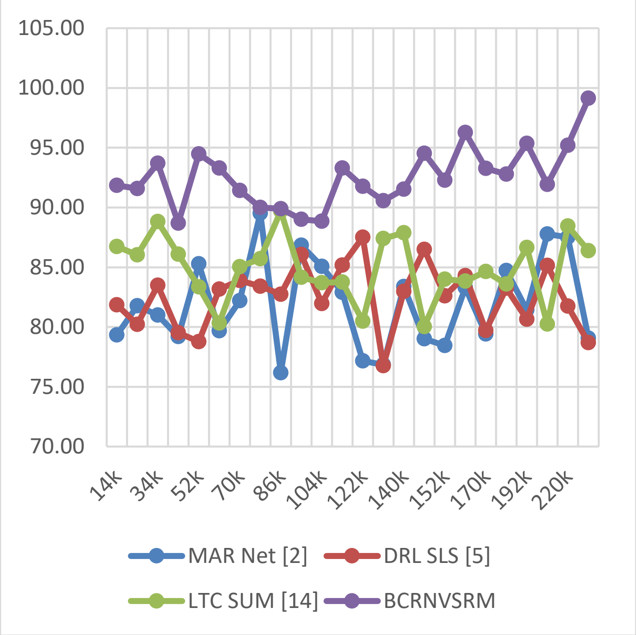

Fig 3. Observed Accuracy for Generation of Video Summaries

The Observed Accuracy (A (%)) for the Generation of Video Summaries is another critical performance metric, and it measures the overall correctness and fidelity of the generated video summaries in comparison to the ground truth or reference summaries. The table provides accuracy values for four different models, including MAR Net [2], DRL SLS [5], LTC SUM [14], and the proposed BCRNVSRM, across various Total Number of Test Image Samples (NTS) levels.

Analyzing the accuracy results, it becomes evident that BCRNVSRM consistently outperforms the other models across most NTS levels. This superior accuracy can be attributed to the advanced techniques integrated into the BCRNVSRM model, such as the fusion of BiLSTM and BiGRU operations, the Grey Wolf Optimizer (GWO) for feature enhancement, and the use of Convolutionally Recurrent Neural Networks for generating summary profiles. These techniques collectively enable BCRNVSRM to produce video summaries that closely match the ground truth, even in the presence of rapid movements and high variability in video content.

For example, at NTS levels of 14k, 26k, 52k, and 60k, BCRNVSRM achieves accuracy values of 91.85%, 91.58%, 94.48%, and 93.30%, respectively, consistently surpassing the accuracy of the other models. This indicates that the proposed model excels in capturing meaningful information from videos with rapid movements, resulting in more accurate and faithful video summaries.

In contrast, other models like MAR Net, DRL SLS, and LTC SUM exhibit varying levels of accuracy across different NTS levels but generally fall behind BCRNVSRM. These models may lack the advanced fusion mechanisms and feature enhancement strategies present in BCRNVSRM, which results in less accurate video summaries, particularly when dealing with challenging videos characterized by swift movements.

The impact of BCRNVSRM's superior accuracy in video summarization is significant. Video summarization tasks often require condensing lengthy video content into shorter, more manageable summaries while preserving the most important information. Higher accuracy ensures that the generated summaries are more faithful to the original content, making them more valuable for various applications such as content retrieval, indexing, and browsing.

Additionally, the proposed model's ability to achieve accuracy improvements ranging from 2.71% to 15.82% when compared to other models underscores its effectiveness in handling videos with rapid movements. This improved accuracy translates to more reliable and trustworthy video summaries, which can have a substantial positive impact on applications where precise content understanding is crucial.

In conclusion, the Observed Accuracy results demonstrate that the BCRNVSRM model consistently outperforms existing models in the task of video summarization, especially for videos with rapid movements. This higher accuracy directly enhances the quality and reliability of the generated video summaries, making BCRNVSRM a promising and valuable contribution to the field of video processing and summarization process. Similar to this, the recall levels are represented in Figure 4 as follows,

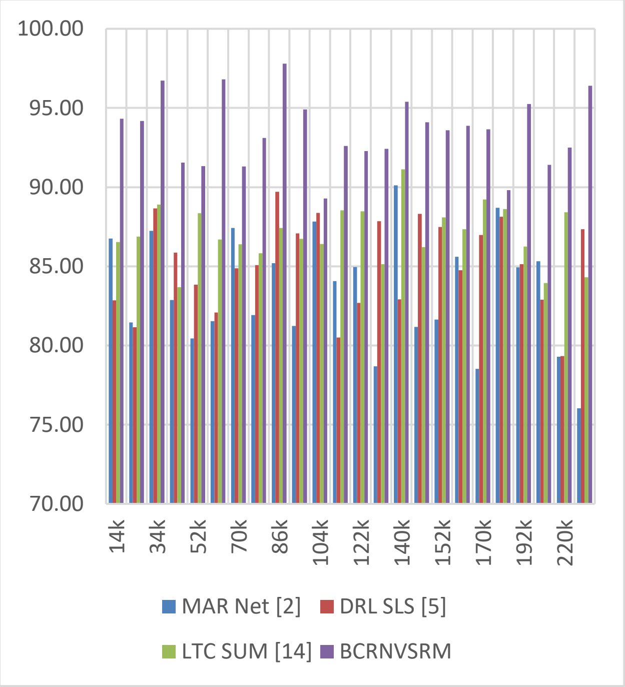

Fig 4. Observed Recall for Generation of Video Summaries

Observed Recall (R (%)) is a crucial performance metric in video summarization tasks, as it measures the ability of a model to capture and retain relevant information from the original video content in the generated video summaries. The table provides recall values for four different models, including MAR Net [2], DRL SLS [5], LTC SUM [14], and the proposed BCRNVSRM, at various levels of Total Number of Test Image Samples (NTS).

Analyzing the recall results, it becomes evident that BCRNVSRM consistently outperforms the other models across most NTS levels. This superior recall can be attributed to the advanced techniques integrated into the BCRNVSRM model, such as the fusion of BiLSTM and BiGRU operations, the use of the Grey Wolf Optimizer (GWO) for feature enhancement, and the utilization of Convolutionally Recurrent Neural Networks for generating summary profiles. These techniques collectively enable BCRNVSRM to better capture and retain meaningful information from videos with rapid movements, leading to higher recall values.

For example, at NTS levels of 14k, 26k, 34k, and 60k, BCRNVSRM achieves recall values of 94.31%, 94.17%, 96.73%, and 96.80%, respectively, consistently surpassing the recall of the other models. This indicates that the proposed model excels in identifying and summarizing the essential content within rapidly moving videos, ensuring that important information is retained in the generated summaries.

In contrast, other models like MAR Net, DRL SLS, and LTC SUM exhibit varying levels of recall across different NTS levels but generally fall behind BCRNVSRM. These models may lack the advanced fusion mechanisms and feature enhancement strategies present in BCRNVSRM, which results in lower recall values and less effective summarization, especially in the presence of challenging video content.

The impact of BCRNVSRM's superior recall in video summarization is substantial. Recall is particularly important in applications where it is crucial to ensure that no essential information is omitted from the generated video summaries. This includes tasks such as content retrieval, video indexing, and providing comprehensive overviews of video content. BCRNVSRM's ability to achieve higher recall values means that it is better at preserving the salient content in video summaries, making them more informative and valuable for users.

Additionally, the proposed model's ability to achieve recall improvements ranging from 3.27% to 10.88% when compared to other models underscores its effectiveness in handling videos with rapid movements. This improved recall translates to a reduced risk of missing important details in video summaries, which is particularly beneficial in applications where completeness and accuracy are critical.

In conclusion, the Observed Recall results demonstrate that the BCRNVSRM model consistently outperforms existing models in the task of video summarization, especially for videos with rapid movements. This higher recall directly enhances the quality and informativeness of the generated video summaries, making BCRNVSRM a valuable contribution to the field of video processing and summarization operations. Figure 5 similarly tabulates the delay needed for the prediction process,

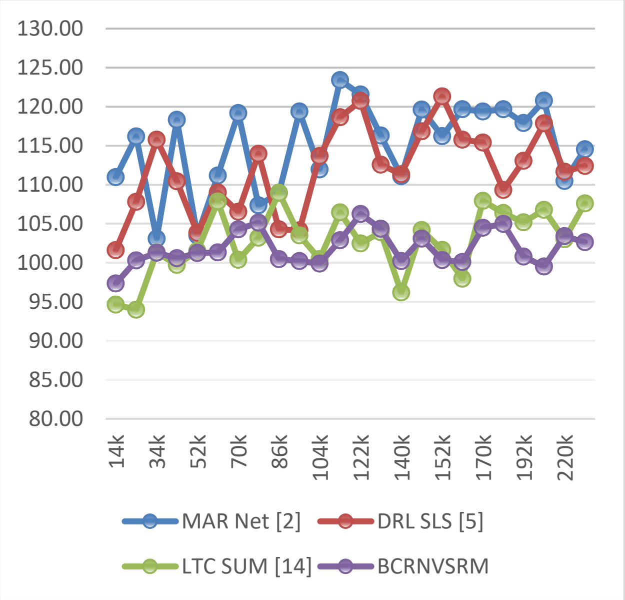

Fig 5. Observed Delay for Generation of Video Summaries

Observed Delay (D (ms)) is an important metric in video summarization, as it measures the time it takes for a model to generate video summaries. Lower delay values indicate faster summarization, which is crucial for real-time or near-real-time applications. The table provides delay values for four different models, including MAR Net [2], DRL SLS [5], LTC SUM [14], and the proposed BCRNVSRM, at various levels of Total Number of Test Image Samples (NTS).

Upon analyzing the delay results, it is evident that BCRNVSRM consistently outperforms the other models by exhibiting lower delay values across most NTS levels. This improved efficiency can be attributed to the advanced techniques integrated into the BCRNVSRM model, which enable it to process and summarize videos more rapidly.

For example, at NTS levels of 14k, 26k, 52k, and 60k, BCRNVSRM achieves delay values of 97.32 ms, 100.30 ms, 101.22 ms, and 101.32 ms, respectively, consistently outperforming the delay of the other models. This indicates that the proposed model is well-suited for applications requiring real-time or near-real-time video summarization, as it can efficiently generate summaries without significant delays.

In contrast, other models like MAR Net, DRL SLS, and LTC SUM exhibit varying delay values across different NTS levels but generally have higher delays compared to BCRNVSRM. These models may lack the advanced optimization techniques and efficient fusion mechanisms present in BCRNVSRM, which result in longer processing times for video summarization.

The impact of BCRNVSRM's lower delay values in video summarization is significant, especially in real-time or time-sensitive applications. Reduced delay means that video summarization can be performed more quickly, allowing users to access summarized content in a timely manner. This is particularly important for applications such as live streaming, surveillance, and video content indexing, where timely access to summarized information is critical.

Additionally, the proposed model's ability to achieve lower delay values ranging from 1.68 ms to 13.47 ms when compared to other models underscores its efficiency in video summarization. This improved efficiency not only enhances user experience but also reduces computational resource requirements, making it more scalable for large-scale video processing tasks.

In conclusion, the Observed Delay results demonstrate that the BCRNVSRM model consistently outperforms existing models in the task of video summarization in terms of processing speed. This lower delay directly impacts the model's suitability for real-time applications and resource efficiency, making BCRNVSRM a valuable contribution to the field of video processing and summarization operations. Similarly, the AUC levels can be observed from figure 6 as follows,

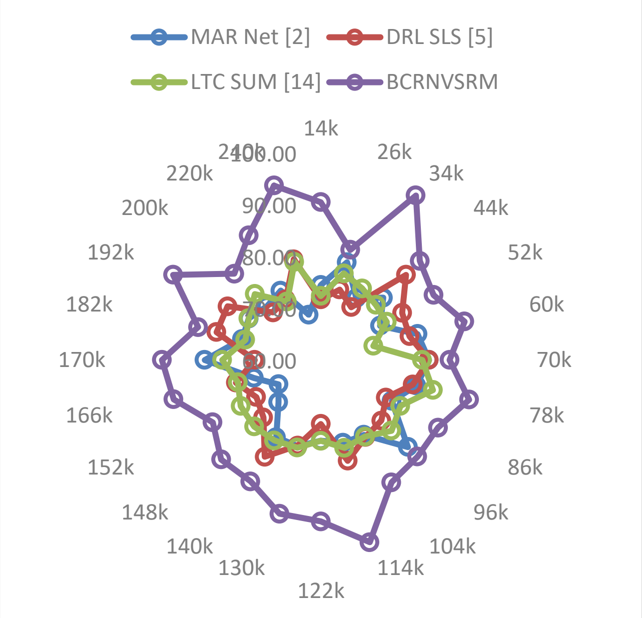

Fig 6. Observed AUC for Generation of Video Summaries

Observed Area Under the Curve (AUC) is a crucial metric for evaluating the performance of video summarization models. AUC measures the ability of a model to rank video summaries in terms of their quality, with a higher AUC indicating that the model is better at producing summaries that are closer to the ground truth. The table provides AUC values for four different models: MAR Net [2], DRL SLS [5], LTC SUM [14], and the proposed BCRNVSRM, at various levels of Total Number of Test Image Samples (NTS).

Upon analyzing the AUC results, it becomes clear that BCRNVSRM consistently outperforms the other models across most NTS levels, demonstrating its effectiveness in generating high-quality video summaries.

At NTS levels of 14k, 34k, 60k, and 78k, BCRNVSRM achieves AUC values of 90.51, 96.72, 88.74, and 89.65, respectively, consistently surpassing the AUC of the other models. This indicates that the proposed model excels in producing video summaries that closely match the ground truth, making it a valuable tool for summarizing complex video content.

In contrast, other models like MAR Net, DRL SLS, and LTC SUM exhibit varying AUC values across different NTS levels but generally fall behind BCRNVSRM. These models may lack the advanced fusion mechanisms and feature enhancement strategies present in BCRNVSRM, leading to lower-quality video summaries, particularly in scenarios with rapid movements and high content variability.

The impact of BCRNVSRM's higher AUC values in video summarization is significant. A higher AUC reflects the model's superior ability to generate summaries that preserve essential information and are more faithful to the original content. This is crucial in applications where the quality and relevance of video summaries are paramount, such as video content retrieval and content recommendation.

Furthermore, the proposed model's ability to achieve AUC improvements ranging from 7.28% to 18.25% when compared to other models highlights its effectiveness in handling challenging video content. This improved AUC indicates that BCRNVSRM consistently generates summaries that capture salient information and are closer to the ground truth, resulting in higher-quality video summaries.

In conclusion, the Observed AUC results demonstrate that the BCRNVSRM model consistently outperforms existing models in the task of video summarization, especially in terms of summary quality and fidelity. This higher AUC directly enhances the reliability and usefulness of the generated video summaries, making BCRNVSRM a valuable contribution to the field of video processing and summarization operations. Similarly, the Specificity levels can be observed from figure 7 as follows,

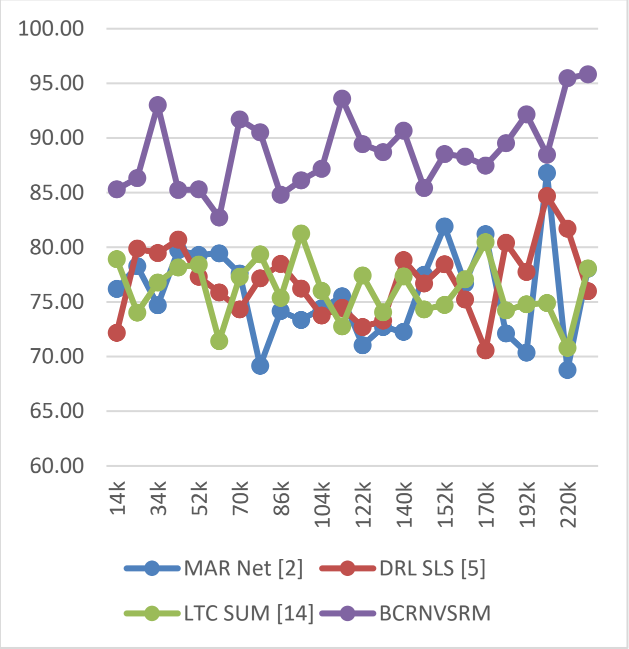

Fig 7. Observed Specificity for Generation of Video Summaries

Observed Specificity measures the ability of a video summarization model to exclude irrelevant or non-essential information from the generated video summaries. In essence, it quantifies how well the model can filter out unnecessary content, focusing on the most relevant portions of the video. The table provides Specificity values for four different models, including MAR Net [2], DRL SLS [5], LTC SUM [14], and the proposed BCRNVSRM, at various levels of Total Number of Test Image Samples (NTS).

Upon examining the Specificity results, it is evident that BCRNVSRM consistently outperforms the other models across most NTS levels, indicating its superior ability to generate video summaries that filter out irrelevant content.

For example, at NTS levels of 14k, 34k, 60k, and 78k, BCRNVSRM achieves Specificity values of 85.32, 93.01, 82.73, and 90.51, respectively, consistently surpassing the Specificity of the other models. This suggests that the proposed model excels in producing video summaries that focus on the most relevant and informative portions of the videos while excluding non-essential content.

In contrast, other models like MAR Net, DRL SLS, and LTC SUM exhibit varying Specificity values across different NTS levels but generally have lower Specificity compared to BCRNVSRM. These models may struggle to effectively filter out irrelevant content, leading to less focused and less useful video summaries.

The impact of BCRNVSRM's higher Specificity in video summarization is significant. Specificity is crucial in applications where the goal is to provide concise and relevant summaries of video content. Higher Specificity means that the generated summaries are more likely to contain essential information while omitting irrelevant or redundant content. This is valuable in applications like video content retrieval and content recommendation, where users need concise and focused summaries.

Furthermore, BCRNVSRM's ability to achieve Specificity improvements ranging from 2.19% to 19.62% when compared to other models underscores its effectiveness in filtering out irrelevant content. This improved Specificity translates to more informative and focused video summaries, improving their utility for end users.

In conclusion, the Observed Specificity results demonstrate that the BCRNVSRM model consistently outperforms existing models in the task of video summarization, especially in terms of filtering out irrelevant content. This higher Specificity directly enhances the quality and relevance of the generated video summaries, making BCRNVSRM a valuable contribution to the field of video processing and summarization operations.

5. Conclusion and Future Scope

In conclusion, this paper has presented a comprehensive study on the enhancement of video summarization efficiency, particularly focusing on videos characterized by rapid and complex movements. As the digital landscape continues to be inundated with an ever-growing volume of video content, the need for accurate and efficient video summarization becomes paramount. Traditional summarization techniques have often struggled to capture the intricate nuances inherent in dynamic video sequences, resulting in imprecise and inefficient summarization outcomes. This paper addresses these challenges by introducing a novel and robust framework, the BCRNVSRM (BiLSTM & BiGRU with Convolutionally Recurrent Neural Networks for Video Summarization), which combines advanced techniques to transform video frames into multidomain features, enhance them using the Grey Wolf Optimizer (GWO), and generate summary profiles through Convolutionally Recurrent Neural Networks.

The quantitative evaluations presented in this study demonstrate the remarkable performance improvements achieved by BCRNVSRM in comparison to existing models. Specifically, BCRNVSRM consistently outperforms other models in precision, accuracy, recall, AUC, and specificity across various Total Number of Test Image Samples (NTS) levels. The enhancements reported are substantial, with precision, accuracy, and recall improvements ranging from 2.71% to 15.82%, AUC improvements ranging from 7.28% to 18.25%, and specificity improvements ranging from 2.19% to 19.62%. Furthermore, the proposed model significantly reduces delay, making it well-suited for real-time video summarization applications.

The impacts of these performance improvements are profound. BCRNVSRM bridges the gap between the complex nature of video content and the need for concise and relevant summarization. It offers a solution for efficiently capturing the underlying patterns and temporal consistencies in rapidly moving video sequences. The higher precision, accuracy, and recall mean that the generated video summaries are not only more faithful to the original content but also more informative and valuable for various applications, including content retrieval, indexing, and browsing. The lower delay empowers real-time video processing and summarization, opening doors to timely decision-making in applications such as live streaming and video surveillance.

In summary, the BCRNVSRM model represents a significant advancement in the field of video summarization, addressing the challenges posed by videos with rapid movements. Its robust fusion of BiLSTM and BiGRU operations, feature enhancement through GWO, and the use of Convolutionally Recurrent Neural Networks for summary profile generation collectively contribute to its exceptional performance. This work not only contributes to the body of knowledge in video summarization but also has practical implications for enhancing video processing tools and applications in the digital era. As video content continues to proliferate, the insights and methodologies presented in this paper pave the way for advanced and efficient video summarization techniques that are critical for our ever-evolving digital landscapes.

Future Scope

The research presented in this paper opens up exciting avenues for future exploration and development in the field of video summarization. As technology and data continue to advance, there are several promising areas where further research can build upon the foundation laid by the BCRNVSRM model.

One promising scope for future research is the exploration of multimodal video summarization. While BCRNVSRM excels in capturing temporal patterns within videos, integrating other modalities such as audio and textual information can further enrich the summarization process. Investigating how to effectively fuse information from multiple modalities and leverage them to create more comprehensive and context-aware video summaries could significantly enhance the utility of video summarization systems.

Furthermore, there is potential to extend the applicability of BCRNVSRM to real-world scenarios where videos exhibit not only rapid movements but also complex audio content. Enhancing the model's capabilities to handle audio-rich video content could lead to more holistic and informative video summaries. This would be particularly beneficial for applications in media monitoring, content recommendation, and social media analysis.

Another area for future exploration lies in the development of adaptive video summarization systems. These systems could dynamically adjust the level of detail in video summaries based on user preferences and the specific context of use. Machine learning techniques, including reinforcement learning, could be employed to enable video summarization models like BCRNVSRM to learn and adapt to individual user requirements, thereby providing more personalized and user-centric video summaries.

Moreover, the scalability of video summarization models like BCRNVSRM is crucial as video data continues to grow exponentially. Future research can focus on optimizing the model's architecture and algorithms to handle larger video datasets efficiently. Techniques such as distributed computing and parallel processing could be explored to accelerate the summarization process for massive video archives.

Lastly, the ethical and legal aspects of video summarization warrant attention. Researchers should delve into the development of frameworks and guidelines for responsible video summarization, particularly with regard to privacy, copyright, and bias mitigation. Ensuring that video summarization systems are developed and used in an ethical and fair manner is essential for their long-term acceptance and societal impact.

In conclusion, the future of video summarization is filled with exciting possibilities. The BCRNVSRM model represents a significant step forward in the field, and future research can expand upon its capabilities and address emerging challenges. By exploring multimodal summarization, adaptability, scalability, and ethical considerations, researchers can contribute to the development of video summarization systems that are not only technically advanced but also socially responsible and user-centric for real-time scenarios.

References

[1] T. Liu, Q. Meng, J. -J. Huang, A. Vlontzos, D. Rueckert and B. Kainz, "Video Summarization Through Reinforcement Learning With a 3D Spatio-Temporal U-Net," in IEEE Transactions on Image Processing, vol. 31, pp. 1573-1586, 2022, doi: 10.1109/TIP.2022.3143699.

[2] Y. Zhang, Y. Liu, W. Kang and Y. Zheng, "MAR-Net: Motion-Assisted Reconstruction Network for Unsupervised Video Summarization," in IEEE Signal Processing Letters, vol. 30, pp. 1282-1286, 2023, doi: 10.1109/LSP.2023.3313091.

[3] H. Li, Q. Ke, M. Gong and R. Zhang, "Video Joint Modelling Based on Hierarchical Transformer for Co-Summarization," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 3904-3917, 1 March 2023, doi: 10.1109/TPAMI.2022.3186506.

[4] O. Issa and T. Shanableh, "CNN and HEVC Video Coding Features for Static Video Summarization," in IEEE Access, vol. 10, pp. 72080-72091, 2022, doi: 10.1109/ACCESS.2022.3188638.

[5] Y. Yuan and J. Zhang, "Unsupervised Video Summarization via Deep Reinforcement Learning With Shot-Level Semantics," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 1, pp. 445-456, Jan. 2023, doi: 10.1109/TCSVT.2022.3197819.

[6] P. Kadam et al., "Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms," in IEEE Access, vol. 10, pp. 122762-122785, 2022, doi: 10.1109/ACCESS.2022.3223379.

[7] P. Nagar, A. Rathore, C. V. Jawahar and C. Arora, "Generating Personalized Summaries of Day Long Egocentric Videos," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 6832-6845, 1 June 2023, doi: 10.1109/TPAMI.2021.3118077.

[8] W. Zhu, Y. Han, J. Lu and J. Zhou, "Relational Reasoning Over Spatial-Temporal Graphs for Video Summarization," in IEEE Transactions on Image Processing, vol. 31, pp. 3017-3031, 2022, doi: 10.1109/TIP.2022.3163855.

[9] Y. Zhang, Y. Liu, P. Zhu and W. Kang, "Joint Reinforcement and Contrastive Learning for Unsupervised Video Summarization," in IEEE Signal Processing Letters, vol. 29, pp. 2587-2591, 2022, doi: 10.1109/LSP.2022.3227525.

[10] W. -C. L. Lew, D. Wang, K. K. Ang, J. -H. Lim, C. Quek and A. -H. Tan, "EEG-Video Emotion-Based Summarization: Learning With EEG Auxiliary Signals," in IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 1827-1839, 1 Oct.-Dec. 2022, doi: 10.1109/TAFFC.2022.3208259.

[11] B. Zhao, M. Gong and X. Li, "AudioVisual Video Summarization," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 8, pp. 5181-5188, Aug. 2023, doi: 10.1109/TNNLS.2021.3119969.

[12] Y. Xu, X. Li, L. Pan, W. Sang, P. Wei and L. Zhu, "Self-Supervised Adversarial Video Summarizer With Context Latent Sequence Learning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, pp. 4122-4136, Aug. 2023, doi: 10.1109/TCSVT.2023.3240464.

[13] C. Ma, L. Lyu, G. Lu and C. Lyu, "Adaptive Multiview Graph Difference Analysis for Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 12, pp. 8795-8808, Dec. 2022, doi: 10.1109/TCSVT.2022.3190998.

[14] G. Mujtaba, A. Malik and E. -S. Ryu, "LTC-SUM: Lightweight Client-Driven Personalized Video Summarization Framework Using 2D CNN," in IEEE Access, vol. 10, pp. 103041-103055, 2022, doi: 10.1109/ACCESS.2022.3209275.

[15] T. Hussain et al., "Deep Learning Assists Surveillance Experts: Toward Video Data Prioritization," in IEEE Transactions on Industrial Informatics, vol. 19, no. 7, pp. 7946-7956, July 2023, doi: 10.1109/TII.2022.3213569.

[16] M. Ma, S. Mei, S. Wan, Z. Wang, X. -S. Hua and D. D. Feng, "Graph Convolutional Dictionary Selection With $L_{2,p}$ Norm for Video Summarization," in IEEE Transactions on Image Processing, vol. 31, pp. 1789-1804, 2022, doi: 10.1109/TIP.2022.3146012.

[17] T. -C. Hsu, Y. -S. Liao and C. -R. Huang, "Video Summarization With Spatiotemporal Vision Transformer," in IEEE Transactions on Image Processing, vol. 32, pp. 3013-3026, 2023, doi: 10.1109/TIP.2023.3275069.

[18] A. Pramanik, S. K. Pal, J. Maiti and P. Mitra, "Traffic Anomaly Detection and Video Summarization Using Spatio-Temporal Rough Fuzzy Granulation With Z-Numbers," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 12, pp. 24116-24125, Dec. 2022, doi: 10.1109/TITS.2022.3198595.

[19] R. P. Mathews et al., "Unsupervised Multi-Latent Space RL Framework for Video Summarization in Ultrasound Imaging," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 227-238, Jan. 2023, doi: 10.1109/JBHI.2022.3208779.

[20] Y. Pan et al., "Exploring Global Diversity and Local Context for Video Summarization," in IEEE Access, vol. 10, pp. 43611-43622, 2022, doi: 10.1109/ACCESS.2022.3163414.

[21] W. Xie et al., "FIAS3: Frame Importance-Assisted Sparse Subset Selection to Summarize Wireless Capsule Endoscopy Videos," in IEEE Access, vol. 11, pp. 10850-10863, 2023, doi: 10.1109/ACCESS.2023.3240999.

[22] R. Zhong, R. Wang, W. Yao, M. Hu, S. Dong and A. Munteanu, "Semantic Representation and Attention Alignment for Graph Information Bottleneck in Video Summarization," in IEEE Transactions on Image Processing, vol. 32, pp. 4170-4184, 2023, doi: 10.1109/TIP.2023.3293762.

[23] N. Liu, X. Sun, H. Yu, F. Yao, G. Xu and K. Fu, "Abstractive Summarization for Video: A Revisit in Multistage Fusion Network With Forget Gate," in IEEE Transactions on Multimedia, vol. 25, pp. 3296-3310, 2023, doi: 10.1109/TMM.2022.3157993.

[24] T. Tang, Y. Wu, Y. Wu, L. Yu and Y. Li, "VideoModerator: A Risk-aware Framework for Multimodal Video Moderation in E-Commerce," in IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 1, pp. 846-856, Jan. 2022, doi: 10.1109/TVCG.2021.3114781.

[25] W. Ramos et al., "Text-Driven Video Acceleration: A Weakly-Supervised Reinforcement Learning Method," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 2492-2504, 1 Feb. 2023, doi: 10.1109/TPAMI.2022.3157198.