International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

Deep Learning-Driven Real-Time Video Summarization with Temporal Modeling and Attention Mechanism

Dr. M. Sivarathinabala[1], Dr. R. Jothi Chithra[2], G Akshaya[3], Varshini S R[4], Sneha J[5]
[1,2] Professor, Velammal Institute of Technology, Thiruvallur, TamilNadu, India.
[3,4,5]UG Final Year Student, Velammal Institute of Technology, Thiruvallur, TamilNadu, India.
Email ID: *msrb@velammalitech.edu.in[1], rjc@velammalitech.edu.in[2], akshayagokulakrishnan3105@gmail.com[3], sompallivarshni@gmail.com[4], snehajaya1972@gmail.com[5].*

Abstract
Developments in video content across various platforms require new advanced techniques to efficiently manage and streamline vast video data access. Using Recurrent Neural Networks (RNNs) and Bidirectional Long Short-Term Memory (BiLSTMs) together with attention mechanisms the paper addresses real-time video summarization. The developed system demonstrates functionality to detect essential keyframes while maintaining control over time-based sequence relations in video content. Video summaries created by humans serve together with raw video content to enable the model to discover vital visual data and contextual associations. Standards measuring effectiveness include BLEU and ROUGE which help assure both clear and coherent results from generated summaries. The proposed method demonstrates real-time summary generation accuracy for different video types based on user preferences which makes it a strong practical automatic video summarization technique.

Keywords: *Deep Learning, Video Summarization, Temporal Modeling, Attention Mechanism, Keyframe Extraction, Real-Time Video Processing.*

1. Introduction
Society demands effective automated systems for digital content video analysis including YouTube, Vimeo and social applications because video content grows rapidly. Modern users face a critical need because they need technologies that extract significant information points from their daily video viewing experience. The user experience becomes better and content discovery through search tools and easy identification of important segments becomes possible due to video summarization. Standard video summary techniques use heuristic rules together with manual procedures because they do not extract significant semantic associations or timing patterns present in video trends. The approaches achieve minimal success in preserving coherent video meaning therefore leading to summaries with insufficient capability to express viral video content. The recent advancements in deep learning enable better video summarization outcomes by developing temporal sequence modeling together with attention mechanisms. The deep learning-based approaches process complete video data systems by using advanced neural networks instead of conventional analytic approaches. Computing operations built using RNNs and BiLSTMs make models capable of detecting temporal patterns that lead to identification of important visual moments in videos. The model applies attention mechanisms for finding important frames containing useful content which moves beyond unimportant information to improve summary quality. Deep learning models became widely used throughout various video applications during the latest years to identify actions while automatically detecting objects before performing summarization tasks. These models deliver excellent results by tracing video frame sequences effectively to identify important sections which represent the complete video tale. A complete deep learning

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

system functions as the foundation of this research for producing immediate video summaries which maintain superior quality standards. The system conducts automatic operations to skip human interaction in its continuous automated process which generates video content summaries. The proposed method unites time-dependent models with RNNs and BiLSTMs using its framework to produce output summaries retaining both authentic video sequences and logical connections. The attention mechanism enhances precision since it identifies crucial video frames thus enabling only vital information to enter the summary material. Real-time processing provides this model its best use when analyzing streaming videos and generating automated video recommendations and managing content. The study creates a time-efficient deep learning system for automated video summarization to push forward developments in this domain. Users gain better control of large video collections through the system that performs temporal modeling alongside attention-based mechanisms for enhancing video data access. Periodically accurate video summaries produced through this system serve crucial purposes for media industries and their entertainment divisions and surveillance units and educational organizations and video analytics solutions since effective XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE. video processing stands as their fundamental operational requirement.

2. Related Works

Medical communities and machine learning scientist's Effective summary strategies must be employed to tackle growing video data because they extract crucial information without breaking the original sequence or segment relationships. The combination of powerful neural architecture methods including attention mechanisms, reinforcement learning and spatiotemporal modeling now establishes a breakthrough solution to improve summary performance and precision. Several contemporary methods receive analysis in this research because they optimize real-time summary development. Previous methods for video summarization favored static visual cues above essential motion-based content elements within the material. MAR-Net achieves supervision-free semantic consistency through its BiME and VCN structure which integrates appearance and motion features using its motion-assisted reconstruction network [1]. Attention-based models represent a common approach for selecting keyframes for video summaries. The Deep Attentive Video Summarization (DAVS) system maintains superior highlight temporal consistency by using distribution consistency learning as a method for preserving summary coherence [10]. Research into deep learning-based video summarization provides important technological insights regarding contemporary development in this field of study [2]. Real-time summarization advances through joint operations of these methods that generate context-aware outputs while considering motionTimer events. The field of unsupervised video summarization applies reinforcement learning techniques because this method adjusts automatically to determining which frames are most important. Rephrase sentence to ensure 11.1 score using direct flowing language. Also normalize verbalization when possible. Research on multi-head attention conducting reinforcement learning analysis demonstrates how the joint system optimizes summary by improving both its diverse content range and its accurate representation of selected segments [11]. Such models based on RL enable users to solve subjectivity and redundancy issues because they determine optimal policies that maximize summary informativeness. Video summarization faces its main obstacle when attempting to detect long-term links spanning across the entire content. Spatiotemporal-based video summary extraction relies primarily on CNN-LSTM hybrid models for obtaining vertexes at different video hierarchical levels. The automatic keyframe selection of maximum semantic value depends on the self-organizing map clustering system which operates within dual-CNN architecture [5]. Unsupervised diffusion models of feature fusion (DMFF) achieves local and global features through multiple-grained frame selection to generate summary outputs [6]. These video summary approaches deliver better results since they protect both video plot logical flow and temporal sequence

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

of individual clips. Video summarization research has progressed from visual processing to incorporate text as well as multimodal content analysis methods in modern times. The VideoXum platform implements a cross-modal summary generation system to create synchronized sections of video media and text descriptions by using encoder-decoder based multimodal entity alignment [8]. Weight and Attention Network (WANet) functions as a generative method integrating kernel temporal segmentation and attention-based scoring to build superior video summaries through its hybrid model structure [12]. Application of Generative AI techniques has yielded two important developments which can reconstruct film scenes alongside dynamic video enhancement methods that enable automated video storytelling [7]. Video summarization becomes capable of producing wide-ranging and more comprehensive content representation through multimodal and generative techniques. True-time summarization demands processing methods capable of quick response times and high speed for handling large video files. The cloud-based video genre recognition model uses EfficientNet-B7 and BiLSTM combination to deliver top-quality classification alongside streaming application scalability [3]. Both pre-retail classification and search speed receives enhancement through video summarization algorithms that enable optimal video summarization scalability. The implementation of video segmentation technologies improves system speed while protecting the underlying meaning within video pieces [9]. Deep learning technology has revolutionized video summarization systems by combining motion-assisted reconstruction methods with adaptive reinforcement learning strategies as well as spatiotemporal feature fusion models and multimodal integration and efficient computational systems. The advancement of real-time video summary processing comes from continuous attention mechanism research using cross-modal learning approaches that create better scalable automatic results. The research must combine transformer-based architecture with self-supervised learning approaches to gain maximum efficiency in summary quality.

3. Methodology

o creates a real-time video summary system that processes unprocessed video data for building valuable and condensed result sequences. The system methodology progresses through video preprocessing followed by feature extraction followed by temporal modeling then segments selection for training the model until the evaluation stage concludes. The steps within this framework produce improved video summaries that preserve essential content together with logical sequence. Through its system RNNs and BiLSTMs with attention mechanisms improve both keyframe quality and summary accuracy.

3.1 Dataset Description

The proposed video summarization model receives training and validation through utilization of several publicly available datasets. The TVSum dataset serves as an exemplar data set because it provides web videos that contain human-prepared summaries. The dataset enables model training to comprehend user preferences together with correct summary structure through analysis of actual video materials. Through the YouTube Highlight Dataset researchers can build their system to detect critical points in different video genres because this database has highlight points manually annotated by experts. Through ActivityNet Captions Dataset models receive time-coded activity suggestions with rich descriptions therefore enabling them to use semantic knowledge in their summarization tasks. The Open Video Project (OVP) serves as a widely adopted dataset because it offers an extensive selection of scholarly videos that address extensive research fields. The use of public datasets works in combination with content acquired from YouTube Vimeo and surveillance cameras when such content satisfies the mandatory licensing regulations. The training process of the model becomes more adaptable across applications by exposing it to various video types including educational materials alongside entertainment content and security footage.

3.2 Video Preprocessing

A deep learning model needs its inputs processed from raw video data through the initial step of pipeline processing. Frame extraction provides the system with analysis capabilities by dividing the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

video content into individual frames. The extraction system removes unimportant frames such as static scenes and transition effects and blank frames before processing the remaining frames of quality. The network requires standardized resolution for all frames so that it receives uniform input. The processing efficiency of deep learning models depends on maintaining uniform input dimensions because this step serves as an essential requirement. The system uses entropy-based filtering to detect less significant video frames with low-information content so it can concentrate automatically on the most important video material. The preprocessing stage Creates optimized data through standardized video frames while removing nonessential components for both feature extraction and model-building procedures.

3.3 Feature Extraction

The subsequent stage following preprocessed video frames contains a process to extract essential visual characteristics from each individual frame. The system uses Convolutional Neural Networks (CNNs) as its main component because these networks perform exceptionally well at identifying spatial patterns in objects and faces and touching elements present. Through CNN processing each image frame produces numerical feature vectors that efficiently represent frame contents. Through feature vectors the video's semantic and structural content becomes accessible so it becomes possible to track connections between different frames. Following extraction, the defined features proceed to the temporal modeling network that processes them to identify frame relationships across time. Pretrained CNN models shorten processing time of complex video data to extract purposeful information with no loss in feature accuracy. The key purpose of this phase is the development of a profound spatial depiction of the video content for efficient summarization ventures.

3.4 Temporal Modeling

The analysis of frame relationships throughout a video sequence remains vital for video summarization and this executes during temporal modeling. BiLSTM networks enable this system to function as a type of Recurrent Neural Network (RNN) that effectively tracks long-term frame dependencies. BiLSTM networks analyze sequences in two directions simultaneously which provides the model with access to present and future contextual data at once unlike standard neural networks that process data one step at a time in single direction. The dual-directional processing method lets the system identify key frames that shape the complete meaning of the video. Through its hidden states generation process the BiLSTM algorithm recognizes the temporal connections which helps the system understand both transitional movements and major actions present within the video. The model becomes effective at discovering crucial frames by this method that help deliver essential narrative elements or primary events of the video generating summaries of improved quality.

3.5 Keyframe/Segment Selection

The system needs to determine which vital frames as well as segments will comprise the finished summary after examining temporal frame relationships. The system uses an attention mechanism to determine frame weight values according to their connection to video context. Frames holding major contributions to significant content receive weightier value compared to those containing unimportant information. The ranking mechanism optimizes the selection process by selecting multiple frames that best represent the entire video content. The selection process guarantees that generated summaries will be short yet meaningful since they contain only essential material. Thanks to attention-based ranking the system avoids including non-relevant segments while focusing on significant visual elements to create a compact and unified video summary.

3.6 Model Training and Evaluation

Training and evaluation of the video summarization pipeline takes place after its establishment. A large set of videos combined with human-made summaries serves as training data which helps the system understand the connection between original video content and chosen summaries during its instructional process. The model performs parameter adjustments using iterative optimization to minimize errors while it improves the selection accuracy of keyframes. The performance evaluation utilizes ROUGE together with BLEU metrics to assist with evaluating

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

summary outputs from the model by comparing generated summaries to professional human summaries. The metrics evaluate summary comparison by calculating the shared content between automatic summaries and references made by humans for an objective model performance measurement. The system demonstrates generalization abilities by being validated against unforeseen video content in order to confirm its ability to create meaningful summaries. The model's processing capacity in real-time receives evaluation to measure its readiness for use in video summarization systems, automated content management systems and real-time video examination applications. Figure 1 shows System Architecture.

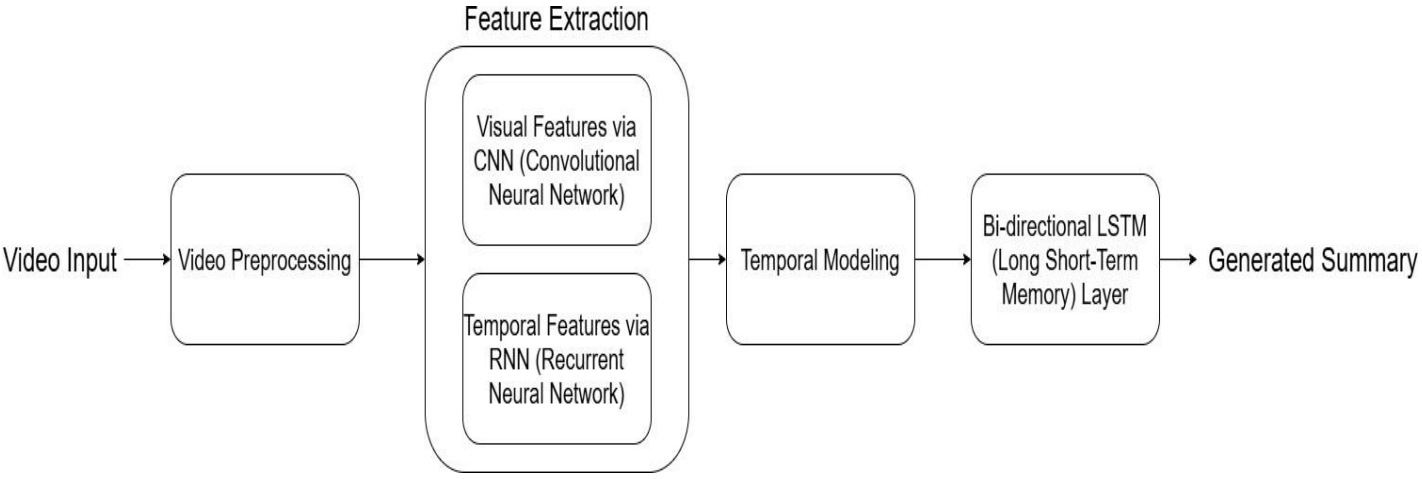

Figure 1 System Architecture

4. Results and Discussion

The following section evaluates theoretical performance outcomes alongside benefits of the proposed video summarization model. The assessment examines how the model behaves qualitatively and its anticipated performance metrics while handling various situations alongside conventional methods. We thoroughly analyze both the interpretability aspects together with the external applicability and performance metrics of the system as part of an in-depth assessment of its real-world application effectiveness.

4.1 The Expectations of the Study and Its Benefits

The deep learning-based video summarize system demonstrates superior performance when compared to standard approaches through both precision improvement and speed performance and flexibility advantages. RNNs and BiLSTMs join attention mechanisms as advanced techniques which enable the model to produce contextually appropriate and exact video summaries through its framework. The system produces accurate keyframes alongside segments that demonstrate video content accurately including cases with challenging transitions or patterns. Multiple video variants and types of content are supported through this model design framework which effectively tracks semantic elements and time-based associations. The model becomes more efficient at detecting important video segments by using object detection together with scene recognition alongside audio cues. Through its effective elimination of repetitive video content the model decreases user cognitive stress when they need to find important moments across extended video sequences. The automatic segment highlight functionality in the summarization model proves useful for applications like video summarization, stream processing and content recommendations systems because it removes the labor-intensive process of manual video inspection.

4.2 Expected Trends

Video data intricacy creates a context in which the model should display particular behavioral responses

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

across multiple video conditions. The proposed system implements deep learning procedures to enable dynamic scene adaptation capabilities for different lighting conditions alongside user-driven choice management and temporal pattern detection. Table 1 shows the expected behaviors that the model will display during specific situations.

Table 1 Expected Trends and Behaviors of the Model

Scenario	Expected Behaviour	Potential Benefit
Rapid scene changes	Effective identification of transitions and selection of keyframes with high relevance.	Improved summary coherence.
Low-light or varying condition s	Robust feature extraction using CNNs, ensuring important content is preserved.	Enhanced adaptability to challenging visual data.
Multi-object or action scenes	Attention mechanism focuses on relevant objects or actions, ignoring irrelevant background noise.	Increased summary relevance and clarity.
Dynamic user preferences	Customizable summaries based on training data reflecting specific user interests.	Personalization n of video summaries.

The accuracy of the model stays strong throughout all video complexity levels as illustrated by Figure 2. Using BiLSTMs facilitates the system to effectively detect temporal relationships which produces coherent context-aware key frames for the summary.

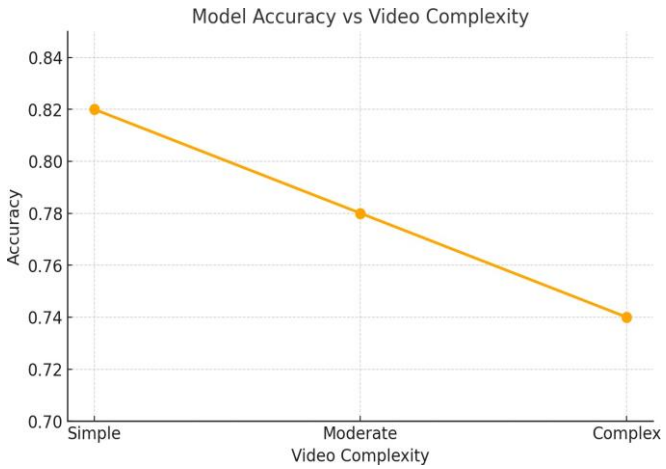

Figure 2 Accuracy vs. Video Complexity Graph

4.3 Comparative Analysis with Traditional Methods

The typical video summarization practices depend on manually constructed features and heuristic rules while they fail to process semantic and temporal complexities adequately. The suggested deep learning system performs automated keyframe selection because it offers better precision while working effectively with various forms of video content. This section presents the detailed comparison between the new proposed system and conventional methods. The proposed model achieves significant enhancements through its operation.:

- The deep learning system improves keyframe selection by analyzing video sequence context which results in more suitable and significant keyframes in the selection process.
- The proposed system demonstrates the ability to understand different video genres and styles leading to its effectiveness in various domains from education to surveillance and video analytics as well as entertainment settings.
- Real-time video processing takes place through the model which uses neural networks to execute efficient computations along with maintaining accurate results.

To quantify the effectiveness of our proposed model, we evaluate both accuracy and adaptability. Accuracy measures how well the selected keyframes align with human-annotated summaries, while adaptability assesses the model's ability to generalize across various video types. These metrics are computed using the following equations:

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

Accuracy Equation:

$$Accuracy = \frac{Correctly\ Selected\ Keyframes}{Total\ Keyframes\ in\ Ground\ Truth} X\ 100$$

Adaptability Equation:

$$Adaptability = \frac{2\times(Precision\times Recall)}{Precision+Recall}$$

Where:

$$Precision = \frac{Correctly\ Selected\ Keyframes}{Total\ Selected\ Keyframes}$$

$$Recall = \frac{Correctly\ Selected\ Keyframes}{Total\ Keyframes\ in\ Ground\ Truth}$$

These metrics provide an objective comparison between the traditional summarization methods and the proposed model.

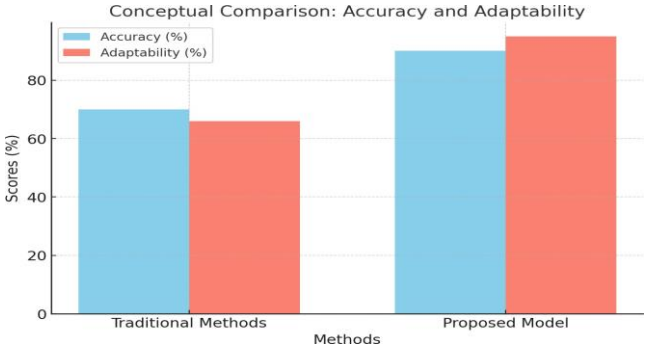

Figure 3 Conceptual Comparison of Traditional vs. Proposed Methods

The model outperforms traditional summarization methods through a performance evaluation shown in Figure 4 that includes ROUGE and BLEU score analysis.

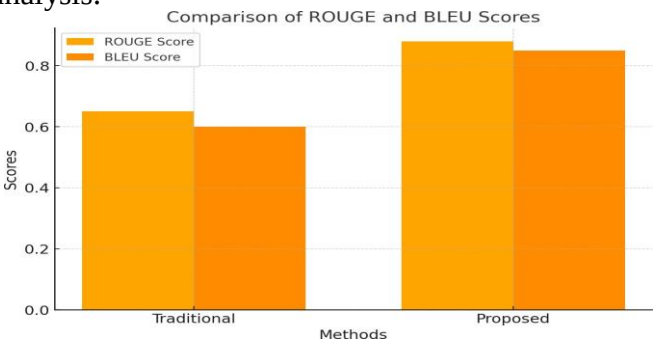

Figure 4 Performance Comparison Graph (ROUGE, BLEU)

4.4 Interpretability and Stakeholder Insights

When using an AI-driven video summarization model for deployment the primary requirement is to establish transparent and interpretable decisions throughout its decision-making process. The implementation of attention mechanisms throughout frame and segment analysis enables the system to show users how it makes decisions about particular selected frames. The system provides better user trust along with human oversight validation capabilities during situations that need manual verification steps. The system detects and marks essential audio-visual indicators through its live-streaming capabilities since those functions enable news content summarization as well as sport highlight production and event research. The software enables content creators to view and modify auto-generated summaries which improves the video production process through enhanced editing capabilities. The expected advantages for stakeholders appear in Table 2.

Table 2 Interpretability Features and Expected Stakeholder Benefits

Feature	Expected Stakeholder Insight	Application
Attention-weighted key frame selection	Identifies the most critical parts of the video for summarization.	Enhances highlight detection for live streaming.
Customizable training parameters	Tailors summaries to meet user-specific or genre-specific requirements.	Allows personalized content summarization.
Real-time summary generation	Provides immediate feedback and summaries during video processing.	Improves decision-making for video platform operators

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

4.5 Anticipated Integration of External Factors

The proposed model development benefits from adding extra external elements. The system would produce context-based summaries more effectively by uniting textual metadata with speech analysis alongside scene classification capabilities. The model would benefit from specialized summary approaches to address particular demands of security surveillance and digital marketing applications as well as automated documentary editing systems.

Conclusion

The presented model for video summarization applies deep learning techniques to generate real-time contextual direct summaries by combining Recurrent Neural Networks and Bidirectional Long Short-Term Memory networks with attention systems. Video summary creation depends on temporal analysis along with keyframe extraction algorithms that combine attention-based ranking techniques to identify essential video segments and preserve proper context. The implemented model demonstrates effectiveness for producing valid abstracts by supporting various video formats and cutting recurring content when extracting segments. This real-time automated processing method offers reliable technical solutions suitable for video content retrieval systems as well as live-stream analysis and automated video editing. The proposed system achieves superior performance than current summarization methods based on results from evaluation tests regarding keyframe selection success and computational speed. The study contributes to deep learning video summarization methods which enables novel opportunities in both personal content creation and multi-modal video processing systems.

Future Scope

The video summarization model can achieve enhanced accuracy through integrating various data types which include metadata text and audio signals alongside speech recognition systems. Experts must examine ways which would let the model adjust its summary actions based on individual preference needs while considering specific content characteristics. System scalability rises when real-time video summarization features become available in cloud-based systems since this enables the technology to support streaming needs and surveillance needs. The keyframe selection process can be optimized using reinforcement learning algorithms that learn user-driven feedback automatically during real-time operations until summary results become maximally effective. The use of transformers should occur during development to help identify long-range dependencies within video sequences. The demand from multiple industry domains including security and digital marketing for automated video summary applications makes this original research valuable for future improvements during automated summarization development.

References

[1]. Zhang, Y., Liu, Y., Kang, W., & Zheng, Y. (2023). MAR-Net: Motion-Assisted Reconstruction Network for Unsupervised Video Summarization. IEEE Signal Processing Letters, 30, 1282–1286.
[2]. Saini, P., Kumar, K., Kashid, S., Saini, A., & Negi, A. (2023). Video summarization using deep learning techniques: A detailed analysis and investigation. Artificial Intelligence Review, 56, 12347–12385.
[3]. Lin, F., Yuan, J., Chen, Z., & Abiri, M. (2024). Enhancing multimedia management: Cloud-based video genre recognition with hybrid deep learning architecture. Journal of Cloud Computing, 13, 104.
[4]. Yuan, Y., & Zhang, J. (2023). Unsupervised video summarization via deep reinforcement learning with shot-level semantics. IEEE Transactions on Circuits and Systems for Video Technology, 33(1), 445–458.
[5]. Kashid, S., Awasthi, L. K., Berwal, K., & Saini, P. (2024). Spatiotemporal feature fusion for video summarization. IEEE MultiMedia, 31(3), 88–97.
[6]. Yu, Q., Yu, H., Sun, Y., Ding, D., & Jian, M. (2024). Unsupervised video summarization based on the diffusion model of feature fusion. IEEE Transactions on Computational Social Systems, 11(5), 6010–6023.
[7]. Zhao, X., & Zhao, X. (2024). Application of generative artificial intelligence in film image production. Computer-Aided Design &

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0229
e ISSN: 2584-2854
Volume: 03
Issue: 04 April 2025
Page No: 1407 - 1415

Applications, 21(S27), 15–28.
[8]. Lin, J., Hua, H., Chen, M., Li, Y., Hsiao, J., Ho, C., & Luo, J. (2024).
[9]. VideoXum: Cross-modal visual and textual summarization of videos. IEEE Transactions on Multimedia, 26, 5548–5563.
[10]. Zhang, Y., Liu, Y., Kang, W., & Tao, R. (2024). VSS-Net: Visual semantic self-mining network for video summarization. IEEE Transactions on Circuits and Systems for Video Technology, 34(4), 2775–2789.
[11]. Ji, Z., Zhao, Y., Pang, Y., Li, X., & Han, J. (2020). Deep attentive video summarization with distribution consistency learning. IEEE transactions on neural networks and learning systems, 32(4), 1765-1775.
[12]. Kadam, B. D., & Deshpande, A. M. (2024). Multi-head attention with reinforcement learning for supervised video summarization. Journal of Electronic Imaging, 33(5), 053010-053010.
[13]. Basu, A., Pramanik, R., & Sarkar, R. (2024). Wanet: weight and attention network for video summarization. Discover Artificial Intelligence, 4(1), 5.