A Multi-Model Video Summarization Framework Integrating Feature Extraction, Embedding and Transformer-Based Learning

Sahaya Sakila V[1], Chitransh Nishad[2], Muthangi Shashank[3] and Tarun Prithi Gopinath[4]
{sahayasv2@srmist.edu.in[1], ca8446@srmist.edu.in[2], mm5473@srmist.edu.in[3], tg4593@srmist.edu.in[4]}

Department of Computer Science and Engineering, SRM Institute of Science and Technology,
Ramapuram Campus, Chennai, Tamil Nadu, India[1,2,3,4]

Abstract. Video summarization is important for managing large volumes of videos across various domains such as media, education, surveillance and so on. Traditional approach for summarization includes keyframe selection and clustering which fails to capture the temporal dependencies and semantic context that leads to incomplete or redundant summaries. To address these limitations faced by existing systems, the proposed method: VidSynape introduces a multi model video summarization that combines frame level analysis with insights from transcript. It uses deep feature embeddings for visual content representation and efficient similarity-based indexing to enhance scalability and speed. Using multi model techniques this approach improves summary, contextual coherence and computation efficiency. The system is tested on datasets such as Sum Me and TVSum with results showing better performance over other methods. It effectively generates quality summaries while reducing processing time making it a solution for real world video analysis application

Keywords: Video summarization, deep learning, OpenAI Whisper, FAISS, PyTorch, ResNet50, semantic embedding, benchmarking.

1 Introduction

In today's digital era, video content has become an important part of communication, entertainment and surveillance. There has been a rapid increase of videos generated through social media, security cameras and online platforms which demand efficient video summarization techniques. Video summarization (VS) is the process of creating a concise and compact information of a video containing the most important content of the video. This process benefits various sectors including education, medical, security, content creation and so on. [2][3][4]

The primary beneficiaries of video summarization are for researchers, students, security analysts, digital content creators, and professionals who work with large video datasets. By generation concise and meaningful summaries VS helps users to gain information from video quickly and efficiently. It is particularly helpful for students as they can generate notes from summaries, law and enforcement where they can get key events of hours long surveillance footage, media production where editor can create highlight for movies, sports or news and content creators as they can generate small content of key events from their videos.[5][6][7.

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357767

Various video summarization methods have been developed over the years, including clustering-based techniques, deep learning approaches, and graph-based models. Clustering methods, such as k-means and spectral clustering[8][9], focus on grouping similar video frames together to identify representative keyframes. However, they often fail to consider the temporal relationships between frames, leading to summaries that may not preserve the video's contextual flow. Various deep learning based approach like Long short term memory (LSTM) networks and Convolutional bureau networks (CNN) had made the summarization better by capturing sequential dependencies in a videos. Despite of their effectiveness these models still require a high computational resources. Graph based methods have also been explored which leverages scene detection and relationship modeling to generate better structured summary but they suffer from high computational complexity and redundant outputs.[1][2][3][4]

An advanced Video Summarization (VS) framework is proposed to overcome these challenges which integrates spatio temporal extraction, attention-based learning and graph-based refinement. The method takes advantage of bidirectional LSTM with attention mechanisms to dynamically assign importance to video frames to ensure that most relevant and important content is retained. Also, spatio temporal analysis allows to capture both motion dynamic and static visual cues which enhances overall representation. The integration of Graph Convolutional Networks (GCN) further improves the process by modelling inter shot dependency and removing redundant frames which results in a more informative summary. [5][6][7]

The proposed system addresses shortcomings of existing methods by improving efficiency and contextual accuracy. Framework is evaluated on SumMe and TVSum dataset which demonstrate increase in F1-score, precision and recall compared to traditional methods. The improvement in the effectiveness of the framework in generating quality summaries that are accurate and computationally efficient. The framework improves the summarization process which allows quicker browsing, reduced storage requirements and better content accessibility across various industries.[8][9].

Video content is now part of every communication, entertainment and surveillance with the emergence of digital times. With the increased amount of videos created through social media, surveillance and online sites leads to demand for better video summarization techniques. VS is the process of creating a lengthy video into a condensed version comprising the most important part of the content. This technology can be useful in various domains such as security for monitoring, medical diagnostic for analyzing and content generation for long multimedia.[1][2][3][4].

The main beneficiaries of vs are students, content producers, data scientists, security experts and researchers. With the creation of informative and short summary VS enables users to go through long video contents fast and efficiently. This is helpful in use cases such as official has to sit through long surveillance footage, for important events, important content for media production for producers, note taking from education video for students.[6][7]

Various video summarization techniques have been proposed over the years such as deep learning based, clustering based methods and graph-based models. K-means and spectral clustering are clustering methods which are directed towards aggregating similar frames in a video into groups in order to detect keyframes. These usually neglect the temporal relationship among frames thus generating summaries that are not assured to maintain the contextual flow of the video.Summarization has been enhanced through deep learning based techniques like LSTM and CNN that learns sequential dependencies in a video. Although effective they take high computational power and are not very efficient at handling long range dependence. Spatio temporal feature extraction and graph based modelling applied with relationship modelling and scene detection has been used to create better structured summaries but they also have high computational complexity as well as redundant outputs.

This video summarization framework proposes a solution to these challenges. It combines spatio temporal feature extraction, attention based deep learning and graph based refinement. These approaches use Bi-LSTM with attention mechanisms to dynamically attribute weights to the frames of video in such a way that the most important part will be preserved. Also the spatio temporal analysis allows us to capture the dynamics of static and motion visual information. Such an approach helps in enhancing the overall representation of content. The combination of GCN which is used to represent relationships between the shots and remove irrelevant frames which make summary more coherent and meaningful.

The proposed system improves efficiency, scalability and contextuality compared to earlier methods.It has been tested on benchmark datasets such as TVSum and SumMe reveals a performance boost in F1-score, precision and recall over the conventional techniques. Through improved summarization, methods makes browsing the right videos quicker, decrease the need of storage and enhances content accessibility in various sectors.

The following section will explore the challenges associated with the existing video summarization techniques, discuss in detail about the methodology and a comparative analysis with state-of-the-art methods. This research aims to contribute to the advancement of video summarization technologies and their practical applications in the real world.

2 Literature Survey

DSNet: A Flexible Detect to Summarize Network for Video Summarization (VS)[1] proposed a method that formulate VS as temporal interest detection problem.[1] Various subset selection for supervised VS introduces sequential determinantal point process model which ensures diversity and informativeness in selected frames while preserving sequential consistency.[2] Video Summarisation: A Conceptual Framework Survey present a broad framework for VS , technique for classifying into internal, external and hybrid methods. It highlights the need for incorporating external information to link semantic gao and improve relevance in summary. [3]

AudioVisual Video Summarization expands traditional VS by incorporating audio cues. It proposes AudioVisual Recurrent Network (ARVN) which combines audio and visual modalities using a two stream LSTM and self-attention mechanism which gives more semantically rich summary [4] Reconstructive Sequence-Graph Network for Video Summarization focuses on leveraging graph-based approaches to model inter-shot dependencies. The model integrates reconstructive learning techniques to enhance video representation and eliminate redundancy. [5] Video Summarization with Attention Based Encoder Decoder Networks explores the use of an attention-based encoder decoder framework to dynamically assign importance scores to video frames. IBy focusing most informative content it improves keyframe selection.[6]

Video Summarization and Scene-Detection by Graph Modeling employs graph-based models to segment and summarize videos based on scene structures. The approach enhances video coherence by leveraging inter-frame relationships. [7] Video Summarization Using Deep Learning Techniques provides an overview of various deep learning techniques applied to video summarization, including CNNs, LSTMs, and reinforcement learning-based approaches. [8] Video Summarization for Spatio Temporal Features indicates the role of spatio temporal feature extraction in summarization. The model combines audio and visual content analysis to improve quality of summary. [9]

A Reinforcement Learning Approach for Video Summarization proposes a reinforcement-based framework for optimizing summaries of video. The method defines summarization as a sequential decision making process which uses reward function to enhance information and diversity present in a summary. [10] Multi view Video Summarization introduces a framework that generates summaries by combining multiple camera viewpoints. The method employs a probabilistic model to select a representative frame while preserving spatial and temporal consistency across views. [11] Online Video Summarization: Predicting Future to Better Summarize Present introduces a predictive framework for real time VS by anticipating future frames to enhance present summary selection. The model takes advantage of deep learning techniques to analyze temporal dependencies which ensures that key moments are selected dynamically as the video progresses. [12].

A Comprehensive Survey of Multi-View Video Summarization provides a vast review of multi view video summarization techniques which categorize based on feature extraction, fusion strategies and summary generation method. The survey examines key challenges such as viewpoint redundancy, temporal alignments and summarization coherence across multiple camera feeds.[13]Video Summarization using Deep Semantic Features proposes a deep learning based approach that leverages semantic features to improve video segments and their corresponding textual descriptions into a shared semantic space using a deep neural network(DNN) .[14]Video Summarization Using Deep Neural Networks: A Survey gives a comprehensive review of deep learning based video summarization techniques focusing on convolutional neural network (CNN), recurrent neural network(RNN), attention mechanisms and reinforcement learning approaches.[15]

Multi Document Summarization via Deep Learning Techniques: A Survey presents a comprehensive review of deep learning based approaches for summarizing multiple documents. The study categorizes methods into extractive, abstractive and hybrid models which analyze the role of neural architecture.[17] V2Xum-LLM: Cross Modal Video Summarization with Temporal Prompt Instruction Tuning presents a novel approach for VS by taking advantage of Large Language models (LLM) and temporal prompt instruction tuning.[18]

M3SUM: A Novel Unsupervised Language Guided Video Summarization provides an unsupervised framework that combines multimodal learning for language guided video summarization. The model takes advantage of textual descriptions to enhance understanding of video which aligns semantic information with visual content. [19] A general framework for edited video and raw video summarization introduces a unified approach to summarize both professionally edited and unedited raw videos. The framework uses deep learning techniques to analyze scene transitions, visuals and temporal structures ensuring that key moments are effectively captures. [20]

These papers mentioned above focused on the development of video summarization techniques which are the consensus of traditional clustering and heuristic approach to refined deep learning, graph based learning and multimodal techniques. By leveraging insights from these papers we proposed a method which combines multiple strategies to enhance summarization.

3 Proposed System

3.1 VidSynapse

VidSynapse is a video summarization framework which combines video, audio and text based features for generating a concise and meaningful summary. The system uses a combination of feature extraction techniques, index mechanism and deep learning models to enhance summarization and computational efficiency. By using the frame level analysis, audio transcription, embedding and summarization techniques, Vidsynapse generates summaries to retain important information while minimizing redundancy.

The proposed system uses the FAIIS vector framework for video summarization which combines multimodal feature extraction, embedding, and deep learning models for generation quality summaries.

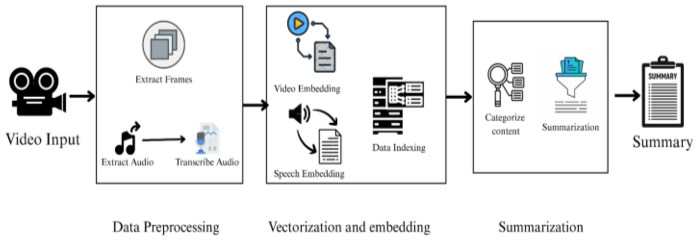

Fig. 1. Working principle of Vidsnype

The above diagram shows a video summarization pipeline architecture, first stage id data processing stage. In this stage audio and frames are extracted from the videos to save visual information. This stage makes sure the data will be accessible in an organized manner for next stages.

In the next stage vectorization and embedding is done from the result of the previous stage. Here the preprocessed data is converted into numerical representations, Frames are converted into video embedding and audio transcript is converted into speech embedding. The data is indexed for better retrieval allowing faster and accurate analysis.

The last stage is for summarization where data is categorized based on its content. Embedding is analyzed to identify key topics and themes. Advanced models are applied to generate summaries containing relevant information.

The Post processing stage then combines the outputs from the different paths and creates a final summary. This approach is to improve efficiency and accuracy.

3.2 Video Processing

Processing the content of the video begins with frame extraction from video. Video consists of a stream of frames and selecting the one which has most information is important. The interval Φ is a parameter to balance capturing key moments and minimizing computational overhead. Mathematically, if a video V has a duration Γ (in seconds) and a frame rate Δ (frames per second), total number of extracted frames Nϕ is given by:

$$N\phi \;=\; (\Gamma x \Delta)/\Phi \qquad (1)$$

After frame extraction each frame undergoes deep feature analysis using pre-trained CNN implemented in pyTorch. These networks function as feature extractors that convert an image into a high-dimensional embedding data.

$$Efi = F(fi) \in R^d \qquad (2)$$

By using the improved system by fine tuning the CNN parameters to enhance representation ensures that spatial details and object level features are preserved. For faster retrieval and elimination of redundancy the generated embeddings are indexed using FAISS The index L is constructed as:

$$I = FAISS(Ef1, Ef2, \dots, EfNf) \qquad (3)$$

For a query embedding Eq, the nearest-neighbor retrieval is performed by:

$$N(Eq) = argEi \in Imin \parallel Eq - Ei \parallel 22 \qquad (4)$$

This addition of FAISS significantly improves the frame selection speed and accuracy during the process.

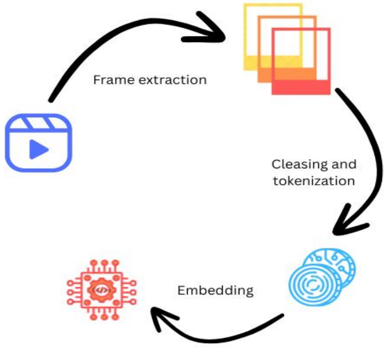

Fig. 2. Video Processing Life Cycle.

This fig 2 illustrates the video processing pipeline within VidSynapse. It begins with frame extraction, where frames are sampled at optimized intervals to ensure efficient processing. These frames then undergo deep feature analysis, where pre-trained CNN models such as ResNet50 or EfficientNet extract meaningful representations. To speed up retrieval, the system utilizes FAISS-based indexing, which organizes the frame embeddings for quick nearest-neighbor searches. Finally, the indexed frames are processed to identify keyframes that are most relevant to the video's content. By demonstrating this life cycle, the fig 1 showcases how VidSynapse ensures accurate and efficient keyframe selection for summarization.

3.3 Audio Processing

The audio is extracted from the video and converted into the transcript at the same time when video processing is going on. Audio is extracted from the video using FFmpeg isolating the raw audio signal A(t). This signal can be represented as a sum of cosine functions:

$$A(t) = n = 0\sum Nancos(2\pi fnt + \phi n) \qquad (5)$$

Where an, fn, and ϕ denote the amplitude, frequency, and phase of the nth component, respectively.

For converting speech to text, the system uses OpenAi's Whisper model which transcribes the audio into text where A(t) represents the audio signals as a function of time τ and W is the whisper model performing transcription.

$$\Theta = W(A(\tau)) \qquad (6)$$

After the transcription of audio the text is tokenized and cleanses which gives a sequence of words Θ= {ω1, ω2,,,,ωη} where ωi represents individual words, and η is the total number of words in the transcript. The system performs well in noisy environments and ensures that the transcript is accurate.

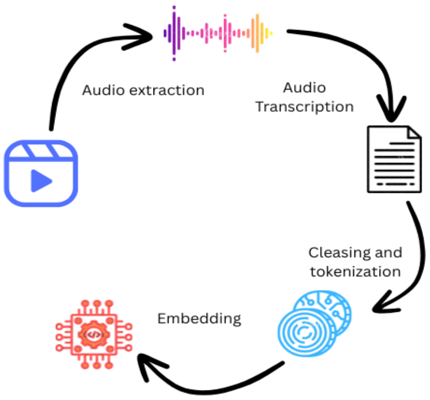

Fig. 3. Audio Processing.

The audio processing in the system starts by extracting the audio from video using FFmpeg. The raw video is then transcribed into text using OpenAi Whisper. After transcription of audio the text goes through tokenization and embedding generation. This step is important as audio and visual data together enhances the summary quality.Th above fig 3 demonstrates the whole process.

3.4 Embedding and Vectorization

After processing both visual and audio content is processed and converted into vector data the next stage converts these modalities into embeddings. Each frame embedding is computed as described earlier:

$$Efi = F(fi) \in Rd \qquad (7)$$

Simultaneously, the transcribed text is encoded using a transformer-based model to obtain transcript embeddings:

$$ET = BERT(T) \in Rd \qquad (8)$$

To achieve a unified representation of the video, to perform multi-modal fusion. This is accomplished by combining the video and transcript embeddings using a weighted sum:

$$Ecombined = \alpha EF + \beta ET \qquad (9)$$

where α and β are adjustable coefficients that determine the contribution of each modality. This also measure the alignment between modalities using cosine similarity:

$$S(Efi, ET) = \| Efi \| \| ET \| \, Efi \cdot ET \quad (10)$$

These fusion techniques allow the model to better understand the relationship between visual scenes and spoken content, improving contextual awareness.

3.5 Summarization Methods

The system applies both extractive and abstractive summarization techniques.

3.5.1 Extractive Summarization:

Key frames are selected based on their importance. Using clustering on the proposed FAISS-indexed embeddings to identify representative frames. A simple selection criterion is:

$$Ck = \{Efi \mid S(Efi, ET) > \theta\} \qquad (11)$$

where θ is a threshold parameter that is tuned to balance between comprehensiveness and brevity. This method removes redundant frames and ensures that only the most representative frames are included.

3.5.2 Abstractive Summarization:

For generating a coherent textual summary, fine-tune a Transformer model (such as T5). The model takes the transcribed text Ξ as input and generates the summary Σ, where Ξ represents the transcribed text, and T5 denotes the T5 model performing the summarization:

$$\Sigma = T5(\Xi) \qquad (12)$$

This abstractive process rephrases the content, ensuring that the summary is not only a collection of key sentences but also linguistically fluent and contextually accurate.

3.6 Training and Optimization

To capture temporal dependencies, using a Bidirectional LSTM (Bi-LSTM) model. This model is designed to process sequential data, ensuring that the narrative continuity of the video is maintained. The hidden state ht at time t is computed as:

$$ht = \sigma(Whht - 1 + Wxxt + bh \qquad (13)$$

where xt is the input at time t (which could be a combined embedding), WhW_hWh and Wx are weight matrices, bh is the bias term, and σ is the activation function. The output distribution over potential summary segments is given by:

$$yt = softmax(Woht + bo) \qquad (14)$$

Parameters such as Wh, Wx, and bh are optimized during training. Additionally, fine-tuning the T5 model on the dataset allows it to adapt to the video-specific linguistic structure, further enhancing the quality of abstractive summaries.

3.7 Post-Processing and Final Summary Generation

After initial summarization the system applies post-processing to refine the output in a proper format. The extracted content is categorized into clusters of similar scenes or topics which is called semantic vector clustering(svc).It emphasizes the extracted features are converted into vector embeddings and then clustered based on semantic similarity:

$$V = \Phi(X), V = \{v1, v2, ..., vN\} \qquad (16)$$

$$C=Fclust(V) \qquad (17)$$

$$C=\{c1,c2,\ldots,cK\} \qquad (18)$$

This categorization helps to organize the summary into coherent segments. Finally the system creates the final summary by combining keyframes and corresponding textual snippets, which is called hierarchical content embedding.

$$S=\{(f1,T1),(f2,T2),\ldots,(fk,Tk)\} \qquad (19)$$

$$vif=\Phi f(fi) \tag{20}$$

$$vit=\Phi t(ti) \tag{21}$$

$$vi=\alpha vif+\beta vit \tag{22}$$

$$C=Fclust(\{v1,v2,...,vK\}) \tag{23}$$

Here each tuple (fi,Ti) represents a keyframe and its associated text. This multi model summary ensures that both visual and narrative aspects of the original video are preserved.

Algorithm 1: Multi-Model Video Summarization Framework

```
Input: V → Input video, F → Frame extraction rate, A → Audio track,
    Mv → Pre-trained visual feature extractor, Ma → Pre-trained audio/text model,
    S → Summarization model, L → Learning rate, λ → Similarity threshold

Extract frames {F1, F2, ..., Fn} from V at rate F
Extract audio A from V and generate transcript T using speech-to-text model Ma
Initialize embedding extraction:
    E_v = Mv(F)  ▻ Extract visual embeddings
    E_a = Ma(T)  ▻ Extract textual/audio embeddings
Perform FAISS-based indexing for similarity search on {E_v, E_a}
Cluster embeddings using cosine similarity with threshold λ
Select keyframes {K1, K2, ..., Km} from clusters
Apply transformer-based model S to generate abstractive text summary S_T
Merge keyframes {K} and summary S_T into final summarized representation
Optimize parameters of Mv and Ma using backpropagation with learning rate L
Output: Final Video Summary {K, S_T}
```

4 Result and Discussion

Experimental evaluation has shown that the system called VidSynapse significantly surpasses video summarization systems based on existing benchmarks datasets such as SumMe and TVSum. The F1 score in the SUMME (SumMe) and TVSum (TVSum) datasets serves as an important evaluation metric for VS, measuring how closely an automatically created summary matches human-provided summaries. It is computed as the harmonic mean of precision and recall, where precision indicates the fraction of chosen frames or shots that are pertinent, and recall represents the fraction of relevant frames or shots that have been selected.

In the SUMME dataset, which contains summaries annotated by humans, the F1 score assesses the correspondence between summaries produced by machines and the actual ground-truth annotations, usually at either the frame-level or shot-level.

Conversely, TVSum features user-assigned importance ratings for various video segments, and the F1 score is utilized to evaluate how the produced summaries align with these user-assigned ratings.

Although both datasets utilize F1 scores for assessment, SUMME focuses on established human summaries, while TVSum relies on multiple user annotations to assess frame significance.

Table.1 Comparison of Multiple Methods.

Method	SumMe	TvSum
Vidsnype	49.8	59.5
Online Motion-AE	37.7	51.5
SUM-FCNunsup	41.5	52.7
DR-DSN	41.4	57.6
EDSN	42.6	57.3
UnpairedVSN	47.5	55.6
PCDL	42.7	58.4
ACGAN	46.0	58.5
SUM-GAN-sl	47.8	58.4
SUM-GAN-AAE	48.9	58.3
SUM-GDAunsup	50.0	59.6
CSNet+GL+RPE	50.2	59.1
CSNet	51.3	58.8
GoogleNet	51.6	64.2
ResNet	52.8	65.0

CLIP-It	54.2	66.3
AC-SUM-GAN	50.8	60.6

The above table 1 provides a performance comparison results of multiple video summarization method with proposed method measured by F1 Scores on the SumMe and TVSum dataset.VidSynapse achieves an F1 score of 49.8 on SumMe dataset and 59.5 on the TVSum dataset shows better score than conventional methods. WHile models such as ResNet, Clip-It and GoogleNet score better than VidSynapse, they rely on pre-training and require high computational resources. The balance between performance and scalability makes the framework an ideal choice for real world application.

The above graph gives a comparative analysis of techniques highlighting their F1 score on the TVSum and SumMe datasets. The proposed framework gives better performance over traditional methods. While advanced models achieve competitive scores, VidSynapse maintains a better balance between accuracy and computational efficiency. By showing the results the graph tells the effectiveness of VidSynapse in generating quality, concise summaries.

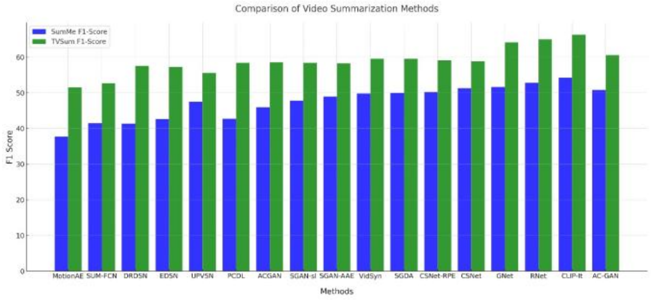

Fig. 4. Comparison of Video Summarization Methods.

Improved F1 score aside the framework also showcases better recall and precision values which imply that the multimodal of visual and transcript embedding led to a better identification of keyframes and critical and narrative elements. The use of similarity during multimodal made sure that the alignment between video frames and textual information is strong, leading to summary that is both less redundant and contextually coherent. Fig 4 Comparison of Video Summarization Methods.

The proposed methods effectiveness can also be attributed to the fine tuning of deep learning models. The bidirectional LSTM captured temporal dependencies between frames, ensuring narrative continuity. The combined use of abstractive and extractive techniques made the system generate summaries that are not only informative but also coherent.

In summary, Vidsynapse takes advantage of multimodal embeddings, efficient indexing, and advanced sequence modeling to overcome the limitations of traditional methods. The improvement in F1 score, precision and recall make the system applicable for real world video analysis tasks.

5 Future Scope

An efficient method for extracting important textual and visual information from the videos has been suggested by the framework. However, a number of fronts still require improvement and research. For example, using multi modal transformers might improve the modelling of text, audio and video interaction for better summarization. Also, a self-supervised learning approach might be used to improve generalization of different video categories and lessen the need for labeled dataset.

Another direction for future research will be real-time video summarization. Development of a framework that can process live video streams may be achieved by improving the computational efficiency of both LSTM based summarization and FAISS based indexing. It will have global applicability which can be achieved by improving the architecture to enable multilingual transcription and summary.Last but not least user engagement and usability would be improved greatly by using personalized summaries accustomed to viewers' choices and interests.

6 Conclusion

Based on the study designed a video summarization approach. It consists of audio transcription, feature extraction and embedding algorithms that combine the summarization models. The framework implements the feature extraction based on deep learning and FAIIS based indexing. It enhances and simplifies the keyframe detection stage and summarization stage is based on the combination of extractive and abstractive summarization so that the system is able to create a relatively short and informative video summary.

The utilization of a multimodal approach which is responsible for the adjustment of visual and textual information to improve the summary. In this manner, the proposed system is capable of capturing both visual and lingual representations of the video content. The results of the experiments show that the approach improves the performance of the video summarization.

In general, this research provides a foundation for further progress in AI-driven video summarization, towards smarter, more scalable and context-aware systems for video analysis.

References

[1] Zhu, W., Lu, J., Li, J., & Zhou, J. (2020). Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, *30*, 948-962.

[2] Gong, B., Chao, W. L., Grauman, K., & Sha, F. (2014). Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, *27*.

[3] Money, A. G., & Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. *Journal of visual communication and image representation*, *19*(2), 121-143.

[4] Zhao, B., Gong, M., & Li, X. (2021). Audiovisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 5181-5188.

[5] Zhao, B., Li, H., Lu, X., & Li, X. (2021). Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(5), 2793-2801.

[6] Ji, Z., Xiong, K., Pang, Y., & Li, X. (2019). Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(6), 1709-1717.

[7] Ngo, C. W., Ma, Y. F., & Zhang, H. J. (2005). Video summarization and scene detection by graph modeling. *IEEE Transactions on circuits and systems for video technology*, *15*(2), 296-305.

[8] 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*(11), 12347-12385.

[9] Laganière, R., Bacco, R., Hocevar, A., Lambert, P., Païs, G., & Ionescu, B. E. (2008, October). Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop* (pp. 144-148).

[10] A Reinforcement Learning Approach for Video Summarization

[11] Fu, Y., Guo, Y., Zhu, Y., Liu, F., Song, C., & Zhou, Z. H. (2010). Multi-view video summarization. *IEEE Transactions on Multimedia*, *12*(7), 717-729.

[12] Lal, S., Duggal, S., & Sreedevi, I. (2019, January). Online video summarization: Predicting future to better summarize present. In *2019 IEEE Winter Conference on applications of computer vision (WACV)* (pp. 471-480). IEEE.

[13] Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S. W., & De Albuquerque, V. H. C. (2021). A comprehensive survey of multi-view video summarization. *Pattern Recognition*, *109*, 107567.

[14] Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2017). Video summarization using deep semantic features. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13* (pp. 361-377). Springer International Publishing.

[15] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, *109*(11), 1838-1863.

[16] Alaa, T., Mongy, A., Bakr, A., Diab, M., & Gomaa, W. (2024). Video Summarization Techniques: A Comprehensive Review. *arXiv preprint arXiv:2410.04449*.

[17] Ma, C., Zhang, W. E., Guo, M., Wang, H., & Sheng, Q. Z. (2022). Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, *55*(5), 1-37.

[18] Hua, H., Tang, Y., Xu, C., & Luo, J. (2024). V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*.

[19] Wang, H., Zhou, B., Zhang, Z., Du, Y., Ho, D., & Wong, K. F. (2024, April). M3sum: A Novel Unsupervised Language-Guided Video Summarization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4140-4144). IEEE.

[20] Li, X., Zhao, B., & Lu, X. (2017). A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing*, *26*(8), 3652-3664.