VidWizz: Personalized Video Summarization Framework Using 2D CNN

Prof Mrs.Syeda Shafia
Department of Computer Science and Engineering
M V J College of Engineering
Bengaluru,India
syedashafiasadaf@mvjce.edu.in

B Sai Deepak
Department of Computer Science and Engineering
M V J College of Engineering
Bengaluru,India
1mj20cs036@mvjce.edu.in

Ambika S
Department of Computer Science and Engineering
M V J College of Engineering
Bengaluru,India
1mj21cs013@mvjce.edu.in

M Prajitha
Department of Computer Science and Engineering
M V J College of Engineering
Bengaluru,India
1mj21cs099@mvjce.edu.in

Varun C
Department of Computer Science and Engineering
M V J College of Engineering
Bengaluru,India
1mj21cs240@mvjce.edu.in

***Abstract*—VidWizz is an innovative, lightweight framework that merges AI-driven video analysis with a responsive, personalized chatbot to deliver efficient, customized video summaries. By leveraging 2D Convolutional Neural Networks (CNNs), it processes video content locally, ensuring minimal computational load and reducing bandwidth consumption on mobile and edge devices. The system dynamically generates video summaries and continuously refines them based on real-time user feedback through the integrated chatbot. This interaction allows VidWizz to adapt to individual preferences, providing a personalized video-to-text summary tailored to each user's unique needs and interests. The chatbot interface acts as a conversational guide, enabling users to adjust the summary's focus and level of detail, optimizing the content for different goals such as learning, entertainment, or research. This blend of AI-powered video analysis and responsive conversational interaction ensures that users receive a concise, relevant, and highly customized video summary experience, making VidWizz an ideal solution for personalized content delivery on mobile and edge devices.**

Keywords — Video to Text summarization, CNN, LSTM, Streamlit, OpenAI, LangChain, KNN.

I. INTRODUCTION

The exponential growth of multimedia content, particularly video, has significantly reshaped how information is consumed and shared. With billions of users engaging with video recording devices, powered by advanced computational capabilities, and leveraging social media platforms for content distribution, the volume of video content continues to increase at an unprecedented rate [1]. As this growth accelerates, there is a growing need for efficient technologies that can help users quickly navigate vast amounts of video content and retrieve relevant segments of interest. Autonomous video summarization techniques have emerged as a promising solution to this challenge, enabling the generation of concise summaries that capture the essential moments of a video. These summaries allow users to gain an overview of the video without needing to view the entire content. For instance, a 90-minute soccer match can be condensed into a few minutes, focusing on critical events such as goals, penalties, and key plays. Such techniques are particularly valuable in contexts where users are pressed for time or looking for specific information, such as in educational videos, news content, and entertainment media. However, existing video summarization methods often face two key challenges: computational efficiency and personalization. Many traditional methods require significant computational resources, making them unsuitable for deployment on resource-constrained devices like smartphones or edge devices. Additionally, these techniques typically provide generic summaries that fail to account for individual user preferences, offering little in terms of customization or relevance to the viewer's specific interests. The intersection of NLP and video summarization, specifically for YouTube content. With a vast and ever-increasing repository of videos on YouTube, summarization is becoming an essential tool for efficient navigation and consumption of content. The review emphasizes the application of NLP in condensing video information by processing textual elements such as transcripts, captions, comments, and metadata. The paper highlights novel approaches to extracting meaningful insights while addressing the problem of processing diverse and voluminous data by focusing on techniques like text summarization, sentiment analysis, and deep learning architectures. In addition, the incorporation of evaluation metrics and common datasets used further provides a structured approach to assessing the performance of summarization systems. Besides this, [1] difficulties related to applying NLP for video summarization and how it deals with challenges in processing multilingual data along with multi-contextual and ever-evolving AI technologies. It puts an emphasis on supervised and unsupervised learning methods and compares them based on their efficiency and constraints. In other words, it provides an overall source for researchers and practitioners by summarizing the recent advancements, pointing out gaps, and providing direction to future work. The study, through NLP, envisions better accessibility, content discovery, and user engagement with the vast content of YouTube, making it more manageable and user-friendly. One of the pioneer applications of NLP[2] is video-to-text summarization, where a short summary of video content can be obtained in text format. It uses a set of techniques and tools to ensure accuracy and efficiency in the process. It starts with extracting audio from video files through MoviePy and further transcription by APIs such as Google's Speech Recognition. The transcription text undergoes summarization models which include extractive approaches which focus on key sentence selection and abstractive approaches that generate new content. Advanced models such as transformer architectures, for instance, as used in Facebook's model BART, are deployed to improve the summarization performance. The process also incorporates the technique of Named Entity Recognition, using resources like SpaCy to identify and classify important entities, which include names, dates, and geographic locations. It presents a framework for building customizable chatbots using large language models (LLMs) for document summarization and question answering that can help in the reduction of the problem of information overload by extracting insights efficiently from long documents through technologies like OpenAI, LangChain, and Streamlit.[3] The paper discusses the architecture, implementation, and practical applications of the framework and provides a step-by-step guide for developers to create end-to-end solutions for document summarization and query response. This research therefore indicates the ability of such systems to increase productivity and ensure better information retrieval. The comprehensive survey focuses on personalized video summarization, with a focus on tools and methods for efficient video analysis and synopsis creation, tailored to user preferences. Key methodologies include feature-based techniques (e.g., object recognition, motion analysis), machine learning models (CNNs, supervised/unsupervised learning), and hybrid approaches combining internal and external data. Categories include static, dynamic, hierarchical, and multi-view summaries, each specifically for different applications like sport highlights, customized trailers, egocentric videos, among others. Methods used here mainly include audiovisual cues- keyframes, and video segments, and so-called advanced methods like deep ranking models, transfer learning, and adversarial networks. Datasets play a critical role in the evaluation of these methodologies, with 37 datasets reviewed for various use cases. Techniques address challenges such as redundancy and domain specificity while leveraging tools such as clustering [4], keyframe selection, and semantic analysis to optimize user-centric summaries. Applications range from sports to movies and personalized search engines, reflecting the diversity and utility of personalized video summarization. Future directions underscore integrating adaptive frameworks and enhanced user profiling to tackle emerging challenges. The proposed client-driven architecture will be used for customized video summarization based on 2D convolutional neural networks. It is focused on reducing computational requirements and on privacy because it does the summarization on the client, either smartphone or laptop, instead of relying on some central server. Thumbnail-centric video representations strongly reduce computational requirements and thus enable the processing of video materials on resource-poor devices.[5] The use of 2D CNNs ensures important features are extracted from video thumbnails that enable high-quality summaries. LTC-SUM is apt for applications where privacy is concerned and users require video summaries customized to their personal preferences, such as personal video archiving and privacy-preserving video summarization. The framework is experimented with in a variety of scenarios to establish its capabilities to generate succinct, user-specific summaries within computationally efficient frameworks. It is used for personal video archiving, privacy-preserving video summarization. This emphasizes the growing complexity related to managing large amounts of video data generated through multiple platforms such as social media. It focuses on video summarization techniques that aim to reduce long videos into short informative summaries, thus saving time and enhancing the retrieval process of information. The paper discusses static and dynamic approaches to video summarization. [6] It focuses mainly on Multi-View Video Summarization (MVS) in order to deal with complex, multi-view data. Furthermore, it discusses the technical challenges related to video summarization, particularly in fields like security and surveillance. Lastly, the research highlights the need for further advancement in these techniques to achieve more efficiency and scalability in real-world applications. The methodology of video summarization uses[15] Fully Convolutional Sequence Networks (FCSN), inspired by methods of semantic segmentation in the image processing domain. This does not depend on complex temporal models such as Recurrent Neural Networks (RNNs);[7] video frames are processed directly as sequential data. This model predicts keyframes that reflect important parts of the video, thereby serving as a scalable and efficient method. This characteristic makes it specifically useful for real-time video summarization applications, such as video surveillance and movie trailers where large amounts of video content need to be collapsed into coherent summaries.Two-stage end-to-end text-to-video generation: Use a conditional VAE to generate the ”gist” of the video, capturing static background and object layout; then, content and motion in the video can be generated by conditioning both on the gist and input text. For this purpose, the authors employ image filter kernels, derived from the text, and apply them to the gist to enrich the representation of motion and detailed content. In this paper, the proposed approach is compared with the traditional GANs and presents progress in the production of dynamic video content. The proposed architecture combines both static properties and dynamic motion within the video generation domain.[8]

II. RELATED WORKS

Video summarization aims to produce short synopses by extracting keyframes or segments. It thus allows people to process and enjoy long video content efficiently. Deep learning methods play a major role, including supervised, unsupervised, and weakly-supervised strategies. Supervised methods rely on annotated datasets in training models that determine the importance of frames, whereas unsupervised methods employ generative models, such as GANs, for creating summaries where no labelled data is available. Weakly-supervised methods fill the gap by taking fewer annotations or metadata and, hence, improving summary quality. Multimodal approaches further enhance performance with the use of textual metadata in aligning summaries to semantic content. Video-to-text summarization uses complex deep learning architectures to give meaningful text descriptions of video content.[9] Techniques such as transformer models and convolutional networks analyze and encode visual features, while semantic alignment processes generate coherent textual outputs. Evaluation protocols, including F-scores and overlap metrics, serve to benchmark performance in comparison to human-generated summaries. In addressing challenges such as subjectivity in summaries and the limited availability of annotated datasets, contemporary research emphasizes scalable and adaptive methodologies that incorporate various modalities to enhance the accuracy and relevance of summarization. The technology of chatbots has dramatically changed human-computer interactions, from rule-based systems to sophisticated AI-driven conversational agents. The rule-based chatbots work with established scripts and are very effective for simple inquiries. AI-based chatbots make use of machine learning, understanding intent and context in order to create dynamic and human-like conversations.[10] Modern chatbot designs are often combined with NLP and neural networks, such as Transformers and RNNs, which greatly improve the quality of generated responses. The applications of these technologies are broad, from customer service and healthcare to education, e-commerce, and even redundant tasks, personalized assistance, and optimization of user experience. Video-to-text summarization is one of the most innovative applications in the conversational AI domain, converting video content into[11] coherent textual summaries using NLP models. VidWizz makes use of a highly sophisticated approach towards video summarization through semantic analysis and sequence-to-sequence models that are engineered to draw out and encode all important features from the video content into coherent text summaries. It thus combines multimodal data: audio, visual, and text components to help the system attain greater levels of accuracy and relevance compared to other methods. Leveraging the power of multimodal information. The framework applies 2D Convolutional Neural Networks[16] (CNNs) for keyframe extraction; this is to ensure an efficient processing of visual features. Further, advanced models such as Transformer-based architectures and Recurrent Neural Networks (RNNs) are applied for the sequence learning process, thus helping in the accurate representation of temporal dependencies in video content. These methods are coupled with Named Entity Recognition (NER) techniques and attention mechanisms, which help identify critical entities and prioritize important segments in the video. The system is equipped with NLP tools like OpenAI's GPT and LangChain,[17] making it an effective tool in generating coherent and concise text outputs.

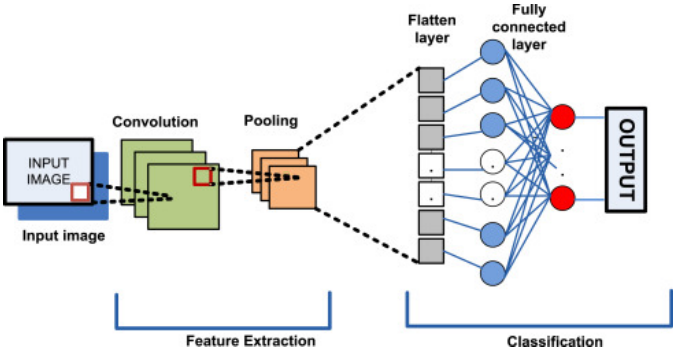

Fig. 1. 2D-CNN

III. RESULT AND DISCUSSION

VidWizz is an innovative framework designed for personalized video summarization, utilizing 2D Convolutional Neural Networks (CNNs) for efficient keyframe extraction and a responsive AI-powered chatbot for dynamic user interaction. By processing videos locally on resource-constrained devices like smartphones and laptops, VidWizz minimizes computational demands, reduces bandwidth usage, and ensures privacy. The system dynamically generates concise video-to-text summaries tailored to user preferences, refining them in real-time through conversational feedback via the chatbot. VidWizz integrates advanced AI and machine learning tools such as OpenAI's GPT, LangChain, and frameworks like TensorFlow and PyTorch to provide highly customized summaries. It leverages semantic analysis, natural language processing (NLP), and multimodal data to improve accuracy and relevance. The chatbot enables users to interactively adjust the focus and level of detail in summaries, enhancing usability for applications like education, entertainment, and personal archiving. Key methodologies in VidWizz include transfer learning, clustering, and semantic alignment, ensuring optimized content delivery while addressing challenges like scalability, redundancy, and domain specificity. By enabling efficient summarization of long videos into digestible segments, VidWizz serves diverse needs, including sports highlights, movie trailers, and privacy-preserving archiving. Its user-centric design and resource-efficient processing make it a standout solution for personalized content delivery in an era of growing multimedia consumption.

IV. CONCLUSION

VidWizz is the innovative platform designed to serve personalized video summaries through seamless interplay between advanced video processing and user interaction technologies. It fundamentally relies on using 2D Convolutional Neural Networks to analyze content in a video, extract keyframes from it, and identify features from those keyframes to prepare a summary. This enables users to get customized short video outputs according to the preferences of the user. The application's functionality is greatly enhanced by the inclusion of a powerful AI chatbot, allowing users to communicate with each other in natural, conversational dialogue that fine-tunes the summaries in real-time.Future developments in context-aware mechanisms and multimodal input handling will further enhance its capabilities. Implementation of advanced machine learning tools like TensorFlow or PyTorch to use 2D CNNs and NLP models like OpenAI's GPT to allow the functionality of the chatbot. Databases like MongoDB or PostgreSQL are essential to store user profiles and video information. Front-end tools such as React.js or Angular build an interactive user interface. Video processing tools like OpenCV and FFmpeg handle video reading and frame extraction. All these technologies work together to help VidWizz provide personalized, scalable, and efficient video summarization with smooth user interaction.

REFERENCES

1. Peronikolis, M., Panagiotakis, C. (2024). Personalized Video Summarization: A Comprehensive Survey of Methods and Datasets. Applied Sciences, 14(11), 4400.

2.Pokhrel, S., Ganesan, S., Akther, T., Karunarathne, L. (2024). Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit. Journal of Information Technology and Digital World, 6(1), 70-86.

3. Faluyi, J. O., Akinwonmi, A. E. (2024). A systematic literature review and existing challenges of chatbot models. IEEE-SEM Proceedings, 12(5), 52–82

4.Son, J., Park, J., Kim, K. (2024). CSTA: CNN-based Spatiotemporal Attention for Video Summarization. IEEE Transactions on Multimedia, 26, 105-116. arXiv.

5.Mishra, Prerna Garg, Kartik Rathi, Naveen. (2023). Video-to-Text Summarization using Natural Language Processing. International Journal of Advanced Research in Science, Communication and Technology.

6. Murthy, N. S. R., Prasanna, S. S., Gupta, P. K. (2023). Automatic Video Summarization and Classification by CNN Model. IEEE Transactions on Multimedia, 25(3), 678-688. IEEE Xplore.

7. Gudakahriz, Sajjad Jahanbakhsh, Amir Masoud Eftekhari Moghadam, and Fariborz Mahmoudi. ”Opinion texts summarization based on texts concepts with multi-objective pruning approach.” The Journal of Supercomputing 79, no. 5 (2023): 5013-5036.

8. Singh, Y., Kumar, R., Kabdal, S., Upadhyay, P. (2023). YouTube Video Summarizer using NLP: A Review. International Journal of Performability Engineering, 19(12), 817.

9. 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.

10.Gudakahriz, Sajjad Jahanbakhsh, Amir Masoud Eftekhari Moghadam, and Fariborz Mahmoudi. ”Opinion texts summarization based on texts concepts with multi-objective pruning approach.” The Journal of Supercomputing 79, no. 5 (2023): 5013-5036.

11. Mujtaba, G., Malik, A., Ryu, E. S. (2022). LTC-SUM: Lightweight client-driven personalized video summarization framework using 2D CNN. IEEE Access, 10, 103041-103055.

12. H. Abdulla, A. M. Eltahir, S. Alwahaishi, K. Saghair, J. Platos and V. Snasel, ”Chatbots Development Using Natural Language Processing: A Review,” 2022 26th International Conference on Circuits, Systems, Communications and Computers (CSCC), Crete, Greece, 2022

13.P. Kadam et al., ”Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms,” in IEEE Access, vol. 10, pp. 122762-122785, 2022.

14. 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).

15. M. Agrawal and D. S. Niranjan, ”Video Summarization using Machine Learning Mechanism: A Comprehensive Review,” 2021 Conference on Advances in Technology, Management Education (ICATME), Bhopal, India, 2021

16.Y. Jiang, K. Cui, B. Peng and C. Xu, ”Comprehensive Video Understanding: Video Summarization with Content-Based Video Recommender Design,” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019

17. Rochan, M., Ye, L., Wang, Y. (2018). Video Summarization Using Fully Convolutional Sequence Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 8199-8208.