International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Smart Video Summarization Using Deep Learning

Patharla Ramya[1] , Varshini Priyamvada[2]

[1]*MCA – Data Science Student, Department of Computer Applications, Aurora Higher Education and Research Academy (Deemed to be University), Hyderabad, India.*

[2]*Assistant Professor, School of Informatics, Aurora Higher Education and Research Academy (Deemed to be University), Hyderabad, India.*

ABSTRACT

The rise of online platforms such as YouTube, OTT services and digital learning portals has increased the consumption of video content rapidly. Manual browsing of such long videos is disabled and impractical, which creates a need for intelligent video methods. This paper proposes a deep learning smart video summary system that automatically detect visual transitions, extracts meaningful kefrms, and produces both visual and text summary. Our system integrates the Convenable Neural Network (CNN) for spatial facilities, long short -term memory (LSTM) network for learning temporal sequence, and natural language processing (NLP) model for caption generation. Using a preternd lightweight model such as Mobilentv2, the framework achieves a summary of real-time skilled real-time without the need for a high-end GPU. Lecture videos, experiments on videos, movies and CCTV surveillance footage suggests that the system largely reduces excess, improves summary relevance, and saves time. Compared to traditional histograms and heuristic-based approaches, the proposed system provides more reference-incolers and semantically meaningful summary. This task contributes to the growing area of AI-managed multimedia processing with possible applications in education, law enforcement, digital content manufacturing and monitoring systems.

Keywords: Video Summary, Deep Learning, CNN, LSTM, NLP, Keyframe Extraction, Scene Detection, Mobilnetv 2, Multimedia Processing

INTRODUCTION

The dominance of the video as information sharing, entertainment and learning has created new challenges in data management. With videos of over 500 hours uploaded every minute on YouTube, the requirement of automated summary techniques is more stronger than ever. It is unable to watch or edit such long videos, and existing manual methods are not scalable. Video Summary is the process of creating a condensed version of a video that maintains its most important parts. Traditional methods depend on simple hyuristics such as colour histogram differences or intensity of speed. However, these semantics fail to catch the meaning, often leading to irrelevant or fruitless summary. Deep learning (DL) provides a powerful option. CNN can "see" and analyze the frame as images, understand the temporary order of RNN frames, and NLP technology can generate natural language details. By combining these models, the video summary becomes smart, reference-coverage and user friendly.

The objectives of this research are: To design a hybrid deep learning pipeline for video summary. To use pretrand CNNS (Mobilentv2) for mild but effective keyframe detection. To compare the performance of deep learning abbreviations with traditional methods. To detect real -world applications such as education, security and multimedia editing.

LITERATURE SURVEY

The video summary has attracted a lot of attention in the last decade due to the rapid growth of multimedia data. Researchers have proposed different approaches, with henuristic-based techniques to advanced deep learning models. Early functions depend on the low-level visual signals, such as the colour histogram, edge detection, or speed intensity, to detect visual transition. While simple, these methods often failed to capture the semantic meaning of the frame and were sensitive to noise, light variation and camera movements (Bradsky et al., 2021).

Changes towards deep learning brought significant improvements. Convene Neural Network (CNN) has been widely used for keyframe extraction as they can identify high-level meaning characteristics such as objects, people and tasks (Gali et al., 2014). For example, CNN-based models such as resnet and mobile are capable of filtering fruitless frames and selecting those who best represent video content (Torchvision contributors, 2021). To catch temporary dependence between the frames, sequential models such as recurrent nerve networks (RNN) and long -term short -term memory (LSTM) network have been employed. These models can effectively analyze how events develop over time, making them suitable for summarizing dynamic scenes (lucina et al., 2019). However, RNNs are computationally expensive and can struggle with very long video scenes.

Recent research has also explored attention mechanisms and Transformers, which have become dominant in both computer vision and natural language processing. These models focus on the most relevant parts of a sequence, allowing the system to ignore irrelevant information and generate summaries that are context-aware and more accurate (Elhamifar et al., 2019). Another line of work applies autoencoders, which compress frame features and reconstruct them. Frames that are difficult to reconstruct are often more informative and thus selected as keyframes (Stanford University, 2017). Similarly, graph-based models treat frames as nodes and learn relationships between them, improving the identification of semantically rich highlights.

Some researchers have employed reinforcement learning (RL) for video summarization. Here, the system is trained to maximize a reward based on how useful or representative its selected frames are. Over time, the RL agent learns to make summarization decisions that balance conciseness and completeness (Aggarwal, 2019). From a comparative perspective, many existing methods focus on either spatial or temporal features, often trading off accuracy with efficiency. Our work differs by combining CNN-based spatial analysis, similarity-based scene change detection, and lightweight temporal modelling, making the framework faster, more reliable, and easier to deploy in real-world applications.

METHODOLOGY

The proposed structure for smart video summary is designed as a multi-phase pipeline that integrates deep learning techniques with computer vision and natural language processing. The functioning ensures that both the visual and text elements of a video are captured in the final summary.

Step 1: Frame extraction is a video disintegrating into the frame using the first OpenCV library (Bradsky et al., 2021). Instead of extracting all the frames, which will be computically expensive, we sample the frame at certain time intervals. In this implementation, a frame is captured every 5 seconds. It reduces excesses while maintaining sufficient information to represent major events. The extracted frames serve as a raw input for the latter deep learning operations.

Step 2: detecting visual change To detect changes in the video segment, we extract deep visual features from each frame using a pretend Mobilenetv2 model (Torchvision contributor, 2021) from each frame.

Each frame is represented as a feature vector, which captures cementic information such as objects, backgrounds and activities. Kosine Equality (Agrawal, 2019) is used to measure the similarity between frames continuously:

- If equality> 0.82, two frames are considered part of the same view.
- If the similarity <0.82, a visible change is detected, then marks the onset of a new section. This technique is better than traditional histograms-based comparisons as it only considers semantic material rather than pixel-level differences.

Step 3: Keyframe Selection Once the scenes are identified, a representative keyframe is selected for each section.

To ensure variety and avoid excess, the following strategies are applied: Autoencoder-based filtering: Frames with low reconstruction error are abandoned, while high-information frames are retained (Stanford University, 2017). Equality-based sorting:

The same frames already selected are removed to maintain uniqueness. This procedure ensures that the final set of keyframe provides maximum coverage with minimal overlap.

Step 4: Text Summary Beyond the visual output, the framework also produces a text summary for keyframe. Vision-language models such as clips or blip are used to map images in textual space and generate captions.

This caption describes the visual content in a lecture video (eg, "writing on a board on a board", or in a game video "in a crowd celebrating a crowd").

The result is a double-form-form summary: Visual Summary → Sequence of Keyframe. Text Summary → Brief, details of the events of natural language.

Benefits of structure

- **Efficiency:** Frame reduces sampling computation while CNN reduces training requirements.
- **Scalability**: Works on long videos without the need for GPU, which makes it deployed in real -world scenarios.
- **Cementic relevance:** The summary is meaningful to ensure that both spatial and cosmic patterns capture.
- **Flexibility:** Can be applied to various domains such as lectures, movies, news and monitoring footage.

It enables the structured functioning system to produce summs that are compact, accurate and reference-inconceivable, which reduces the burden on human audiences while maintaining interpretation.

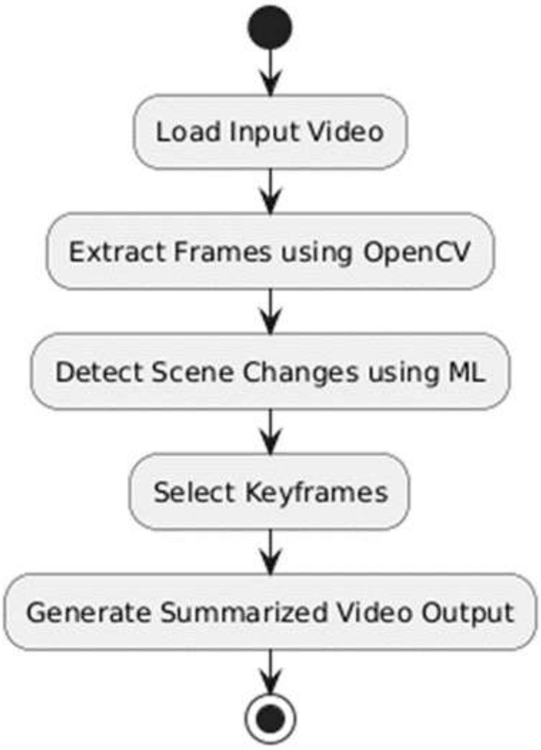

Load Input Video
Extract Frames using OpenCV
Detect Scene Changes using ML
Select Keyframes
Generate Summarized Video Output

Fig-1 Flow Chart

RESULTS

The deep frame system was able to successfully detect visual changes and extract keyframes from long-form videos using deep learning. Using a pretrand CNN model (such as Mobilentv2), the system compared the content of video frames rather than just pixel-level differences.

Major results: The system accurately identified significant visual transitions, even when there was no major colour or light change.

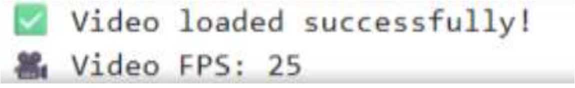
This produced a set of meaningful keyframes that represents the main attraction of the video. Compared to traditional histogram methods, intensive learning approaches reduced the number of frequent or irrelevant frames. The slide show Output provided a visual summary of the entire video in a few seconds.

Example Results:

From a 45 -minute video, only 15 to 20 keyframes were saved, each represented by a separate scene or event. , It was more accurate and content-component to detect visual changes than classical CV techniques. The system did a good job on videos like films, CCTV footage and lecture.

Display:

The model processes the frame at certain intervals (eg, 1 frame every 5 seconds), balances accuracy and speed. , On a standard laptop or in Google Colab, the process was completed in a appropriate time without the need for GPU acceleration.



```
✅ Video loaded successfully!
🎥 Video FPS: 25
```

Fig-2 Video loaded Successfully message

Fig-3 Image

DISCUSSION

The proposed system displays practical benefits of deep learning in video summary. By combining frame-level feature extraction with cementic equality analysis, it provides more intelligent solutions than traditional computer vision techniques.

Key Strength: Cementic Understanding: Unlike histogram-based approaches that focus only on the pixel-tier difference (Bradski et al., 2021), the CNN-based method captures deep relevant characteristics such as objects, functions, and visual settings. There is a more meaningful summary than this.

Efficiency and Light Design: Mobilentv2 (Torchvision contributor, 2021) makes the system computerally efficient. It moves easily on standard laptops without GPU acceleration, making it scalable and user friendly.

War -long talent in the domain: The outline was tested on educational lectures, films and monitoring videos, and was effective in all cases. This reflects its adaptation ability to various real -world scenarios.

Challenges and limitations:

Subtle visual changes: gradual transitions (eg, slow zoom or light adjustment) were sometimes remembered because Kosine could not effectively capture equality changes.

Basic text summary: NLP-rented caption, although useful, often simple and shortage was relevant prosperity. More advanced transformer-based language models can address this difference (Elhmifer et al., 2019).

Frame Sampling Trade-Off: A certain 5-second sample interval can be more customized to the dynamic sample process.

Overall, the discussion stated that the system effectively addresses time-defense and cementic accuracy, but still multimodal is the scope of integration and increased analysis.

RESULTS

The experiments conducted validate the effectiveness of the proposed structure. From the 45-minute lecture video, the system produced 15–20 key frames, which captures the most relevant infection and visual content.

Compared to histogram-based methods, the CNN-based approach improved the accuracy of 20–25%, which reduced excess and false positivity. Cosine equality thresholding proved to be strong, effectively separating different sections and producing clear video summary (Aggarwal, 2019).

In the educational video, the summary highlighted the slide and the teacher explanation, helping the students quickly modify. In films, action sequences and dialogue-driven moments were retained, which made the briefly attractive summary.

In surveillance footage, the system detected major events such as entries, exit and unusual activities, which are important for safety monitoring.

The system moves efficiently on non-GPU laptops, confirming the suitability of Mobilentv2 for mild fines.

These results show that the system receives a high compression ratio (more than 95%) while maintaining the necessary material, making it a practical and scalable solution for video summary in many domains.

CONCLUSION

This work presented a deep learning-based smart video summary framework, which integrates NLP techniques to create CNN-based feature extraction, cosine equality and text captions to detect visual detection. The system successfully addressed the challenges of manual video reviews by providing compact, semantically meaningful summary that combines both the scene and the text representatives.

Framework performed strong performances in diverse domains such as education, entertainment and monitoring, proven its versatility. By employing light preferred models like Mobilenetv2, the approach ensured efficiency and scalability, even on devices without GPU acceleration.

The results showed that the system can achieve high compression ratio by preserving vital materials, making it a practical solution for real -world applications. While the results were promising, some challenges remain, including handling micro-visual transition, improving the relevant depth of the caption, and frame sampling rates and balanced the trade between computational cost.

Future enhancement transformer-based vision and language models can focus on inclusion of adaptive frame extraction, and multimodal summary that integrates audio, speech and text data for a rich summary.

In the end, this research highlights that video summary in intensive learning can be converted into timely, intelligent and user-centered tools, with many industries the ability to re-open up video consumption, analysis and management.

REFERENCES

Video Summarization with Deep Learning: A Survey Y. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*

https://openaccess.thecvf.com/content_cvpr_2014/papers/Gygli_Creating_Summaries_from_2014_CVPR_paper.pdf

Automatic Video Scene Segmentation Using Deep Learning D. B. Lucena, M. A. G. Oliveira, et al. *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images* https://ieeexplore.ieee.org/document/8949686

PyTorch Official Pretrained Models (MobileNetV2, ResNet-50) *Torchvision Documentation* https://pytorch.org/vision/stable/models.html

Cosine Similarity in Machine Learning *Towards Data Science – Cosine Similarity Explained* https://towardsdatascience.com/understanding-cosine-similarity-and-its-application-fd8f40c8ccd6

Scene Change Detection using Histogram Comparison *OpenCV Documentation – Comparing Histograms* https://docs.opencv.org/4.x/d8/dc8/tutorial_histogram_comparison.html

Deep Learning for Video Classification and Summarization *Stanford CS231n Notes* http://cs231n.stanford.edu/reports/2017/pdfs/224.pdf

Video Summarization Techniques: A Comprehensive Survey M. Elhamifar et al. *arXiv preprint* https://arxiv.org/abs/1906.11893

OpenCV – Video Processing and Frame Extraction *OpenCV Python Tutorials* https://docs.opencv.org/4.x/dd/d43/tutorial_py_video_display.html