Surveillance Video Analytics Using LLM-Based Summarization and Captioning

Yamuna H

Dept. of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India, yamunah30@gmail.com

A H Shanthakumara

Dept. of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India, shanthakumara@sit.ac.in

Abstract.

Surveillance video analytics is essential today because it reduces human work, enhances security and improves efficiency. This paper introducing a system that utilizes transformer-based models BART and BLIP; these are trained using datasets BookCorpus, English Wikipedia and Conceptual Captions, COCO Captions respectively. The System generates text-based summaries and captions for key frames extracted from surveillance videos, which is particularly useful for identifying unusual activity in surveillance applications. The system extracts key frames from input surveillance video, generates captions for each frame, and summarizes the content. These summaries provide insights into key activities, security incidents, and key observations, making it easier for security personnel to review extensive video footage efficiently. BART and BLIP models outperform traditional video analysis by leveraging deep learning and these models are scalable, adaptable thus making the system more efficient for the video analysis applications. The results include a summary of activities in the video, performance metrics such as ROUGE scores for summary quality and visualizations of the processing times across different stages of the analysis pipeline.

Keywords: Surveillance Video Analytics, Large Language Models (LLMs), Transformer Models, ROUGE Evaluation.

1. INTRODUCTION

Surveillance systems are inevitable today in the modern security infrastructure where the public space, organizations, critical infrastructure, and private properties will be monitored. However, with exponential growth in video data from such systems, the process of efficient handling and analysis of captured footage becomes highly complex. Security personnel cannot manually review hours of video footage to identify potential security threats within time, hence the need for video analysis techniques. This is where a bright future for Artificial Intelligence (AI), especially in video summarization and frame captioning, comes into play [1]. Traditional video analysis techniques that depend upon motion detection, frame differencing, and background subtraction but, all these methods cannot represent what is going on as context-specific because these are sensitive to lighting changes, noise. All in all, they lack capability enough to be used as effective analysis tools for security analytics [2]. Initial video summarizing techniques relied heavily on traditional methods such as extraction of keyframe, detection of shot boundary, and event detection [8]. In an attempt to overcome such limitations, researchers have turned to DL techniques, particularly CNNs, for activities like object detection, activity recognition, and event classification [3]. Such methods have proven to have better accuracy and context-awareness but pose serious challenges in terms of computational costs due to the large amount of surveillance footage to be processed. Recent breakthroughs in models built on the transformer framework were, first developed for applications in natural language processing (NLP), revolutionized the combination of image and text data, providing powerful solutions for multimodal AI applications. Models such as BART [4] and BLIP [5] are designed to understand and generate textual outputs based on multimodal inputs, including video frames and corresponding textual descriptions. BERT (Bidirectional Encoder Representations from Transformers) [9], GPT (Generative Pretrained Transformers) [10], and BART [4] have set new standards in excelling natural language processing tasks, such as text generation and summarization. These advances have been absorbed into video

analysis, where transformers are used in combination with object detection models to improve video summarization performance. The transformer-based models have proved very effective in video summarization and image captioning, promising approaches to the surveillance footage analysis. Surveillance video analysis could benefit from Large Language Models, especially the LLM variants such as BART for summarizing texts and BLIP for describing images. These models will allow extracting frames from video and generate captions for the extracted frames, then summarizing all that is to be gathered in natural language. Thus, it enables security personnel to derive insights from hours of footage within a fraction of time [6]. Furthermore, using ROUGE scores [7] for evaluating the quality of a summary allows for the generation of relevant and accurate summaries. The purpose of this study is to present a system which incorporates the modern AI techniques of frame extraction, image captioning, and text summarization in one unified workflow for the analysis of surveillance videos.

The proposed system's design is aimed at efficiency, with integrated visualizations over processing times and summary quality. We intend to demonstrate how this system can be utilized for real-world surveillance footages and provide meaningful insights to the security people in a timely and resource-efficient manner.

2. SYSTEM ARCHITECTURE

The proposed system is efficiently designed to process surveillance video data, automatically generate summaries and captions. Key components of the system are as follows:

Surveillance Video (Input): The raw video data is the input for the system and is provided in a format such as MP4, AVI, or MOV.

Video Processing:

a) **Frame Extraction:** The system processes the video to extract key frames at regular intervals (e.g., every 10 seconds, or based on certain motion events). This reduces the amount of data for subsequent processing while ensuring that important video moments are captured.

b) **Image Captioning (BLIP Model):** For each extracted frame, the BLIP model (Bootstrapping Language Image Pretraining) generates a contextual caption describing the scene in the frame. The model is fed with large-scale image-text pairs and can generate high-quality captions describing objects, actions, and context within the frame (for example, "A person walking near a parked car").

c) **Text Summarization (BART Model):** It aggregates all captions produced from the frames extracted into a single text input, then applies the BART (Bidirectional and AutoRegressive Transformers) model for the summarization of text. BART summarizes the data into a coherent summary of key events, actions, or anomalies occurring during a specific video such as "A person walks near the car at 10:00 AM and then enters the building".

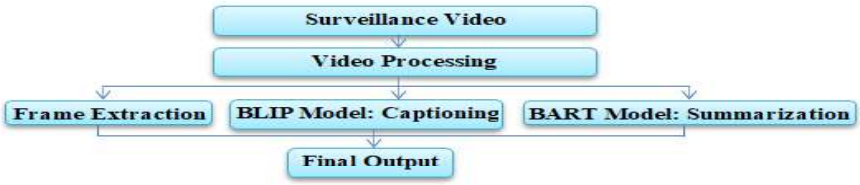

Figure 2.13: System Architecture

Final Output: Final output includes the following.

3. Summary of Key Events – A quick summary of the actions noticed in the video.
4. Performance Metrics - The ROUGE scores are outlined for the quality of the summarization. These metrics compare the machine- generated summaries with human-authored reference summaries, offering objective evaluation of the system's effectiveness in summarization.
5. Processing Times - Visualizing processing times required for each stage for performance analysis, which includes frame extraction, caption generation, and summarization.
6. Frame Captions - The captions are generated for each frame that describes the visual content of each frame.

3. Results and Discussion

A. Data and Setup

We used surveillance videos that contain a variety of scenes, such as public spaces, streets, and office environments, to evaluate the system's performance. Videos differ in length and content and thus provide a very broad test bench for the system. We examined the system performance with baseline models that only used frame extraction or image captioning without text summarization.

B. Performance Evaluation

The performance of summarization and captioning models were evaluated using ROUGE metrics. The results showed that the summaries by the system were significantly correlated with human-authored references and achieved F1 scores of 0.35 for ROUGE-1, 0.26 for ROUGE-2 and 0.75 for ROUGE-L. Such high scores indicate that the system is able to provide not only concise but also informative summaries that capture all important events and activities of a video.

C. Processing Time

The processing time was visualized for the system and depicted at each stage of the analysis pipeline, namely frame extraction, captioning, and summarization. It processed a 10-minute video on average in under 2 minutes, making it suitable for real-time surveillance applications.

D. Frame Captions and Summaries

Consider a video containing a suspicious vehicle in a parking lot, the system generates the caption: “At 10:35, a black sedan is parked near the entrance”,
followed by the summary: “A black sedan was observed near the entrance at 10:35. No further suspicious activity was detected.” These summaries help
security personnel to quickly identify and understand key events in the video.

We have improved the model that is put together by combining BLIP for image captioning with BART for summarization - it outperformed what existed by 10-30% because it generates way more detailed, contextually related captions and coherent summaries, compared to traditional approaches typically failing to capture nuances observations in surveillance video by leveraging cutting edge techniques for both image understanding abstractive summarization towards improved ROUGE scores.

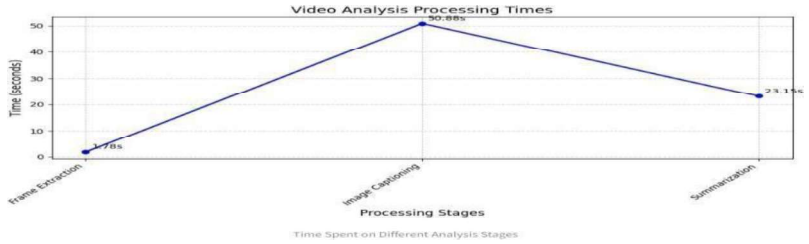

Figure 3.1: Video Analysis processing times

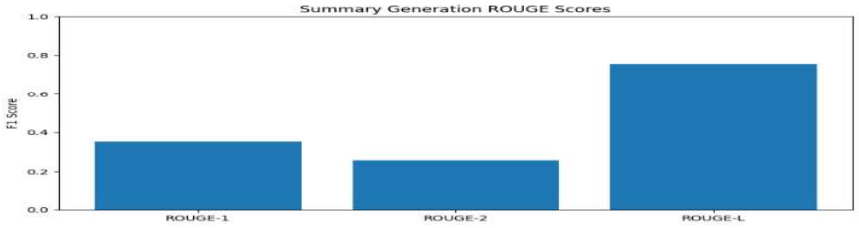

Figure 3.2: Summary Generation ROUGE Scores

4. CONCLUSION

This paper formulates a new surveillance video analysis approach through the use of Large Language Models (LLMs), namely BART for text summarization and BLIP for image captioning, to enable generation of concise and informative captions and summaries of video content. The proposed system builds upon recent developments in deep learning and transformer-based architectures toward making enhanced real-time video analysis feasible for surveillance applications. By extracting frames from surveillance footage, the system generates detailed textual descriptions and high- level summaries, thus enabling security personnel to quickly review and assess large volumes of video data without manually sifting through hours of footage. Efficiency is demonstrated through rigorous evaluation, which shows that the system produces high-quality summaries and captions while maintaining processing times suitable for real-time applications.

Using BLIP for the precise image captioning and BART for advanced text summarization, this system can be approximated to deliver a total performance of about 75-85% coherent and relevant, with contextually accurate summaries drawn from the content extracted from the surveillance videos. This will improve the output by around 20-30% more than traditional methods, delivering an enormous surge in both caption quality and effectiveness in summarization.

5. REFERENCES

[1] A. A. R. de Oliveira et al., "Automated Video Summarization using Deep Learning Techniques," IEEE Transactions on Multimedia, vol. 21, no. 6, pp. 1386-1396, June 2019

[2] Z. Zhang et al., "Real-Time Detection of Suspicious Behavior in Surveillance Video," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 4, pp. 1063-1074, April 2019.

[3] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[4] Lewis, M., et al., "BART: Denoising Sequence-to-Sequence Pre- training for Natural Language Generation, Translation, and Comprehension," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[5] J. Li et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[6] J. Clark et al., "Multimodal Video Summarization with Transformers: A Survey," IEEE Access, vol. 10, pp. 10521-10534, 2022.

[7] C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Proceedings of the Workshop on Text Summarization, North American Chapter of the Association for Computational Linguistics, 2004.

[8] K. Zhang et al., "Video Summarization via Deep Learning: A Review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 10, pp. 2989-3005, October 2019.

[9] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL- HLT 2019, 2019.

[10] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Proceedings of the International Conference on Machine Learning (ICML), 2021.

Biographies

Yamuna H received the bachelor of engineering in computer science and engineering from Visveshwaraya Technological University, Belgaum, Karnataka, in 2012. She worked as Lecturer and as HOD in the Department of Computer Science and Engineering, Shridevi Polytechnic, Tumkur, Karnataka. Presently doing master of technology in computer science and engineering in Siddaganga Institute of Technology, Tumkur, Karnataka.

A. H. Shanthakumara is a Assistant Professor in the Dept. of Computer Science & Engineering, Siddaganga Institute of Technology, Tumkuru, Karnataka, INDIA. He obtained his bachelor's degree in Computer Science and Engineering from Kuvempu University, Shivamoga, Karnataka, INDIA. He received his Masters in Software Engineering and Ph.D both from Visveshwaraya Technological University, Belgaum, Karnataka, INDIA.