Deep-Learning Powered Video Analysis and Scene Summarization

Session: BSc. Spring 2024

Project Supervisor: Dr Muhammad Imran Shehzad

Co-Supervisor: Dr Shoaib Azmat

Submitted By

Shehryar Khattak

FA20-BCE-004

Department of Electrical and Computer Engineering

Comsats University Islamabad Abbottabad Campus

Certification

This is to certify that **Shehryar Khattak FA20-BCE-004** has successfully completed the final project **Deep Learning Powered Video Analysis and Scene Summarization,** at the **Comsats University Islamabad Abbottabad Campus** to fulfill the partial requirement of the degree **Bachelors in Computer Engineering.**

External Examiner

[Name of Examiner]

[Designation]

Project Supervisor

Dr Muhammad Imran Shehzad

Assistant Professor

Dr Shoaib Azmat

Assistant Professor

Chairman

Department of Electrical and Computer Engineering,

Comsats University Islamabad Abbottabad Camous

Abstract

In the ever-evolving digital landscape, video content has surged exponentially, becoming a primary medium for capturing and conveying information. This thesis presents a deep learning-powered system designed to address the challenge of efficiently analyzing and summarizing video data . The system employs advanced object detection and video captioning techniques to understand and describe video content, specifically focusing on human activities within the videos. It provides real-time identification of objects and activities, along with concise summarizations that highlight critical events. This approach will allow for rapid analysis of extensive video footage, ensuring crucial moments are captured and reviewed without the need to watch hours of footage. Applications range from enhancing security and surveillance understanding consumer behavior in retail environments, healthcare monitoring and content management in entertainment. Ultimately, this project lays the groundwork for more advanced AI-driven video analysis tools, paving the way for intelligent Video Analysis solutions that are both time-efficient and cost-effective.

Undertaking

I certify that the project **Deep Learning Powered Video Analysis and Scene Summarization** is our own work. The work has not, in whole or in part, been presented elsewhere for assessment. Where material has been used from other sources it has been properly acknowledged/ referred.

Shehryar Khattak

FA20-BCE-004

Acknowledgement

We truly acknowledge the cooperation and help made by my supervisors **Dr Muhammad Imran Shehzad and Dr. Shoaib Azmat** who provided assistance throughout the development of this Project**.** They have been a constant source of guidance throughout the course of this project. We would also like to thank committee members, **Dr Shahid Mustafa and Dr. Ali zahir** who offered guidance and support throughout this project.

We are also thankful to our friends and families whose silent support led us to complete our project.

Table of Contents

Figures

Acronyms

LSTM	Long-ShortTermMemory
S2VT	Sequence to Sequence Video to Text
CNN	Convolutional Neural Network

List of Equations

Chapter 1 Introduction

1.1 Deep Learning based Video Analysis

Today's world is flooded with videos. From security cameras in cities to videos shared on social media, we are recording moments all the time. This huge amount of video data has a lot of important information. But, with so much to watch, we often miss out on key details. Imagine if we had an Intelligent Computer program that could watch these videos for us and detect certain human actions that we are looking for as well as inform us about all the important parts of the video. That's where this project comes in. The goal of this project is to create a system that uses deep learning to analyze videos and detect and understand the main human actions in them and generate a short video clip which shows only the important parts of the video. This way, we can easily search through the extensive video data and find out the most important information about what's happening in them without having to watch everything.

1.2 Motivation and Aim

In our fast-paced world, efficiency is key. As videos become a primary source of information, we find ourselves overwhelmed with content. Manually analyzing videos takes a lot of time and energy, and there's always a risk of missing out on crucial details. My motivation lies in addressing this challenge. By using deep learning to automate video analysis, I aim to automate the video analysis process without the need to watch every second of footage. This not only saves time but ensures that vital events or activities aren't overlooked.

The potential uses for a deep learning-powered video analysis and summarization system are vast and varied:

Security and Surveillance: Monitoring public spaces, transportation hubs, and sensitive locations to quickly identify unusual or suspicious activities.

Retail Management: Understanding customer behaviors in stores, like which aisles they visit most, or identifying potential theft situations.

Healthcare: Observing patients in hospitals or elder-care facilities to ensure their safety and well-being.

Entertainment: Quickly categorizing and tagging content, making it easier for viewers to find videos of interest or for platforms to offer targeted recommendations.

These are just a few examples. The ability to quickly and accurately analyze video content opens the door to countless other possibilities across various industries and personal interests.

1.3 Problem Statement and Proposed Solution

The surge in video content across sectors presents a challenge. Manual analysis of these videos is not only tedious and error-prone but also inefficient. From security to retail to entertainment, the need is clear: a faster, reliable way to extract vital details from vast video data.

Proposed Solution:

My solution is a deep learning-driven video analysis system that has the following capabilities:

• **Object Detection:** Utilizes advanced models, like YOLO, to detect humans and other objects in

videos.
• **Video Captioning:** Employs specialized sequence based models which take in and process video data to detect and describe human actions, creating descriptions of the human activities.
• **Video Summarization:** Generate a short summary of the video, highlighting the crucial human activities for a quick overview.

My goal is a system that quickly and efficiently delivers the essential details from a large amount of video data.

1.4 Objectives / Outcomes

My main goal is to Develop an AI-driven tool that can analyze, highlight, and summarize video contents, focusing on human actions, to offer users a quick grasp of the important details without needing to watch the entire video.
Specific Goals:
Real time Analysis: The system should process videos in real time or nearly so, giving fast results and summaries.
Accuracy: The system must reliably detect and recognize human actions, minimizing any false detections.
Compatibility: Our system should smoothly work alongside other software, devices, or platforms, broadening its use.
Expected Benefits:
Efficiency: The tool will cut down the time and effort currently used in manual video analysis, leading to cost savings.
Real time Alerts: The system will be able to generate alerts once it detects certain human activities
Summarization: The system will generate a short summarized video clip which will contain all the clips involving important human activities.
Building Block for Future: This Project could serve as a foundation, making way for more refined AI applications and expanding its detection spectrum.
Conclusively, reaching these targets will promote more widespread use of AI in video analysis, catering to varying industry and personal needs.

1.5 Block Diagram

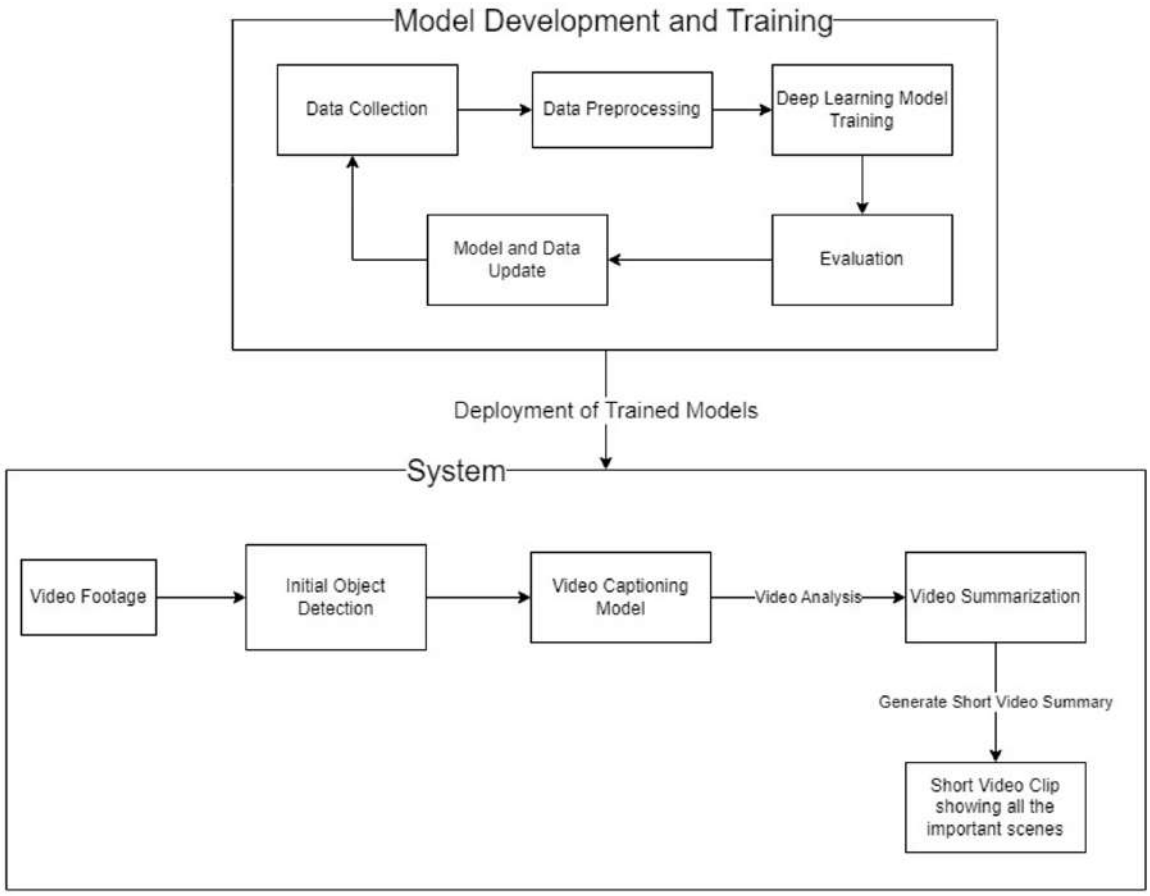

Fig 1.1 Block Diagram

Chapter 2 Literature Review

2.1 Methodology

I have reviewed the extensive literature of the many architectures and Models for Deep-Learning based Video Analysis, The Architectures which were reviewd all had 3 common Processes for the Video Captioning Task.

Preprocessing: Preparing raw video data for analysis.

Feature Extraction: Leveraging CNN models to identify and extract information from Video Data.

Sequence Processing: Utilizing sequence based Deep Learning models like LSTM(Long Short Term Memory), GRU(Gated Recurrent Unit), or Transformers to generate video captions.

2.2 Metrics for Evaluation

1. **METEOR (Metric for Evaluation of Translation with Explicit Ordering):**
   A metric that was originally created for Machine Translation but has been adapted for Capton Generation. It checks how well the generated text matches the meaning of a reference text, considering synonyms and sentence structure. Aimed at detecting similarity in meaning between True Output and Predicted Output.

2. **BLEU (Bilingual Evaluation Understudy):**
   A standard measure for comparing a generated text against a reference translations.Focuses on how many words and phrases in the generated text appear in the reference texts.
3. **CIDEr (Consensus-based Image Description Evaluation):**
   Developed Specifically for scoring the descriptions of images or videos. Looks at the relevance and uniqueness of the generated descriptions compared to a set of reference captions.Places more weight on terms that better capture the essence of the image or video.

2.3 Datasets

2.3.1. Dataset Exploration and Description:

In this study, we have utilized several publicly available video datasets, including MSVD (Microsoft Video Description Corpus), UCF Crime, and WorldExpo'10. These datasets were selected for their diversity in content, encompassing various scenarios and activities, which are essential for developing a robust video analysis model. The MSVD dataset, for instance, includes videos with a wide range of human activities, making it suitable for testing the performance of scene summarization algorithms.

2.3.2. Data Characteristics:

The videos in these datasets vary in length, resolution, and content type. The MSVD dataset consists of short clips, usually about 10-25 seconds long, with annotations in multiple languages.

UCF Crime contains longer videos, often several minutes in duration, depicting various criminal activities,The variability in these datasets provides a comprehensive test bed for evaluating the effectiveness of deep-learning models for video Analysis.

2.4 Literature Review

2.3.1 Sequence to Sequence – Video to Text [1]

This research introduces S2VT model which utilizes a CNN-LSTM hybrid architecture for converting videos to textual descriptions of the events in the video.
In this Architecture, Feature Extraction was done using VGG-16 and Alexnet and Caption Generation using LSTM. The datasets used were MSVD, MPII-MD and M-VAD. The evaluation metric used is METEOR which was showing a 29.2% score.

2.3.2 Crowd Video Captioning [2]

In this work, a method for crowd behaviour analysis through video captioning is introduced utilizing a CNN-LSTM hybrid approach.
The CNN Architecture Inception V3 used to extract Frame level features while C3D (which is primarily used for Video Classification) is used to extract Video level features.
The features are then fed into an LSTM network in sequence, after the encoding stage is finished, The decoding stage will begin and the Model will start generating the captions in sequence. The dataset that was used for model training and evaluation was the Shanghai WorldExpo'10 dataset which was for crowd behaviour analysis. The metric for evaluation used was METEOR and CIDEr giving the results of
METEOR = 58.38%
CIDEr =72.07%

2.3.3 Captionomaly: A Deep Learning Toolbox for Anomaly Captioning in Social Surveillance Systems [3]

In this research paper we are introduced to a system for real-time video captioning system for detecting human activities in surveillance videos. The system combines anomaly detection with video captioning models to create descriptive reports of detected anomalies. It utilizes UCF Crime dataset to train the Anomaly detection model and the newly created UCF-Crime Video Description (UCFC-VD) dataset for Video Captioning. The tool is designed to be used on surveillance video to detect human activities both normal and anomalous.
The system is comprised of two parts:
1)Anomaly Detection:
The video clip is divided into segments.
It utilizes a 3d Conv Net to extract spatial and temporal features from the video clip segments
The features are fed into a Multiple Instance Learning (MIL) framework which will assign anomaly scores to each segment.
Then the segments with the highest anomaly scores are passed to the Video Captioning Model.

2)Video Captioning:

The anomalous parts of the video clip is passed to the Resnet XT feature Extractor and the extracted features are then passed in a sequence to a GRU Network which will produce the captions in sequence.

The metric used for evaluating the outputs generated by the video captioning model are the cap_score which is the unified form of the Bleu-4, METEOR, CIDEr,

and ROUGEL scores using the formulae.

$$cap\_score = \frac{1}{4}\left(\frac{B4_i}{B4_b} + \frac{C_i}{C_b} + \frac{M_i}{M_b}\right)$$

The highest capscore_score achieved was 96.02%.

2.3.4 SAVCHOI: Detecting Suspicious Activities using Dense Video Captioning with Human Object Interactions [4]

In this research, a method for detecting human activites in videos is devised by using the two stage Process of human-object Interactions recognition and dense event captioning.

The system takes both video and audio data as input.

For Training and evaluation of the models, The datasets UCF-Crime and HICO-DET(which is used for Human Object Interactions) were used.

It uses QAHOI (Query-based Anchors for Human Object Interaction Detection) based on a SWIN transformer to extract the features of RGB frames and Optical flow frames from the video data. And the Audio features are extracted from the audio data using the VGGish Model which was pretrained for audio classification,

Both of these features are then passed to the Bi modal transformer for video captioning, which will output the captions,

Finally, a fine-tuned BERT model classifier analyzes the video captions generated, classifying them as suspicious and non-suspicious. The METEOR score was used to score the results and the final results gave a score of 15.05%.

2.3.5 Real Time Crime Detection by Captioning Video Surveillance Using Deep Learning [5]

In this work, a real time video captioning system is introduced whose main application is detecting human activities in Surveilance footage.

Methodology

- Frames are extracted from the video
- They are passed through the VGG-16 architecture to extract features from them.
- The features are passed in sequence to the LSTM layer ,after the features have all been input into the network, the LSTM will start to output the captions for the video clip.
- The generated captions are encrypted and stored for security.

Datasets used:

- MSVD
- UCF Crime

Results were not published

Chapter 3 Proposed Solution

3.1 Proposed Solution

The methodology which I implemented for the Project has two parts.

- **Object Detection using YOLO**
  For the initial Object Detection, I have chosen the YOLO (You Only Look Once) Model due to it's efficiency in Speed and Accuracy. It divides the input image into a grid. Each grid cell predicts confidence scores and other data for objects in those cells. Each box contains predictions for position, size, and class probabilities. The confidence score signifies the presence of an object The model combines the predictions from all grid cells and applies non-maximum suppression to filter overlapping boxes, selecting the one with the highest confidence. This single-pass prediction contrasts with traditional methods that perform separate passes for object localization and classification, making YOLO exceptionally fast.

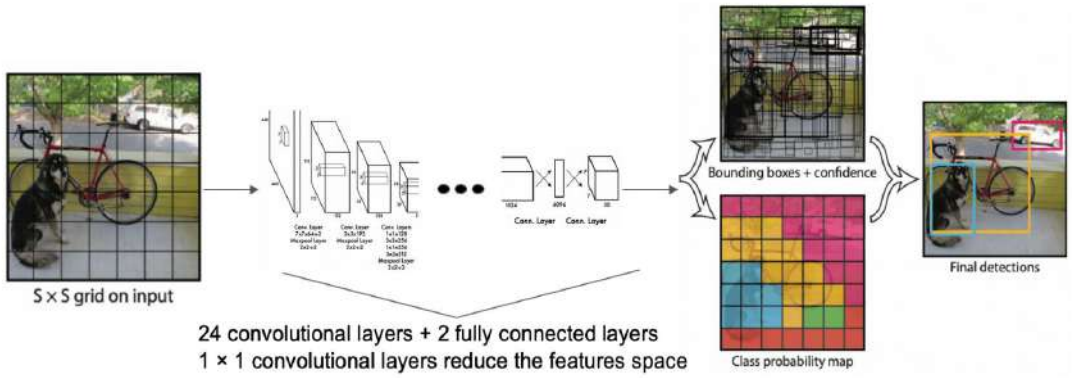

Fig 3.1 YOLO Architecture

- **Video Captioning using S2VT Architecture**
  For the Video Captioning Task, I have implemented the S2VT Architecture which uses CNN-LSTM hybrid architecture to generate captions for input video clips.

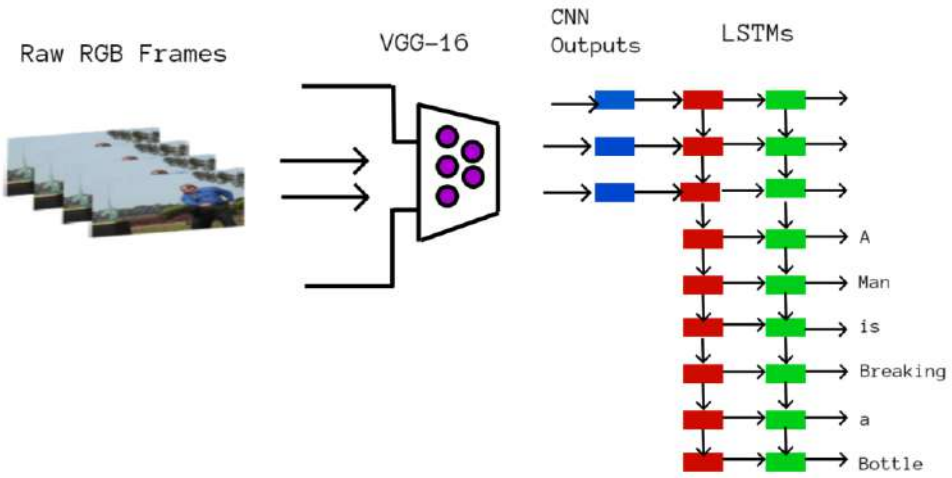

Fig 3.2 S2VT Architecture

VGG-16

This Architecture uses a VGG-16 Network Pre-Trained on the Object Classification Task to extract the features. The final 3 Fully-Connected layers have been removed.

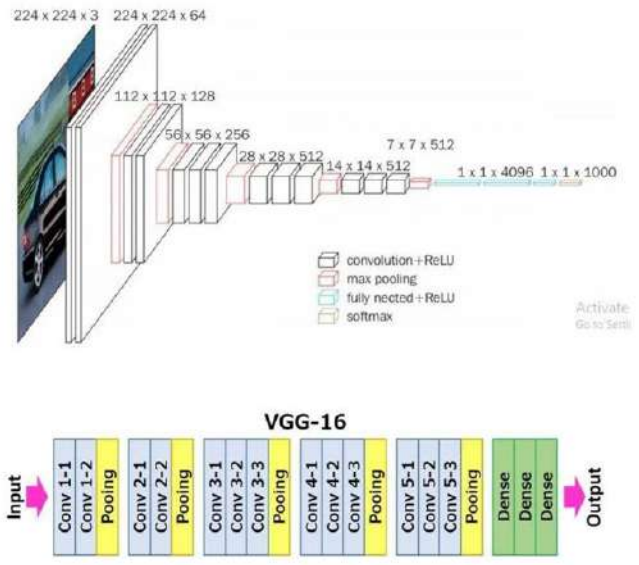

Fig 3.3 VGG-16 Architecture

LSTM

It passes the extracted features to the LSTM Model with an Encoder and Decoder Stage to generate the Captions.

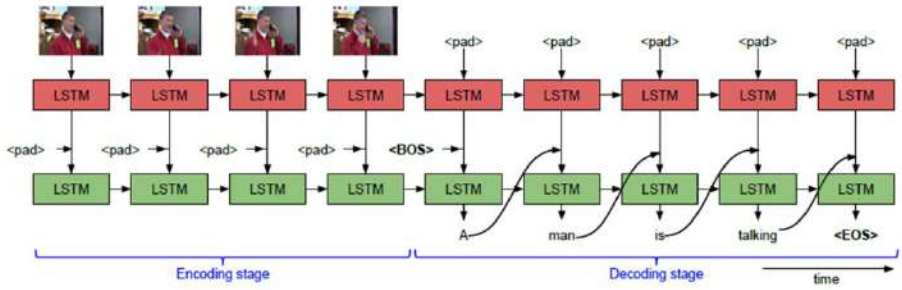

Fig 3.4 LSTM Network

3.2 Simulation and Testing

For my Models Testing, I have utilized Google Colab to gain access to GPU. To load and test the S2VT Model, I have used the Tensorflow Framework. To load and test the YOLO Model, I have used OpenCV Library.I have used the Test Set of the MSVD Dataset to test the Performance of the Pre-Trained S2VT Model.I have gathered and used random Surveilance Videos to Test my Object Detection Model.

Chapter 4

4.1 Object Detection Results

The YOLO (You Only Look Once) model was employed for object detection tasks. The unique architecture of YOLO allowed for real-time processing of videos while maintaining high accuracy. Through the use of the YOLO architecture, the model demonstrated exemplary performance in detecting objects with high confidence scores and generating precise bounding boxes. Representative outputs are included, showcasing the bounding boxes around detected objects, along with their class probabilities.

Fig 4.1 Output showing bounding boxes with Class Probabilities

4.2 Video Captioning Results

The results from the Sequence to Sequence – Video to Text (S2VT) architecture used for video captioning are presented in this section. A Pre-Trained S2VT Model was used for Experimentation which was downloaded from Github, and the Test Set of the MSVD Dataset was used.

Sample outputs from the video captioning model are presented, comparing the ground truth (GT) captions with the model-generated captions. For instance:

Fig 4.2 GT: a man is playing a violin
Model: a man is playing

Fig 4.3 GT: a woman is cooking eggs
Model: a woman is cooking

Fig 4.4 GT: a man is riding a motorcycle
Model: a man is riding a car

Fig 4.5 GT: : someone is pouring rice into rice cooker
Model: a person is pouring water in a bowl

Chapter 5 Discussions

5.1 Assessment of Object Detection Model

The object detection module, powered by the YOLO model, has demonstrated its capability to identify and localize multiple objects within a single frame accurately. The bounding boxes and class probabilities presented in the results attest to the model's precision in real-world scenarios. the results suggest that the model is robust in detecting various objects with high confidence, which is crucial for practical applications like surveillance.

5.2 Assessment of Video Captioning Model

The video captioning results demonstrate the model's ability to comprehend and describe video content. Although the captions generated by the model are less detailed compared to the ground truth, they capture the essence of the scenes. The model's success in generating relevant captions underlines its potential utility in creating descriptive captions for videos, aiding in automated Video Analysis applications.

5.3 Implications for Practical Application

The object detection results have shown potential for deployment in real-time systems, where rapid and accurate identification is crucial. Similarly, the video captioning model, even in it’s initial Development Stage has shown great promise in Video Analysis and in generating descriptions for human activities in videos.

5.4 Limitations and Challenges

During the experimentation and testing, several limitations of the models were discovered, in the case of Object Detection, The model had difficulty detecting objects at a longer distance from the camera, and sometimes it resulted in false positives. While in the case of the Video Captioning Model, the Output Captions Generated by the model sometimes did not match with the Ground Truths for the videos from the Test Dataset and produced wrong descriptions for the Video Content.

5.5 Recommendations for Future Research

Future research should focus on enhancing the models to cover the weaknesses demonstrated by the tested models as shown above. Enhancements in the models could include fine-tuning with a more extensive dataset and incorporating better data preprocessing techniques to make the models more focused on human activities. More research should also be conducted to discover other Techniques in Computer Vision and Natural Language Processing to increase the Accuracy and Efficiency of the Models.

Chapter 6 Summary

6.1 Summary and Future work

This project developed a deep-learning-powered system for video analysis and scene summarization. The primary aim was to enable efficient analysis and summarization of video content, particularly focusing on human activities. The methodology employed involved advanced object detection using the YOLO model and video captioning using the Sequence to Sequence Video to Text (S2VT) architecture. Testing on diverse datasets demonstrated the system's capability to accurately detect objects and generate concise video captions that summarize crucial human activities.

The findings from this thesis suggest several avenues for further research. Firstly, enhancing the accuracy of both object detection and video captioning models could involve integrating more sophisticated deep learning architectures or employing larger and more varied training datasets. Additionally, exploring real-time processing capabilities would be vital for applications requiring immediate analysis, such as surveillance and emergency response systems. Future studies could also expand the system's applicability to other areas like traffic management, wildlife monitoring, and automated content generation for media. Addressing the challenges of scale and environment variability will further refine the system’s utility and robustness.

Chapter 7 Conclusion

7.1 Conclusion

The research undertaken in this thesis successfully demonstrates the potential of a deep-learning-powered system to revolutionize video analysis. The developed system not only supports rapid and accurate video content Analysis and Summarization but also ensures that critical moments are effectively captured without requiring exhaustive manual review. This system holds promise for significant improvements in fields ranging from security and surveillance to healthcare and retail, where quick and reliable video analysis is paramount. Continued advancements in this domain could lead to more adaptive, intelligent systems capable of functioning across diverse environments and fulfilling the growing demand for automated video analysis solutions.

References

[1] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue,Raymond Mooney,Trevor Darrell, Kate Saenk(2015) Sequence to Sequence – Video to Text

[2] Liqi Yan1, Mingjian Zhu1, Changbin Yu (2019). Crowd Video Captioning

[3] Adit Goyal, Murari Mandal, Vikas Hassija, Moayad Aloqaily, Vinay Chamola (2023). Captionomaly: A Deep Learning Toolbox for Anomaly Captioning in Social Surveillance Systems

[4] Ansh Mittal, Shuvam Ghosal, Rishibha Bansal (2022). SAVCHOI: Detecting Suspicious Activities using Dense Video Captioning with Human Object Interactions

[5] Nagesh Nayak, Shlesha Odhekar, Sapna Patwa, Sukanya Roychowdhury (2022). Real Time Crime Detection by Captioning Video Surveillance Using Deep Learning