Video Summarization

Submitted By

Parthkumar Patel

21MCEC16

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2023

Video Summarization

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering (Computer Science and Engineering)

Submitted By

Parthkumar Patel

(21MCEC16)

Guided By

Dr. Vishal Parikh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2023

Certificate

This is to certify that the major project entitled **"Video Summarization"** submitted by **Parthkumar Patel (Roll No: 21MCEC16)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering (Specialization in title case, if applicable) of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-I, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Vishal Parikh
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sudeep Tanwar
Professor,
Coordinator M.Tech - CSE (Specialization)
Institute of Technology,
Nirma University, Ahmedabad

Dr. Madhuri Bhavsar
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. R. N. Patel
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Parthkumar Patel**, Roll. No. **21MCEC16.**, give undertaking that the Major Project entitled "**Video Summarization**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering (Computer Science & Engineering)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student

Date:

Place:

Endorsed by

Dr. Vishal Parikh

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Vishal Parikh**, Assistant Professor, Computer Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Madhuri Bhavsar**, Hon'ble Head of Computer Engineering/ Information Technology Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. R. N. Patel**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Parthkumar Patel
21MCEC16.

Abstract

In today's society, a lot of information is released every day in various formats, including text, video, and images. Day to day YouTube is generating large amount of data. In July 2015, YouTube revealed that it receives over 400 h of video content every single minute, which translates to 65.7 years worth of content uploaded every day. Since then, we are experiencing an even stronger engagement of consumers with both online video platforms and devices (e.g., smartphones and wearables) that carry powerful video recording sensors and allow instant uploading of the captured video on the Web Video summarization technologies aim to create a concise and complete synopsis by selecting the most informative parts of the video content.Generally LSTM is used for this video summarization.In this project, we present an automated system for extracting specific person's footage from large surveillance videos. The system employs computer vision techniques, including person detection and recognition algorithms, to analyze each frame of the video and identify the desired persons of interest. The person detection algorithm utilizes the YOLOv3 object detection model to locate persons in the frames, while the person recognition algorithm utilizes a pre-trained face recognition model to verify the identity of the detected persons. By combining these algorithms, the system identifies frames that contain the target persons and extracts them for further analysis. The extracted frames are then used to create a short video comprising the selected footage. The system offers a convenient and efficient solution for surveillance video analysis, allowing for the isolation of relevant footage and reducing the need for manual inspection.

Abbreviations

LSTM	Long Short-term Memory
DL	Deep Learning.
NLP	Natural Language Processing.
YOLO	You Only Look Once, Version 3.

–

Contents

List of Figures

Chapter 1

Introduction

1.1 Genral Introduciton

Network of surveillance cameras are everywhere nowadays major problem is to figure out how to extract useful information from the videos captured by these cameras. The process of extracting useful information from the videos is called as Video Summarization. The goal of this project is to build a model which can create summary of video more accurately. Above is a case, there are many other applications are there. [1].

With the increasing prevalence of surveillance cameras in public spaces, there is a growing need to efficiently analyze the vast amount of video data generated by these systems. One particular challenge is extracting specific persons' footage from large surveillance videos, which can be a labor-intensive and time-consuming task. Manual review of the entire video content is impractical, necessitating the development of automated methods to streamline the process.

1.2 Introduction of Video Summarization

In recent years, the demand for efficient video summarization techniques has increased due to the widespread availability of digital media and the abundance of video content. Video summarization involves condensing videos into shorter representations while preserving their essential content and overall meaning. This process allows users to quickly grasp the main ideas and key aspects of a video without watching it in its entirety.

Applications for video summarization may be found in many fields. It speeds up the study of lengthy film by investigators in the surveillance and security industries, making it easier to spot important events or suspicious activity. Summarization facilitates quicker content browsing and targeted retrieval from huge video collections for video search and retrieval. Additionally, video summarising enhances user experience in journalism, social media, and online video platforms by giving succinct summaries of lengthy videos, increasing content discovery, and lowering information overload.

Handcrafted elements like as colour histograms, motion vectors, or keyframe selection techniques are frequently used in traditional video summary approaches. However, the advent of deep learning has transformed this sector. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are deep learning algorithms that can learn meaningful representations directly from raw video data. CNNs excel in extracting spatial information from video frames, but RNNs represent temporal relationships effectively by taking into account the sequential character of video material. Video summarising algorithms that use deep learning can extract rich spatiotemporal information, interpret context, and provide succinct summaries that emphasise the most essential regions of the video.

The use of deep learning in video summarising opens up possibilities for improving the quality, efficiency, and flexibility of summarization systems. To increase the performance of video summarization systems, researchers have suggested several architectures and methodologies, such as attention mechanisms, reinforcement learning, and generative models. Furthermore, the availability of large-scale video collections and assessment measures has permitted benchmarking and comparative analysis, promoting progress in the field.

However, video summary remains difficult. Videos include visual, aural, and textual information, and properly combining these modalities to provide complete summaries is difficult. To manage computational restrictions, real-time or near-real-time summarization necessitates efficient processing algorithms. Addressing these issues and enhancing the accuracy, variety, and interpretability of video summarising systems are ongoing research projects.

The purpose of this study is to give a complete overview of video summarising techniques, with a special emphasis on deep learning approaches. It looks at current accomplishments, problems, assessment measures and benchmark datasets, and emerging trends and future directions. Understanding the present state of the art will enable academics and practitioners to take use of deep learning's capabilities and create more effective video summarization systems, unleashing the potential of video data in a variety of disciplines.

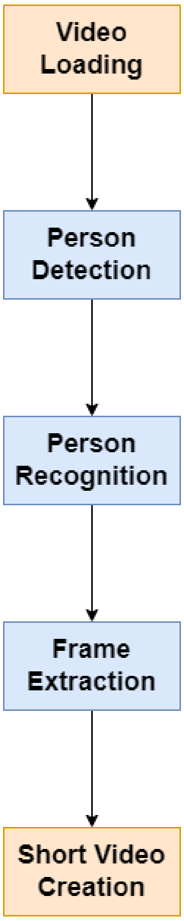


Figure 1.1: Block Diagram

1. Video Loading: The input video file is loaded for processing.

2. Person Detection: YOLO (You Only Look Once) model is used to detect persons in each frame of the video.

3. Person Recognition: The detected faces are extracted and compared with known persons using a face recognition model.

4. Frame Extraction: Frames containing the target persons are stored for creating a short video.

5. Short Video Creation: The extracted frames are combined to create a short video.

1.3 Introduction of YOLOv3

In this project, we propose an automated system that addresses this challenge by leveraging computer vision techniques. Our system combines person detection and recognition algorithms to identify and extract frames containing the footage of specific persons of interest. By utilizing state-of-the-art deep learning models, we can accurately locate persons in each frame and verify their identities.

The first step of our system involves person detection, where we employ the YOLOv3 object detection model. This model is trained to identify various objects, including persons, in real-time. By applying this model to each frame of the surveillance video, we can quickly and accurately detect the presence of individuals.

Once persons are detected in the frames, the system proceeds to the person recognition step. Here, a pre-trained face recognition model is employed to match the detected persons with known individuals. This model has been trained on a dataset of known persons' images, enabling it to assign identities to the detected faces. By establishing a match, we can determine if the detected persons are the specific individuals we are interested in.

By combining the outputs of the person detection and recognition algorithms, the system identifies frames that contain the target persons. These frames are then extracted and compiled to create a concise and focused video containing the desired footage. This automated approach saves significant time and effort compared to manual review, enabling efficient analysis of surveillance videos and facilitating targeted investigations.

In summary, our system offers an automated solution for extracting specific persons' footage from surveillance videos. By integrating person detection and recognition algorithms, we can accurately identify and extract frames of interest, eliminating the need for manual inspection of the entire video content. This automated approach has the potential to enhance the efficiency and effectiveness of video analysis in various applications, including security, law enforcement, and forensic investigations.

The YOLOv3 architecture, with its multi-scale detection and anchor box mechanism, allows for the detection of objects at different sizes and aspect ratios in real-time. It balances speed and accuracy, making it well-suited for applications such as object detection in surveillance videos, including person detection in the context of the above project.

With the increasing prevalence of surveillance cameras in public spaces, there is a growing need to efficiently analyze the vast amount of video data generated by these systems. One particular challenge is extracting specific persons' footage from large surveillance videos, which can be a labor-intensive and time-consuming task. Manual review of the entire video content is impractical, necessitating the development of automated methods to streamline the process.

In this project, we propose an automated system that addresses this challenge by leveraging computer vision techniques. Our system combines person detection and recognition algorithms to identify and extract frames containing the footage of specific persons of interest. By utilizing state-of-the-art deep learning models, we can accurately locate persons in each frame and verify their identities.

The first step of our system involves person detection, where we employ the YOLOv3 object detection model. This model is trained to identify various objects, including persons, in real-time. By applying this model to each frame of the surveillance video, we can quickly and accurately detect the presence of individuals.

Once persons are detected in the frames, the system proceeds to the person recognition step. Here, a pre-trained face recognition model is employed to match the detected persons with known individuals. This model has been trained on a dataset of known persons' images, enabling it to assign identities to the detected faces. By establishing a match, we can determine if the detected persons are the specific individuals we are interested in.

By combining the outputs of the person detection and recognition algorithms, the system identifies frames that contain the target persons. These frames are then extracted and compiled to create a concise and focused video containing the desired footage. This automated approach saves significant time and effort compared to manual review, enabling efficient analysis of surveillance videos and facilitating targeted investigations.[2]

In summary, our system offers an automated solution for extracting specific persons' footage from surveillance videos. By integrating person detection and recognition algorithms, we can accurately identify and extract frames of interest, eliminating the need for manual inspection of the entire video content. This automated approach has the potential to enhance the efficiency and effectiveness of video analysis in various applications, including security, law enforcement, and forensic investigations.

1.4 Introduction of ResNet

ResNet (Residual Neural Network) is a powerful deep learning architecture that has primarily been used for tasks like image classification and object detection. However, it can also be leveraged for video summarization, which aims to condense long videos into shorter, more concise summaries while preserving important information. Here's a high-level overview of how ResNet can be applied to video summarization:

1. Video Representation: Each frame of the input video is fed into a pre-trained ResNet model, which extracts high-level features from the frames. ResNet architectures, such as ResNet-50 or ResNet-101, are commonly used due to their depth and strong representation learning capabilities.

2. Feature Extraction: The ResNet model processes each frame individually, generating a feature vector for each frame. These feature vectors capture rich visual representations of the frames, highlighting salient information.

3. Temporal Modeling: To capture temporal dependencies and understand the video as a sequence of frames, recurrent neural networks (RNNs) or 3D convolutional neural networks (CNNs) can be employed. The feature vectors from ResNet are fed into the temporal modeling network, which models the temporal relationships between frames.

4. Importance Scoring: The temporal modeling network outputs importance scores for each frame based on its relevance to the overall video content. These scores indicate the significance of each frame in summarizing the video. Frames with higher scores are considered more important and likely to be included in the final summary.

5. Summary Generation: Frames with high importance scores are selected to create the video summary. These selected frames are either concatenated together to form a shorter video summary or presented as keyframes representing essential moments in the original video.

6. Evaluation and Refinement: The generated video summary is evaluated based on different metrics, such as content coverage, diversity, and user preferences. The

system can be refined by incorporating user feedback or employing reinforcement learning techniques to optimize the summarization process.

ResNet provides strong feature extraction capabilities, allowing the system to capture and represent important visual information from the input video frames. By combining ResNet with temporal modeling techniques, the system can effectively summarize videos by selecting frames that best represent the overall content and story. This approach leverages the strengths of deep learning and ResNet's ability to learn complex visual representations, enabling the creation of informative and concise video summaries.

1.5 Introduction of VGG Face

VGG Face is a popular deep learning model that was specifically designed for face recognition tasks. While it excels in facial analysis, it is not directly applicable to video summarization, which aims to condense videos into shorter summaries while preserving important content. However, we can discuss an alternative approach to video summarization using VGG Face features. Here's an outline of how VGG Face features can be utilized for video summarization:

1. Video Preprocessing: The input video is preprocessed by extracting frames at a certain frame rate. These frames serve as the basis for subsequent analysis.

2. Face Detection and Tracking: Using a face detection algorithm, faces are identified and localized within each frame. Tracking algorithms can be employed to associate faces across frames, enabling the system to recognize and track individuals throughout the video.

3. Face Representation: For each detected and tracked face, VGG Face is utilized to extract deep facial features. VGG Face is a convolutional neural network that has been trained on a large-scale dataset of facial images, enabling it to capture high-level facial representations.

4. Temporal Modeling: To capture the temporal dynamics of the video, temporal modeling techniques can be employed. Recurrent neural networks (RNNs) or 3D convolutional neural networks (CNNs) can be utilized to model the sequential information from the VGG Face features.

5. Importance Scoring: The temporal modeling network assigns importance scores to the face representations based on their significance in summarizing the video. These scores reflect the relevance of each face in capturing the key moments or individuals in the video.

6. Summary Generation: Faces with higher importance scores are selected to create the video summary. These selected faces can be presented as keyframes or concatenated together to form a shorter video summary that highlights the significant individuals and moments in the original video.

7. Evaluation and Refinement: The generated video summary is evaluated using various metrics such as content coverage, diversity, and user preferences. User feedback and reinforcement learning techniques can be incorporated to refine the summarization process and improve the quality of the generated summaries.

While VGG Face was not specifically designed for video summarization, by utilizing its facial recognition capabilities and combining them with temporal modeling techniques, we can leverage the power of facial features to generate video summaries that focus on the key individuals and moments within the video.

1.6 Objective of Research

In this research we are trying to define our research objectives of this project and how we are going to do our research according our specified objectives are mention below :-

- To create summary of video without missing any important information. There are many methods and many research articles available. Single view & multi-view cameras available.
- There are many surveillance cameras are everywhere which are generating more and more data daily.
- It required lots of space, money and maintenance.
- That's why we need to summarize the data.
- And create a single video from large videos.

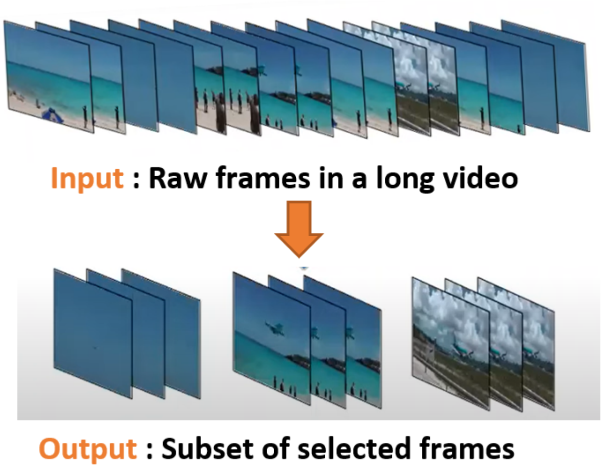

Input : Raw frames in a long video
Output : Subset of selected frames

Figure 1.2: Frame selection

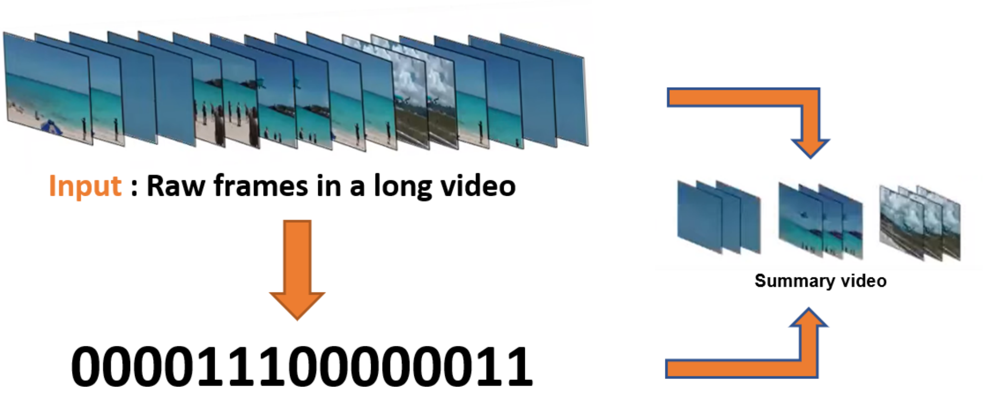


Figure 1.3: Binary Frame selection

1.7 Applications

Below are some area of application were we can use our implementation.

1. Video Surveillance: The system can be employed in surveillance applications to automatically detect and track individuals, bolstering security measures and alerting authorities to potential threats.

2. Crowd Monitoring: In crowded settings like public events or transportation hubs, the system can be utilized to monitor and analyze people's movements, identifying specific individuals or tracking crowd flow.

3. Content Filtering: By automatically detecting and recognizing individuals, the sys-

tem can assist in content filtering tasks, helping to identify and flag content that violates guidelines or policies.

4. Access Control: The system can be integrated into access control systems, using biometric-based recognition to replace traditional identification methods and ensure secure authentication.

5. Marketing and Audience Analysis: Utilized in marketing research and audience analysis, the system can gather data on demographics and behaviors, enabling targeted advertising, product placement, and customer preference analysis.

6. Law Enforcement: The system can aid law enforcement agencies in identifying suspects or wanted individuals from video footage, expediting investigations and enhancing identification accuracy.

7. Social Media Analysis: By automatically identifying individuals in shared videos on social media platforms, the system can improve content indexing, recommendation systems, and social media analytics.

8. Human-Computer Interaction: Integrated into interactive systems, such as smart homes or virtual reality applications, the system can enable personalized experiences by recognizing individuals and adapting system behavior accordingly.

9. Retail Analytics: The system can be utilized in retail environments for customer analysis and behavior understanding. By recognizing individuals, it can track customer movements, analyze shopping patterns, and provide valuable insights for improving store layout, product placement, and personalized marketing strategies.

10. Retail Analytics: The system can be utilized in retail environments for customer analysis and behavior understanding. By recognizing individuals, it can track customer movements, analyze shopping patterns, and provide valuable insights for improving store layout, product placement, and personalized marketing strategies.

11. VIP Recognition: The system can be employed to identify and provide personalized services to VIPs or high-profile individuals in various settings such as hotels, airports, or exclusive events. It can enhance customer satisfaction and streamline VIP management processes.

12. Attendance Tracking: The system can be used in educational institutions or workplaces to automate attendance tracking. By recognizing individuals, it eliminates the need for manual attendance marking, reducing administrative tasks and improving efficiency.

13. Healthcare Applications: In healthcare facilities, the system can aid in patient identification and monitoring. It can help healthcare professionals access patient records, track patient movements within the facility, and enhance patient safety and care.

14. Visitor Management: The system can be integrated into visitor management systems in corporate offices, government buildings, or event venues. It can accurately identify and register visitors, enhancing security protocols and streamlining check-in processes.

15. Personalized Entertainment: In entertainment venues such as amusement parks or concert halls, the system can offer personalized experiences. By recognizing individuals, it can tailor entertainment options, provide personalized recommendations, and create interactive and immersive experiences.

16. Customer Service Enhancement: The system can be utilized in customer service applications to provide personalized assistance. By recognizing individuals, customer service representatives can access customer profiles, preferences, and history, delivering a more tailored and satisfactory customer experience.

17. Smart Transportation: In transportation systems, such as airports or train stations, the system can assist in passenger flow management, crowd control, and personalized services. By recognizing individuals, it can optimize passenger movement and provide real-time information and assistance.

18. Sports Analysis: The system can be used in sports analysis to track player movements and provide valuable insights into player performance, team strategies, and game analysis. It can contribute to data-driven decision making in sports coaching and scouting.

19. Personalized Advertising: By recognizing individuals, the system can enable personalized advertising in digital signage or targeted marketing campaigns. It can

deliver tailored content based on individual preferences, demographics, or past interactions.

Chapter 2

Literature Survey

2.1 Video summerization:-

Table 2.1: Literature Review on Video Summarization Algorithms

No.	Title	Year	Algorithm
1	A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system [3]	2022	Deep CNN
2	An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning [4]	2022	OOI (Object of interest) selection done based on YoloV3
3	Deep Reinforcement Learning for Video Summarization with Semantic Reward [5]	2022	SV-DSN
4	Context-Adaptive Online Reinforcement Learning for Multi-view Video Summarization on Mobile Devices [6]	2023	COORS (context-adaptive online reinforcement learning multi-view video summarization framework)
5	Video Summarization by Removing Duplicate Frames from Surveillance Video Using Keyframe [7]	2017	SVD
6	Intelligent video surveillance: a review through deep learning techniques for crowd analysis [8]	2019	YOLO & VGG-16 Net

Table 2.2: Literature Review on Video Summarization Algorithms

No.	Title	Year	Algorithm
7	An Efficient Method for Underwater Video Summarization and Object Detection Using YoLoV3 [1]	2022	YOLOv3
8	Face Recognition Model Based On MTCNN And Facenet [9]	2022	MTCNN and Facenet
9	Face Recognition System using Facenet Algorithm for Employee Presence [10]	2020	FaceNet
10	Deep Learning Based Automated Sports Video Summarization Using YOLO [11]	2021	YOLOv3
11	A New Action Recognition Framework for Video Highlights Summarization in Sporting Events [12]	2021	YOLOv3
12	Patrolling Automated Guided Vehicle Enhanced with Object and Face Recognition Functions [13]	2021	YOLO
13	FaceNet with RetinaFace to Identify Masked Face [14]	2021	FaceNet
14	Student Abnormal Behavior Detection Using Dlib Combined With YOLO Models [15]	2022	YOLO

- In paper[3], I have analyse that The practical use of the suggested framework in the field of video summarization, particularly on campuses, is highlighted in the paper's conclusion. The study's findings show how well the technique produces concise summaries of academic activity recorded by the campus surveillance system. These summaries provide a succinct overview of significant events that happened recently, allowing administrators and stakeholders to make choices fast and find requested material more rapidly.

  The report notes that the unwillingness of organisations to share video data owing to privacy concerns is one of the major obstacles in this research area. As a result, there isn't much research being done in this complex application domain. However, the study described in this paper acts as a first step in resolving this problem and promoting excellent research in the area. The authors stress the need of creating generalised trained models and identifying difficulties for related issues for video analytic researchers.

The study highlights future approaches by seeking to create frameworks for producing customised video summaries utilising Long Short Term Memory (LSTM) and 3D-CNN architecture in addition to the framework that is currently being provided. These developments would help to improve video summarising methods and broaden their scope of use.

- In paper[4] paper's conclusion focuses on the Object of Interest (OoI) and the usefulness and efficiency of the suggested video summarization (VS) framework. The framework performs better than other cutting-edge techniques in terms of both speed and accuracy. The suggested system uses YOLOv3 for object detection, demonstrating accurate and effective detection of a variety of items, making it dependable and adaptable for producing pertinent video summaries.

  Extensive tests are run on three datasets—VSUMM, TVSum, and an internal dataset to verify the framework's efficacy. On the VSUMM dataset, the findings demonstrate outstanding performance with a 99.6% accuracy rate and a total time savings of 82.8%. On the TVSum dataset, a summarization rate of 78.8% and an accuracy of 99.9% are attained. The findings from the own dataset are likewise encouraging, with a 99.3% accuracy rate and a total time savings of 87.86%.

  The creation of a desktop programme with user-friendly features and personalised item selection is also mentioned in the report. The authors propose improving the model's applicability in further work by enlarging the vocabulary and training it for new Objects of Interest. The platform also has the potential to be used in real-time to capture condensed footage for different crime scenes, expanding its applicability in real-world situations.

- In paper[5], The conclusion of the paper addresses the limitations of existing video summarization models in semantically rich scenarios and presents a novel approach called SV-DSN (Semantic Video Summarization with Deep Reinforcement Learning). The proposed model introduces a semantic reward mechanism to enhance the quality of generated summaries. It leverages subtitle information, which is often underutilized, to improve the summarization process.

  The SV-DSN model consists of a strong representational summarization network that serves as an auxiliary model. This auxiliary model combines image summarization and caption summarization, effectively fusing visual and textual information to generate comprehensive and semantically meaningful summaries.

By incorporating deep reinforcement learning techniques and the semantic reward mechanism, the proposed model aims to address the deficiencies of existing video summarization models in semantically rich scenarios. The utilization of subtitle information and the introduction of semantic rewards contribute to the production of high-quality summaries that capture important content.

Overall, the SV-DSN model presents a promising approach for video summarization, emphasizing the significance of considering semantic aspects in the summarization process. The proposed model's effectiveness and potential for improving summarization quality make it a valuable contribution to the field of video summarization in semantically rich scenarios.

- In paper[6], The conclusion of the paper introduces a scene adaptive online multi-view video summarization system called COORS, designed specifically for resource-constrained IoT surveillance networks. The proposed system addresses the challenges of limited resources and real-time processing in these networks.

  COORS operates by first segmenting multi-view videos into shots using a lightweight multi-intelligence collaborative target detection method. This step helps identify relevant scenes within the videos. Next, domain-independent features are extracted from the shots to generate summaries. Notably, the summary model is adaptively fine-tuned in response to scene changes, ensuring that the summarization remains effective and up-to-date.

  To achieve real-time performance, the paper employs representation learning and migration learning to accelerate the model retraining speed. These techniques optimize the system's efficiency without compromising the quality of the summaries.

  The effectiveness of COORS is validated through extensive experiments, which demonstrate the advantages of the proposed system. By addressing the challenges of resource constraints and real-time processing in IoT surveillance networks, COORS provides a practical and efficient solution for multiview video summarization in such environments.

- In paper[1], In order to meet the objectives of marine researchers, this work proposes an automated deep learning-based system for underwater video processing. The suggested method involves extracting keyframes from underwater films using the Perceived motion energy (PME) method, then improving these frames to remove blurriness. The next step is to analyse ecosystems and identify diverse underwater animals, such as crabs, jellyfish, and fish of varying sizes, using the YoLoV3 object detection algorithm. Pre-trained DarkNet53 weights are used to fine-tune the YoLoV3 model for precise object recognition. The framework has shown encouraging findings and has the potential to help marine scientists with their study. In order to further improve performance, future work will entail extending the framework to include more object identification techniques, various backbone networks, and attention algorithms.

  Proposed methodology and used model:

  1. Keyframe Extraction Using PME

  2. YOLOv3

- In paper[9], The facial recognition model presented in this research makes use of Facenet and MTCNN to overcome the drawbacks of conventional systems. Facenet collects face feature vectors from photos, whereas the MTCNN model, which consists of three convolutional neural networks, extracts faces from images. In order to accurately identify features based on Euclidean spatial distances, the model uses Triplet Loss as the loss function. The accuracy obtained in experiments using the LFW dataset was 86.0%. Overall, the suggested model outperforms conventional facial recognition algorithms in terms of accuracy and speed, making it a useful tool in a variety of industries.

- In paper[10], two deep learning models for a detection system that uses facial recognition—FaceNet and Openface—are tested. The most accurate facial recognition system is FaceNet, created by Google researchers. Openface is a variant of FaceNet, trained on smaller datasets but with comparable performance. Preprocessing the employee's face photographs for the research entails their detection, cropping, and resizing. Utilising FaceNet and Openface, facial characteristics are then retrieved into 128 dimensions. For classification, Support Vector Machine (SVM) is utilised, and 5-fold cross-validation is used for model validation. While Openface obtains an accuracy of 93.33%, the FaceNet model reaches a flawless accuracy of 100%. With a threshold probability of 0.25, FaceNet is also implemented with a 100% accuracy. In conclusion, the FaceNet model with SVM confirms the baseline correctness of each model and outperforms Openface in terms of accuracy. Further ensuring flawless accuracy is the deployment of FaceNet with a threshold probability of 0.25.

- In paper[11], The automated key event extraction and video summarization of sporting events utilising scoreboard recognition are presented in this study as a low-cost approach. The method locates and crops scoreboards, reduces noise, and extracts scores by combining YOLO object identification, image processing methods, and OCR. Timestamps for significant occurrences are generated by a rule-based algorithm based on the game. The average F1 Score for the suggested strategy across various sports leagues is 0.979. Future work will focus on expanding video format compatibility, generalising the scoreboard detection methodology, and enhancing OCR. The model is appropriate for sports analysis and provides exact event timestamps.

- In paper[12], The study on separately cutting sports video highlights using machine learning techniques' significant findings and achievements are highlighted in paper.

  The article proposes a high-accuracy framework based on two popular open-source structures, YOLO-v3 and OpenPose, and a three-level prediction method. With only a little quantity of training data, the approach shows precise sports activity highlights clipping, exceeding earlier algorithms.The creation of two systems for automating sports video editing and highlights production is highlighted in the conclusion.

  It emphasises how the YOLO-based technique under the suggested framework outperforms the OpenPose-based approach. Comparing the three-level prediction system to conventional frame-by-frame recognition and removal techniques, the summarization accuracy is improved dramatically.

  Furthermore, the auto-clipping systems created in this study work satisfactorily despite employing a minimal quantity of training data.

  Overall, the study provides possible applications in video summarising and match analysis systems within the sports industry as well as a viable foundation for automated sports video summary.

Chapter 3

Implementation

3.1 Introduction of Implementation

In this implementation, we utilize two powerful deep learning models, YOLOv3 and FaceNet, to achieve an efficient video summarization process. Firstly, YOLOv3 is employed to detect individuals within the video, enabling the extraction of frames featuring a specific person of interest. Subsequently, utilizing FaceNet, we perform face recognition to ensure the accurate identification of the desired individual. By generating a summary video composed of these extracted frames, we effectively condense the original video while focusing solely on the specific person, facilitating easy retrieval of relevant content.

3.2 Methodology

Below listed used methodology in implementation.

1. YOLOv3
2. FaceNet by Google

3.2.1 YOLOv3

In the project we tried YOLOv3 model for for face recognition, The architecture of YOLOv3 used in the above code consists of a deep convolutional neural network (CNN) with several layers. YOLOv3 is a popular object detection algorithm known for its real-time performance and accuracy. Here is a detailed overview of the YOLOv3 architecture:

- Input Layer:

  The input layer takes an image as input, typically with a size of 416x416 pixels.

- Backbone Network:

  YOLOv3 uses a DarkNet-53 architecture as its backbone network. It consists of 53 convolutional layers, including shortcut connections, which enable the network to capture and represent complex features in the input image.

- Feature Extraction:

  The backbone network processes the input image and extracts high-level features through a series of convolutional and pooling layers. This helps in learning important spatial and semantic representations.

- Detection Layers:

  YOLOv3 has multiple detection layers that are responsible for predicting bounding boxes and class probabilities at different scales. Each detection layer is connected to the last layer of the backbone network and performs object detection at a specific scale. The detection layers use 1x1 convolutional layers to reduce the number of channels and extract features at different resolutions. Each detection layer predicts bounding box coordinates, objectness scores (confidence), and class probabilities using a set of convolutional layers and activation functions.

- Anchor Boxes:

  YOLOv3 utilizes anchor boxes to handle objects of different sizes and aspect ratios. Anchor boxes are pre-defined bounding boxes of various dimensions that act as priors for object detection. The detection layers predict offsets for anchor boxes to accurately localize objects in the image.

- Output:

  The final output of the YOLOv3 architecture is a set of bounding boxes, confidence scores, and class probabilities for the detected objects. These predictions are generated at different scales and are then processed using non-maximum suppression (NMS) to eliminate overlapping and redundant detection.

- Post-processing:

  After applying NMS, the remaining bounding boxes with high confidence scores are selected as the final detection. These detection can be further filtered based on a confidence threshold to control the precision of the object detection.

The YOLOv3 architecture, with its multi-scale detection and anchor box mechanism, allows for the detection of objects at different sizes and aspect ratios in real-time. It balances speed and accuracy, making it well-suited for applications such as object detection in surveillance videos, including person detection in the context of the above project.

3.2.2 FaceNet for face recognition

Face recognition is critical in video summarising because it allows for the identification and tracking of persons throughout a film. FaceNet has emerged as a very important deep learning-based approach for face recognition in recent years, revolutionizing the field and enabling substantial advances in video summarising techniques.

FaceNet, created by Google researchers, provided a game-changing deep neural network architecture capable of learning highly discriminative face embeddings. It employs a triplet loss function that learns to map face pictures into a high-dimensional feature space, where distances between embeddings of the same person's face are minimized and distances between embeddings of different people's faces are maximized. FaceNet can build compact yet semantically relevant representations of faces thanks to this learning goal.

FaceNet's facial recognition skills are very useful in the context of video summarization. Condensing lengthy films into shorter summaries that contain the main material is what video summarising is all about. FaceNet allows video summarization algorithms to reliably detect and recognize faces, allowing for the selection and inclusion of critical events involving specific persons.

FaceNet incorporation into video summarising pipelines allows for the detection of crucial events or interactions involving specific persons, resulting in more informative and personalized video summaries. FaceNet, for example, can assist summarise footage in surveillance applications by emphasizing crucial moments when certain persons appear, such as possible threats or suspicious actions. FaceNet can help in summarising films in social media or vlogging situations by concentrating on significant persons or social interactions, improving overall knowledge and engagement with the summary.

The advantages of utilizing FaceNet for facial recognition in video summary go beyond just accurate identification. FaceNet's compact face embeddings provide efficient storage, retrieval, and comparison of face representations, allowing for quicker processing and increased scalability in video summarization systems.

Furthermore, FaceNet's resistance to changes in position, lighting, and facial emotions leads to the effectiveness of face identification in video summarization, even in difficult real-world circumstances. Its ability to handle a wide range of faces while maintaining high accuracy makes it ideal for summarising films with variable material and persons.

Finally, FaceNet has had a big impact on the field of face identification in video summarization. Its deep learning-based architecture and face embedding capabilities allow it to accurately and efficiently identify persons throughout a video, providing for more informative and personalized video summaries. FaceNet incorporation into video summarising pipelines improves the identification of crucial events involving specific persons, hence boosting the overall quality and relevance of the output summaries. FaceNet is a valuable technique for providing robust and successful face identification in the context of video summarising as face recognition and video summarization continue to advance.

3.2.3 Flow Diagram of Implementation

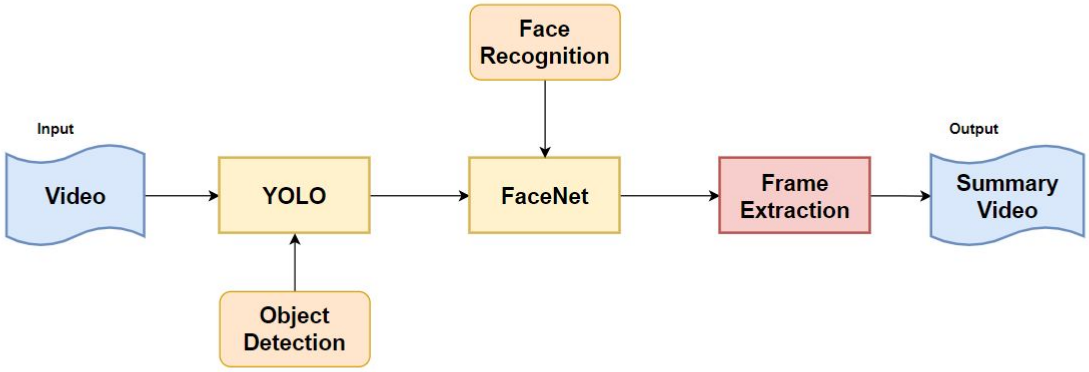


Figure 3.1: Flow Diagram

Above fig. 3.1 is flow diagram of our implementation.

3.3 Result and Analysis

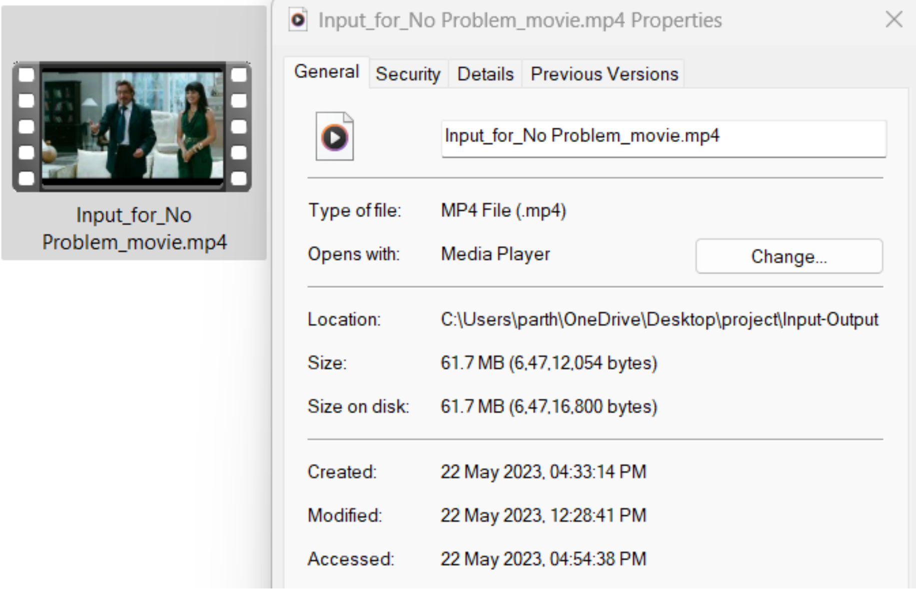
In this scenario, we have chosen a video from YouTube. To accomplish our goal, we will utilize YOLO (You Only Look Once), a popular object detection algorithm. The first step is to detect individuals within the video using YOLO's person detection capabilities.

Once we have successfully detected the individuals, we will employ FaceNet, a state-of-the-art face recognition model developed by Google. FaceNet is renowned for its accuracy in identifying and distinguishing between different faces. We will leverage this model to specifically identify a particular person of interest within the video.

Finally, utilizing the detections made by FaceNet, we will generate an output video that comprises frames featuring the specific person we are interested in. This output video will be tailored to showcase only the frames containing the individual we seek, thus effectively extracting and presenting the desired person's appearances throughout the original video.

3.3.1 Input Video

In Fig. 3.2 shown the size of input video named as Input_for_No Problem_movie.mp4.



Input_for_No Problem_movie.mp4 Properties

Input_for_No Problem_movie.mp4

Type of file: MP4 File (.mp4)

Opens with: Media Player

Location: C:\Users\parth\OneDrive\Desktop\project\Input-Output

Size: 61.7 MB (6,47,12,054 bytes)

Size on disk: 61.7 MB (6,47,16,800 bytes)

Created: 22 May 2023, 04:33:14 PM

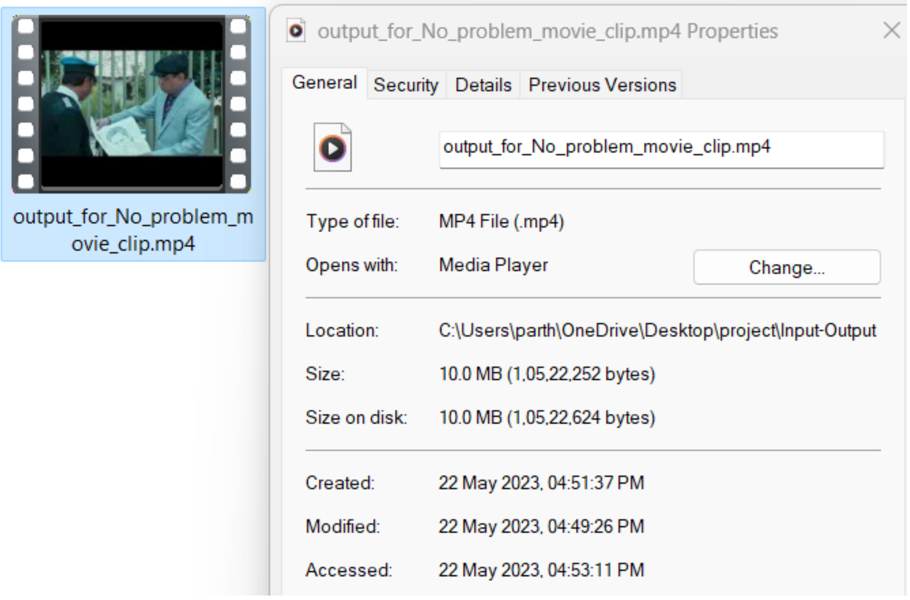
Modified: 22 May 2023, 12:28:41 PM

Accessed: 22 May 2023, 04:54:38 PM

Figure 3.2: Input Video

3.3.2 Output Video

In Fig. 3.3 shown the size of output video named as output_for_No Problem_movie_clip.mp4. This video contains frames of having paresh raval as our input recognized image.


output_for_No_problem_movie_clip.mp4
output_for_No_problem_movie_clip.mp4 Properties
General
Security
Details
Previous Versions
Change...

output_for_No_problem_movie_clip.mp4

Type of file: MP4 File (.mp4)

Opens with: Media Player

Location: C:\Users\parth\OneDrive\Desktop\project\Input-Output

Size: 10.0 MB (1,05,22,252 bytes)

Size on disk: 10.0 MB (1,05,22,624 bytes)

Created: 22 May 2023, 04:51:37 PM

Modified: 22 May 2023, 04:49:26 PM

Accessed: 22 May 2023, 04:53:11 PM

Figure 3.3: Output Video

Chapter 4

Conclusion and Future work

4.1 Conclusion

In this research project mentioned above uses the Google FaceNet facial recognition model and the YOLOv3 object detection model to identify specific person in videos. The project takes still images from a movie, uses YOLOv3 to locate person, and then uses FaceNet to identify the person it has found. Target individuals are identified by the project by comparing their face embeddings to those of known individuals. The appropriate frames are kept to make a quick movie if a target individual is found. The project offers a helpful foundation for automatic person detection and recognition in videos, which may be used for a number of purposes including content filtering, security, and surveillance. So, basically we have selected a video clip of a movie and find frames which contain actor paresh raval and make summery of that video.

4.2 Future work

The implementation described above lays the foundation for future work and potential enhancements. Here are some directions for future improvements and extensions:

- Real-time Processing: The current implementation processes videos frame by frame, but further optimization can be done to achieve real-time performance. This could involve exploring hardware acceleration techniques, parallel processing, or model compression to improve speed without sacrificing accuracy.

- Multi-person Tracking: Enhancing the system to perform multi-person tracking would enable the tracking of individuals across consecutive frames, providing richer information about their movements and interactions. This could involve integrating tracking algorithms such as Kalman filters or deep learning-based trackers.

- Occlusion Handling: Dealing with occlusions where individuals are partially or fully obstructed by objects or other people remains a challenge. Future work could explore techniques to handle occlusions, such as leveraging contextual information or utilizing depth sensors for better depth-based segmentation.
- Continuous Learning: Implementing a continuous learning mechanism would allow the system to adapt and improve over time. This could involve periodically retraining the face recognition model with new face samples to enhance recognition accuracy and adapt to changes in appearance due to aging, hairstyle variations, or accessories.
- Robustness to Environmental Factors: Enhancing the system's robustness to challenging environmental factors, such as variations in lighting conditions, pose changes, or low-resolution videos, would improve its performance in real-world scenarios. This could involve data augmentation techniques, model regularization, or domain adaptation methods.
- Large-scale Deployment: The system could be scaled up for deployment in larger and more complex environments. This would involve efficient distribution of processing tasks across multiple computing units or leveraging cloud-based infrastructure to handle increased computational requirements.
- User Interface and Integration: Developing a user-friendly interface and integrating the system into existing platforms or applications would make it more accessible and practical for end-users. This could involve building a web-based dashboard, APIs for integration, or developing mobile applications for on-the-go usage.
- These future directions for the project can enhance the system's performance, scalability, privacy, and usability, making it more effective and applicable in real-world scenarios. By addressing these areas, the system can be further advanced to meet evolving needs and challenges in person detection and recognition.

Bibliography

[1] F. A. J. S. Y. K. Mubashir Javaid, Muazzam Maqsood, "An efficient method for underwater video summarization and object detection using yolov3," *Intelligent Automation & Soft Computing*, vol. 35, no. 2, pp. 1295–1310, 2023.

[2] M. Kini M. and K. Pai, "A survey on video summarization techniques," vol. 1, pp. 1–5, 2019.

[3] "Muhammad w, ahmed i, ahmad j, nawaz m, alabdulkreem e, ghadi y. 2022. a video summarization framework based on activity attention modeling using deep features for smart campus surveillance system. peerj computer science 8:e911 https://doi.org/10.7717/peerj-cs.911,"

[4] "Jour, qazi, nadeem,asif, muhammad,ashraf, rehan, mahmood, toqeer,an effective video summarization framework based on the object of interest using deep learning, 2022/05/12,,"

[5] H. Sun, X. Zhu, and C. Zhou, "Deep reinforcement learning for video summarization with semantic reward," pp. 754–755, 2022.

[6] J. Hao, S. Liu, B. Guo, Y. Ding, and Z. Yu, "Context-adaptive online reinforcement learning for multi-view video summarization on mobile devices," in *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 411–418, 2023.

[7] J. Zhou and L. Lu, "Wide and deep learning for video summarization via attention mechanism and independently recurrent neural network," pp. 407–407, 2020.

[8] "Sreenu, g.,saleem durai, m. aintelligent video surveillance: a review through deep learning techniques for crowd analysis,2019, journal of big data.,"

[9] “S. qi, x. zuo, w. feng and i. g. naveen, ”face recognition model based on mtcnn and facenet,” 2022 ieee 2nd international conference on mobile networks and wireless communications (icmnwc), tumkur, karnataka, india, 2022, pp. 1-5, doi: 10.1109/icmnwc56175.2022.10031806.,”

[10] “F. cahyono, w. wirawan and r. fuad rachmadi, ”face recognition system using facenet algorithm for employee presence,” 2020 4th international conference on vocational education and training (icovet), malang, indonesia, 2020, pp. 57-62, doi: 10.1109/icovet50258.2020.9229888.,”

[11] “Chakradhar guntuboina , aditya porwal , preet jain, hansa shingrakhia , deep learning based automated sports video summarization using yolo , 2021,”

[12] “C. yan, x. li and g. li, ”a new action recognition framework for video highlights summarization in sporting events,” 2021 16th international conference on computer science education (iccse), lancaster, united kingdom, 2021, pp. 653-666, doi: 10.1109/iccse51940.2021.9569708.,”

[13] C.-C. Chang and H.-R. Wang, “Patrolling automated guided vehicle enhanced with object and face recognition functions,” pp. 199–203, 2021.

[14] S. E. Sitepu, G. Jati, M. R. Alhamidi, W. Caesarendra, and W. Jatmiko, “Facenet with retinaface to identify masked face,” pp. 81–86, 2021.

[15] M. A. E. Alkhalisy and S. H. Abd, “Student abnormal behavior detection using dlib combined with yolo models,” pp. 1–7, 2022.

Parth

ORIGINALITY REPORT

6%	4%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, Ioannis Patras. "Video Summarization Using Deep Neural Networks: A Survey", Proceedings of the IEEE, 2021 Publication	1%
2	Jingyi Hao, Sicong Liu, Bin Guo, Yasan Ding, Zhiwen Yu. "Context-Adaptive Online Reinforcement Learning for Multi-view Video Summarization on Mobile Devices", 2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS), 2023 Publication	1%
3	qmro.qmul.ac.uk Internet Source	1%
4	arxiv.org Internet Source	<1%
5	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018 Publication	<1%
	deepai.org	

6	Internet Source	<1 %
7	Weiqin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar. "Machine Vision-Enabled Traffic Controller for Safer and Smoother Traffic Flow Around Construction Sites", 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019 Publication	<1 %
8	repository.tudelft.nl Internet Source	<1 %
9	Siyao Qi, Xinyu Zuo, Weijia Feng, I G Naveen. "Face Recognition Model Based On MTCNN And Facenet", 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), 2022 Publication	<1 %
10	Submitted to Liverpool Hope Student Paper	<1 %
11	www.researchgate.net Internet Source	<1 %
12	Submitted to Aston University Student Paper	<1 %
13	Submitted to Segi University College Student Paper	<1 %

14	Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, Jon Froehlich. "Tohme", Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14, 2014 Publication	<1 %
15	ieee-icpads.net Internet Source	<1 %
16	link.springer.com Internet Source	<1 %
17	thescholarship.ecu.edu Internet Source	<1 %
18	Wasim Muhammad, Imran Ahmed, Jamil Ahmad, Muhammad Nawaz, Eatedal Alabdulkreem, Yazeed Ghadi. "A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system", PeerJ Computer Science, 2022 Publication	<1 %
19	"Program and Abstracts Book", 2022 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE), 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		