VIDEO ANALYSIS THROUGH PROMPTING

Hemanth A.K IDK20CS033
Safeedha M.K IDK20CS049
Vaishnavi D.A IDK20CS060
Vaishnav Vijayan IDK20CS061

Dept.of Computer Science and Engineering
Govt.Engineering College Idukki

December 5, 2023

Guide : Prof.Anish Abraham

Introduction

- Video analysis is a difficult undertaking that frequently calls extreme manual work to extract particular objects or events from long videos.
- Our project is to automate and simplify video content analysis, reducing the need for human intervention.
- Users can effortlessly specify their desired events for extraction within our web application, streamlining the process of gathering valuable insights from videos.

Design Issues

- **Laborious Video Annotation:** Manually annotating specific events or objects in lengthy videos is a time-intensive process, hindering efficient video content analysis.
- **Limited AI Accessibility:** The lack of accessible AI tools and models restricts the widespread use of advanced video analysis techniques.
- **Inefficient Information Retrieval:** Current methods for extracting relevant content from videos are often inefficient, impeding effective knowledge extraction.
- **Slow Model Training:**Training deep learning models for video analysis is computationally demanding and time-consuming, delaying the development of efficient solutions.

Problem Statement

To solve the time consuming problem of identifying specific events from a lengthy video using deep learning techniques

Objective

- Automate video analysis to extract specific events or objects efficiently.
- Utilize advanced deep learning models, including Large Language Models (LLM), for accurate event recognition.
- Create an intuitive user interface for event specification.
- Improve overall efficiency in video content analysis, reducing manual effort.

Literature Survey

Sl No.	Paper	Author	Content
[1]	A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition(Published at Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 8323962).	Viet-Tuan Le , Kiet Tran-Trung , and Vinh Truong Hoang .	This article provides a comprehensive overview of recent deep learning techniques for human action recognition using RGB video data. The article divides the methods into five categories: 2D CNNs, RNNs, 3D CNNs, multistream approaches, and convolution-free architectures.
[2]	MiniGPT-4:Enhancing Vision-Language Understanding with Advanced Large Language Models(Published at arXiv:2304.10592v2 [cs.CV] 2 Oct 2023).	Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny.	This article introducing MiniGPT-4, a novel vision-language model. MiniGPT-4 combines a static visual encoder with an advanced, large language model called Vicuna. The article demonstrates that MiniGPT-4 achieves high-level vision-language capabilities comparable to GPT-4 and explaining unusual visual phenomena .

Literature Survey Cont.

Sl No.	Paper	Author	Content
[3]	Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering(Published at 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)).	Zhenwei Shao,Zhou Yu,Meng Wang,Jun Yu.	This paper focuses on Knowledge-based Visual Question Answering (VQA), where questions about images are answered using external knowledge. The paper introduces "Prophet," a framework that guides GPT-3 with answer heuristics. These heuristics include answer candidates and answer-aware examples. Prophet outperforms existing methods on challenging VQA datasets like OK-VQA and A-OKVQA using minimal computational resources.
[4]	Improved Residual Networks for Image and Video Recognition(Published at 2020 25th International Conference on Pattern Recognition (ICPR) Milan, Italy, Jan 10-15, 2021).	Ionut Cosmin Duta, Li Liu, Fan Zhu, Ling Shao.	This paper presents enhanced Residual Networks (ResNets) for image and video recognition. It focuses on improving information flow, residual blocks, and projection shortcuts. The authors introduce a stage-based network architecture with various residual block types and an improved projection shortcut to reduce information loss. They also introduce a new building block with increased spatial channels. Experimental results demonstrate consistent improvements across six datasets, including image classification, object detection, and video action recognition. Remarkably, the authors achieve success in training very deep networks, such as a 404-layer network on ImageNet and a 3002-layer network on CIFAR-10 and CIFAR-100.

Literature Survey Cont.

Sl No.	Paper	Author	Content
[5]	Self-Supervised Learning for Videos:A Survey(Published at 13 July 2023 Published in CSUR Volume 55, Issue 13s).	Madeline C. Schiappa,Yogesh S. Rawat,Mubarak Shah.	This Paper is a survey on self-supervised video learning. It categorizes methods into four types: pretext learning (predicting video properties), generative learning (reconstructing or generating video frames), contrastive learning (distinguishing between positive and negative video pairs), and cross-modal agreement (aligning different video modalities). These approaches aim to extract useful representations from unlabeled video data for downstream tasks.
[6]	Survey: Transformer based Video Language pre-training (Published at AI Open, Volume 3, 2022,).	Ludan Ruan, Qin Jin .	This Paper is a survey paper that explores the latest developments in transformer-based pre-training for video-language learning. It delves into the fundamental components, including the transformer architecture, the pre-training and fine-tuning process, video-language tasks, pre-training methods, and the distinction between single-stream and multi-stream structures.

Literature Survey Cont.

Sl No.	Paper	Author	Content
[7]	VQA: Visual Question Answering(Published at 2015 IEEE International Conference on Computer Vision (ICCV)).	Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol , Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh .	This Paper focuses on Visual Question Answering (VQA), a task where questions about images are answered using both visual and textual information. It highlights the VQA Dataset, an extensive collection of images, questions, and answers, emphasizing its diversity and complexity. The page also includes VQA Analysis, which explores question and answer types within the dataset, and VQA Methods, evaluating approaches using text and image features for answer generation, comparing them with human performance.
[8]	Visual Question Answering using Deep Learning: A Survey and Performance Analysis (Published at Computer Vision and Image Processing. CVIP 2020. Communications in Computer and Information Science, vol 1377. Springer, Singapore).	Srivastava Y., Murali V., Dubey S.R., Mukherjee S.	This research paper provides a comprehensive overview of Visual Question Answering (VQA), a field at the intersection of computer vision and natural language processing. It covers various VQA datasets, state-of-the-art models, and comparative results. The paper also identifies current challenges like bias and ambiguity and suggests future research directions for enhancing VQA systems.

Literature Survey Cont.

Sl No.	Paper	Author	Content
[9]	Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition(Published at 2017 IEEE International Conference on Computer Vision Workshops (ICCVW)).	Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh .	This research paper focuses on 3D Residual Networks, a type of convolutional neural network designed for action recognition in videos. It delves into the network architecture, which extends convolutional kernels and pooling dimensions based on ResNets. The paper also outlines the training procedure using the ActivityNet and Kinetics datasets. In terms of evaluation, it compares the performance of 3D ResNets to other methods like C3D and I3D, demonstrating that 3D ResNets outperform shallow networks and compete favorably with very deep networks on datasets like ActivityNet and Kinetics.
[10]	Video analytics using deep learning for crowd. analysis: a review (Published at Multimed Tools Appl 81, 27895–27922 (2022)).	Bhuiyan,Abdullah,Junaidi, Al Farid, Noramiza, Hashim, Fahmid	This review article discusses deep learning for crowd analysis, with a specific focus on the challenging context of the Hajj pilgrimage. Crowd analysis involves understanding large group behavior from video data, finding applications in public safety, event management, and urban planning. Deep learning, including FCNN, has shown effectiveness in this field. Crowd analysis techniques can be network-based or image-based, both utilizing FCNN for density or count estimation. The article also highlights the significance of crowd analysis datasets like UCF, World Expo, and Shanghai Tech for research and development.

Comparisons of techniques

Table 1: Comparison of Different Approaches and Architectures (Part 1)

Approach/Architecture	Advantages	Disadvantages
2D CNNs [1]	Efficient for image classification and feature extraction.	Limited in handling spatio-temporal information for video analysis. May not capture long-range dependencies in sequences.
RNNs [1]	Suited for sequential data with variable lengths.	Prone to vanishing gradient problems. Limited parallelism and computational efficiency.
3D CNNs [1][9]	Explicitly model spatio-temporal information for video analysis.	Computationally expensive and may require large datasets. Limited applicability to tasks primarily based on spatial features.

Comparisons of techniques Cont.

Table 2: Comparison of Different Approaches and Architectures (Part 2)

Multistream Approaches[1]	Combine multiple modalities for enhanced understanding.	Increased model complexity and resource requirements. Challenging integration of different data types.
ResNet [4]	Mitigate vanishing gradient problems, allowing for deeper networks.	Increased model complexity can lead to overfitting. May require more data and computational resources.
3D ResNets [9]	Extend ResNets to spatio-temporal data, improving performance in video analysis.	Increased computational demands compared to 2D ResNets. Sensitive to hyper-parameters and dataset size.
FCNN [10]	Well-suited for dense prediction tasks like semantic segmentation.	May not capture long-range dependencies in sequential data. Not ideal for tasks requiring hierarchical feature

Comparisons of techniques Cont.

From these different approaches and techniques we choose

- **3D ResNets :** These extend the capabilities of ResNets to spatio-temporal data, making them a strong choice for video analysis. They can mitigate vanishing gradient problems and perform well in video understanding tasks.

Applications

- Media Content Management
- Security and Surveillance
- Education Enhancement
- Market Research and Insights

Block Diagram

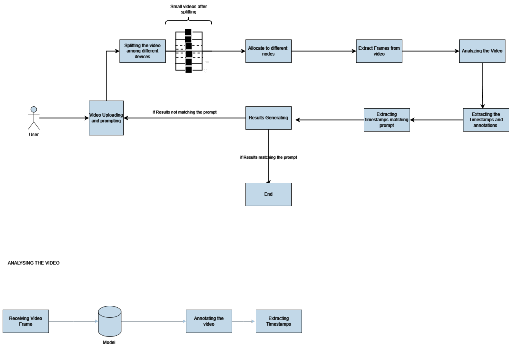


Figure 1: Block Diagram

Usecase Diagram

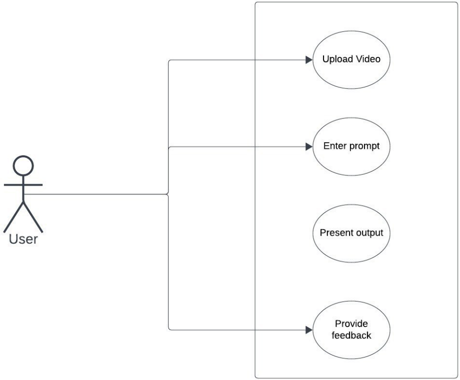


Figure 2: Usecase Diagram

Sequence Diagram

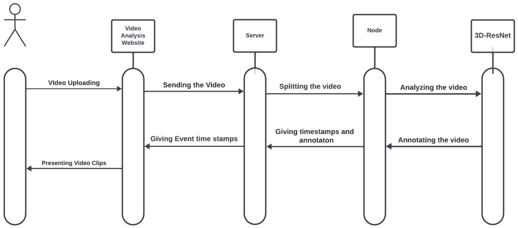


Figure 3: Sequence Diagram for video uploading

Sequence Diagram

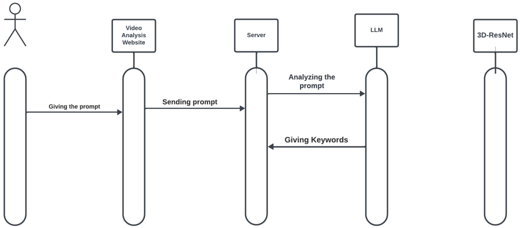


Figure 4: Sequence Diagram for prompt

Model Development

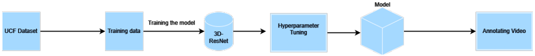


Figure 5: Model Development

Datasets

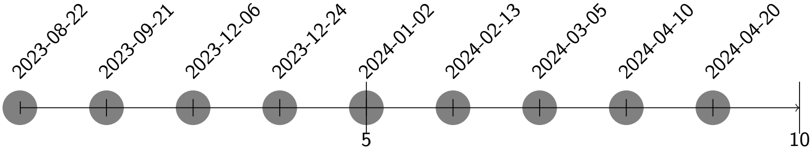
UCF Crime Dataset

- The UCF-Crime dataset is a large-scale dataset of 128 hours of videos that can be used for crime detection.
- It consists of 1,900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies.
- Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, Vandalism.

Tools Required

- Python
- FastApi
- Next.js
- vs code
- Yolo

Gantt Chart



Project Initiation	2023-08-22
Data Collection	2023-09-21
Web App Design	2023-12-06
Model Development	2023-12-24
Implementation	2024-01-02
Testing and Debugging	2024-02-13
User Interface Refinement	2024-03-05
Documentation	2024-04-10
Project Completion	2024-04-20

Conclusion

- Event extraction will be made simple by effectively automating video analysis.
- For accurate event recognition, we will use cutting-edge deep learning models like Large Language Models(LLM).
- An easy-to-use user interface for simple event specification will be developed.
- We will save a lot of time and resources when doing video analysis.

References

- [1] Aiswarya Agarwal.; Jiasen Lu.; Devi Parikh.2022.VQA:Visual Question Answering.In proceedings of the Sixteenth AAAI Conference on Artificial Intelligence.
- [2]Ludan Ruan,Qin Jin.2022 Survey:Transformer based video-language pre-traininig.In proceedings of the school of information,Renmin University of China,Beljing,China.
- [3]Deyao Zhu,Jun Chen,Xiang Li,2 0ct 2023,Enhancing Vision-Language Understanding With Advanced Large Language Models.In proceedings of the King Abdullah University and Technology.

References

- [4]Viet-Tuan Le , Kiet Tran-Trung , and Vinh Truong Hoang, A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition,Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 8323962.
- [5] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models,arXiv-preprint:2304.10592v2 [cs.CV] 2 Oct 2023.
- [6]Ionut Cosmin Duta, Li Liu, Fan Zhu, Ling Shao. *"Improved Residual Networks for Image and Video Recognition"* Published at 2020 25th International Conference on Pattern Recognition (ICPR) Milan, Italy, Jan 10-15, 2021.
- [7] Zhenwei Shao,Zhou Yu,Meng Wang,Jun Yu , Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering,2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

References

[8] Srivastava Y., Murali V., Dubey S.R., Mukherjee S,Visual Question Answering using Deep Learning: A Survey and Performance Analysis,Computer Vision and Image Processing. CVIP 2020.Communications in Computer and Information Science, vol 1377. Springer Singapore.

[9]Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh,Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition,2017 IEEE International Conference on Computer Vision Workshops (ICCVW).

[10]Bhuiyan, Md Roman,Abdullah, Junaidi,Hashim, Noramiza, Al Farid, Fahmid,Video analytics using deep learning for crowd analysis: a review,Multimed Tools Appl 81, 27895–27922 (2022).


Thank
you