TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING, PULCHOWK CAMPUS

A Project Proposal On
Human Action Recognition In Video

Submitted By:

Anil Shrestha (078BCT009)
Ashok Prasad Neupane (078BCT021)
Jeevan Neupane (078BCT097)

Submitted To:

Department of Electronics and Computer Engineering
Pulchowk Campus, Lalitpur

November 29, 2024

Acknowledgement

First and foremost, we would like to express our sincere gratitude to the Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, for providing the necessary support and resources throughout the course of this project. Special thanks to the Head of the Department, Deputy Head of the Department, and the Project Management Team, including IC Chair Asst. Prof. Santosh Giri and Asst. Prof. Bibha Sthapit, for their guidance and constant encouragement.

We would like to extend our heartfelt appreciation to the Department for providing the opportunity to undertake this collaborative project. This has greatly contributed to our practical understanding and application of knowledge, allowing us to develop a project of our own as part of the Minor Project for Third Year. The experience gained from this endeavor will certainly enhance our academic and professional skills.

We are also grateful to our respected seniors who have generously shared their knowledge, experience, and suggestions, helping us navigate challenges during the project. Our sincere thanks to all our friends who have assisted us, both directly and indirectly, throughout the course of this project.

Lastly, we would like to express our deep gratitude to our family members for their unwavering support and constant source of inspiration. Without their encouragement and belief in us, this project would not have been possible.

Any suggestions or constructive feedback will be greatly appreciated and acknowledged.

Anil Shrestha (078BCT009)
Ashok Prasad Neupane (078BCT021)
Jeevan Neupane (078BCT097)

Abstract

This project aims to develop a system for human action recognition in RGB video by leveraging pretrained video foundation models to classify actions in short clips featuring single-person scenarios. We are using the Kinetics dataset and exploring different video foundation models such as SlowFast, VJEPA, and Sapiens, we will fine-tune models and create pipeline designed for accurate and efficient action classification with minimal preprocessing. The system will serve as a foundation for extending functionality to multi-person action recognition, integrating human detection for practical applications like theft detection in supermarkets. Our approach prioritizes precision, scalability, and computational efficiency, paving the way for real-world deployment.

Keywords: Human Action Recognition (HAR), Computer Vision, Video Analysis, Deep Learning, Convolutional Neural Networks (CNN), Action Classification, Video Surveillance.

Contents

List of Figures

List of Abbreviations

HAR Human Action Recognition

CNN Convolutional Neural Network

SOTA State-of-the-Art

LSTM Long Short-Term Memory

SVM Support Vector Machine

MAE Masked Autoencoder

NAS Neural Architecture Search

SSL Self-Supervised Learning

FAIR Facebook AI Research

JEPA Joint Embedding Predictive Architecture

VJEPA Video Joint Embedding Predictive Architecture

API Application Programming Interface

1 Introduction

1.1 Background

Human Action Recognition (HAR) is a critical area of research in computer vision and artificial intelligence, focused on the automatic detection and classification of human actions from visual inputs such as video or images. The ability to recognize human actions has wide-ranging applications across various domains, including surveillance, healthcare, sports analysis, human-computer interaction, and robotics.

Historically, early attempts at HAR were based on hand-crafted features, such as motion or appearance descriptors, which required significant domain expertise and were often ineffective in handling complex, dynamic environments. These methods struggled with challenges like occlusions, variations in human posture, background noise, and changes in lighting conditions. However, with the advent of deep learning, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), significant progress has been made in addressing these challenges. CNNs are effective in capturing spatial features from images, while RNNs, especially Long Short-Term Memory (LSTM) networks, excel in modeling the temporal dependencies of human actions in videos.

In recent years, HAR has benefited from large-scale datasets, which provide diverse examples of human activities and help train robust models. Advanced techniques such as pose estimation and 3D convolutional networks have been integrated into HAR models to improve the precision and robustness of action recognition across different contexts and environments. Despite this progress, HAR remains a challenging problem due to issues like intra-class variability (e.g., different individuals performing the same action differently), inter-class similarity (e.g., similar actions that are difficult to distinguish), and real-time processing requirements.

This research aims to develop and improve upon existing HAR models by leveraging deep learning techniques and large annotated datasets to address the challenges of accuracy, real-time processing, and adaptability in dynamic environments. The focus is on creating models that can effectively recognize human actions across a wide variety of contexts, with applications spanning from security and surveillance to interactive human-robot systems.

1.2 Problem Statement

Human Action Recognition (HAR) is a challenging task in computer vision that involves detecting and classifying human actions from video data. Despite significant advancements in the field, there remain several issues that limit the effectiveness of existing models, particularly in real-world scenarios. These challenges include the complexity of accurately recognizing a wide range of human actions, especially in dynamic environments with variations in pose, scale, lighting, and occlusions. Furthermore, traditional methods often struggle to model temporal dependencies in video sequences, which are crucial for distinguishing between similar actions that differ only in the sequence or timing of

movements.

The primary objective of this project is to develop a robust and efficient system for recognizing human actions in real-time video streams. The system should be able to accurately classify actions from diverse datasets, while overcoming issues related to motion variations, background noise, and incomplete or noisy data. It also aims to explore the integration of advanced deep learning techniques, such as Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for modeling the temporal dynamics of human actions.

The goal is to contribute a solution that not only improves the accuracy and robustness of HAR systems but also enables their deployment in practical applications such as surveillance, healthcare monitoring, virtual reality, and human-robot interaction.

1.3 Objectives

The primary objectives of this minor project are:

- **To understand the fundamental principles of human action recognition and its applications:** This objective focuses on exploring the basic principles behind human action recognition, including techniques like CNNs and RNNs, as well as understanding the challenges faced in the field. It also involves investigating the various applications of HAR in domains such as healthcare, security, sports, and human-computer interaction.
- **To design and implement multimodal SOTA models in PyTorch:** This objective involves designing and implementing advanced, state-of-the-art (SOTA) multimodal models in PyTorch for human action recognition. By combining different data sources like video, audio, and sensor inputs, the goal is to improve the accuracy and robustness of the recognition system using deep learning architectures such as CNNs and RNNs.
- **To evaluate the performance of the system on different datasets and scenarios:** This objective aims to assess the performance of the designed HAR model across various datasets and real-world scenarios, analyzing factors such as accuracy, precision, and recall. The evaluation will help determine the model's generalizability and reliability under different conditions and environments.
- **To explore potential applications and improvements for future research:** This objective focuses on identifying the future directions of human action recognition by exploring potential applications in new fields and suggesting improvements. It includes understanding how emerging technologies, such as multimodal fusion or self-supervised learning, could enhance HAR systems.
- **To handle the challenges of real-time human action recognition:** This objective aims to address the challenges involved in performing human action recognition in real-time. This includes optimizing the model for low-latency predictions

and ensuring that the system can handle video streams or sensor data on the fly, which is crucial for applications like surveillance or robotics.

- **To improve model accuracy through transfer learning and fine-tuning:** This objective focuses on improving the performance of the HAR model by leveraging pre-trained models and applying transfer learning techniques. By fine-tuning models on domain-specific data, the aim is to enhance the accuracy of the model for particular action classes or environments.

1.4 Scope

The scope of this project on Human Action Recognition (HAR) is focused on the development and implementation of a video-based action recognition system using deep learning techniques. The project will begin with the collection of a diverse set of video data containing various human actions, which will serve as the primary input for the system. These videos will be preprocessed to remove noise, standardize frame rates, and ensure uniformity across the dataset. The scope also includes the extraction of meaningful features from the video data to enable accurate action recognition, utilizing techniques such as optical flow, skeleton-based methods, and CNN-based feature extraction.

Additionally, the project will explore potential applications of the HAR system in real-world scenarios, such as security surveillance, healthcare (e.g., fall detection), and sports analysis. The implementation will also investigate the challenges of real-time action recognition, focusing on improving the computational efficiency and minimizing latency to make the system feasible for practical applications. Finally, the project will outline opportunities for future improvements, including handling more complex actions, integrating multimodal data sources, and enhancing the generalization capabilities of the system.

2 Literature Review

HAR in videos has witnessed significant advancements over the past decade, driven by the evolution of the deep learning field, fueled primarily by the availability of powerful GPUs and the advent of transformer architectures [8, 11]. This study began by exploring classical methods, such as those using optical flow combined with Support Vector Machines (SVMs), and progressed through the development of more sophisticated techniques in computer vision. A notable trend in recent research involves self-supervised pretraining of foundation models on large datasets, followed by supervised fine-tuning for specific tasks [1]. Self-supervised learning (SSL) for visual representation, even without additional supervision, has demonstrated competitive performance when fine-tuned for action recognition tasks [Bardes et al.].

Recent approaches to HAR have shifted towards end-to-end models, in contrast to earlier methods that relied on intermediate tasks like pose estimation for skeleton-based modeling [12] and optical flow analysis [16]. Multimodal data integration, including depth information along with raw video, has also been explored for action recognition [16]. Moreover, many authors have utilized ResNet models pretrained on large-scale datasets like ImageNet to enhance HAR performance.

In the deep learning era, various models have been proposed to capture the spatiotemporal information inherent in videos. Both discriminative and generative models have been employed. For example, 3D Convolutional Neural Networks (3D CNNs) are effective in classifying shorter video clips but are computationally expensive for longer sequences [5, 7]. Multi-stream neural networks, such as AssembleNet [14, 13] and two-stream convolutional networks [15], combine raw video inputs with preprocessed data (e.g., optical flow or pose estimations). These architectures often utilize Neural Architecture Search (NAS) to optimize connections between multistream blocks. Similarly, Long Short-Term Memory (LSTM) networks have been integrated with CNNs for temporal modeling in action classification tasks [3]. The SlowFast network is another innovative approach, leveraging low frame rates to capture spatial semantics and high frame rates to model temporal dependencies effectively [6].

The emergence of transformers and self-attention mechanisms has transformed the field of visual recognition [4]. Transformer-based models not only outperform traditional CNN-based methods but also enable novel approaches in self-supervised learning. For instance, Masked Autoencoders (MAE) [17, 19, 18] and Joint-Embedding Predictive Architectures (JEPA) [Bardes et al., 1] have been utilized to harness the power of SSL for HAR tasks.

The availability of open-source datasets and foundation models from leading research labs like FAIR has opened new avenues for experimentation and innovation in HAR. Recent advancements, such as Sapiens [10], provide tools like real-time pose estimation with 308 key points, accurate segmentation, normal estimation, and depth estimation. These capabilities expand the toolkit for researchers, allowing for the design of neural architectures that capture rich temporal signals to complement the spatial modeling strengths of foundation models.

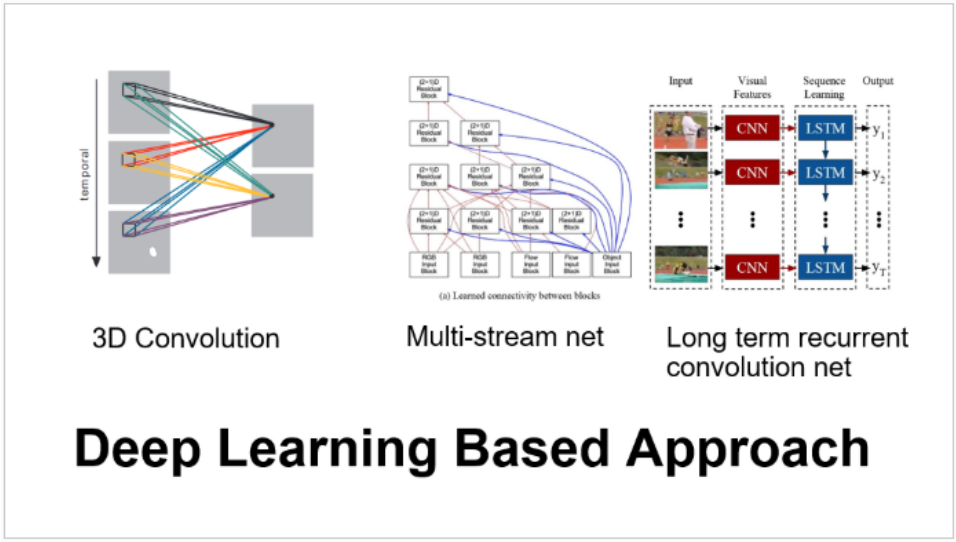

Figure 1: Recent Deep Learning Approaches for Video Understanding

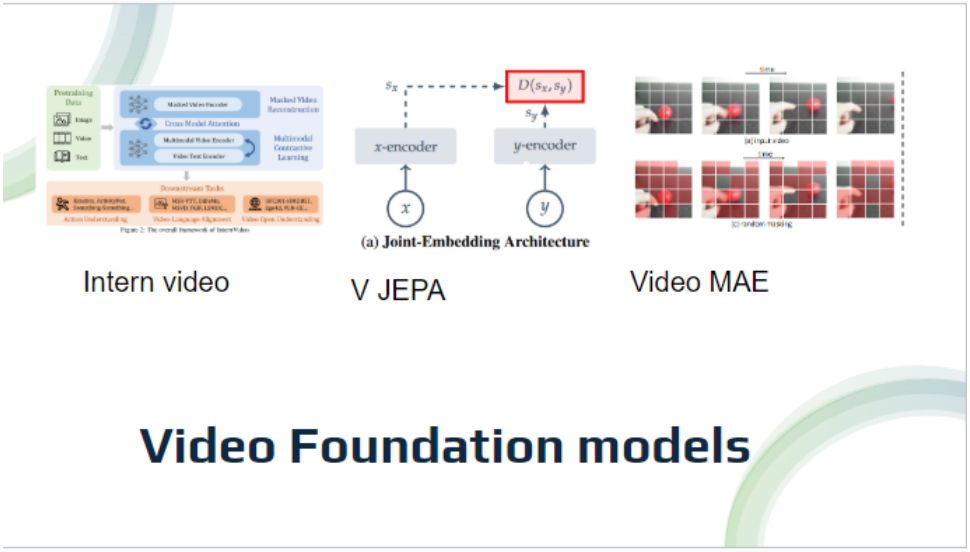

Figure 2: Video Foundation Models

Datasets play a crucial role in training state-of-the-art models. Widely used datasets include Kinetics [9], UCF101, and others, as shown in Figure 3. These datasets provide diverse and large-scale video collections, enabling the development of robust and generalizable HAR models.

ActivityNet [50]	2015	28,000	-	203	-	RGB	Uncontrolled
YouTube-8M [4]	2016	8,000,000	-	4,716	-	RGB	Uncontrolled
Charades [230]	2016	9,848	2	157	-	RGB	Controlled
NTU-RGB+D [229]	2016	56,680	-	120	106	RGB+D+IR+Skeleton	Controlled
PKU-MMD (Phase 1) [29]	2017	1076	3	51	66	RGB+D+IR+Skeleton	Uncontrolled
PKU-MMD (Phase 2) [29]	2017	2000	3	49	13	RGB+D+IR+Skeleton	Uncontrolled
NEU-UB	2017	600	-	6	20	RGB-D	Controlled
Kinetics [100]	2017	500,000	-	600	-	RGB	Uncontrolled
AVA [72]	2017	57,600	-	80	-	RGB	Uncontrolled
20BN-Something-Something [70]	2017	108,499	-	174	-	RGB	Uncontrolled
SLAC [330]	2017	520,000	-	200	-	RGB	Uncontrolled
Moments in Time [178]	2017	1,000,000	-	339	-	RGB	Uncontrolled
EPIC-Kitchens [35]	2018	90,000+	-	397	32	RGB	Uncontrolled
COIN [253]	2019	11,827	1	180	–	RGB	Uncontrolled
HACS Segments [328]	2019	50,000+	1	200	–	RGB	Uncontrolled
HAA00 [28]	2021	10,000	-	500	-	RGB	Uncontrolled
MultiSports [146]	2021	3200	-	4	-	RGB	Uncontrolled

Figure 3: Datasets Used in HAR Research

3 Proposed Methodology

We will be exploring and testing differeng video foundation models and finding out which ones balance the perfomace vs efficiency trade off.

Our approach can be summarized by a block diagram in figure 5. We will fine tune different foundation models including VJEPA, Sapiens, and InternVideo in popular datasets. Then we will compare and contrast between theses approaches and select one that suits best for our purposes. We will adapt the selected model for detecting shoplifting in supermarket. The models will be served via API and web application.

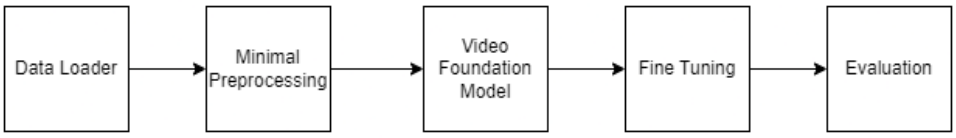

Figure 4: Block Diagram of Our Approach

4 Proposed Experimental Setup

4.1 Dataset Preparation

We will use the **Kinetics dataset** for training and testing. The data will be preprocessed by sampling frames, resizing them to 224×224 pixels, and normalizing pixel values. Data augmentation techniques such as random cropping, flipping, and rotation will be applied to introduce diversity and improve model generalization. For finetuning to detect shoplifting, we will utilize different free dataset such as https://www.kaggle.com/datasets/mateohervas/dcsass-dataset.

4.2 Model Selection

We will explore pretrained models such as **SlowFast**, **VJEPA**, and **Sapiens** for human action recognition. These models will be fine-tuned on the Kinetics dataset to improve performance.

4.3 Model Architecture

The models we experiment with will utilize **Video Masked Auto Encoder**, **JEPA** and others. We will finetune the foundation models trained for general video understanding task to classify actions accurately.

4.4 Training Configuration

The models will be loaded and finetuned in Pytorch using AdamW optimizers. Our project will use different types of loss functions for different foundation models but for the action classification task we will be using catergorical cross entropy. Nvidia GPUs with cudaNN will be utilized to speed up the finetuning process.

4.5 Evaluation Metrics

The performance of the models will be evaluated using the following metrics: Accuracy, Precision, Recall, F1 score, and Confusion matrix.

4.6 Real-Time Action Recognition

We will explore ideas to make real-time action recognition using techniques such as **quantization** and **pruning**. Inference latency will be measured to ensure the system is suitable for real-time deployment.

4.7 Testing Scenarios

The primary focus will be on **single-person** action recognition. Additionally, we will extend the model for **multi-person** action recognition by incorporating human detection

techniques to track and classify multiple individuals in the same video.

4.8 Deployment and Applications

We will test the system for real-world applications such as Theft detection in supermarkets, Fall detection in healthcare environments.

5 Proposed System design

5.1 Block Diagram of Application

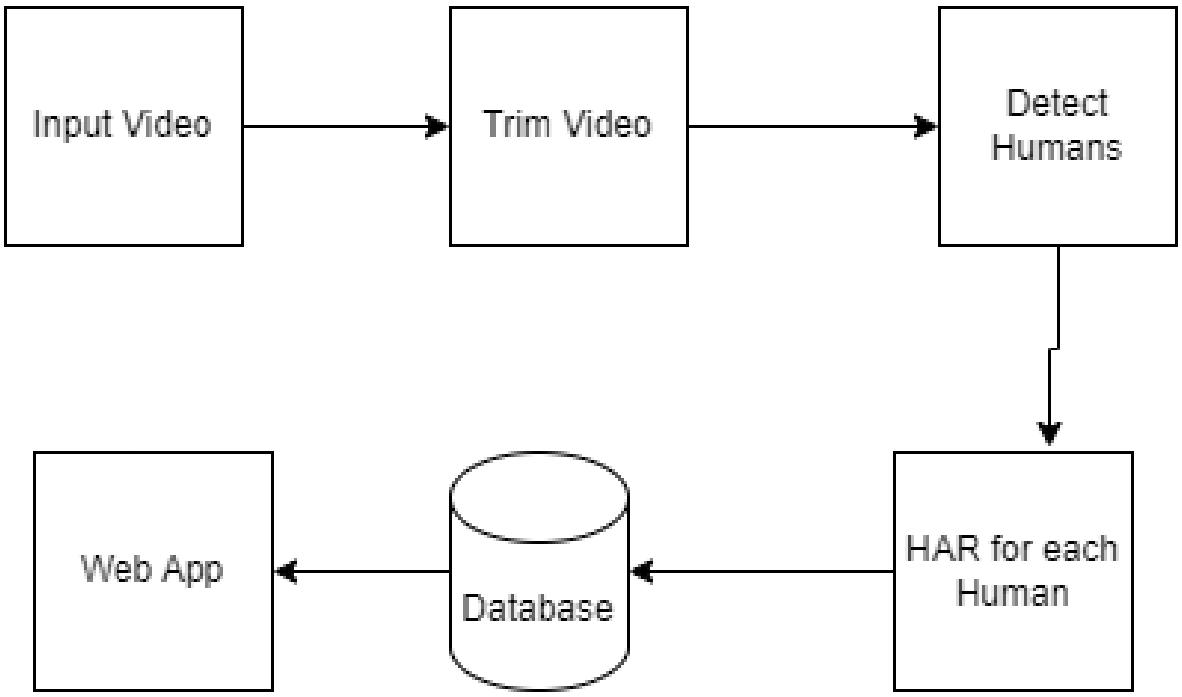

Figure 5: Block Diagram of Application

5.2 Use Case Diagram

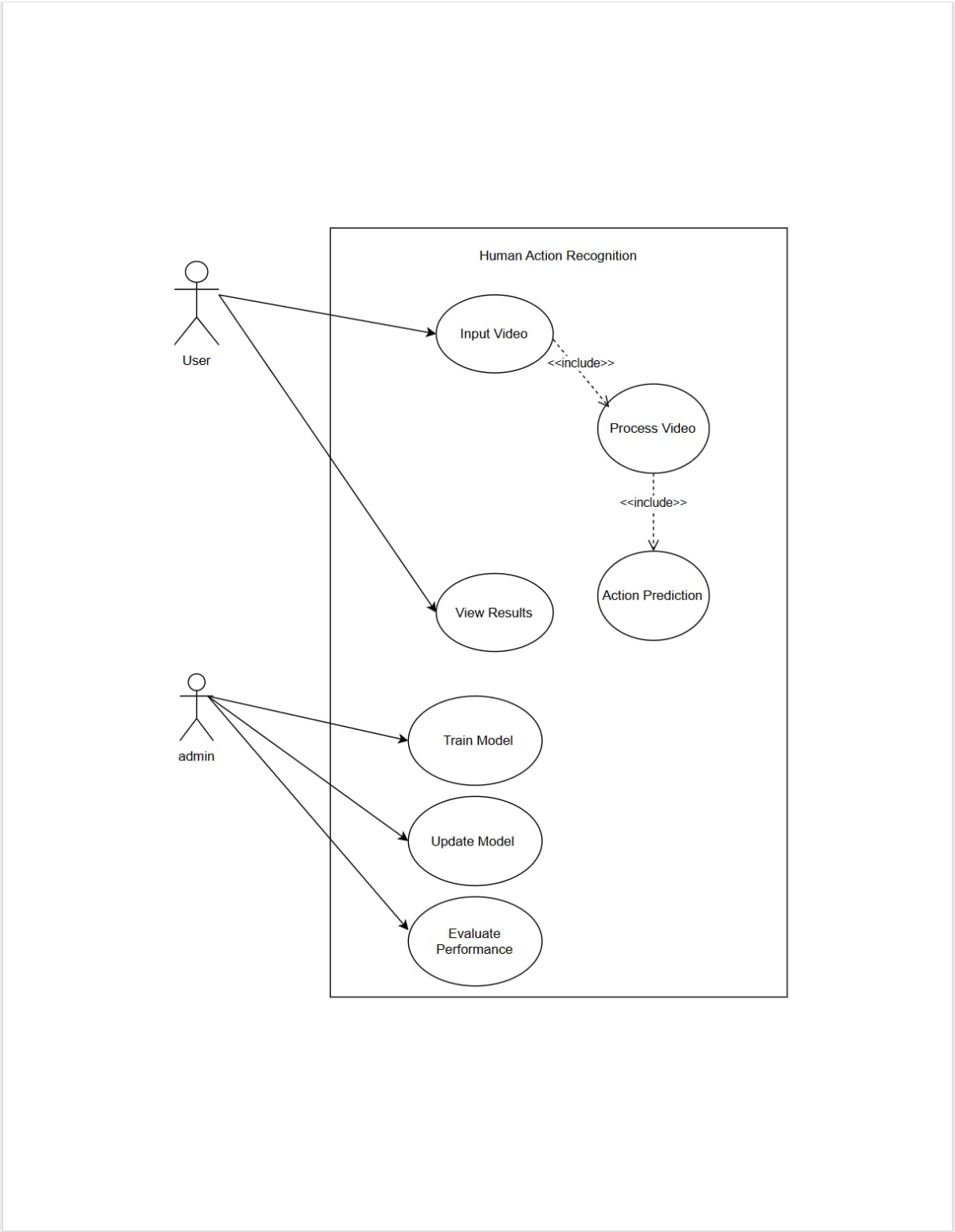

Figure 6: Use Case Diagram

6 Timeline

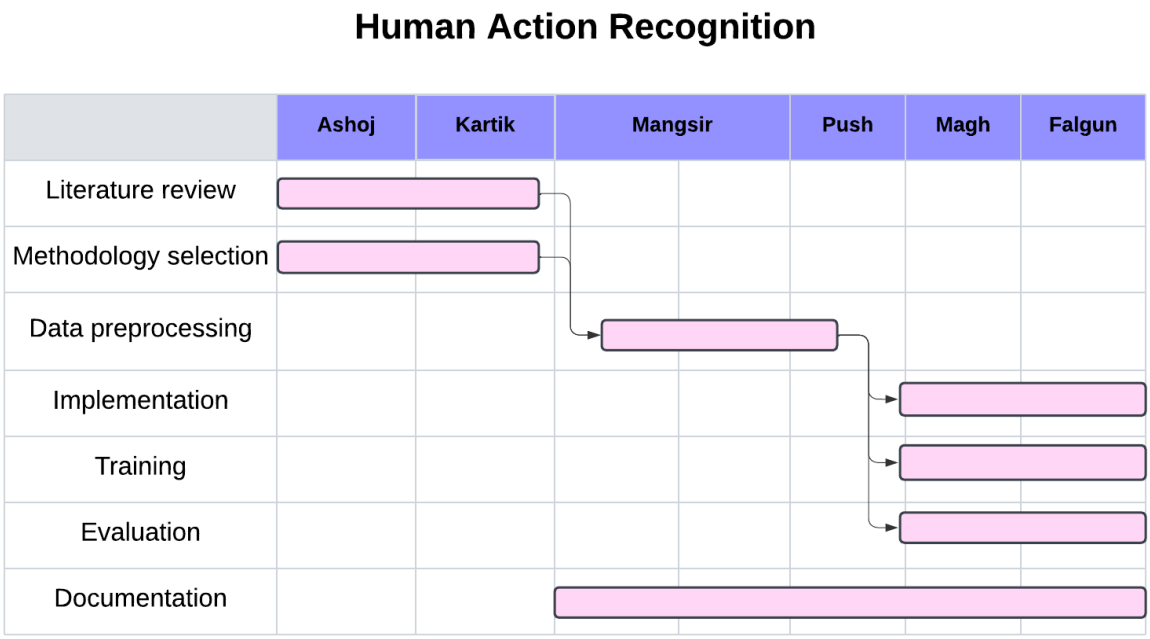

Figure 7: Proposed Gantt chart of the project.

7 Resources Required

- **Computational Resources:** Access to high-performance computing systems, including GPU or CPU clusters, for training and testing machine learning models.
- **Dataset:** A high-quality dataset that includes relevant features such as historical crop yields, weather conditions, soil properties, and farming practices. Public datasets like those from *Kaggle*, *FAO*, or local agricultural agencies may be utilized.
- **Papers and Literature:** Access to relevant research papers and publications through platforms like *Google Scholar*, *IEEE Xplore*, or university libraries to stay updated on state-of-the-art methods in AI and agriculture.
- **Guidance:** Mentorship from faculty advisors and experts in the fields of machine learning, data science, and agricultural studies to guide the research process, evaluate results, and refine methodologies.

8 Expected Outcomes

- **Accurate Action Classification:** A trained machine learning model capable of accurately classifying various human actions based on video footage.
- **Fine-Tuned Model for Supermarket Theft Detection:** A specialized version of the model fine-tuned to detect and identify suspicious behaviors indicative of theft in a supermarket environment.
- **Web Application Deployment:** Deployment of the final model in a user-friendly web application for real-time monitoring and analysis, enabling easy access and usability for end-users.

Figure 8: Expected Outcome

References

[1] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. arXiv:2301.08243.

[Bardes et al.] Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting Feature Prediction for Learning Visual Representations from Video.

[3] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2016). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv:1411.4389 [cs].

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

[5] Du, X., Li, Y., Cui, Y., Qian, R., Li, J., and Bello, I. (2021). Revisiting 3D ResNets for Video Recognition. arXiv:2109.01696 [cs, eess].

[6] Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). SlowFast Networks for Video Recognition. arXiv:1812.03982.

[7] Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

[8] Karim, M., Khalid, S., Aleryani, A., Khan, J., Ullah, I., and Ali, Z. (2024). Human Action Recognition Systems: A Review of the Trends and State-of-the-Art. *IEEE Access*, PP:1–1.

[9] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*.

[10] Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., and Saito, S. (2024). Sapiens: Foundation for Human Vision Models. arXiv:2408.12569.

[11] Kong, Y. and Fu, Y. (2022). Human Action Recognition and Prediction: A Survey. arXiv:1806.11230 [cs].

[12] Liu, M., Liu, H., and Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362.

[13] Ryoo, M. S., Piergiovanni, A. J., Kangaspunta, J., and Angelova, A. (2020a). AssembleNet++: Assembling Modality Representations via Attention Connections. arXiv:2008.08072 [cs].

[14] Ryoo, M. S., Piergiovanni, A. J., Tan, M., and Angelova, A. (2020b). AssembleNet: Searching for Multi-Stream Neural Connectivity in Video Architectures. arXiv:1905.13209.

[15] Simonyan, K. and Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. arXiv:1406.2199 [cs].

[16] Song, L., Yu, G., Yuan, J., and Liu, Z. (2021). Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055.

[17] Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. arXiv:2203.12602.

[18] Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Yan, Z., Zheng, R., Xu, J., Wang, Z., Shi, Y., Jiang, T., Li, S., Zhang, H., Huang, Y., Qiao, Y., Wang, Y., and Wang, L. (2024). InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. arXiv:2403.15377 [cs].

[19] Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., and Qiao, Y. (2022). InternVideo: General Video Foundation Models via Generative and Discriminative Learning. arXiv:2212.03191 [cs].