Project Proposal including background research summary

Topic: DroneAI: Humans interacting with Drones

Course Code: FIT3161

Team: MCS14

Submitted by:

Rahul P. Rajendran (32912617)

Lim Zheng Haur (32023952)

Yap Jit Feng (32898339)

Thehara Nikhila Goonewardena (32312512)

Submission Date: 30/10/2023

Word Count: 8808

Table of Contents

Introduction

Due to the constantly evolving realm of digital technology, the need to find more innovative solutions that cater to modern-day challenges has never been more pressing. Recognizing this need, our team embarked on a journey to conceptualise and develop a project that promises not only to acknowledge the significance of these issues but also to bring forth a transformative change in its respective domain.

The essence of our endeavour stems from “DroneAI : Humans interacting with Drones”. This project primarily focuses on applying accurate and efficient in depth analysis of human activities, behaviours, and patterns through drone-captured data. Our mission is not only to address the significance of the issue but to improve upon the methods of approaching and understanding the problem.

As we steer into the upcoming semester, our comprehensive project plan, detailed in the subsequent sections of this report, illustrates a roadmap delineated with milestones, strategies, and anticipated outcomes. This plan is not merely a testament to our team's commitment but also a beacon directing our efforts towards achieving our shared vision.

The structure of this report has been meticulously structured to provide a holistic view of our project. Following this introduction, a thorough literature review will shed light on previous works and the inspiration behind our approach. Subsequent sections dive deep into the intricacies of our project management plan, design methodologies, and testing strategies, among others. The report culminates in a cohesive conclusion that ties together our collective insights and aspirations.

As for our team (Rahul, Lim, Yap, and Thehara), we comprise a diverse group of individuals, each bringing a unique set of skills and perspectives to the table. United by a shared passion for innovation and a commitment to excellence, it serves as the main driving force that enhances our passion. We eagerly anticipate not just the challenges that lie ahead, but more importantly, the potential impact and value our project promises to deliver.

Literature Review

The integration of Artificial Intelligence(AI) and drone technology have revolutionised the industry of computer vision and surveillance. The effective combination of drones or unmanned aerial vehicles (UAV) and machine learning algorithms has expanded endless amounts of possibilities in applications such as search and rescue, security, dynamic motion animations and many more. The capabilities of detecting human movement from an aerial point of view unlocks potentials in bringing safety and efficiency to military systems.

The concept of allowing drones to detect human actions requires multiple layers of machine learning models working simultaneously to produce an accurate prediction of a human action. To allow machine learning models to detect human interactions, the video footage must first be converted into information recognized by the model. In this case, video footage will be converted into frames which are essentially multiple sequences of still images that, when played in rapid succession, creates an illusion of the video. These frames will then be fed into the machine learning algorithm to determine the location of the human and the action performed by the human.

In order to detect human movements/actions, the ability to detect humans in a given frame in a video footage is essential. This is important as the learning models might mistakenly classify random objects detected in a given frame performing an action. To eliminate this event from happening, an object detection learning model such as You only Look Once(YOLO), Fast R-CNN and SSD will be used. These object detection algorithms are efficient in identifying wide variants of objects and in this context of study it will mainly be used to detect the presence of a human. In this literature review, these algorithms will be undergoing a comparative analysis upon their models architecture, efficiency and accuracy.

After the location of the human in a given video frame is determined, the information will be passed onto a pre-trained action detection model such as Inflated 3D-CNN, ResNet (2+1)D (R(2+1)D) and Temporal Shift Module. These algorithms are pre trained using large varieties of labelled dataset that consist of at least 100 classes such as walking, punching, running and many more. This is to improve the accuracy of the learning algorithms. These algorithms will mainly be used to predict the action performed by the human detected by object detection model. By combining both object detection model and action recognition mode, these models can identify the individuals in a video frame and effectively predict the action performed accurately.

Object Detection Model

There are multiple different object detection algorithms that are used for detecting human movements on drones. The main notable ones are the You Only Look Once model (YOLO), Fast R-CNN and Single Shot MultiBox Detector (SSD). These models are the initial step in

identifying the humans in the given frame. Without these algorithms, the action detection algorithm's prediction accuracy score will be affected negatively.

YOLO model is a real time object detection machine learning algorithm that treats objects as a regression task. It is unique as it has the fastest algorithm in recognizing objects that are shown in an image and video. This is because YOLO adapted a method that only requires a single pass of the data through a neural network which makes it efficient, quick and reliable. YOLO accomplishes this by dividing the input image/frame into a grid such that each grid is responsible for predicting the object within the grid. If the objects are located in multiple different grids, then the grids will be combined and given a prediction (Ahmad et al, 2022; Jiang et al, 2022). This grid base approach allows the algorithm to simultaneously predict multiple objects in the same image. Besides that, the recent YOLO model also uses a bounding boxes method that identifies the object in the image by placing a bounding box around the object. These bounding boxes will be given the location of the object and the confidence level of the predicted output. There are a large number of different sized predefined bounding boxes to be used on the image. This is to ensure that the model can predict objects with different sizes. This method is known as anchor boxes. By detecting the location of objects in the grid, the algorithm will use anchor boxes to determine the size of the object (Ahmad et al, 2022). Then the detected image will be passed inside the neural network to provide an prediction for each predefined anchor box (Jiang et al, 2022). The highest confidence level for the object will be the predicted output for the given object in the grid.

Single Shot Multi Detector (SSD) is similar to the YOLO model as both of them provide real time detection and provide single pass features that allow it to detect multiple objects in an image. Unlike YOLO, SSD uses convolutional neural networks as its initial layers. This CNN layer is responsible for extracting different features from the input image, transforming it into sets of feature maps (Liu et al, 2016). These feature maps capture diverse ranges of characteristics about the object such as the patterns, texture, edges, object parts and many more. After the features map has been acquired, SSD uses anchor boxes to determine the size of the object within the feature map (Liu et al, 2016). Each of these anchor boxes, the boxes will be given coordinates and a confidence score. Finally, the anchor box with the highest confidence score will be selected as the primary output for the object. This approach with SSD allows the algorithm to handle wide ranges of object sizes efficiently. Anchor boxes also allow detection of multiple objects at once, which makes it a versatile model.

Fast Region convolutional neural network (Fast R-CNN) is the improved version of R-CNN network . One of its key advancements is using a single forward pass network that improves its speed and efficiency. The algorithm uses a method called the selective search algorithm that helps identifying regions of interest in an image (Jiang & Miller, 2017). Selective Search Algorithm starts by dividing the image into different segments that share similar texture, colour and shape (Jiang & Miller, 2017). These regions will be combined together to generate a proposed region that contains a potential object. Each of these regions are represented with bounding boxes with coordinates that shows the object location. To prevent regions from

overlapping each other, the model adopts a method called the non maximum suppression (NMS). NSM is responsible for taking all the regions and sorts them based on regions with the highest confidence score. NSM will use intersection over union(IOU) to check for overlapping regions and remove the region with the lower confidence score (Jiang & Miller, 2017).

Valarmathi et al. (2023) has conducted an experiment of comparing multiple different object detection models, which ultimately decided that YOLO is the most efficient and accurate. The reasoning is the model computational efficiency. Unlike the SSD and Fast R-CNN model that heavily relies on convolutional neural networks, YOLO uses grid based detection. The design has effectively reduced the computational demands which makes it suitable for the project. YOLO is able to detect more object classes than other algorithms on drones with lower camera specification (Valarmathi et al, 2023). This makes YOLO a reliable choice where versatility is important when facing limited resources. Besides that, YOLO is also flexible in customisation such that it can cooperate with other techniques such as gradient boosting, R-CNN, and many more (Ahmad et al, 2022; Valarmathi et al, 2023). YOLO is also the fastest object detection algorithm as its design model is efficient and reliable making it reliable for real time detection. However, the performance of the algorithm will be affected by the altitude in which the drone captures data. The higher the angle in which the drone captures the data, the worse the prediction accuracy of the algorithm (Ahmad et al, 2022). In terms of accuracy, fast R-CNN produced the best results. This high accuracy comes with a trade off with a much slower speed and higher computational power (Samma & Sama, 2023; Valarmathi et al, 2023).

In Summary, YOLO is the fastest object detection algorithm that is computationally less demanding and provides effective real time detection that is feasible for any application. Fast R-CNN excels in providing the best accuracy scores but it requires a long time and high computational power. SSD is the middle of the pact as it provides a balance between speed and accuracy. For this project, YOLO will ultimately be the best algorithm as it is proven by several researches that its speed and accuracy provide the best solution for Drone AI Human action detection.

Action Detection Model

Once the object detection model highlighted above has identified the subject(s) of interest, these subject(s) are then treated as input to our action recognition model which analyses these subject(s) over a period of time and outputs a prediction that tries to correctly classify the action(s) that are being carried out by the subject(s).

Human action recognition is a major part of our application and there are many methods to achieve this. According to Archana & Hareesh (2021), the main consideration to solve this standard problem in computer vision, is by utilising feature extraction, where in a given input RGB data, the spatial and temporal features are extracted and analysed for prediction and

activity recognition. With the recent developments and advancements in deep learning techniques, Convolutional Neural Networks (CNN) have emerged as the frontrunner on real-time human activity recognition.

Deep 2D convolutional neural networks originally led the breakthrough of image classification, however such models are only able to handle 2D inputs at any given moment and lack temporal modelling. As such, 2D CNN on individual frames are not able to model temporal information well, for instance, distinguishing between opening and closing a box will give opposite results if the order of input was reversed (Lin et al., 2019). Due to innovations in the convolutional layers of these models, these issues with neural networks were mitigated and have become the linchpin of action recognition models, primarily by performing 3D convolutions effectively turning these 2D CNNs to 3D CNNs (Archana & Hareesh, 2021). Hence, 3D CNNs are set apart compared to other deep learning models for their innate capacity to decipher spatiotemporal patterns within video sequences, enabling them to automatically learn and predict complex actions.

One such deep learning model that we are considering implementing is the Two-Stream Inflated 3D ConvNet (I3D) proposed by Joao Carreira and Andrew Zisserman in 2017. The I3D model extends upon image classification architectures by inflating their filters and pooling kernels - endowing them with an additional temporal dimension (Carreira & Zisserman, 2017). According to Carreira and Zisserman (2017), the I3D model takes inspiration from the Inception-v1 model, a 27-layer deep CNN, and adapts it by combining the output of two 3D CNNs with one I3D network trained on RGB inputs and the other processing a group of optical flow predictions among consecutive RGB inputs. As a result, Carreira and Zisserman (2017) examined that their I3D model is able to compete with other state of the art deep learning models (at the time) posting a Top-5 accuracy of 88.0% and Top-1 accuracy of 68.4% for RGB inputs after training and testing on the Kinetics 400 dataset beating out other deep learning architectures such as Long-Short Term Memory (LSTM), 3D-ConvNet (C3D) and Two-Stream Networks. On average, the Two-Stream I3D model also outperformed the models mentioned above on the UCF-101 and HMDB-51 dataset with an accuracy of 98.0% when pretrained on the Kinetics dataset and 80.9% when pretrained on both the Kinetics and Imagenet datasets respectively (Carreira & Zisserman, 2017). In their paper on violence detection in videos, Selvaraj and Anuradha (2022) implemented the I3D ConvNet model on preprocessed input data and observed an accuracy of 80%, a drastic improvement over other architectures such as LSTM, Xception and Fight-CNN. This further proves the effectiveness of the I3D model in action classification over other state of the art options.

While the Two-Stream Inflated 3D ConvNets (I3D) model, with its 3D spatiotemporal analysis capabilities are a significant leap forward in terms of action recognition, the downside to this is the demand for substantial computational resources, which can be detrimental when trying to perform real-time or near real-time analysis of RGB data.

One such solution to decreasing the computational cost without compromising the model's performance is by implementing the Temporal Shift Module (TSM). TSM is not a standalone model on its own, but an additional component that is integrated into existing neural networks to improve the temporal modelling of data. TSM is able to achieve the performance of a 3D CNN but maintain the complexity of a 2D CNN (Lin et al., 2019). Furthermore, Lin et al, (2019) mentioned that TSM can be seamlessly integrated into 2D CNNs, allowing for efficient temporal modelling with no extra computational load or parameters and extended into the online setting opening up the door for real-time, low-latency video and action recognition. TSM is able to accomplish effective temporal modelling by adapting the convolution operation to involve shifting in the temporal dimension by one time step (forward and backward) and in the context of real-time online video analysis, a unidirectional TSM can be employed. According to Lin et al. (2019), backbone 2D CNNs such as ResNet-50, once equipped with TSM had their performance and accuracy improve by double digits over their baseline counterparts on datasets like Kinetics, UCF101, HMDB51, Something-Something V1 and V2 and Jester. When compared to other state of the art models, TSM records a Top-1 accuracy of 45.6% on the Something-Something V2 dataset and 74.1% accuracy on the Kinetics dataset and when compared with the I3D model, the TSM method is faster by an order of magnitude and 1.8% higher accuracy with much less computations. Lin et al. (2019) also measured the accuracy of offline and online TSM with the online version recording a latency of 4.8ms and similar accuracy to its offline counterpart on the datasets mentioned above.

Another deep learning architecture that we considered to implement which allows a 2D CNN to capture spatiotemporal features effectively is the ResNet(2+1)D model that was proposed by Tran et al. (2018). According to Tran et al. (2018), spatial and temporal modelling can be broken down into 2 different steps. As such, Tran et al. (2018) proposed that a "(2+1)D convolutional block that separates 3D convolutions into 2D spatial convolutions and 1D temporal convolutions can be integrated into a 2D CNN to approximate the accuracy of a 3D CNN. The advantages of this approach is the increase in non-linear transformations in the model with the same number of parameters allowing the model to represent more complex functions as well as a potential to simplify the optimisation process that results in lower training and testing losses. When training and testing the R(2+1)D model with a backbone consisting of a 2D ResNet with 18 layers, the proposed model outperforms all other 2D CNN models for action recognition accuracy on the Kinetics dataset. Furthermore, when comparing with other state of the art CNNs, R(2+1)D is observed to outperform C3D by 10.9% on the Sports-1M dataset, outperforms the I3D model by 4.5% on the Kinetics dataset and almost on par with the I3D model on the UCF101 and HMDB51 datasets. Huang et al. (2021) was not only able to replicate these numbers but also improve upon them in their research paper on R(2+1)D-based Two-Stream CNN for human activity recognition. In the temporal stream on input data, the optical flow images are extracted before being inputted into (2+1)D CNN allowing higher efficiency in learning temporal features leading to an improvement in the performance of the model and a 94.97% accuracy on the UCF101 dataset.

In summary, TSM and R(2+1)D are both computational efficient and highly accurate action recognition models that can be used for our application. The I3D model is just too computationally expensive to justify being implemented considering the other three proposed models are on par in terms of accuracy with the I3D model. For this project, we have chosen to implement the TSM with 2D ResNet backbone to be used as our action recognition model not only due to its efficiency but also due to an online version of the model being trained, tested and proven to be accurate on various datasets (Lin et al., 2019). Thus, we believe this will be the best solution to our action recognition problem in our project.

Project Management Plan

1. Project overview

This project aims to utilise and to enhance the latest technologies in live video analysis adapted to drone captured video data. The focus of this project is on human analysis, specifically action identification, and additionally include facial analysis and emotion recognition. The ultimate end product is a delivery of drones that not only observe and process human behaviours but also be able to demonstrate interaction with humans in diverse environments. This transformative capability will be realised through integrations of state-of-the-art computer vision algorithms, taking a significant leap forward in the application of drone technologies.

This initiative involves several core components:

1. Live video analysis
2. Object identification (Human)
3. Human action identification
4. Facial analysis
5. Computer vision algorithms

Overall, this ambitious project is pushing the boundaries of drone technology by the integration of different advanced artificial intelligence techniques to enable drones to not just capture video data, but to actively interact with humans by responding to humans actions and behaviours.

2. Project scope

a. Project scope

The project scope of this project is to develop a system that is able to process live drone video footage by analysing human and facial data. This creates the opportunity to extract valuable insights from drone video data to establish new interaction between drones and humans. To identify priority tasks and non-priority tasks, we have divided our tasks into in-scope and out-of-scope items. The in-scope tasks are to apply object identification algorithms to identify humans and also to apply action identification algorithms to recognise actions and behaviours of humans through live drone captured video feed and allowing drones to interact with humans with it. Our out-of-scope items include integrating facial analysis on top of action recognition to better identify the behaviour of humans, to develop a user-friendly and intuitive user interface, and to develop our own drones for better compatibility and optimisation. This project aims to utilise cutting-edge technology in video

processing, and could be implemented to enhance various applications such as security, surveillance and crowd monitoring.

b. Product characteristics and requirements

To ensure the product meets its intended purpose and satisfies users needs, it is essential to define the product characteristics and requirements for both the physical device and software application. These characteristics are used as a blueprint for the development and evaluation of the product. The requirement traceability matrix is attached to the appendix as **item A1**. The requirement traceability matrix helps in managing our requirements by providing a clear view of the source, the category, and the type of each requirement which ensures that all the requirements are adequately addressed throughout the project phases.

c. Product user acceptance criteria

The user acceptance criteria for each requirement is tabulated into a table as **item A2** in the appendix. These acceptance criteria set up a clear standard to ensure that the requirements meet the benchmark. We would be able to evaluate and confirm that each requirement could be accepted and delivered. Setting up the user acceptance criteria also increases our group transparency and allows for better communication between our team, stakeholders and end users as everyone would have a clear understanding of the expectations which foster a more effective collaboration. Our team would also use these as a framework for testing each of our deliverables to reduce rework and misunderstandings within the development team.

3. Project organisation

a. Process Model

We will be adopting the predictive project life cycle model for this project. The rationale behind choosing this approach is to provide a structured and systematic framework to manage the complex deliverables of this project. Due to the nature of this project, which involves advanced techniques of computer vision analysis, a predictive model allows us to define milestones upfront which is ideal for this project where the objective is concrete with likely no changes. The predictive model enables our team to establish a detailed project management plan with clear milestones and deliverables.

b. Project Responsibilities

In a predictive process model, roles are essential for thorough planning, requirement specification, defining deliverables, and adhering to the appointed deadlines as per project schedule. Hence, each role has their individual responsibilities in ensuring a predictive process model is able to be executed successfully, where the project is able to progress systematically to meet predefined milestones and objectives within the initial constraints. However, it is important to note that all group members equally contribute to the development of the project regardless of their roles similarly to an Agile process model.

1. Project Manager (Lim):

   The project manager will play a critical role in planning, executing, and monitoring the project. Project managers are responsible for ensuring the deliverables are on schedule, and the project is kept within other constraints such as the budget. The project manager will also be important in ensuring that the team is cohesive, where every member shares the same goal and to ensure that the productivity and efficiency of the team is optimised. Risk assessment and mitigation is also crucial for the stability of the project where decisive decisions are to be made.

2. Technical Lead (Rahul, Yap):

   The technical lead serves as the anchor of the group that oversees and develops the critical aspects of the project. Their main responsibilities are making key decisions in the designing of the overall system architecture, including both hardware and software components. The technical lead will also be responsible for leading the development and integration of the hardware and software components. This evolves coordination with other team members, ensuring that the project's requirements and objectives are met.

3. Quality Assurance (Thehara):

   The quality assurance is responsible for rigorously testing the hardware and software components and ensuring that these components function as intended and meet the quality standards. It is also important for quality assurance to be responsible for assessing the security risks and implement proper security measures and protocols.

4. Management Process

a. Risk management

The initial step in the risk management process is to identify the risks. Our team identifies risks by looking at various aspects of the projects from the technical perspective, the operational perspective, and external factors. Examples of technical risks are such as drone hardware failure, and software bugs, operational risks are such as weather conditions on testing drone performance, while external risks are such as changes in regulations for drone usages. The next step in risk management is to analyse the risks that were identified, the likelihood of each risk occurring, and the impact of each risk are analysed to determine the priority and significance of each risk. Furthermore, mitigation and contingency strategies are formulated for each risk and tabulated into our risk register which may include creating redundant systems for critical components. Finally, in the operation of the project, risks are monitored and controlled where occurrences of risks are reported and the mitigation strategies are executed and assessed for their effectiveness. Our team risk register is attached at the appendix as **item A3**.

b. Stakeholder analysis and communication plan

Stakeholders are essential to any project and it is important to stakeholders to be informed and engaged in the project. Therefore, a communication plan is crucial to ensure that stakeholders are regularly informed of concerns to the project. In the analysis of the stakeholders, first the stakeholders to the project are identified and for each stakeholder, their roles and interests are determined. By evaluating their roles and interest, the communication plan, such as the method and frequency of reports are decided. The communication plan is located in **item A4** in the appendix,

c. Monitoring and controlling mechanisms

Effective communication is vital to every project's success. Our team has established a communication plan, which outlines how information is communicated within our team. This communication plan could be found in the appendix as **item A5**. The main communication platform within our team would be online meeting platforms such as Zoom and messaging platforms such as WhatsApp.

Task allocation would be handled by the Technical Lead, and also to each team member's personal skills and preferences. The tasks would ideally be equal among team members, but we are ready to provide assistance to other

team members that require aid to ensure that deliverables are completed on time and each milestone of the project is accomplished.

The project manager would be responsible for managing the project schedule and ensuring that deliverables are delivered on time, as well as milestones being accomplished on schedule. In any case where the project is behind schedule, the project manager could take action according to the risk register where the risk of a delay in the project schedule has been identified. The project manager would also identify the cause of the problem and to take actions to find the solution to the issue.

A versioning control system is set up for auditing purposes and also allows for review of the progress of the team. The quality assurance carries the responsibility of ensuring the quality of the product meets the standard of the team, and also other important aspects of the project such as documentation are fulfilled.

5. Schedule and Resource Management

a. Work Breakdown Structure (WBS)

Phase-Based Work Breakdown Structure

The work breakdown structure (WBS) above breaks down our project into smaller, more manageable tasks which act both as progression checkpoints and major milestones throughout the project's duration.

b. Schedule

Throughout this project, we have a total of 3.5 months which is about 14 weeks to develop the user interface and test the performance of the object detection model combined with the action detection model. In order to manage the progress of the project efficiently, a project scheduler was used to schedule the tasks based on its importance. In this case, Gantt charts were used as our project scheduler template. The project scheduler can be found in the appendix as **item A6**.

c. Resource

Time: 30 weeks

Personnel: 5

Roles: 1 Project Manager, 2 Technical Lead, 1 Quality Assurance, 1 Project Supervisor

Budget: RM5000 for drone + RM1000 for online tools subscription + RM1000 for utilities

External Design

The main objective of the project is being able to use a drone to detect human presence and predict the action they are doing. In order to see the results, a user interface will be created to allow users to interact with the drone and observe the predicted action generated from the algorithm based on the live feed captured on the drone.

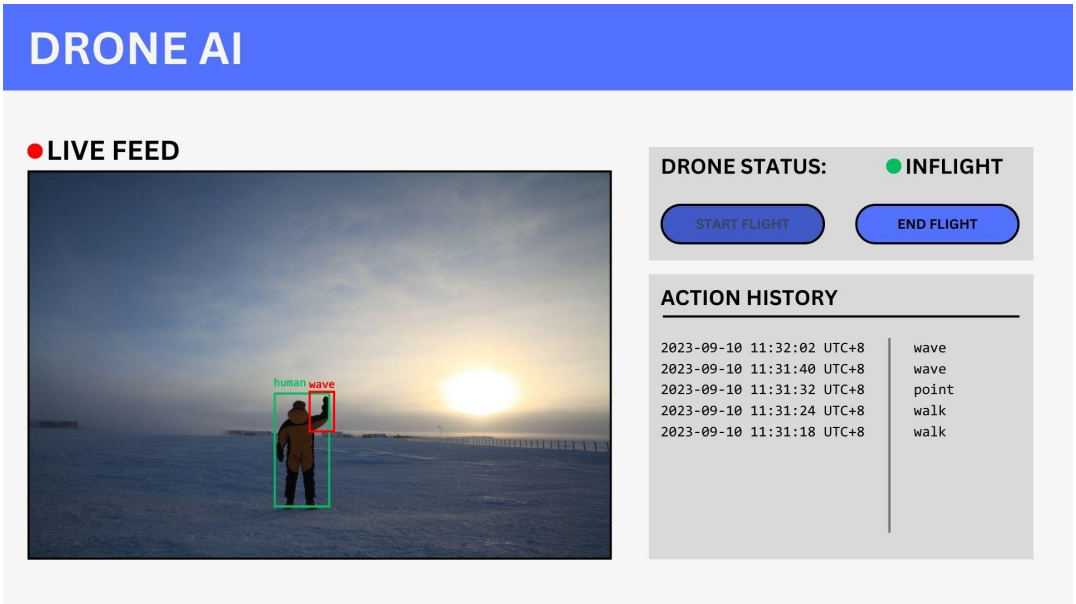

Figure 1: Web User Interface Mockup

The figure above demonstrates a glimpse of a web-based user interface mockup that plays a huge role in providing real time communication between the user and the drone. The interface serves as a base station, providing users with the tools to manage and monitor drone controls. In the User interface there are four features, each designed to provide control and comprehensive information. On the top right of *Figure 1*, users will find two buttons which are start flight and end flight. When the user clicks on the start flight button, the backend of the ui will initialise the server to load all the necessary models/algorithms. After all the necessary systems are booted, then the drone itself will start to commence flight and hover at a predefined altitude while activating its camera. Once the drone reaches its fixed altitude, the live feed will start to display on the user interface.

The central element of the user interface is the live feed display. This feature allows users to access real time video captured by the drone's camera. The live feed display comes with bounding boxes that surround any human detected within the frame. This bounding box is generated by the machine learning algorithm within the user interface. It is also a tool to make it visible for the user to see the location of the detected human. Within each bounding box, users will see labels that show the corresponding predicted action. In situations where there are multiple human presence in a given frame. Distinct colour will be assigned to bounding boxes. This will eliminate the confusion and allow users to establish the actions caused by different humans.

On the bottom right of *Figure 1*, it consists of an “action history” section. This section shows users the actions predicted over time and offers valuable historical information. All the detected actions in each frame will be recorded and displayed. This is to allow users to analyse and review the performance of the algorithms over time. The information will be displayed in a queue format where it removes the old prediction and replaces them with the new prediction.

To stop the drone operations, users can click on the end flight button. Once the button is clicked, the user interface will prompt the drone to commence a slow descent from its hovering status. This is to ensure the safety of the drone. Simultaneously, all the machine learning algorithms responsible for human detection and action detection will terminate effectively. All recorded action predictions will automatically be stored in the cloud storage for future reference.

Methodology

A. Toolset and Approaches

- Programming Language: We've chosen Python as the backbone for the droneAI project due to its versatility and robustness, especially in data analytics and machine learning tasks. Python's extensive libraries cater precisely to our requirements.

- Libraries:

    - PyTorch: A deep learning framework known for its flexibility in computing dynamic graphs. In our droneAI system, it's pivotal for the HumanDetectionModel and ActionRecognition modules, ensuring fast and accurate outcomes.

    - OpenCV: This library aids in image processing tasks, making it easier to preprocess and analyse footage from drones.

- Visualisation Tools: We use tools such as Tableau and PowerBI for visual representation of our analytical results.

- Data Management and Storage:

    - Structured Data:
        - SQL Databases: Offers efficient querying and robust storage solutions for structured data.
    - Unstructured Data:
        - AWS S3 Buckets: Utilised for securely storing and managing large quantities of raw drone footage.

  For efficient storage and retrieval of drone-captured data, we employ a combination of SQL databases and cloud storage solutions like AWS S3 and Google Cloud Storage for video files. This ensures optimal data availability and security.

- Dataset:

    - UAV-Human Dataset

  Both the training and testing of our action recognition model will be done on the UAV-Human Dataset. The dataset contains 67,428 multi-modal video sequences and 119 subjects for action recognition, 22,476 frames for pose

estimation, 41,290 frames and 1,144 identities for person re-identification, and 22,263 frames for attribute recognition (Tianjiao et al., 2021). The reason we have chosen this dataset is because the videos in other common datasets such as Kinetics, are collected on ground cameras which do not take into account motion blurs and differing resolutions of subjects due to the fast vertical and horizontal motion of the drone. The videos on the UAV-Human Dataset take these and much more into consideration such as a large sample size to prevent overfitting, a diverse number of subjects of varying ages, gender and clothing, videos under changing weathers in different time periods and all of these over different drone viewpoints and altitudes.

B. Version Control System

Git & GitHub: To maintain a smooth development lifecycle and allow multiple developers to work simultaneously, we've integrated Git for version control, with GitHub as our repository platform. This ecosystem fosters collaboration and ensures our code remains organised and accessible.

C. Algorithms and Pseudocode

The essence of droneAI lies in its ability to accurately detect and analyse human actions. Our algorithm is meticulously designed to execute the following steps:

1. Initialization: Activate drone's camera and sensors.
2. Data Capture: Obtain real-time footage and telemetry data.
3. Preprocessing: Enhance footage quality through noise reduction, colour normalisation, and stabilisation.
4. Object Detection: Our chosen state of the art object detection model detects and highlights the objects of interest on individual frames of the input data.
5. Action Recognition: The object(s) of interest determined by the object detection model are then analysed by our human action recognition model over a period of time and predict/classify the action that is being performed in the video clip.
6. Analysis: Delve deeper into patterns, behaviours, and trajectories.
7. Data Storage: Push prediction results into the database(s) for further reference and analysis.
8. Visualisation: Prediction results are then displayed to the end user(s) who are currently using the application. Generate comprehensive reports and dashboards to depict findings based on prediction results.

D. Pre-processing Steps and Quality Assurance

- Noise Reduction: Ensure clarity in the footage by removing grain and noise.
- Colour Normalisation: Standardise the colour palettes to maintain consistency.
- Stabilisation: Counteract any shaky footage due to wind or drone movement.

E. Class Diagram

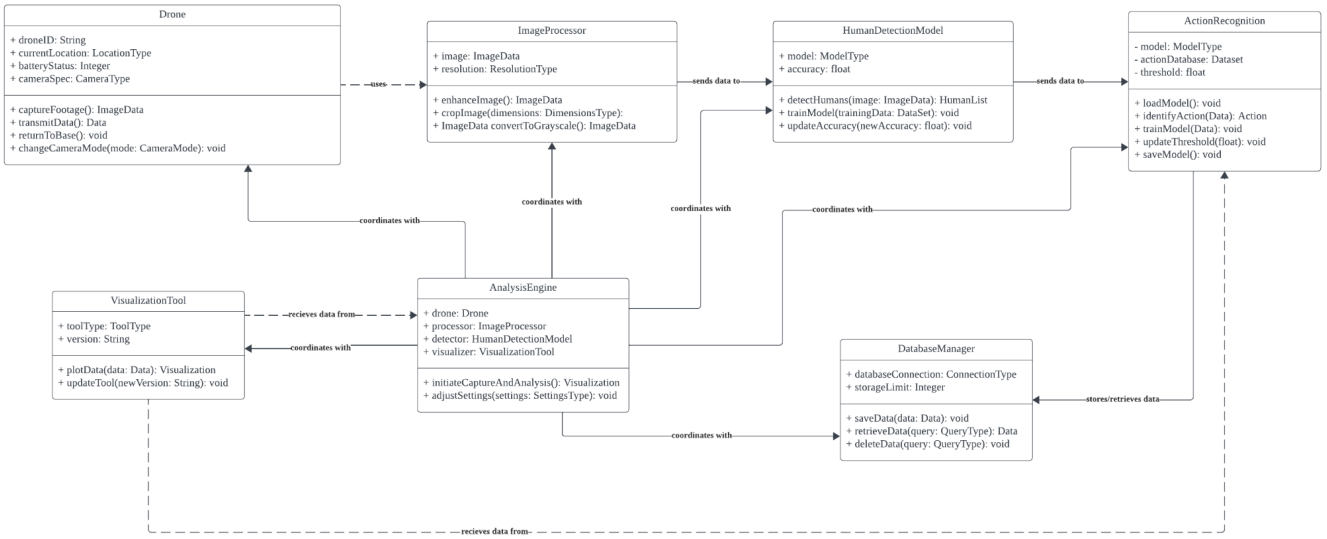

Figure 2: Class Diagram for DroneAI

The figure above demonstrates the class diagram and flow of the program and how each class communicates with each other.

The Drone class is a pivotal component of our system. It is specifically designed to capture high-resolution images or video footage from varying altitudes and angles. It comes equipped with state-of-the-art camera features that allow users to adjust settings to best fit the environment, ensuring clarity and precision in every capture. Moreover, its navigation features are enhanced to ensure that if ever the drone strays too far or encounters an obstacle, it can autonomously navigate back to its base or designated location.

Once our drone captures the visual data, it is handed over to the ImageProcessor class. This class is our first line of data refinement. It employs a series of advanced algorithms to clean and enhance the raw images. This includes adjusting the brightness, contrast, and even removing any noise or distortions that might be present, ensuring the images are primed for detailed analysis.

As the name suggests, the HumanDetectionModel class is tailored for one key function: human detection. It dives deep into the processed images, sifting through every pixel, and leveraging cutting-edge AI or machine learning algorithms to spot and accurately identify human figures amidst a multitude of elements in the visual data.

Building upon the foundation laid by the HumanDetectionModel, the ActionRecognition class takes our analysis a step further. Not only does it recognize the presence of humans, but it also intricately analyses the postures, movements, and gestures of these detected individuals. It then categorises them into recognizable actions or activities, providing an additional layer of insight into the captured data.

Understanding that raw data can sometimes be overwhelming, the VisualizationTool class comes into play to provide a user-friendly interface. It translates the complex data analytics results into intuitive visual representations, making it easier for users to understand and interpret. Users can see marked locations where humans are detected, alongside visual indicators that highlight the actions or activities they are engaged in.

The heart and brain of this entire setup is undoubtedly the AnalysisEngine class. Think of it as the conductor of an orchestra, ensuring every instrument, or in this case, class, operates in harmony. It effectively manages the flow of data and tasks between the drone, image processing, human detection, action recognition, and visualisation components. It ensures that every class gets the required data at the right time and that the overall system operates without hitches.

Finally, the DatabaseManager class acts as our system's data custodian. Recognizing the importance of data integrity and management, this class employs stringent measures to store, organise, retrieve, and, when necessary, delete data. Whether it's the raw images captured by the drone, the refined data post-analysis, or any other relevant system information, the DatabaseManager ensures it's systematically archived and easily accessible.

In essence, each class in this system brings its unique value, working in tandem to ensure that every piece of drone-captured data is meticulously processed, analysed, and visualised for the end-user's benefit.

F. Data Collection

Our primary data source is the high-resolution footage captured by drones. Additionally, we collaborate with relevant authorities and organisations to access historical and supplementary data sets, further enhancing our analysis quality.

Test Planning

In ensuring the reliability and robustness of our droneAI software system, it is crucial to carry out meticulous testing. The testing phase aims to validate the correctness, performance, and security of our software. Here is a preliminary overview of our test plan:

1. Unit Testing:

- Purpose: Validate each piece of the software performs as it was intended.
- Method: Implement tests for individual functions and methods, ensuring they return expected outputs for given inputs.
- Areas Covered: Image processing, human detection algorithms, action recognition methods, and database operations.

2. Integration Testing:

- Purpose: Validate that different modules or services of the application work together without any issues.
- Method: Test the interfaces between the classes in our class diagram, such as how the Drone class integrates with the ImageProcessor, or how the HumanDetectionModel class works in tandem with the ActionRecognition class.

3. System Testing:

- Purpose: Ensure that the complete system, as a whole, functions as intended.
- Method: Simulate real-world scenarios where the drone captures images, processes them, and visualises results to ensure all components interact seamlessly.

4. Performance Testing:

- Purpose: Ensure that the system performs effectively under load.
- Method: Simulate scenarios with large datasets, or rapid consecutive image captures, to see how the software responds.

5. Security Testing:

- Purpose: Validate that our system is protected against common vulnerabilities and threats.
- Method: Perform vulnerability scans and penetration testing, especially on the DatabaseManager class, to ensure data integrity and confidentiality.

6. Usability Testing:

- Purpose: Ensure that the VisualizationTool offers an intuitive user experience.
- Method: Solicit feedback from a group of users who will interact with the VisualizationTool, making note of any areas of confusion or difficulty.

Testing Schedule & Resources:

- Phase 1 - Development: 2 months.
- Phase 2 - Unit and Integration Testing: 3 weeks.
- Phase 3 - System, Performance, and Security Testing: 4 weeks.
- Phase 4 - Usability Testing: 2 weeks. Will require a group of potential users and supervisor to oversee the sessions.

Required Testing Resources:

Time:	Approximately 3.5 months in total.
People:	Software developers, test engineers, cybersecurity experts, and a group of potential users.
Technical Resources:	Servers for data storage and processing, multiple drone units for simultaneous testing, testing software tools, and cybersecurity tools for vulnerability and penetration testing.

Conclusion

In conclusion, the primary objective of this project is to take advantage of drone technology to detect humans and detect their actions. To enhance its usability, user interfaces were developed to act as a command centre for users to interact with the drone and analyse the performance and the efficiency of the machine learning model.

Object detection model plays a vital role within the human action detection system, as it serves a fundamental purpose of the action recognition system to locate the human in the frame of the video feed. Without this element, the action detection model will increase the chance of misclassifying unrelated object detection performing an action. Therefore, a quick and highly efficient object detection algorithm is pivotal for real-time detection. In this context, You Only Look Once (YOLO) models images as the best optimal model for the task. YOLO stood out as the fastest algorithm that demands the least computational power and provides real time detection. Although its accuracy is not as good as Fast R-CNN, its speed compensates for it. This efficiency allows human detection with rapid and accurate succession. Thus, it is a prime candidate for the initial detection task for the action detection model.

Human action recognition has always been a standard problem in computer-vision based applications and breakthroughs in deep learning techniques such as convolutional neural networks (CNN) have made it possible to tackle this issue with incredibly high accuracy. In our literature review, the Temporal Shift Module (TSM) with the backbone of a 2D Residual Network (ResNet) have been proven to be the most suitable model for our project owing to its computational efficiency, accuracy and its ability to perform in an online setting which allows for accurate predictions of actions from live drone videos.

References

Ahmad, T., Cavazza, M., Matsuo, Y., & Prendinger, H. (2022). Detecting human actions in drone images using Yolov5 and stochastic gradient boosting. *Sensors*, *22*(18), 7020. https://doi.org/10.3390/s22187020

Archana, N., & Hareesh, K. (2021). Real-time human activity recognition using ResNet and 3d Convolutional Neural Networks. 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS). https://doi.org/10.1109/access51619.2021.9563316

Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.502

Huang, M., Qian, H., Han, Y., & Xiang, W. (2021). R(2+1)d-based two-stream CNN for Human Activities Recognition in videos. 2021 40th Chinese Control Conference (CCC). https://doi.org/10.23919/ccc52363.2021.9549432

Jiang, H., & Learned-Miller, E. (2017). Face detection with the faster R-CNN. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. https://doi.org/10.1109/fg.2017.82

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of Yolo algorithm developments. *Procedia Computer Science*, *199*, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135

Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., & Li, Z. (2021). UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr46437.2021.01600

Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal shift module for efficient video understanding. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2019.00718

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *Computer Vision – ECCV 2016*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Samma, H., & Sama, A. S. (2023). Optimized deep learning vision system for human action recognition from drone images. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-15930-9

Selvaraj, J., & Anuradha, J. (2022). Violence detection in video footages using i3d convnet. Advances in Intelligent Systems and Computing, 63–75. https://doi.org/10.1007/978-981-19-0475-2_6

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2018.00675

Valarmathi, B., Kshitij, J., Dimple, R., Srinivasa Gupta, N., Harold Robinson, Y., Arulkumaran, G., & Mulu, T. (2023). Human detection and action recognition for search and rescue in disasters using yolov3 algorithm. *Journal of Electrical and Computer Engineering*, *2023*, 1–19. https://doi.org/10.1155/2023/5419384

Appendix

A1 Requirement Traceability Matrix

Req. ID	Description	Category	Type	Source
R1	Real-time human detection	Functionality	FR	Project Scope
R2	Real-time human action recognition	Functionality	FR	Project Scope
R3	Human object datasets to train and test human detection model	Data	FR	Project Scope
R4	Human action datasets to train and test action recognition model	Data	FR	Project Scope
R5	User interface to view drone footage and review algorithms output	UI	FR	Stakeholders
R6	User interface must be simple and user friendly	UI	NFR	Stakeholders
R7	Data must be stored in a secure environment	Security	NFR	Stakeholders
R8	Drone should execute specific actions when certain actions are detected	Functionality	FR	Project Scope
R9	Adhere to regulatory bodies regulations	Compliance	NFR	Regulatory Bodies
R10	The human detection model is accurate	Functionality	FR	Stakeholders
R11	The human action recognition model is accurate	Functionality	FR	Stakeholders

(FR = Functional Requirement, NFR = Non-Functional Requirement)

A2 User Acceptance Criteria

Req. ID	User Acceptance Criteria
R1	➔ The model must be able to detect objects in the video ➔ The model must be able to classify human from non-human objects among the detected objects
R2	➔ The model must be able to detect actions from the detected human in the video ➔ The model must be able to classify actions into 155 different action classes (as specified in the UAV-Human Dataset).
R3	➔ The datasets must encompass diverse human demographics, clothing and environment ➔ Datasets must be legally obtained and comply with regulatory laws
R4	➔ The datasets must encompass diverse human demographics, clothing and environment ➔ The datasets must be classified into at least 100 classes of actions. ➔ Datasets must be legally obtained and comply with regulatory laws
R5	➔ User interface must display the live video feed captured by the drone ➔ User interface has to be integrated with video analysis algorithms to display the output or prediction of said algorithms
R6	➔ User interface should not be cluttered ➔ User interface should be easy to use for end-users with the assumption of having no IT experience.
R7	➔ Data should be encrypted with industry-standard encryption protocols to ensure users' private information is secure. ➔ Access to the input data and prediction results should be restricted ➔ Data storage must be compliance to data privacy laws ➔ Data must be stored with redundancy in mind in the case of disasters
R8	➔ Drone should be able to executed actions when triggered by certain actions from the human ➔ The maximum accepted latency is 2 seconds.
R9	➔ No laws were infringed during the course of the development and deployment of the project
R10	➔ The Top-1 accuracy of the model should be at least 70% and Top-5 accuracy should be at least 90%. ➔ The false positive rate should be below 10%.
R11	➔ The Top-1 accuracy of the model should be at least 70% and Top-5 accuracy should be at least 90%. ➔ The false positive rate should be below 10%.

A3 Risk Register

No.	Risk	Likeli-hood (1-5)	Impact (1-5)	Mitigation Action	Contingency Action
1	**Repository server down** Team member unable to access repository and development will be halted	1	5	Ensure a local up to date copy of the source codes are always available.	Use an alternative software version control system as a new repository to replace the current software version control system provider.
2	**Miscalculated project schedule** In planning the project schedule, we might underestimate or overestimate complex tasks' duration required, which might set unrealistic time goals for team members	3	3	Team members are to share their progress frequently and complex tasks are schedule apart from each other	Make changes to the project schedule and assign more team members on certain tasks to go back on track of the initial scheduling plan and inform stakeholders.
3	**Natural Disaster** Natural disasters such as earthquakes and floods that lead to a complete halt of development due to loss of resources.	1	4	Periodically backup data offsite in the event of data loss to natural disaster.	Notify stakeholders and follow local authority health and safety protocols.
4	**Missing Team Members** The loss of contact of a team member causes work congestion due to dependency on their work.	2	3	Minimise work dependencies and attempt to contact team members as soon as they are missing from any meeting.	Reallocate their responsibilities to other team members and investigate the issue.

5	**Miscommunication** Miscommunication might occur that causes work to progress incorrectly or tasks to be incomplete.	4	2	Ensure that all meetings are documented, and important messages must be passed via writing instead of verbally.	Correct any mistakes that have been made and notify stakeholders if there is any delay to the project schedule.
6	**Error in project design** Mistakes in the design of the project that appear to be impossible or invalid.	2	4	Provide alternative options in the design of the project and backup plans in the case where the design is infeasible	Adopt alternative designs prepared and notify stakeholders about potential project schedule delays.
7	**Health / Epidemic** A rise of a new epidemic/pandemic that halts development.	1	4	Attempt to move as much data as possible to the cloud to allow for flexible working spaces and working areas are to be hygienic to prevent the spread of diseases.	Allow flexible working spaces and switch to alternative hardware solutions such as personal device cameras in the case of the lack of drones for development as hardware could not be shared.
8	**Unable to contact stakeholders** Stakeholders might be busy with other issues in their hand which might lead to negligence to this project	2	3	Conduct meetings periodically with stakeholders on a set frequency to ensure that stakeholders are up to date.	The team works independently for the period of time where stakeholders are unavailable and urgent concerns are to be decided by the Project Manager and other concerns are to be delayed to the next meeting with stakeholders.
9	**Inability to procure drone or drone malfunction**	2	4	Research the availability of the drone to be purchased beforehand and	Procure a different drone with similar internal and external specifications if availability is sparse.

	The availability of the drone in Malaysia might be sparse or the possibility of the drone malfunctioning which leads to difficulties in testing our application			ensure at least 2 drones are procured to ensure that there are replacements.	We may also choose to contact our Project Supervisor to check whether the drones in the Monash University CyPhi AI Lab is available for use.

A4 Stakeholders' Communication Plan

No.	Stakeholder	Reporting mechanism	Report format	Frequency
1	Supervisor	Email & online meeting	Detailed reports	Every major milestone of the project
2	Finance	Online forms	Receipts, transaction slips, statements, etc.	As required
3	Regulatory Authorities	Compliance reports	Regulatory-specific format	As required
4	Data Providers	Data handling reports	Data privacy assessment	As required

A5 Team Communication Plan

No.	Roles	Reporting mechanism	Notes	Frequency
1	Project Manager	Online meeting, in person meeting	Review project status, key issues, deliverables	Bi-weekly
2	Technical Lead	Online meeting, in person meeting	Review task progress, delegate tasks, raise technical concerns	Bi-weekly
3	Quality Assurance	Online meeting, in person meeting	Review deliverables quality standard	Bi-weekly
4	Team members	In person communication, online communication (social messaging applications)	Updates, queries, notifications	Daily

A6 Project Schedule - Gantt Chart

Project Planner - MCS14

Select a period to highlight at right. A legend describing the charting follows.

Period Highlight: 14 | Plan Duration | Actual Start | % Complete | Actual (beyond plan) | % Complete (beyond plan)

	ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
1	Conduct Technology Assessment.	1	2	1	2	100%
2	Elicit project requirements from project supervisor.	3	3	3	3	100%
3	Conduct in-depth research on the project scope.	3	3	3	3	100%
4	Initiate research into concepts and design for DroneAI project.	6	3	6	3	100%
5	Identify potential design patterns and methodologies.	6	3	6	3	100%
6	Synthesize findings from the high-level design phase of the DroneAI project.	9	2	9	2	100%
7	Sequence WBS tasks, establish a timeline, prepare a literature review from prior research, and develop a comprehensive project proposal with timeline, budget, and goals.	11	3	11	3	100%
8	Procure drone(s) for human/facial analysis.	14	3	14	3	0%
9	Familiarize with state-of-the-art computer vision algorithms.	14	3	14	3	0%
10	Develop advanced techniques to enhance baseline results from standard computer vision algorithms based on predefined metrics.	17	4	17	4	0%
11	Test the drone-based prototype on different datasets and/or real-world settings.	21	3	21	3	0%
12	Evaluate the drone prototype's performance and prepare for its final demo, including relevant documentation.	24	3	24	3	0%