

**VIDEO SURVEILLANCE: FRONT-YARD MONITORING**

**BY**

**BRYAN LOW KENG SEONG**

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

**BACHELOR OF COMPUTER SCIENCE (HONOURS)**

Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2023

## REPORT STATUS DECLARATION FORM

**Title:** VIDEO SURVEILLANCE: FRONT-YARD MONITORING

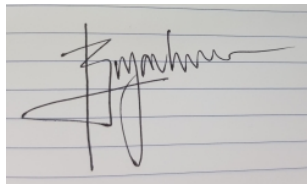
**Academic Session:** Jan 2023

I  
BRYAN LOW KENG SEONG  
(CAPITAL LETTER)

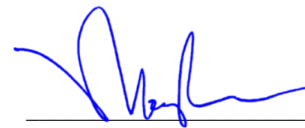
declare that I allow this Final Year Project Report to be kept in  
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

**Address:**

361102, Kemuncak Condo,  
Section 9, Shah Alam, Selangor

Prof. Dr. Leung Kar Hang  
Supervisor's name

**Date:** 19/04/2023

**Date:** 23 April 2023

<b>Universiti Tunku Abdul Rahman</b>			
Form Title : <b>Sample of Submission Sheet for FYP/Dissertation/Thesis</b>			
Form Number: <b>FM-IAD-004</b>	Rev No.: <b>0</b>	Effective Date: <b>21 JUNE 2011</b>	Page No.: <b>1 of 1</b>

**FACULTY/INSTITUTE\* OF INFORMATION AND COMMUNICATION TECHNOLOGY**  
**UNIVERSITI TUNKU ABDUL RAHMAN**

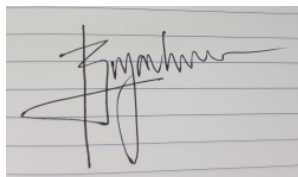
Date: 19/04/2023

**SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS**

It is hereby certified that Bryan Low Keng Seong (ID No: 20ACB01314 ) has completed this final year project/ dissertation/ thesis\* entitled “ \_Video Surveillance: Front-Yard Monitoring\_” under the supervision of Prof. Dr. Leung Kar Hang (Supervisor) from the Department of FICT, Faculty/Institute\* of Information and Communication Technology, and Dr. Ng Hui Fuang (Co-Supervisor)\* from the Department of FICT, Faculty/Institute\* of Information and Communication Technology.

I understand that University will upload softcopy of my final year project / dissertation/ thesis\* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

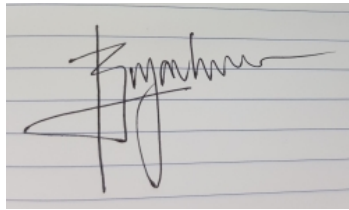
Yours truly,



(Bryan Low Keng Seong)

## DECLARATION OF ORIGINALITY

I declare that this report entitled “**VIDEO SURVEILLANCE: FRONT-YARD MONITORING**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

A handwritten signature in black ink on a light blue lined background. The signature is stylized and appears to read 'Bryan Low Keng Seong'.

Signature :

---

Name : Bryan Low Keng Seong

Date : 19/04/2023

## **ACKNOWLEDGEMENTS**

I sincerely thank and appreciate my supervisor, Prof. Dr. Leung Kar Hang. He let me engage in an exciting computer vision project that has the potential to contribute back to society. A million thanks to you for your patient guidance during the meetings.

Moreover, I must thank my family and friends for their love and continuous encouragement throughout the course. Their support has motivated me always to push forward.

# ABSTRACT

Home intrusion is a severe crime that disrupts the safety and well-being of neighbourhoods. It has been happening at a shocking rate in recent years. Families have tried curbing the issue by implementing CCTVs and motion sensors. However, these approaches have considerable flaws, including over-relying on human supervision and high false positive alarms.

With the latest technological advancements, an approach with computer vision techniques is proposed to aid crime identification. This project aims to deliver an automated front-yard intrusion system. It addresses the significant issues of CCTVs and motion sensors by removing the need for manual video monitoring and reducing false positive cases.

The proposed intelligent surveillance system is expected to have two vital functionalities. The first feature is to give a mild warning if a person has stepped foot into the front yard or the restricted area. Here, the YOLO algorithm will be used for human detection.

When there is a human presence, the second stage checks for excessive motions resembling violence with dense optical flow. The model will be evaluated with several video datasets containing intrusions and violence. It will trigger an alarm when violent activities such as fighting or snatching theft happen.

The system's primary purpose is to reduce victims' potential loss and damage by providing immediate notifications on intrusion without delay.

# TABLE OF CONTENTS

<b>TITLE PAGE</b>	<b>i</b>
<b>REPORT STATUS DECLARATION FORM</b>	<b>ii</b>
<b>FYP THESIS SUBMISSION FORM</b>	<b>iii</b>
<b>DECLARATION OF ORIGINALITY</b>	<b>iv</b>
<b>ACKNOWLEDGEMENTS</b>	<b>v</b>
<b>ABSTRACT</b>	<b>vi</b>
<b>TABLE OF CONTENTS</b>	<b>vii</b>
<b>LIST OF FIGURES</b>	<b>x</b>
<b>LIST OF TABLES</b>	<b>xi</b>
<b>LIST OF SYMBOLS</b>	<b>xii</b>
<b>LIST OF ABBREVIATIONS</b>	<b>xiii</b>
<b>CHAPTER 1 INTRODUCTION</b>	<b>1</b>
1.1 Project Background	1
1.2 Problem Statement	2
1.3 Motivation	3
1.4 Objectives	4
1.5 Project Scope and Direction	4
1.6 Contribution	5
1.7 Report Organisation	5
<b>CHAPTER 2 LITERATURE REVIEW</b>	<b>7</b>
2.1 Previous Works on Abnormal Activity Detection Systems and Human Action Recognition	7
2.1.1 Real-world Anomaly Detection in Surveillance Videos [7]	7
2.1.2 Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviours in Shopping Malls [8]	9
2.1.3 YOLO-based Human Action Recognition and Localization [9]	13

2.1.4	Detection of Abnormal Events via Optical Flow Feature Analysis [10]	16
2.1.5	Enhanced skeleton visualisation for view invariant human action recognition [11]	18
2.1.6	Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features [12]	24
<b>CHAPTER 3 SYSTEM METHODOLOGY/APPROACH</b>		<b>28</b>
3.1	Methodologies and General Work Procedure	28
3.1.1	First-level security	28
3.1.2	Second-level security	29
3.2	Assumptions	30
<b>CHAPTER 4 SYSTEM DESIGN</b>		<b>31</b>
4.1	System Flow Diagram	32
4.2	System Components Specifications:	32
4.2.1	Initialisation and Pre-processing	33
4.2.2	Human and intrusion detection	34
4.2.3	Excessive motion and violence detection	
<b>CHAPTER 5 SYSTEM IMPLEMENTATION</b>		<b>36</b>
5.1	Hardware Setup	36
5.2	Software Setup	36
5.3	System Operation	37
5.3.1	Defining the region of interest	37
5.3.2	Analysis board	38
5.3.3	Monitoring board initialisation	39
<b>CHAPTER 6 SYSTEM EVALUATION AND DISCUSSION</b>		<b>41</b>
6.1	System Testing and Results	41
6.1.1	Scenario A	41
6.1.2	Scenario B	41
6.1.3	Scenario C	42



6.2 Project challenges	44
6.3 Objectives evaluation	44
<b>CHAPTER 7 CONCLUSION AND RECOMMENDATION</b>	<b>46</b>
7.1 Conclusion	46
7.2 Recommendation	46
<b>REFERENCES</b>	<b>47</b>
<b>WEEKLY LOG</b>	<b>50</b>
<b>POSTER</b>	<b>56</b>
<b>PLAGIARISM CHECK RESULT</b>	<b>57</b>
<b>FYP2 CHECKLIST</b>	<b>60</b>

## LIST OF FIGURES

<b>Figure Number</b>	<b>Title</b>	<b>Page</b>
Figure 1.1	Potential signs of front-yard intrusion	4
Figure 2.1	Flow diagram of the proposed anomaly detection approach [7]	7
Figure 2.2	Evolution of score on training video over iterations [7]	8
Figure 2.3	Video pre-processing and human tracking [8]	9
Figure 2.4	Graphical description of the proposed occlusion management method based on visual appearance [8]	10
Figure 2.5	The angular model proposed for the entry and exit trajectories if the entrance line is oriented at 60 degrees [8]	11
Figure 2.6	A real example of an alarm given off by detecting entry and exit events [8]	11
Figure 2.7	A real example of an alarm given off by detecting a loitering event [8]	12
Figure 2.8	Workflow of the algorithm [9]	14
Figure 2.9	The architecture of YOLO [9]	14
Figure 2.10	Recognition and localisation of actions in single frames from the LIRIS dataset [9]	15
Figure 2.11	Normal (a,c) and abnormal frames (b,d) [10]	16
Figure 2.12	Histogram of optical flow orientation (HOFO) feature descriptor on the original or foreground image [10]	17
Figure 2.13	The flowchart of the proposed feature classification-based abnormal detection method [10]	18
Figure 2.14	The pipeline of the proposed method [11]	19
Figure 2.15	Configuration of body joints in the NTU RGB+D dataset [11]	20
Figure 2.16	Illustration of proposed invariant transform [11]	21

Figure 2.17	Comparison between proposed sequence-based transformation and traditional skeleton-based transformation [11]	22
Figure 2.18	Illustration of weighing skeleton joints according to motion energy [11]	23
Figure 2.19	The framework of the proposed DB-LSTM for action recognition. [12]	24
Figure 2.20	Frame-to-frame features representation and changes in a sequence of frames [12]	25
Figure 2.21	The external structure of the proposed DB-LSTM network [12]	26
Figure 3.1	Project Methodology	28
Figure 3.2	Potential intruder present within the restricted area	29
Figure 3.3	Excessive action between humans	30
Figure 4.1	System Flow Diagram	31
Figure 4.2	System initialisation	32
Figure 4.3	Human and intrusion detection	33
Figure 4.4	Excessive motion and violence detection	34
Figure 5.1	Before click on points	37
Figure 5.2	After click on points	37
Figure 5.3	System connects the points and forms a polygon	38
Figure 5.4	Different risk levels of the analysis board	38
Figure 5.5	Initialisation of the whole monitoring board	39
Figure 6.1	System ignoring humans outside the gate	41
Figure 6.2	Intruder jumps over the gate	41
Figure 6.3	Different monitoring stages during a fight	42

## LIST OF TABLES

<b>Table Number</b>	<b>Title</b>	<b>Page</b>
Table 2.1	Results obtained for alarm detection in shopping malls [8]	13
Table 2.2	The architecture of the pre-trained AlexNet model [12]	25
Table 3.1	Comparison of average recognition score of the proposed DB-LSTM for action recognition with state-of-the-art methods [12]	27
Table 5.1	Hardware setup	36
Table 5.2	Software setup	36

## LIST OF ABBREVIATIONS

<i>AUC</i>	Area under the ROC Curve
<i>C3D</i>	Convolutional 3D
<i>CCTV</i>	Closed-circuit television
<i>CNN</i>	Convolutional Neural Networks
<i>GCH</i>	Glocal Colour Histogram
<i>HOFO</i>	Histogram of Optical Flow Orientation
<i>HOG</i>	Histogram of Oriented Gradients
<i>LBP</i>	Local Binary Pattern
<i>LSAP</i>	Linear Sum Assignment Problem
<i>LSTM</i>	Long short-term memory
<i>MIL</i>	Multiple Instance Learning
<i>MT</i>	Mostly tracked
<i>PCA</i>	Principal Component Analysis
<i>ReLU</i>	Rectified Linear Unit
<i>RNN</i>	Recurrent Neural Networks
<i>ROI</i>	Region of Interest
<i>YOLO</i>	You Only Look Once

# Chapter 1

## Introduction

### 1.1 Project Background

Front-yard intrusions or break-ins disrupt the psychological well-being and overall peace of household residents and neighborhoods. Department of Statistics Malaysia [1] shows that Malaysia's "House break-in & theft" crime type within 2019 and 2020 have frequencies of 16,497 and 14,040 cases, respectively. Although there was a minor reduction in the cases, the number is still alarming.

Due to the breakthrough in deep learning and computational advancements in the last decade, it is feasible and ideal to tackle this issue with the assistance of artificial intelligence. Two main areas this project touches on are computer vision and its subset, image processing. These areas of study aid the creation of an intelligent detection system. The system will use human action recognition techniques to identify the abnormal activities involved in intrusions or break-ins.

Gonzalez and Woods [2] explain that digital image processing involves manipulating digital images with the help of computers and can have three types of computerised processes: low-, mid-, and high-level strategies. Firstly, low-level operations deal with primitive functions such as noise reduction, image sharpening, and contrast enhancement. Secondly, mid-level operations deal with tasks like segmentation and classification within images. Thirdly, high-level operations deal with image analysis, where the algorithm tries to make sense of the recognised objects.

On the other hand, Gonzalez and Woods [2] also explain that computer vision is a branch of artificial intelligence that emulates human vision. It includes learning, making inferences, and taking actions based on visual inputs.

## 1.2 Problem Statement

One of the well-known methods of anomalous action detection is video surveillance, widely known as CCTV (closed-circuit television). Video surveillance is monitoring a particular region while looking out for abnormal behaviours of potential offenders. In our case, surveillance of front yards is the primary concern.

The number of surveillance cameras is increasing day by day. According to Keval and Sasse [3], there is an issue of a high camera-to-operator ratio, whereby operators are overwhelmed by the number of cameras they must monitor. Furthermore, Gill M. et al. [4] have shown that there are not enough available screens to display all footage simultaneously. Therefore, operators must switch between footages from time to time, making it prone to missing out on some instances.

In addition, Renata and Philippe [5] and Temple et al. [6] have witnessed a phenomenon called “vigilance decrement,” whereby when guards require long periods of attention and when critical events occur very rarely, the guard’s performance in detecting crime drops and become unreliable. Temple et al. [6] have also found that detection performances can drop after 10 to 12 minutes of monitoring. Besides, the rarity of abnormal occurrences makes the supervision task more challenging.

In this case of front-yard intrusions, neglecting even a single occurrence is not tolerable as it can lead to severe financial losses or injuries.

Another popular solution is motion sensor installation. Several of these embedded systems can be installed throughout the corners of the front yard to detect and measure movements. These sensors give off alarms upon detecting motion.

Although motion sensors have a high recall in picking up most of the motions, their precision may not be reliable. For example, a misplaced sensor might pick up motion signals outside the front-yard gate, which is none of its concern. For instance, the sensor could trigger a false alarm if a food deliverer steps near the front entrance. On the other hand, they may incorrectly flag animal movements, such as stray cats ‘visiting’ the house.

Over and above, there is a lack of anomaly surveillance projects that target front-yard intrusion detection, which is necessary to provide improved protection to homeowners.

To conclude, an approach that minimally relies on humans and has a lower false positive rate is needed to tackle front-yard intrusions. It must capture most illegal attempts, not miss out on them, and be sure about an event before raising the alarm.

### **1.3 Motivation**

An intelligent surveillance system is a reliable solution to addressing the issues in CCTVs and motion sensors. It is essentially a traditional video surveillance system with added intelligent features. These features reduce human supervision as the system can automatically detect criminal activities and instantly give a warning. Therefore, it should be able to capture more cases as human resource is not a limitation compared to CCTVs.

Besides, the intelligent system will have a lower false positive rate than motion sensors. Its innovative features will check for several conditions before triggering an alarm. To elaborate, it will not randomly trigger a warning because of any movement. The system will check the overall context and identify if the action is related to criminal intentions. Also, it will determine whether a human being is performing it. Additionally, the system will focus only on the front yard and ignore whatever movements are happening outside the front gate, significantly reducing false positive rates.

### **1.4 Objectives**

The intelligent surveillance system aims to address the problems faced by traditional CCTVs, mainly over-relying on human resources. Also, to address the high false alarm rates faced by motion sensors. Moreover, this intelligent system reduces the delay of alarms by raising alarms immediately when specific suspicious criteria occur.

The project objectives are as follows:

#### **1. To develop a reliable automated front-yard monitoring system**

As discussed in the problem statement, conventional CCTVs over-rely on human resources. Fatigue and loss of focus of security staff lead to missing out on certain break-ins, which is unacceptable. Automation of restricted area monitoring reduces the required human labour and ensures immediate intrusion notification.



## 2. To develop a two-level intruder detection feature by meeting two sub-tasks:

### i. To detect humans within the front yard

Users can manually input the polygon of the front yard (region of interest) before the monitoring starts. The system should give the first-level notice when it detects humans in the specified restricted area. Ignoring humans outside the gate reduces false positive warnings.

### ii. To detect violent or excessive actions

After the first-level notice, the system checks the human count and looks for criminal activities. Some violent acts include fighting or snatching theft. Video frames are passed to a model to classify them into violent or non-violent classes. The system triggers an alarm when it detects a sequence of frames that indicate violent actions.

## 1.5 Project Scope and Direction

The project scope covers the automation of video monitoring to detect front-yard intrusions. Here, the proposed surveillance system needs to notice human presence within the restricted area, as shown in Figure 1.1 (a), and to look out for violent actions, as shown in Figure 1.1 (b).



(a) Human enters restricted area



(b) Humans showing excessive motion

Figure 1.1 Potential signs of front-yard intrusion

For violent actions, the system will focus on those that are clear and not too far away from the camera or footage.

The UCF crime dataset from Sultani et al. [7], specifically the fighting category, is used while evaluating the model's ability to detect violence.

## **1.6 Contributions**

Our primary target is to reduce the occurrence of front-yard intrusions, which leads to fewer casualties and financial losses in the neighbourhood—ultimately improving residents' safety and living standards.

Homeowners should receive a light notification whenever a person enters the front yard. Then, when the system recognises violent actions, such as fighting or snatching theft, an alarm should go off immediately. It may cause the intruder to flee from the scene without further property damage or loss. At the same time, an emergency signal could be sent to the nearest police station.

One good thing is that it requires fewer human resources. With the help of computer vision, it can cut down on human supervision and monitoring screens with CCTVs as the system can automatically detect abnormal scenes and warn us about illegal intrusions. It significantly reduces human error concerns such as inattention or fatigue.

In addition, due to the condition checking within the system, whenever the intelligent system gives off an alarm, there is a high probability that an intrusion has indeed taken place, and it is not just some random movement in the front yard.

## **1.7 Report Organisation**

The report is organised into seven chapters: Introduction, Literature Review, System Methodology/Approach, System Design, System Implementation, System Evaluation and Discussion, Conclusion and Recommendation.

The first chapter covers the project overview, including the background, problem statement, motivation, objectives, project scope and direction, contribution, and report organisation. Chapter 2 will cover the literature review of papers related to the project, mainly on human action recognition and abnormal activity detection systems. Moreover, the third chapter includes the system methodology/approach, whereby the methodologies and general work procedures will be discussed.

Next, Chapter 4 includes the system design, containing the system flow diagram and block diagrams of different system components. Subsequently, chapter 5 goes through the system implementation, regarding how certain parts of the system are implemented. Then, the sixth chapter shows the experimental results of different scenarios. Lastly, chapter 7 will have the conclusion and recommendation.

# Chapter 2

## Literature Review

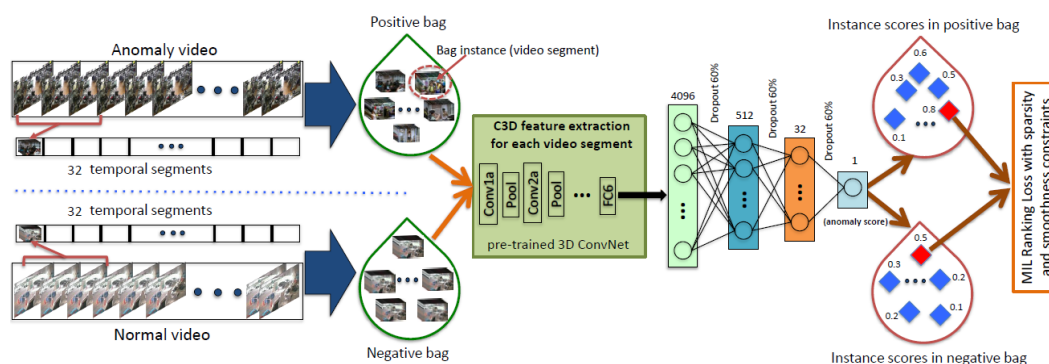
### 2.1 Previous Works on Abnormal Activity Detection Systems and Human Action Recognition

#### 2.1.1 Real-world Anomaly Detection in Surveillance Videos

Sultani et al. [7] proposed a deep learning approach to detect 13 abnormal activities, including fighting, road accident, burglary, robbery, etc., and normal activities. Their system utilises both normal and abnormal videos in the learning process. Depending on different conditions, it helps resolve the ambiguity when an action could be normal or abnormal.

Furthermore, they leveraged weakly labelled training videos for convenient annotation and time efficiency. The training labels (anomalous or normal) are assigned to videos as a whole instead of individual video segments.

Figure 2.1 shows the training process. It implements Multiple Instance Learning (MIL) in differentiating anomalous and normal classes



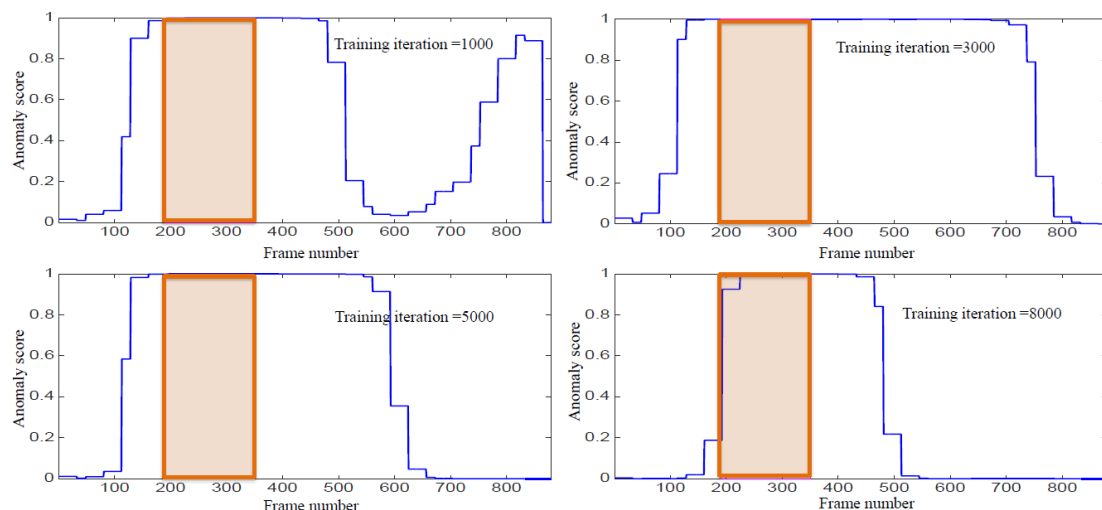
**Figure 2.1** Flow diagram of the proposed anomaly detection approach. [7]

Here, videos are represented as either positive (for anomalous videos) or negative bags (for normal videos). These bags contain each video's temporal segments, represented as instances.

Afterwards, the bags of instances go through a pre-trained 3D Convolutional Neural Network, where the model extracts C3D features from the video segments. MIL learns a deep anomaly ranking model throughout this deep learning process, which computes the ranking loss between the highest-scored instances (shown in red) in the positive and negative bags.

To explain, the segments in positive bags (anomalous videos) with higher anomaly scores most likely contain abnormal activities. On the other hand, segments in negative bags (normal videos) with higher anomaly scores are false positive examples. These false positive examples aid in understanding and resolving some ambiguity.

Figure 2.2 depicts the performance of the model over the training iterations. The coloured windows represent anomalous regions. As the iteration increases, differentiating between abnormal and normal actions improves. As seen in the 8000<sup>th</sup> iteration, abnormal frames have higher anomaly scores, while other normal frames have lower ones.



**Figure 2.2 Evolution of score on training video over iterations [7]**

Given enough positive and negative videos with video-level labels, the network can better understand frames and automatically localise anomalies. A higher number of iterations reflects higher precisions in localising anomalies.

This approach has been compared to two other state-of-the-art methods for the evaluation process. Results show that this approach has the highest AUC of 75.41. The dictionary-based and deep auto-encoder techniques had AUC results of 50.6 and 65.51, respectively.

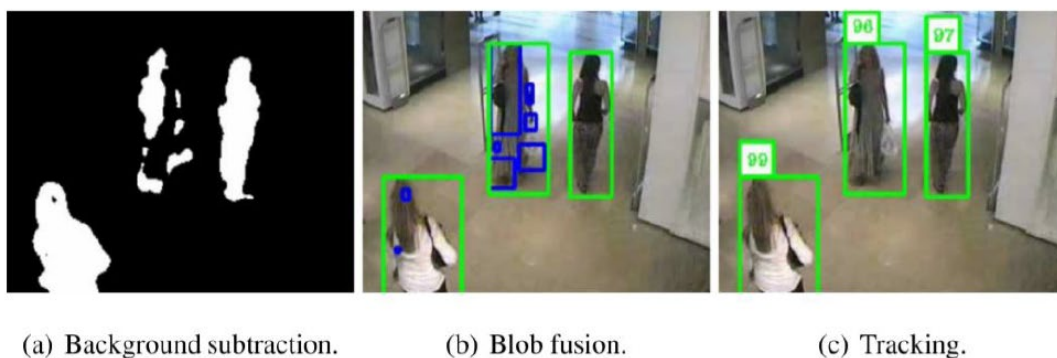
Moreover, this approach also has a lower false alarm rate than other approaches, indicating robustness, possibly due to utilising both normal and abnormal videos and having challenging training datasets that were long and untrimmed with intra-class differences. The proposed method had a false alarm rate value of 1.9. The dictionary-based and deep auto-encoder approaches had false alarm rates of 3.1 and 27.2, respectively.

However, three weaknesses exist in the proposed system. Firstly, it cannot detect anomalous scenes in the dark during nighttime. Second, some false alarms are caused by occlusions by flying insects in front of the camera and even scenarios where people suddenly gather around. Therefore, occlusion management and intensity enhancement techniques can be looked into to tackle these weaknesses and improve robustness.

### 2.1.2 Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviours in Shopping Malls

Arroyo et al. [8] have proposed an expert system for detecting potentially suspicious human behaviours in shopping malls. Human behaviours include shop entry or exit of people, loitering, and unattended cash desk situations.

Figure 2.3 illustrates the initial process of locating the people and following them. Background subtraction techniques are first used to perform image segmentation. During this process, segmentation errors cause the appearance of blobs. Segmentation errors are significantly reduced by their novel blob fusion technique. It groups the partial blobs into final human targets. Then, object tracking was implemented using two previous works – LSAP association and Kalman filtering.



**Figure 2.3 Video pre-processing and human tracking [8]**

When two tracked objects occlude each other, it disrupts the proposed tracking algorithm, causing it to lose the tracked object's identity number. An occlusion management algorithm was submitted to mitigate this weakness of the tracking method based on LSAP association and Kalman filtering.

The occlusion management algorithm considers three appearance features: GCH, LBP, and HOG. Firstly, GCH is the color histogram of the object, and it helps differentiate people by their clothes' colour. Then, LBP is a texture descriptor that looks into a human appearance. As for HOG, it describes the shapes of people through its gradient distribution.

As shown in Figure 2.4, the system saves information on appearance features before the occlusion happens. During occlusion, the system knows it is happening, treating the man (number 1) and the woman (number 2) as a single entity. After the occlusion is over, the system will re-identify the people again by comparing the appearance features with the help of SVM kernels.

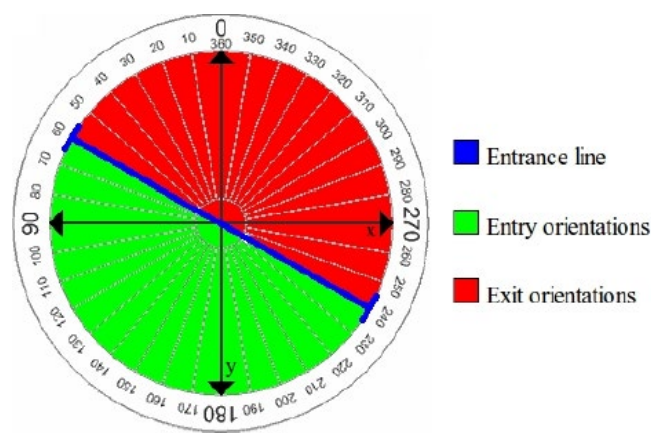


**Figure 2.4 Graphical description of the proposed occlusion management method based on visual appearance. [8]**



Entries and exits of people are used to identify crowded situations when too many people enter or to identify suspicious behaviours of potential shoplifters running away. This project uses tracked trajectories to solve these problems.

The entrance line shown in Figure 2.5 is manually marked by human operators, dividing the line between the exterior and interior of the shop. Classification is done when a tracked object passes this line. Whether or not a tracked object is classified as entering or exiting is based on the object's moving direction or trajectory degree.



**Figure 2.5 The angular model is proposed for the entry and exit trajectories if the entrance line is oriented at 60 degrees. [8]**

Image 1 in Figure 2.6 shows that a man appears but is not marked yet because he has not reached the manually set entrance line. The man gets to the entrance line in image 2 and is classified as exiting (OUT) the store as his trajectory belongs to the exit orientation shown in Figure 2.5. The following pictures depict a lady entering (IN) the store.



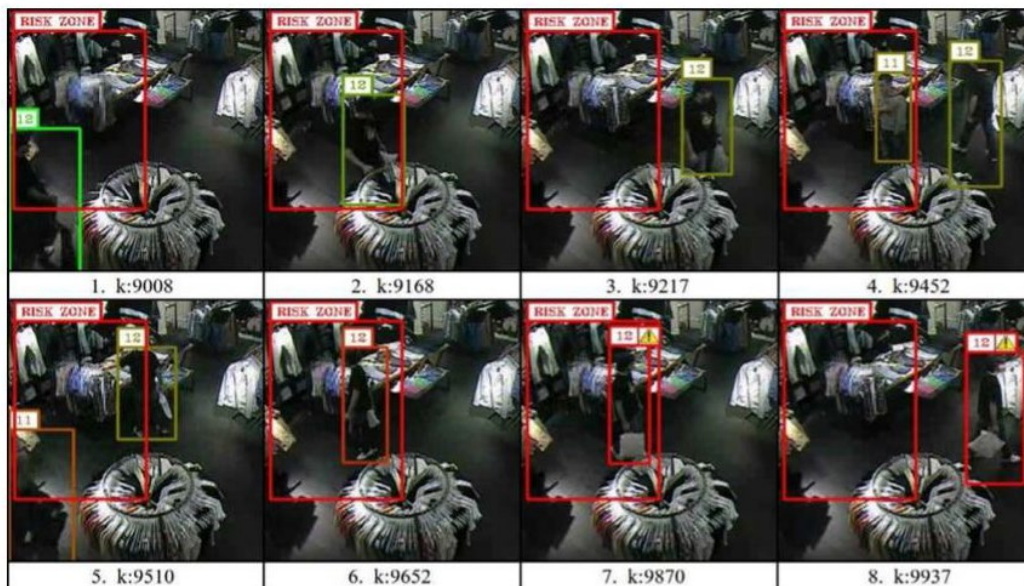
**Figure 2.6 A real example of an alarm given off by detecting entry and exit events. [8]**



Loitering is when an individual hangs around in an area for some time longer than a given time threshold.

In this approach, human security officers must manually specify risky zones (rectangular boxes) to be monitored for loitering. The system then starts to monitor the trajectories of tracked objects within the specified box, checking if they have been staying in the risk zone for over a specific time limit.

Figure 2.7 depicts various levels of suspicion with colour gradients between green and red. Image 7 in the figure shows a loitering alarm marked on person 12 due to staying in the box over the time limit. Even after person 12 leaves the risky zone, they will still be marked as a potential suspect.



**Figure 2.7** A real example of an alarm given off by detecting a loitering event. [8]

In essence, this approach reduces segmentation errors while detecting people with their blob fusion technique, improving overall tracking performances. Also, the proposed occlusion management algorithm outperformed other related works because it considers various visual appearance information.

In terms of a new tracking metric MT (mostly tracked), this approach had the highest MT of 86.7% and a remarkable processing speed of 50 fps.

Table 2.1 displays the considerable performance of the current system in giving off alarms on the desired events.

Video	Alarms				Overall
	Entry	Exit	Loitering	Unattended c. d.	
Entrance A	426/480	435/512	0/0	0/0	861/992 (86.8%)
Entrance B	76/99	41/60	0/0	0/0	117/159 (73.6%)
Interior A	0/0	0/0	22/24	0/0	22/24 (91.6%)
Interior B	0/0	0/0	17/18	0/0	17/18 (94.4%)
Cash desk	0/0	0/0	0/0	1/1	1/1 (100.0%)
<b>Overall</b>	502/579 (86.7%)	476/572 (83.2%)	39/42 (92.8%)	1/1 (100.0%)	

**Table 2.1 Results obtained for alarm detection in shopping malls [8]**

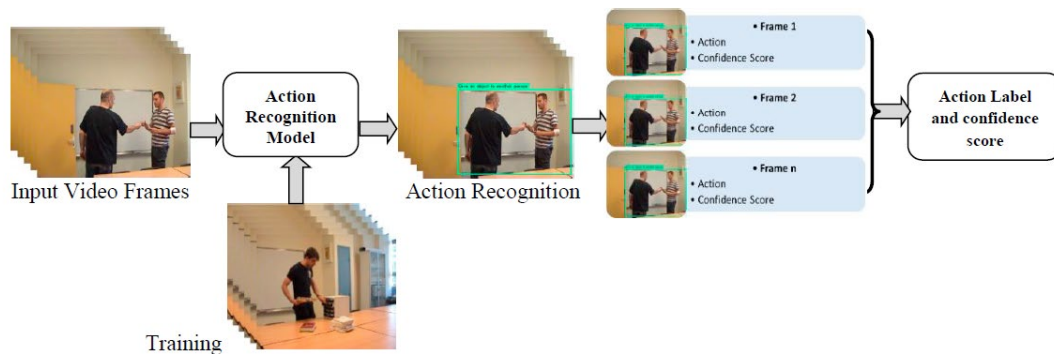
Arroyo et al. [8] pointed out that the usage of monocular vision in this project can be upgraded to stereo camera, to improve the extraction of tridimensional data. Also, the current system requires the supervision of a human operator to manage alarms when they give off—suggesting an improvement with automation. For example, suppose the surveillance system detects someone loitering in an area. In that case, security guards should be notified about the case instantly about the location via a direct message to their smartphones.

### 2.1.3 YOLO-based Human Action Recognition and Localisation

Shinde et al. [9] demonstrate that You Only Look Once (YOLO) effectively detects, localises, and recognises actions very close to real-time performances. The model takes in input frames after a period and can assign an action label only based on a single frame. It helps reduce overall computational time by utilising a minimum number of frames.

It addresses certain downsides of other techniques, including two-stream CNNs and skeleton tracking, which tend to use more information than required.

The YOLO algorithm implements the training and testing processes. Figure 2.8 shows a simple representation of the workflow. The model was trained on the LIRIS dataset of human activities, with frames containing appropriate actions. Here, 30 video frames are selected and used as input to the model.

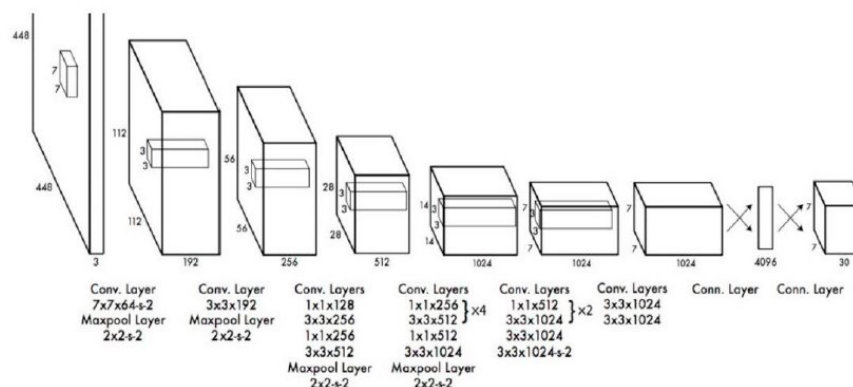


**Figure 2.8 Workflow of the algorithm [9]**

Each frame's action labels and confidence scores are stored in a CSV file during the detection phase. Then, if an action label appears above five times and has a confidence threshold of over 0.5, it will be concluded as the video's action label. Also, it calculates the average confidence score of the action label.

YOLO is an advanced object detector that applies a single CNN to an entire image divided into grids. Each image grid will predict bounding boxes and assign confidence scores. The confidence scores are used to analyse the bounding boxes.

Figure 2.9 depicts the architecture of YOLO, which has 24 convolutional layers and two fully connected layers. The leftmost panel in Figure 2.9 shows that YOLO first resizes an input image into 448×448 pixels. The image will go through multiple network layers and end up with an output of 7×7×30 tensor, shown in the rightmost panel in Figure 2.9.

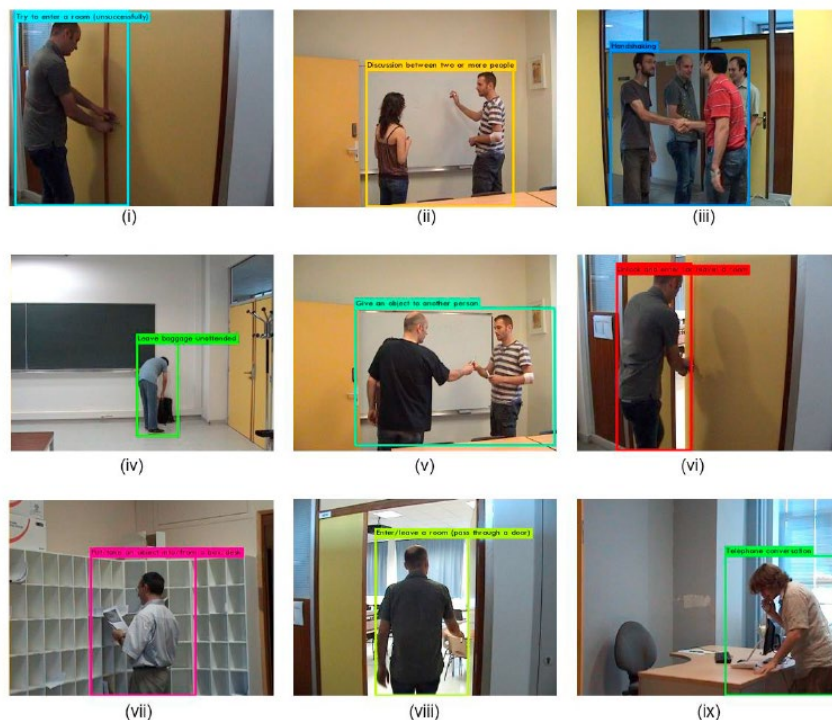


**Figure 2.9 The architecture of YOLO [9]**

The output tensor provides two crucial pieces of information. Firstly, it's the coordinates of the bounding boxes—secondly, the probability distribution over all the trained classes. As for the thresholding process, it removes class labels with confidence scores (probability) below 30% to reduce redundancy or false positives.

This approach's strengths are that it applies a single CNN for classifying and localising objects rather than separating them into two steps. Also, it can process images at 40-90 frames per second, leading to real-time video processing with negligible latency.

Overall experimental results on the LIRIS dataset have been promising and competitive against other state-of-the-art methods. Figure 2.10 provides some experimental examples. The proposed YOLO method achieved the best F-score of 88.358% on the LIRIS dataset compared to different approaches. The main upside of the system is it recognises actions with few frames, sometimes even a single frame.



**Figure 2.10 Recognition and localisation of actions in single frames from the LIRIS dataset [9]**

However, most of these datasets do not consider occluded scenarios. Therefore, some occlusion testing and management is suggested. Some future improvements mentioned are to

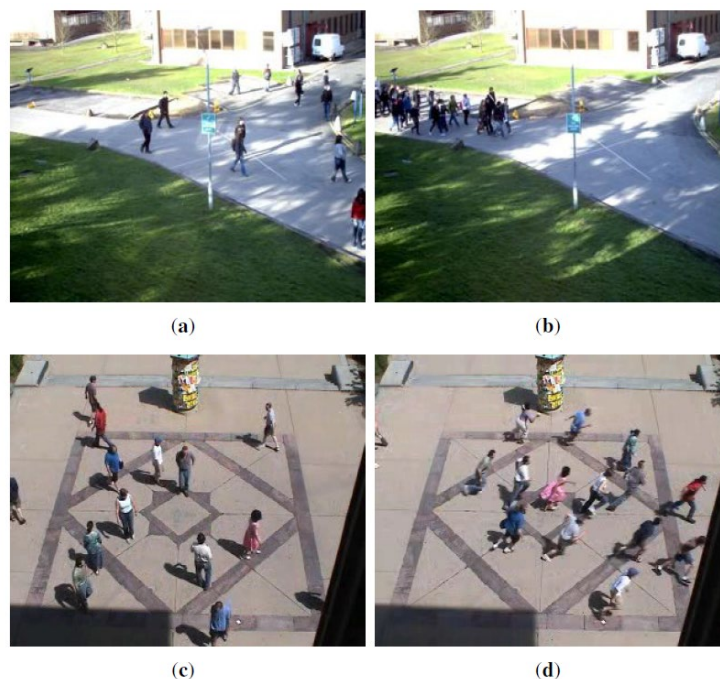
recognise more complex human actions. Also, calculating the euclidean distances between centres of moving objects and humans may provide more information about activities.

#### 2.1.4 Detection of Abnormal Events via Optical Flow Feature Analysis

Wang and Snoussi [10] have proposed an abnormal events detector in video streams based on the histogram of the optical flow orientation descriptor and classification model. These histograms explain the movements of the foreground frame or global video frame.

Throughout the learning process of classifying normal behaviours, one-class support vector machines and kernel principal component analysis methods aid in detecting abnormal events in frames.

Figure 2.11 depicts examples of normal and abnormal frames. Frames a and c are normal as individuals are promenading in different directions. In contrast, frames b and d are abnormal as the individuals move in the same direction. The reason is that multiple movements in the same direction could indicate either an attraction by some event or escaping from a dangerous zone.



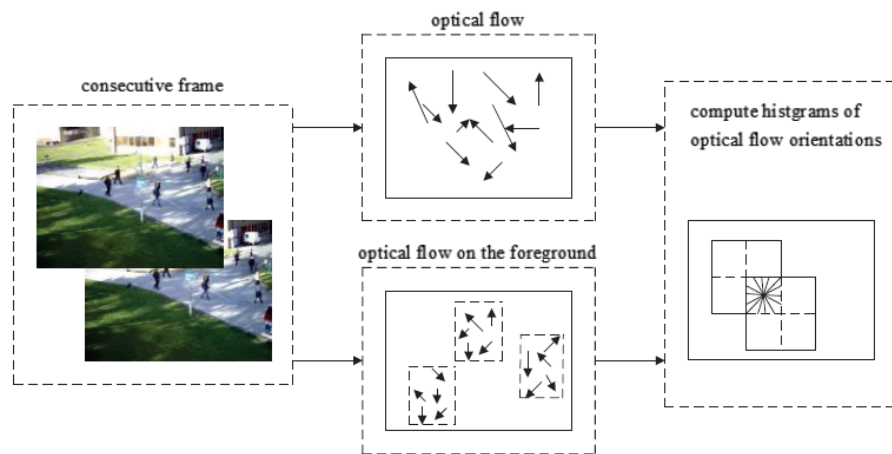
**Figure 2.11 Normal (a,c) and abnormal frames (b,d) [10]**



The detection algorithm is composed of two parts. Firstly is the extraction of visual features without tracking. Secondly comes the classification of the extracted features using one-class support vector machines (SVM) and principal component analysis (PCA).

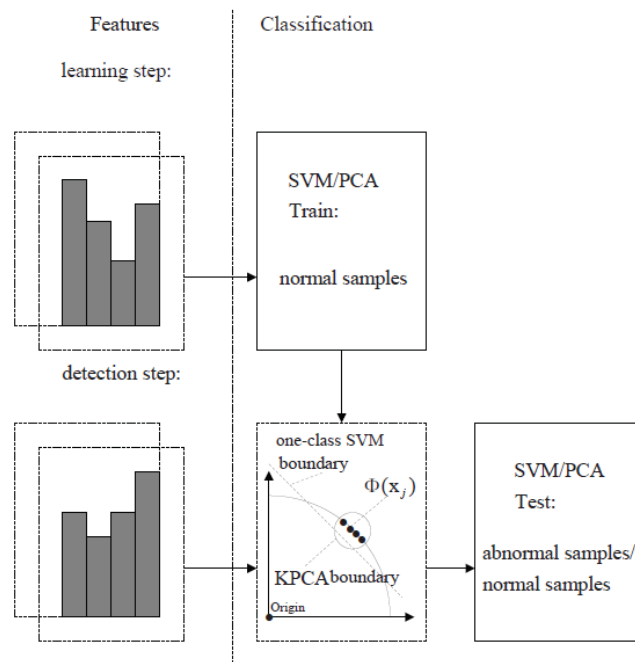
The proposed approach learns normal behaviours, and classifies abnormal ones.

The Horn-Schunck method in optical flow represents the movement information of actions. It extracts images' optical flow orientation features and gathers them into a high-dimensional feature vector. Figure 2.12 shows a 2×2 rectangular cell HOFO descriptor, informing motion information from two consecutive frames.



**Figure 2.12 Histogram of optical flow orientation (HOFO) feature descriptor on the original or foreground image. [10]**

After calculating the histogram of each frame's optical flow orientation (HOFO), as shown in Figure 2.12, comes the process in Figure 2.13. Both the one-class support vector machine and kernel principal component analysis are used to classify the features, and both can counter non-linear problems.



**Figure 2.13 The flowchart of the proposed feature classification-based abnormal detection method. [10]**

Here, a state transition threshold  $N$  is used. When the number of the predicted anomalous frames after a normal video clip exceeds  $N$ , the overall state will be updated from “normal” to “abnormal”.

Experiments on the PET and UMN datasets have performed well in detecting abnormal activities. This approach achieved AUC results of over 90% for all places, including lawns, indoors, and plazas.

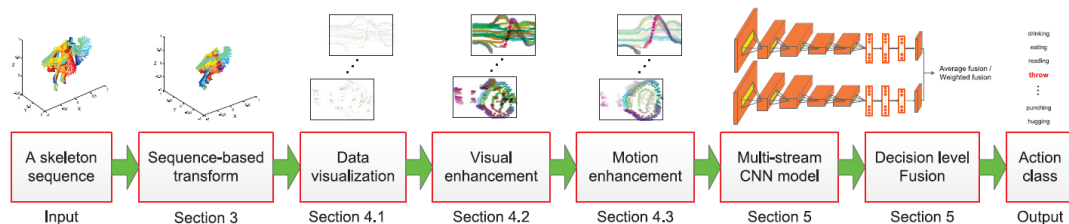
However, suggested improvements include reducing false alarms by capturing more efficient features based on optical flow. Also, implementing online learning can deal with many normal samples. Moreover, this approach only focuses on global abnormal events. Therefore, it is essential to include local abnormal events in future work.

### 2.1.5 Enhanced skeleton visualisation for view invariant human action recognition

Liu et al. [11] have developed an enhanced skeleton visualisation method for human action recognition. According to Liu et al. [11], “the human body can be represented as an articulated system with hinged joints and rigid bones, and human actions can be denoted as movements of skeletons.”

Kinect is a type of depth sensor that resolves several issues of RGB data by providing depth data. Depth data is more robust to lighting changes due to infrared radiation estimations. Also, it is better at subtracting foreground from backgrounds as it ignores irrelevant textures and colour from the cluttered environment. Therefore, this approach uses an RGB-D camera called Kinect to provide depth maps, which provide three-dimensional information on object structures.

Figure 2.14 shows the process of human action recognition with this approach.



**Figure 2.14** The pipeline of the proposed method [11]

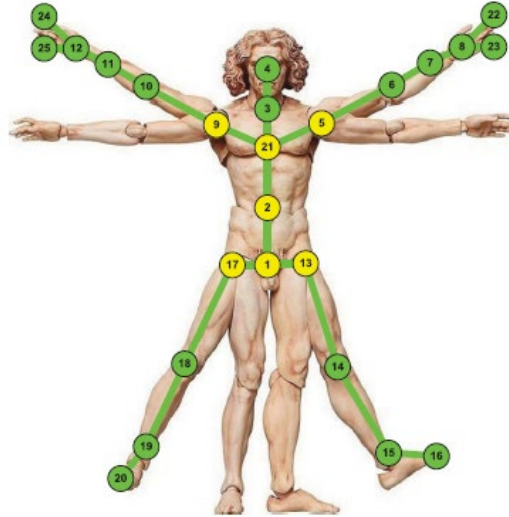
Kinect’s output skeleton sequences are represented in various colours that encode spatial and temporal cues extensively. It contains rich spatio-temporal information on skeleton joints.

Then, to increase the robustness of dealing with various angles of action observations, it goes through a novel transform process that resolves the viewpoint change issues. Here, it preserves more relative motions than traditional methods.

After some enhancements on the visuals and motions are done, it gets fed to the multi-stream CNN model. The weighted fusion method in the model is an improvement over the traditional average fusion method. The model extracts and fuses features from the inputs. Lastly, it outputs the prediction action.



Figure 2.15 shows the body joints dataset used in this project. Here, the torso joints are colored yellow.

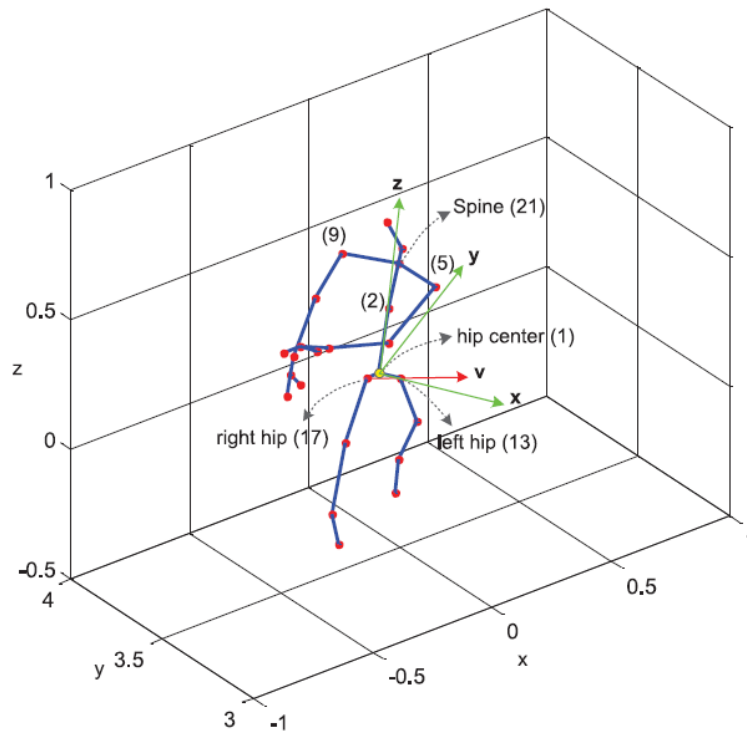


*Figure 2.15 Configuration of body joints in the NTU RGB+D dataset. [11]*

Labels of torso joints include 1-hip center, 2-middle of the spine, 5-left shoulder, 9-right shoulder, 13-left hip, 17-right hip, and 21-spine.

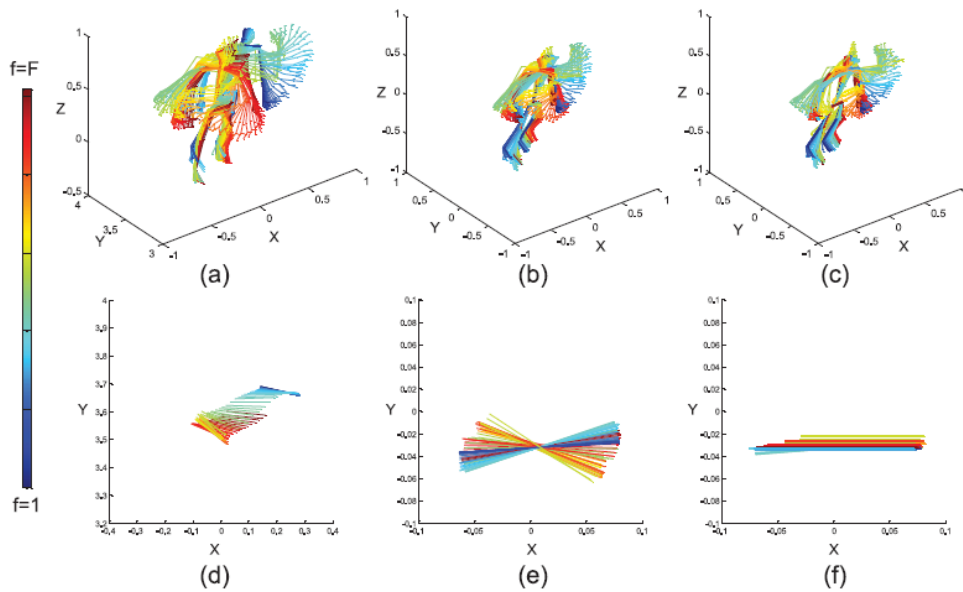
Traditional works have attempted to deal with resolving different viewpoint issues. However, they have downsides of losing out on some partial relative motion. This proposed invariant transform counters the problem by synchronously transforming all skeletons.

Figure 2.16 depicts the proposed invariant transform to counter the viewpoint problems. For new coordinate systems, the hip center acts as the origin, and  $x, y, z$  are all directions of new axes. Arrow  $x$  is vertical to arrow  $z$ , while  $y$  is their product. Arrow  $v$  denotes the direction from “right hip” to “left hip.”



**Figure 2.16 Illustration of proposed invariant transform [11]**

Image (a) in Figure 2.17 is the original “throw” skeleton sequence. Image (b) is the proposed transformation approach, and image (c) is a traditional transformation. For ease of comparison, each skeleton is simplified into a link between the “right hip” joint and the “left hip” joint. Images (d),(e),(f) are the respective motions of hip link rotation of images (a),(b),(c).



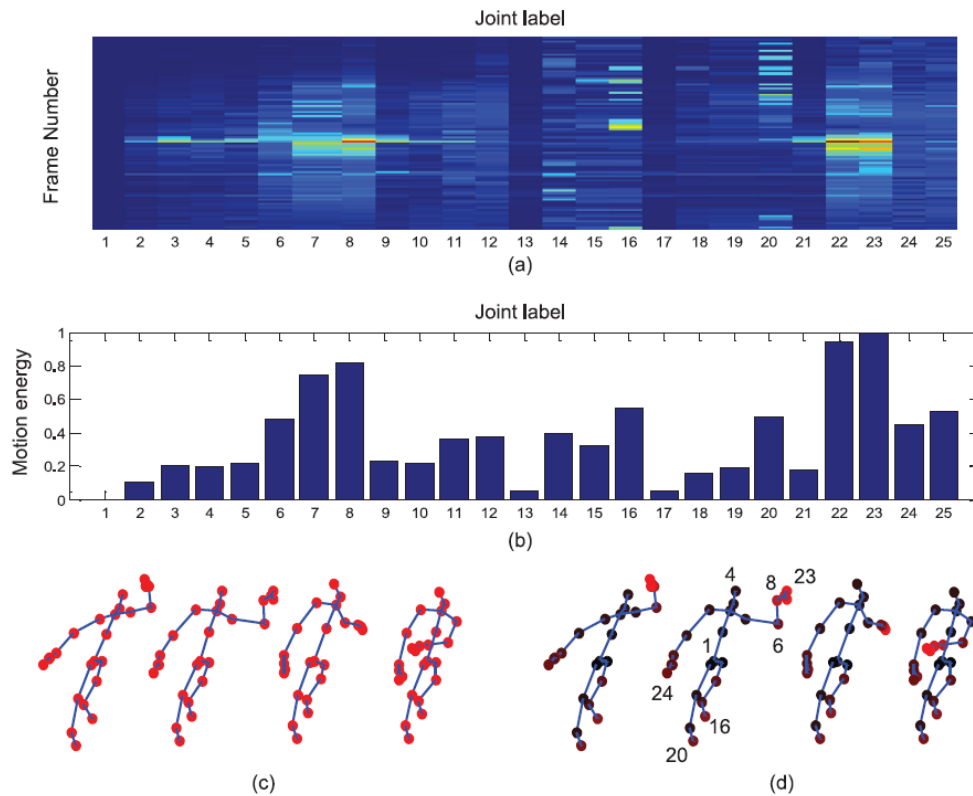
**Figure 2.17 Comparison between proposed sequence-based transformation and traditional skeleton-based transformation. [11]**

Results in image (e) succeed image (f) in capturing the hips' rotation. The proposed method is found to be significantly better at retaining relative motion (e.g., rotation).

The joint labels shown in Figure 2.18 indicates the skeleton joint labels shown in Figure 2.15 from the NTU RGB+D dataset. Individual joints range from number 1 to 25.

In image (b), the nth bar displays the motion energy of the respective nth skeleton joint. Image (d) shows the skeleton joint movements, added with red colour for joints with larger weights.

In the “throw” action example, joint number 23 (left hand) will have immense motion energy as it is most related to the action. It aids in the identification of actions.



**Figure 2.18** Illustration of weighing skeleton joints according to motion energy. [11]

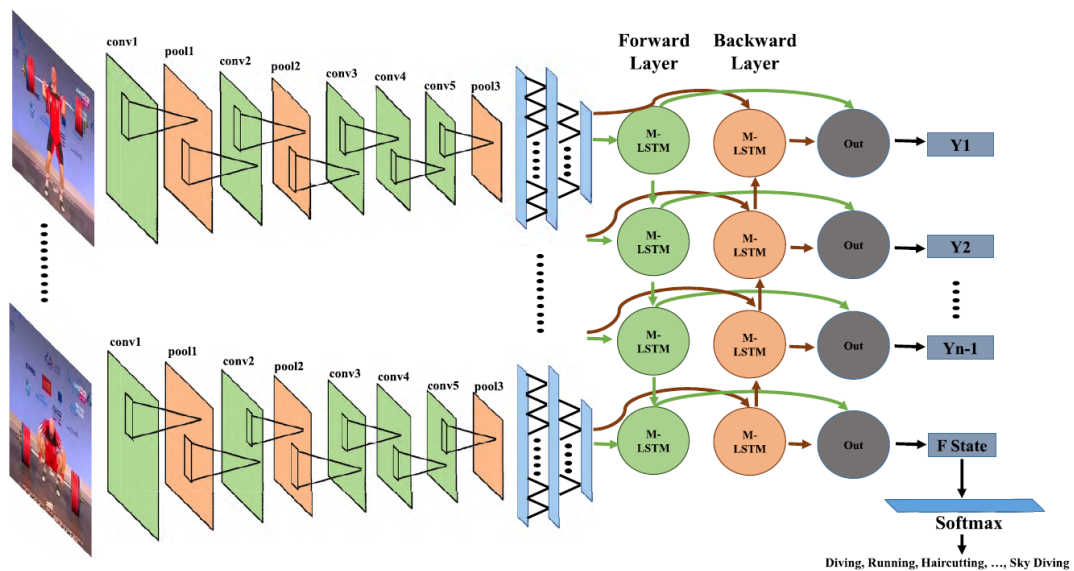
Experiments on several popular datasets work well with arbitrary viewpoints, noisy skeletons, inter-similarities, and intra-varieties. It had 87.21% and 96.62% accuracy for the NTU RGB+D and MSRC-12 datasets.

Improvements suggested are to swap the weighted probability fusion to soft probability fusion to increase the CNN's flexibility. Also, exploration on other fusion methods can be done. Moreover, data augmentation, such as adding gaussian noise to training data, can be done to elevate robustness. Lastly, training the algorithm to detect actions from untrimmed sequences.

### 2.1.6 Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features.

Amin et al. [12] have proposed a novel action recognition method. It combines convolutional neural networks (AlexNet) and deep bidirectional LSTM (DB-LSTM) networks. Whereby the CNN extracts the spatial features, the LSTM identifies temporal relations between frames.

The left portion of Figure 2.19 shows the pre-trained convolutional neural network, AlexNet, for feature extraction. The network has previously been trained on the large-scale ImageNet dataset. Here, CNN represents and classifies video frames individually. Every frame is described in CNN features.



**Figure 2.19** The framework of the proposed DB-LSTM for action recognition. [12]

Besides, the model avoids redundancy and complexity with an optimal frame jump of size six; not every video frame will be processed. Skipping through some redundant frames helps lower computational costs. Moreover, it does not change the action sequence.

The first row in Figure 2.20 shows the frames in a sequence; the second row shows their respective feature maps. Here, the feature maps capture minor player position changes while passing a basketball.



**Figure 2.20** *Frame-to-frame features representation and changes in a sequence of frames.*

[12]

According to Table 2.2, the AlexNet model used in the feature extraction phase has five convolution layers, three pooling layers, and three fully connected layers. A norm and ReLU non-linear activation function follow each layer.

Layers	Conv1	Pool1	Conv2	Pool2	Conv3	Con4	Con5	Pool5	FC6	FC7	FC8
Kernel	11x11	3x3	5x5	3x3	3x3	3x3	3x3	3x3	-	-	-
Stride	4	2	1	2	1	1	1	2	-	-	-
Channels	96	96	256	256	384	384	256	256	4096	4096	1000

**Table 2.2** *The architecture of the pre-trained AlexNet model [12]*

On the other hand, the right portion of Figure 2.19 shows the BD-LSTM. The CNN feeds the extracted in-depth features to the proposed DB-LSTM.

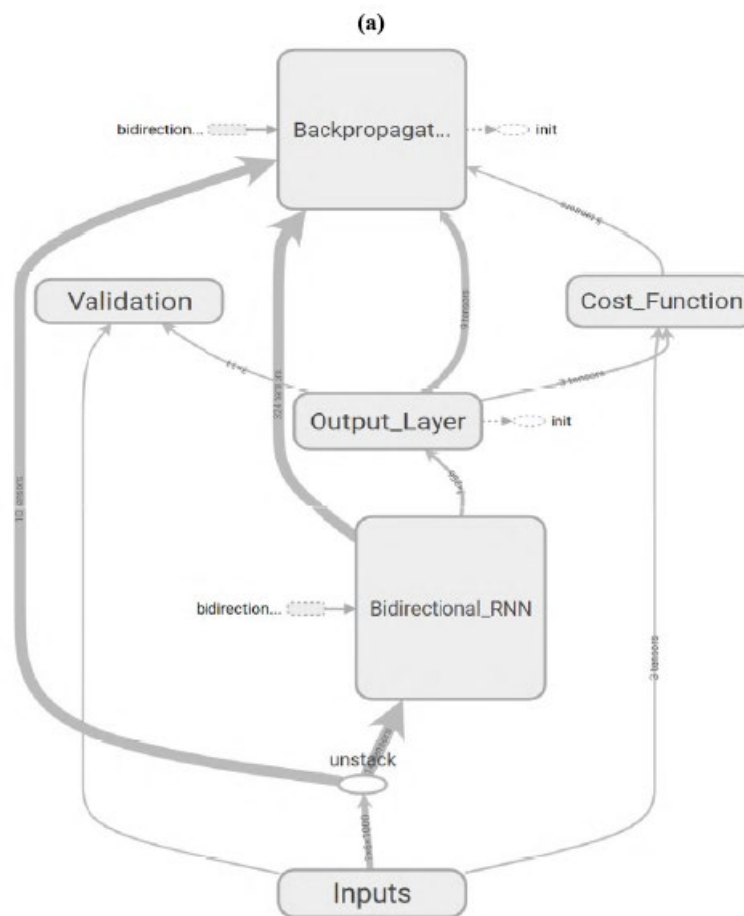
LSTM is a special type of RNN that addresses the vanishing gradient problem found in original RNNs. RNNs tend to forget the earlier inputs of the sequence in long-term sequences. LSTM solves this by keeping information of sequential frames for a long time, increasing overall understanding. LSTM considers information from previous frames when comprehending actions, learns sequential data, and analyses changes between frames.

Due to the extensive data with complex sequence patterns, a single LSTM cell may not be sufficient to identify. Therefore, an ML-LSTM (Multi Layers LSTM) is built by stacking multiple LSTM cells to understand long-term dependencies better. In this approach, two LSTM layers are stacked in the network. During the training process, each layer gets inputs of hidden states from the previous layers.

In addition, this approach brings up the robustness and makes the two LSTM layers bidirectional. An output from time  $t$  depends on both the previous and upcoming frames. In

this algorithm, one LSTM cell moves in a forward direction (forward pass), and another one goes in a backward direction (backward pass). As shown in Figure 2.21, the output layer combines the hidden state of both forward and backward cells in the Bidirectional-RNN (specifically LSTM). The validation and costs are calculated from the outputs. Also, back-propagation adjusts weights and biases.

In short, the output of frame  $t$  is affected by frames  $t-1$  and  $t+1$ , as the layers process frames bidirectionally.



**Figure 2.21** The external structure of the proposed DB-LSTM network [12]

Evaluations have been done on three major datasets: YouTube ACTIONS, HMDB51, and UCF101. Table 2.3 displays the recognition scores of various state-of-the-art action recognition methods and the proposed DB-LSTM. It is apparent from the table that the proposed method dominates the others.

<b>Method</b>	<b>YouTube</b>	<b>HMDB51</b>	<b>UCF101</b>
Multiresolution CNNs [21]	-	-	65.4%
LSTM with 30 frame unroll [6]	-	-	88.6%
Two-stream CNNs [22]	-	59.4%	88.0%
Multiple dynamic images [23]	-	65.2%	89.1%
RLSTM-g3 [32]	-	55.3%	86.9%
Hierarchical clustering multi-task [33]	89.7%	51.4%	76.3%
VideoDarwin [34]	-	63.7%	-
Discriminative representation [35]	91.6%	28.2%	79.7%
Ordered trajectories [8]	-	47.3%	72.8%
Factorized spatio-temporal CNNs [36]	-	59.1%	88.1%
Temporal pyramid CNNs [37]	-	63.1%	89.1%
Adaptive RNN-CNNs [38]	-	61.1%	-
Improved trajectories [39]	-	57.2%	-
Super-category exploration [40]	-	60.8%	-
Multi-layer fisher vector [41]	-	68.5%	-
<b>Proposed DB-LSTM</b>	<b>92.84</b>	<b>87.64</b>	<b>91.21</b>

**Table 2.3 Comparison of average recognition score of the proposed DB-LSTM for action recognition with state-of-the-art methods. [12]**

In short, the CNN layers extract spatial features from the frames and are passed on to the LSTM layer(s) at each time-steps to learn temporal sequence. This way, the hybrid combination model learns spatiotemporal features directly during training, increasing overall robustness.

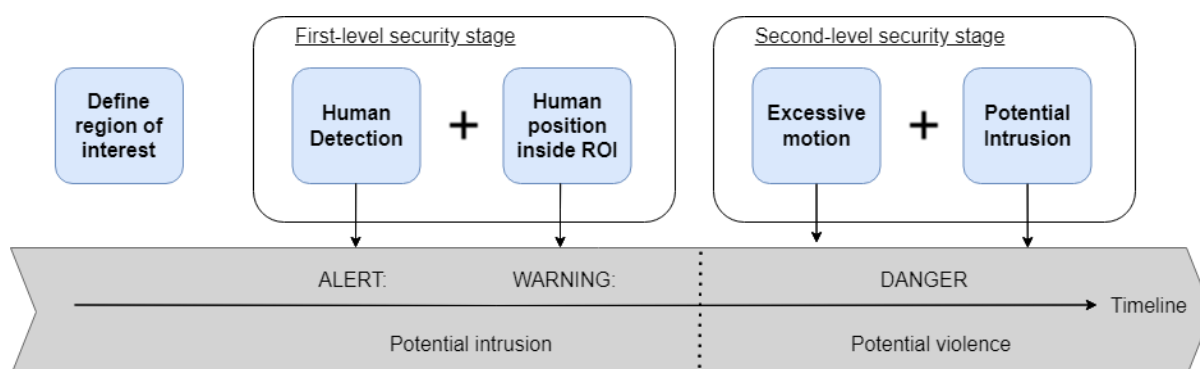
Future works discussed by Amin et al. [12] only analyse relevant regions of the frames for action recognition instead of the whole frame. Moreover, they intend to extend this work for activity recognition other than actions. Finally, the proposed method can be combined with people counting approaches to analyse crowded behaviour and dense situation.



# Chapter 3

## System Methodology/Approach

### 3.1 Methodologies and General Work Procedure



**Figure 3.1 Project Methodology**

#### 3.1.1 First-level security

The pre-processing step of the first security stage is to locate the region of interest and disregard the area beyond the house gate. Here, the system should receive manual input via clicks from the user to specify the polygon (restricted area) to be monitored. The input will be a series of coordinates that forms a polygon covering the front-yard area.

After that, the system begins to perform object detection with the help of the YOLO algorithm. Irrelevant information is reduced by focusing only on objects of the ‘person’ class. Everyday neighbourhood objects such as cars, bicycles, potted plants, etc., are ignored.

When a human is detected, their position will be retrieved and compared with the specified restricted area. If they are outside the gate, there will not be any notice. Once they set foot into the region of interest (the selected polygon), a warning of human presence is

given. Figure 3.2 shows an example frame whereby a human is in the front yard and should trigger a potential intrusion notice.



*Figure 3.2 Potential intruder present within the restricted area*

### **3.1.2 Second-level security**

The system calculates the motion intensity between frames throughout the video frames and displays it with a bar for monitoring.

During a potential intrusion, the system checks the threshold set for the motion intensity. If the current motion intensity exceeds the set threshold, it should trigger an alarm on potential violence.

Figure 3.3 shows an example frame that should trigger the alarm on potential violence.



*Figure 3.3 Excessive action between humans*

### **3.2 Assumptions**

1. The camera needs to be fixed so the footage is not moving around.
2. The entire body of the potential intruders must be present in the scene for the model to detect human presence.
3. Violent and excessive motions such as punching or kicking are assumed to be relatively near to the camera and not too far away.

# Chapter 4: System Design

## 4.1 System Flow Diagram

Figure 3.1 displays the overall system flow. Here, different colours in the figure separate the two security levels. The first security stage (left half) covers the detection of humans within the front yard. Subsequently, the second security stage (right half) checks the video frames for violent actions.

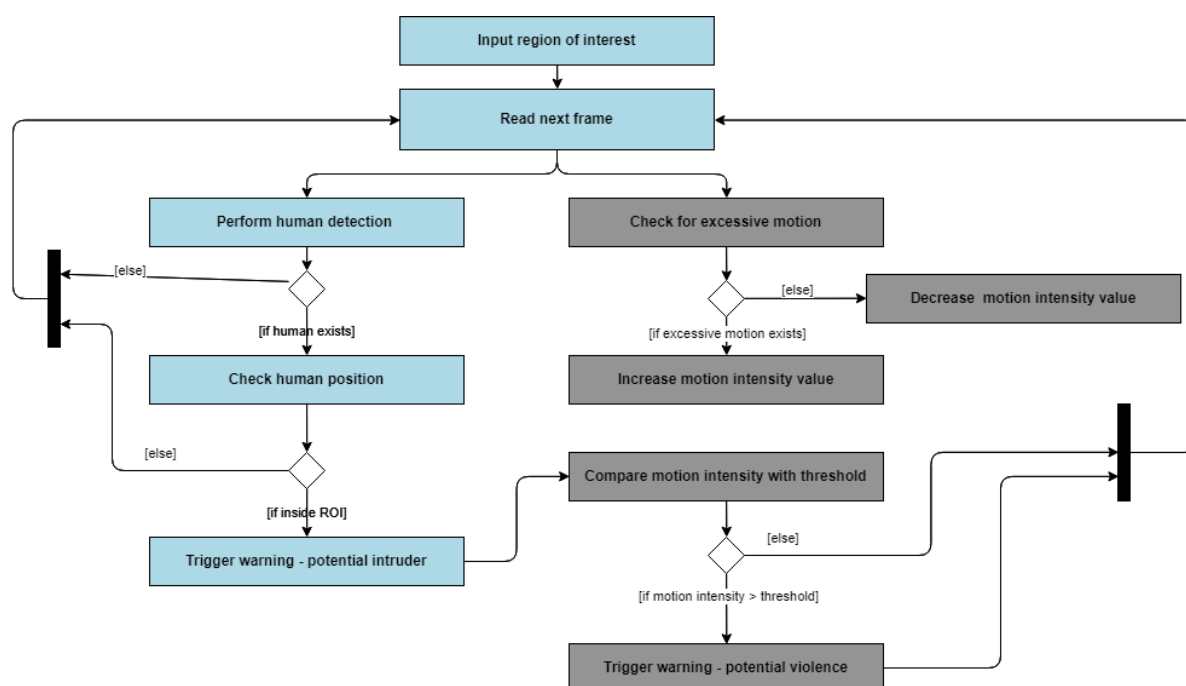


Figure 4.1 System Flow Diagram

Figure 4.1 shows the system flow diagram which processes our monitoring system, consisting of intrusion and violence detection.

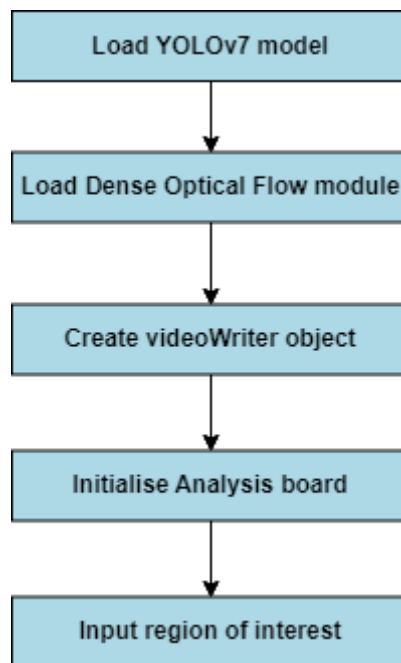
The sequence of frames will go through four main steps, including the pre-trained YOLOv7 model for human presence detection. Then it will pass through the intrusion detection module to detect whether or not the human is trespassing on private property.

At the same time, the dense optical flow module will display the flow calculated between every consecutive frame and update the motion intensity for excessive motion detection. The

violence detection module will check whether the current motion intensity is over the threshold if a potential intruder exists. Throughout these steps, important information is displayed on an analysis board beside the footage for monitoring.

## 4.2 System Components Specifications

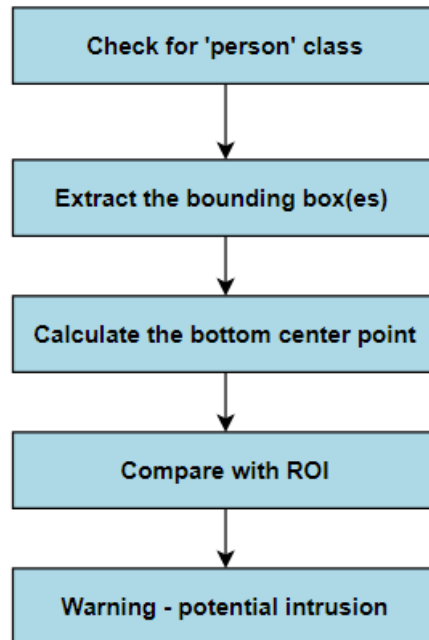
### 4.2.1 Initialisation and Pre-processing



**Figure 4.2 System initialisation**

The pre-trained YOLOv7 model is first loaded into the system in the initialisation phase. The imported model will be filtered to only check on the 'person' class and ignore all the other classes. Then, the dense optical flow and drawFlow function will be loaded into the system, which will be used for excessive motion detection when it checks for motion magnitudes. After that, a videoWriter object will be created to save the model's output. Moreover, the analysis board will initialise important information such as human count, presence, motion intensity, threshold, etc. Lastly, after the user has specified the video path, the system will read the first frame and display it to the user, allowing them to specify the region of interest by clicking on points to form a polygon.

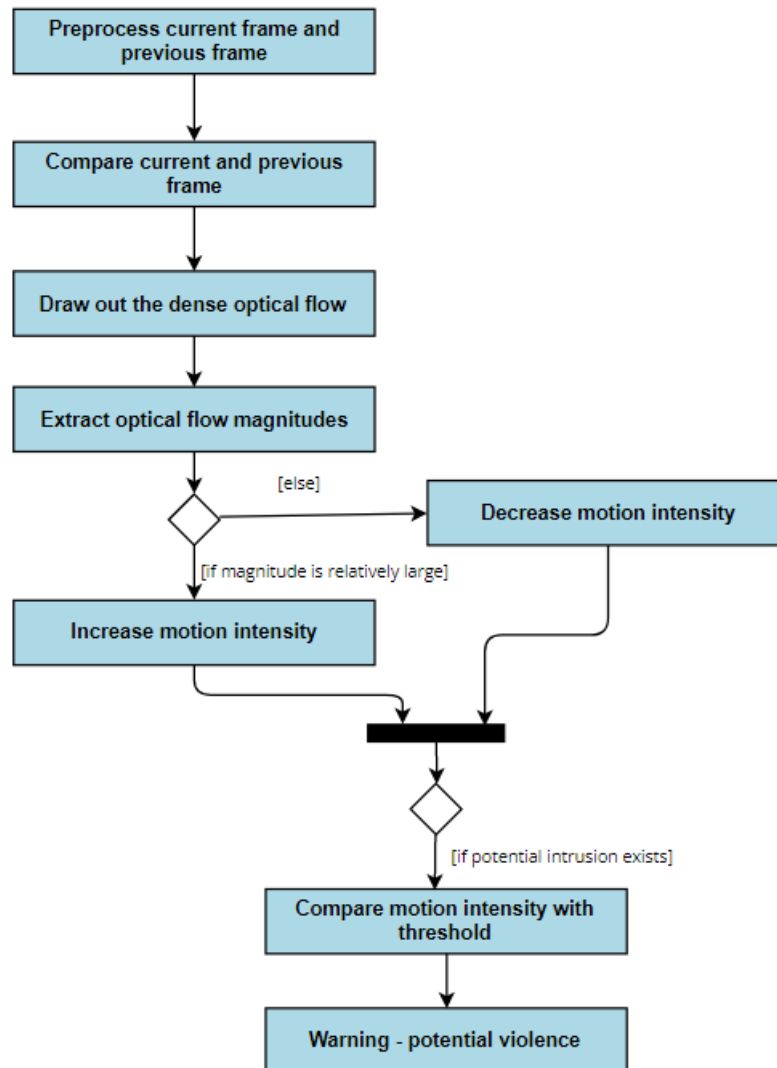
#### 4.2.2 Human and intrusion detection



**Figure 4.3 Human and intrusion detection**

Yolov7 is used to help detect the presence of humans in the frames. When a 'person' class is detected within a frame, the bounding box(es) coordinates will be extracted to calculate the bottom center point of the box(es). Then, this point will be compared with the specified region on interest in the beginning. A warning message for potential intrusion is shown if it resides within the polygon.

### 4.2.3 Excessive motion and violence detection



**Figure 4.4 Excessive motion and violence detection**

After extracting the current and previous frames, they will be converted to greyscale, as the dense optical flow algorithm expects. Then, they are passed to the dense optical flow function to calculate the optical flow generated from the first to the second. These flows will then be passed onto a drawFlow function to display the optical flows on the screen for visual purposes.

Now, the system checks for the optical flow magnitude values to see if there are any large movements between the frames. If there are, the motion intensity bar will increase by a small value. On the other hand, if there were no large movements, the motion intensity bar value would be decreased by a small value.

This is to be sure about an excessive motion if related to violence. As seen in violent scenarios, excessive motions usually happen for many consecutive frames and not only one. The system does not randomly warn about excessive motion whenever it detects one significant movement. Instead, the system checks if significant movements have occurred for several frames. To elaborate, the motion intensity can only pass the set threshold if large movements have occurred for several frames. The intensity slowly increases past the threshold in each frame containing excessive motion. This step further increase the confidence that the excessive motions are related to violence. On the other hand, when there are no large motions, the motion intensity will slowly decrease to zero.

Next, the system will only check the motion intensity bar value whenever there is a potential intrusion and see whether it crosses the set threshold. This step significantly reduces the false positive alarms by ensuring intruders make excessive motions, not random objects or dust flying around. If it does cross the threshold, it displays a warning about potential violence.



# Chapter 5: System Implementation

## 5.1 Hardware setup



The following table shows the hardware specifications used to develop the proposed system.

System	Information
Computer model	ASUS TUF A15
Operating system	Windows 11 Home
Processor	AMD Ryzen 7 4800H
Graphics card	NVIDIA GeForce GTX1660 Ti
CPU	2.90 GHz
Memory (RAM)	16.00 GB
System type	64-bit operating system, x64-based processor

**Table 5.1 Hardware setup**

## 5.2 Software setup

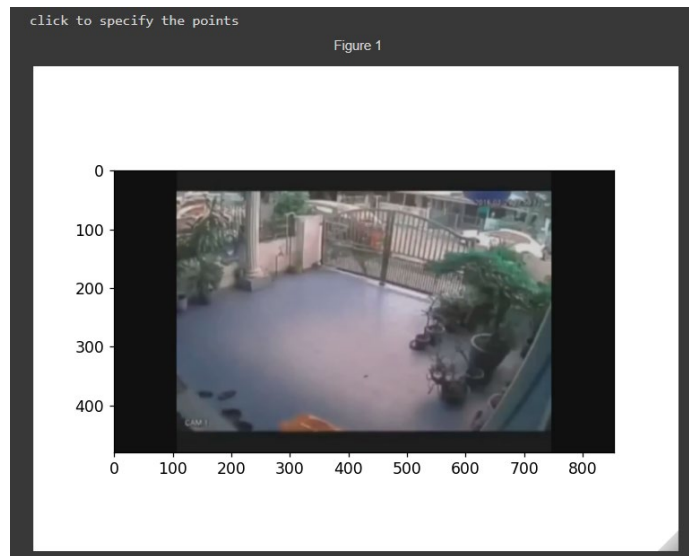
The following table shows the software tools and IDEs used to develop the proposed system.

Software	Description
	OpenCV is an open-source library containing computer vision and machine learning functionalities. It was built to provide an infrastructure for computer vision applications. The library is chosen for its suitability for developing reliable vision applications.
	Google Collaboratory is a web-based IDE for python that Google released in 2017. It is an excellent tool for data scientists to execute Machine Learning and Deep Learning projects with cloud storage capabilities. The IDE is chosen for its accessibility to GPUs and TPUs to anyone who needs to build AI-related models.

**Table 5.2 Software setup**

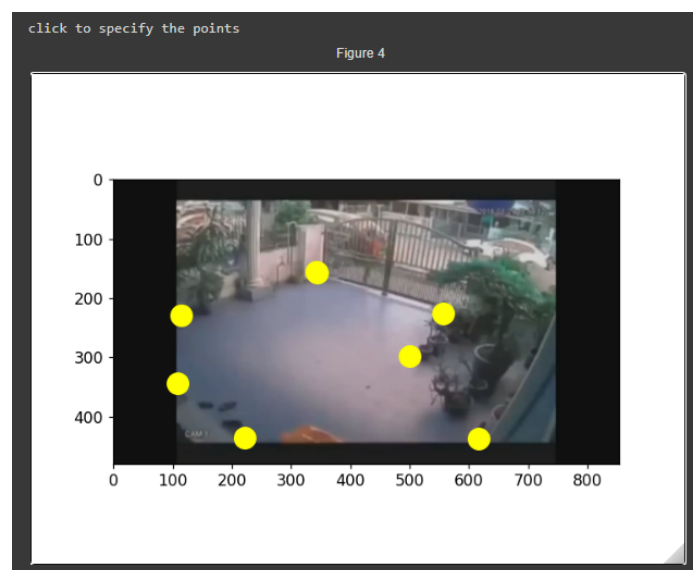
## 5.3 System Operation

### 5.3.1 Defining the Region of Interest



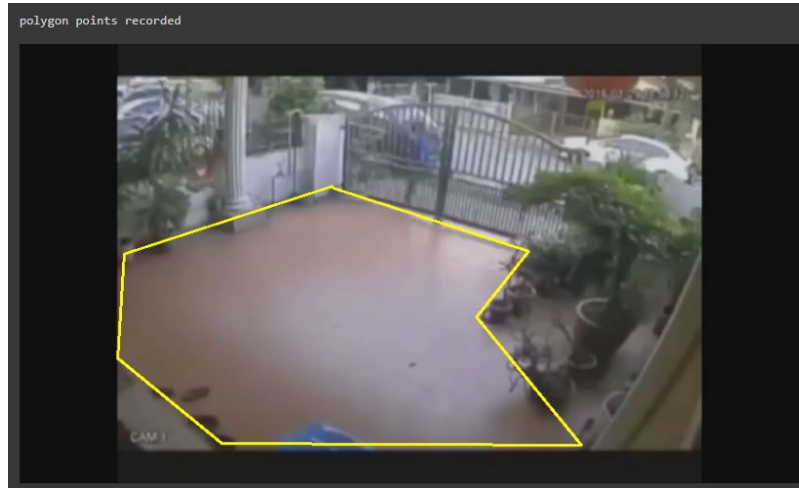
**Figure 5.1 Before clicking on points**

After the user specifies the video path to be monitored, the system reads the first frame and displays it to the user. Here, it expects the user to click on the points to form a polygon surrounding the interest region.



**Figure 5.2 After clicking on points**

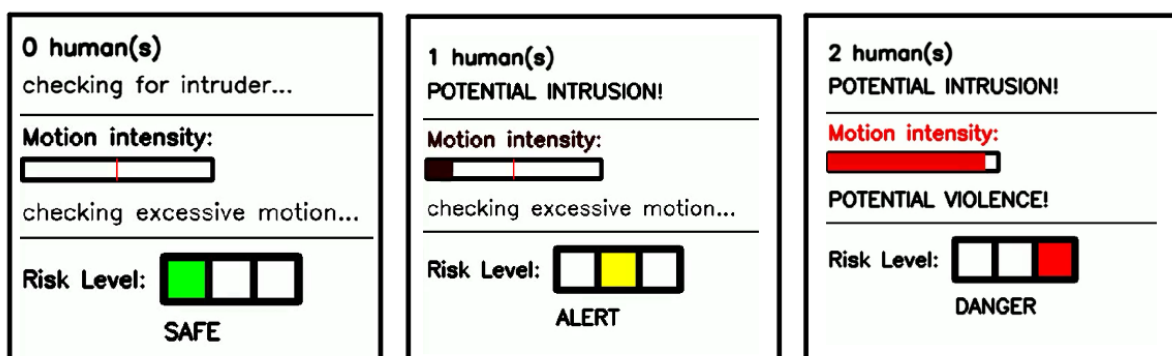
The user clicks on the points in either a clockwise or anti-clockwise manner. After specifying the final point, the user must click on the initial point again. It informs the system that those are all the points needed for the polygon.



**Figure 5.3 System connects the points and forms a polygon**

Next, the system passes all the specified points to the polylines function to sketch the polygon for visualisation. It connects the points accordingly to the order they were provided.

### 5.3.2 Analysis board



**Figure 5.4 Different risk levels of the analysis board**

The analysed risk levels can be classified into 'SAFE,' 'ALERT,' and 'DANGER.'

The first board shows a risk level of ‘SAFE.’ At this point, no humans are present within the frame. Therefore, it is still checking for intruders. Without humans, the motion intensity is practically zero when no movements are going on. Therefore, it is checking for excessive motion. In this stage, the risk level will be labelled with a green box to indicate a safe moment.

The second board shows a risk of ‘ALERT.’ Here, the system has detected an object of the ‘person’ class, and has found that the object is within the restricted area drawn by the polygon. A warning message of ‘POTENTIAL INTRUSION’ is displayed. At this point, the intruder is roaming around the ROI without too many large movements. Therefore the motion intensity is still below the threshold. In this stage, the risk level will be labelled with a yellow box to indicate a suspicious and potentially dangerous moment.

The third board shows a risk of ‘DANGER.’ At this point, there has been a potential intruder within the ROI. On top of that, that system has detected significant motion for a consecutive number of frames. The scenario where large movements happen for a while indicates violent activities such as fighting. The motion intensity bar increases slowly until it surpasses the threshold. When it goes over the line, the system will show a warning message of ‘POTENTIAL VIOLENCE’. In this stage, the risk level will be labelled with a red box to indicate a dangerous moment.

### 5.3.3 Monitoring board initialisation

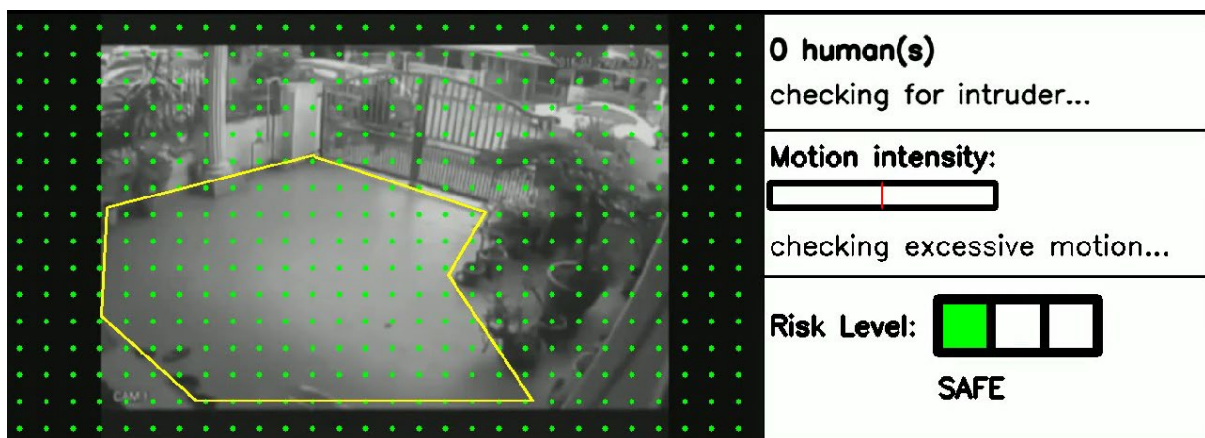


Figure 5.5 Initialisation of the whole monitoring board

Important information discussed in the two previous subsections is displayed on the monitoring board, including the region of interest and the analysis board, which is initially in a 'SAFE' risk level. Besides, the monitoring points for the dense optical flow are printed onto the video frame. These points will display the optical flow between frames and the magnitude of flows via drawing lines, whereby the flow length translates the flow magnitudes. However, initially, no optical flows are being generated yet.

The dense optical flow is computed with Gunnar Farneback's algorithm by Gunnar F. [16], with the `calcOpticalFlowFarneback()` function. The function mainly takes in the prev frame and current frame to compute the dense optical flow.

# Chapter 6: System Evaluation and Discussion

## 6.1 System Testing and Results

### 6.1.1 Scenario A

Category: Multiple people waking outside the gate

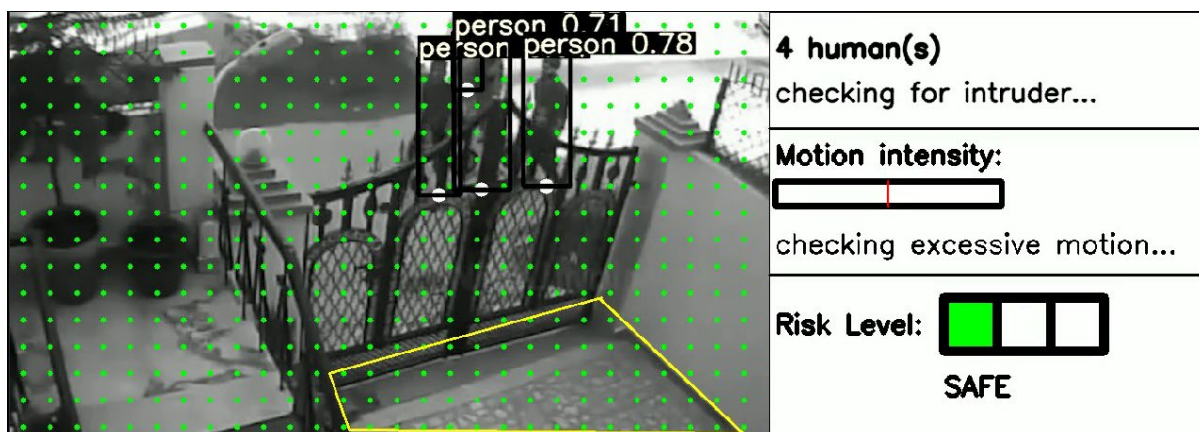


Figure 6.1 System ignoring humans outside the gate

The system was tested for the first level security stage with a part of a video where multiple people are walking past the house. In this part of the video, the people were outside the house.

The system detects the humans, checks their position with the ROI, and finds that they are not within the polygon. Therefore their bottom centre points are white, indicating they have not trespassed on the private property. Here, the risk level remains 'SAFE'.

### 6.1.2 Scenario B

Category: Intruder jumps over the gate

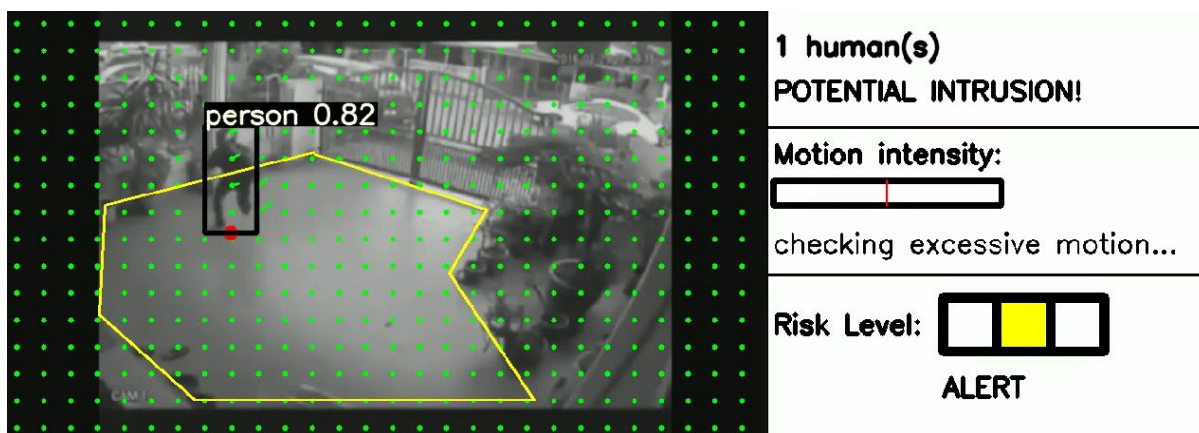


Figure 6.2 Intruder jumps over the gate

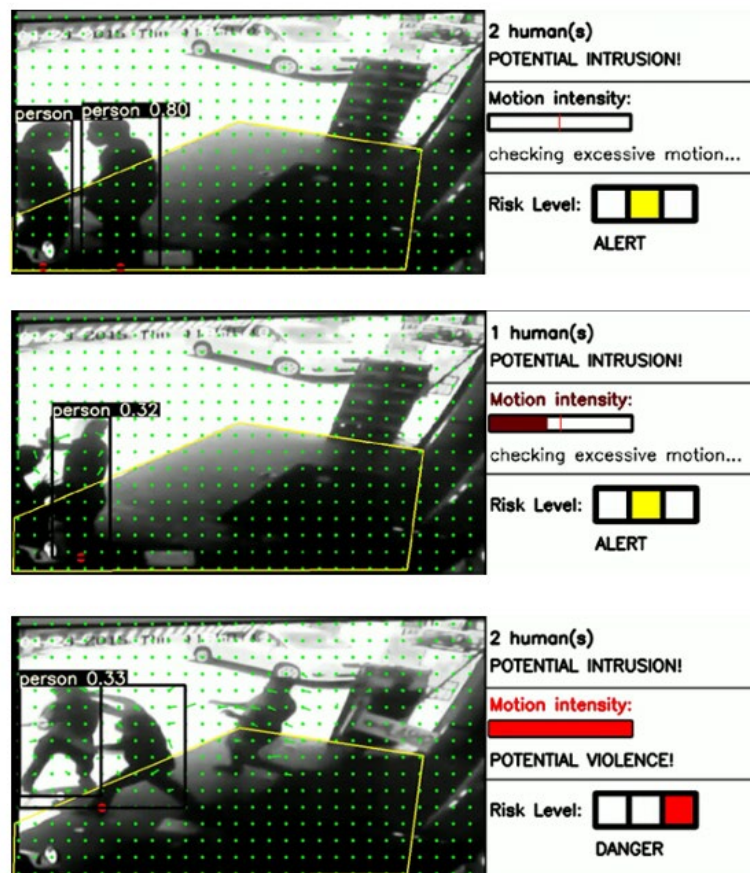


For the first level security stage, the system has been tested with a video containing an intruder entering a private property during the daytime.

Figure 6.1 shows that the system can capture this intruder. It has correctly detected one human and ensures the person is within the polygon. The bottom centre point of the ‘person’ object is red as it is within the ROI. Here, the analysis board displays the correct warning message of ‘POTENTIAL INTRUSION!’ and labels the correct risk level of ‘ALERT’ with a yellow box.

### 6.1.3 Scenario C

**Category: Fighting activity**



**Figure 6.3 Different monitoring stages during a fight**

For the second level security stage, the system was tested with a video containing fighting between a few people. Figure 6.3 shows 3 phases of monitoring potential violence.

The first phase is before the fight begins. The system is notified about people within the ROI. At this point, no significant motions are happening between frames. Therefore, the motion intensity is zero.

The second phase is when the fight first begins. At this point, the system detects large movements between the frames but is unsure whether they are related to violence. Therefore, it will look through more frames to check whether significant movements are still present after a while, which is a good indication of potential violence.

It does this by incrementing a small value to the motion intensity bar if two frames contain a large optical flow magnitude. On the other hand, if there were no significant movement between two frames, the system would also decrease the motion intensity by a small value. In other words, for the motion intensity to cross the threshold, large magnitudes of optical flow must happen for some time or several consecutive frames.

Besides, the system can only detect one person in the second part, as the other person is not entirely in the frame. It is important to note that sometimes during a fight, the human pose may be unusual or overlap with each other. Therefore, the YOLOv7 model may not be able to detect it sometimes.

The third phase is when the fight has been going on for a while, and the system is sure that large magnitudes of optical flow have been happening consecutively for some time. The motion intensity has increased for a while and has already passed the threshold.

At this point, the system knows that there are potential intruders, and large movements have been going on for a while. After checking for the two stages, the system shows the 'POTENTIAL VIOLENCE!' warning.



## 6.2 Project challenges

Due to the lack of computational resource, frames have to be shrunk to smaller resolutions to ensure that it does not take up too much computational power. The reduction of the resolution slightly decreases the performance of YOLOv7. A short personal experiment found that YOLOv7 performs better on frames of resolution 1980 x 1080 when compared to frames of resolution 854 x 480 used in the project.

Another challenge is for the system to detect excessive motions far from the camera. When a fighting activity is happening far from the camera, the two people are relatively small, and the generated optical flow magnitude is tiny too. Therefore, expecting the system to detect them would not be ideal.

## 6.3 Objectives evaluation

### 1. A reliable automated front-yard monitoring system has been developed.

The developed system is fully automated, enabling it to perform human-like monitoring tasks without human intervention. The algorithm helps to reduce reliance on human resources.

### 2. A two-level security stage has been fulfilled:

#### i. To detect people within the region of interest

Users can specify the polygon of the region of interest before the monitoring starts. The system can give a first-level warning when it detects humans in the specified restricted area. Besides, the system ignores humans outside the gate, reducing false positive warnings.

#### ii. To detect violent or excessive actions caused by intruders

The system calculates the magnitude of optical flow generated between every consecutive frame and updates the motion intensity accordingly. The system triggers an alarm when it detects a sequence of frames that indicate violent actions.

# Chapter 7: Conclusion and Recommendation

## 7.1 Conclusion

In conclusion, the frequency of house break-in cases in Malaysia is high. The intelligent surveillance system aims to reduce front-yard intrusions and violent crimes. The system can lead to lower casualties and financial losses to household residents in the long run. Ultimately, it improves residents' safety and living standards.

The system has several perks when compared to traditional approaches. Firstly, it requires less human supervision as manual monitoring is no longer needed. The system can look for abnormalities at all times and immediately notify users to take necessary actions. Moreover, human error, such as inattention or fatigue, is significantly avoided due to the reduced reliance on guards to monitor CCTV footage.

Furthermore, the system reduces false positive alarms. Whenever it warns of 'potential intruder' or 'violence detected', there is a high chance that those cases exist in the video. The reason is that the system checks for criteria before setting off the warning message, unlike the traditional motion sensors that may start ringing after detecting some random motion.

All in all, the proposed system will have two levels of security stages. The first-level security gives notice when there is a potential intrusion into the front yard. Meanwhile, the second-level security stage raises the alarm when violence is detected.

## 7.2 Recommendation

Another tremendous but costly approach is the CNN+LSTM deep learning approach for violence detection to improve. Whereby the CNN learns the spatial information of frames, the LSTM looks through a sequence of forward and backward frames to check for violence. A model trained upon a considerable number of videos can perform even complicated tasks, such as differentiating fighting and dancing activities, which significantly reduces the false positives of the system. This system may misclassify a dancing activity as potential violence, as dancing can create large magnitudes of optical flows.

An attempt has been made at this method. However, the computation costs to train the model on thousands of videos were too high at one point. Henceforth, the dense optical flow was chosen and still performed well. Ensure to have a good GPU and enough resources to try this approach.

## REFERENCES

- [1] Mohd U. "Crime Statistics Publication, 2021." Department of Statistics Malaysia Official Portal.  
[https://www.dosm.gov.my/v1/index.php?r=column/cthemByCat&cat=455&bul\\_id=eHE0eGZWSmNROG1BbHR2TzFvZzZxQT09&menu\\_id=U3VPMldoYUxzVzFaYmNkWXZteGduZz09](https://www.dosm.gov.my/v1/index.php?r=column/cthemByCat&cat=455&bul_id=eHE0eGZWSmNROG1BbHR2TzFvZzZxQT09&menu_id=U3VPMldoYUxzVzFaYmNkWXZteGduZz09) (accessed July 22, 2022).
- [2] Gonzalez R.C. and Woods R.E., "Introduction", in *Digital Image Processing* (4<sup>th</sup> Ed.). Upper Saddle River, New Jersey, Addison-Wesley, 2018
- [3] Keval H. and Sasse M. A., "'Not the Usual Suspects': a study of factors reducing the effectiveness of CCTV." *Security Journal*, vol. 23, no. 2, pp. 134-154, April 2010.
- [4] Gill M., Little R., Spriggs A., Allen J., Argomaniz J., "Assessing the impact of CCTV: The Hawkeye case study." *Home Office Online Report*, vol. 12, no. 5, Jan 2005.
- [5] Renata M. and Philippe L., "When and why threats go undetected: Impacts of event rate and shift length on threat detection accuracy during airport baggage screening." *Human Factors The Journal of the Human Factors and Ergonomics Society*, vol. 58, no. 2, pp. 218-228, Nov 2016.
- [6] Temple J., Joel W., William D., Keith J., Constance M. Matthews G., "The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task." *Human Factors The Journal of the Human Factors and Ergonomics Society*, vol. 42, no. 2, pp. 183-194, Feb 2000.
- [7] Sultani W., Chen C., Shal M., "Real-world Anomaly Detection in Surveillance Videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479-6488, Jan 2018.

- [8] Arroyo R., Miguel L.B., Garcia I.d., “Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviours in Shopping Malls.” *Expert Systems with Applications*, vol. 42, no. 21, pp. 7991-8005, Nov 2015.
- [9] Shinde S., Kothari A., Gupta V., “YOLO based Human Action Recognition and Localization.” *International Conference on Robotics and Smart Manufacturing*, vol. 133, pp. 831-838, Jan 2018.
- [10] Wang T., Snoussi H., “Detection of abnormal events via optical flow feature analysis.” *Sensors*, vol. 15, no. 4, pp. 7156-7171, Mar 2015.
- [11] Liu M., Liu H., Chen C., “Enhanced skeleton visualisation for view invariant human action recognition.” *Pattern Recognition*, vol. 68, pp. 346-362, Aug 2017.
- [12] Amin U., Jamil A., Khan M., Sajjad M., Sung W. B., “Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features.” *IEEE Access*, vol. 6, pp. 1155-1166, Nov 2017.
- [13] Wang C., Bochkovski A., Liao H., “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, July 2022.
- [14] Wong K., “Official YOLOv7”, Github. <https://github.com/WongKinYiu/yolov7> [accessed Nov 2022].
- [15] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T., “Long-term recurrent convolutional networks for visual recognition and description.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625-2634, June 2015.

[16] Gunnar F., “Open CV Video Analysis”, OpenCV.  
[https://docs.opencv.org/3.4/dc/d6b/group\\_\\_video\\_\\_track.html#ga5d10ebbd59fe09c5f650289ec0ece5af](https://docs.opencv.org/3.4/dc/d6b/group__video__track.html#ga5d10ebbd59fe09c5f650289ec0ece5af) [accessed April 2023]

## WEEKLY LOG

### FINAL YEAR PROJECT WEEKLY REPORT

(Project I)

Trimester, Year: 3,3	Study week no.: 2
Student Name & ID: Bryan Low Keng Seong 20ACB01314	
Supervisor: Prof. Dr. Leung Kar Hang	
Project Title: Video Surveillance: Front-yard Monitoring	

#### 1. WORK DONE

- ask for input of the number of points for polygon
- let users click and specify the points of the region of interest (polygon)

#### 2. WORK TO BE DONE

- simplify the polygon input (skip the number of points input)
- stop taking click input when the latest click is close to the first click at the start
- research on proposed violence detection method


#### 3. PROBLEMS ENCOUNTERED

#### 4. SELF-EVALUATION OF THE PROGRESS



Supervisor's signature

9 Feb 2023



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project I)

Trimester, Year: 3,3	Study week no.: 3
Student Name & ID: Bryan Low Keng Seong 20ACB01314	
Supervisor: Prof. Dr. Leung Kar Hang	
Project Title: Video Surveillance: Front-yard Monitoring	

## 1. WORK DONE

- simplify the polygon input (skip the number of points input)
- stop taking click input when the latest click is close to the first click at the start
- initial stages in building the model (CNN+LSTM approach)
- gathered some dataset for violence model training
  - Movie Fight Detection Dataset
  - Hockey Fight Detection Dataset

## 2. WORK TO BE DONE

- study more on the proposed model
- modify model architecture, experiment different parameters

## 3. PROBLEMS ENCOUNTERED

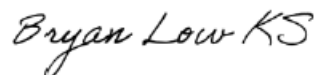
- after training with datasets mentioned, model unable to detect front yard violence well (possible issue with current dataset, or model architecture/parameter)

## 4. SELF-EVALUATION OF THE PROGRESS



Supervisor's signature

27 Mar 2023



Student's signature



# FINAL YEAR PROJECT WEEKLY REPORT

(Project I)

Trimester, Year: 3,3	Study week no.: 5
Student Name & ID: Bryan Low Keng Seong 20ACB01314	
Supervisor: Prof. Dr. Leung Kar Hang	
Project Title: Video Surveillance: Front-yard Monitoring	

## 1. WORK DONE

Further trained the existing model on  
-Real Life Violence Situations Dataset

## 2. WORK TO BE DONE

-find ways to improve model  
-perform further testing, evaluate the model performance/ success metric (include in report)

## 3. PROBLEMS ENCOUNTERED

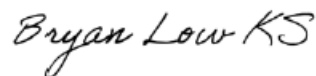
-model does not detect violence that good when slightly futher away from camera, needs improvement

## 4. SELF-EVALUATION OF THE PROGRESS



Supervisor's signature

27 Mar 2023



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project I)

Trimester, Year: 3,3	Study week no.: 6
Student Name & ID: Bryan Low Keng Seong 20ACB01314	
Supervisor: Prof. Dr. Leung Kar Hang	
Project Title: Video Surveillance: Front-yard Monitoring	

## 1. WORK DONE

Research two different optical flows approaches for violence detection (in progress)

- dense optical flow
- sparse optical flow

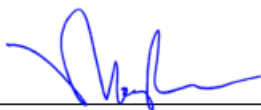
## 2. WORK TO BE DONE

- evaluate the dense optical flow approach
  - > check if the optical flow magnitudes are proportional (normal/slow motion vs fast/excessive motion, flow magnitudes should be different)
  - >further understand the code on optical flow magnitude

## 3. PROBLEMS ENCOUNTERED

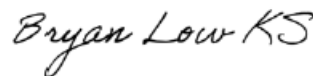
-so far, optical flow only able to detect motion well when movement motions are nearer to the camera (may need to narrow down the scope)

## 4. SELF-EVALUATION OF THE PROGRESS



Supervisor's signature

27 Mar 2023



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project I)

Trimester, Year: 3,3	Study week no.: 9
Student Name & ID: Bryan Low Keng Seong 20ACB01314	
Supervisor: Prof. Dr. Leung Kar Hang	
Project Title: Video Surveillance: Front-yard Monitoring	

## 1. WORK DONE

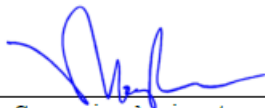
- show intensity of the warning message 'excessive motion'
- range from 0 to 255 (gradually increase and decrease accordingly, cooling effect)

## 2. WORK TO BE DONE

- make the warning message more readable/understandable
- avoid black screen, show the text at all times, find a way to show intensity info better
- report writing

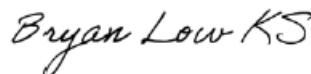
## 3. PROBLEMS ENCOUNTERED

## 4. SELF-EVALUATION OF THE PROGRESS



Supervisor's signature

30 March 2023



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project I)

Trimester, Year: 3,3	Study week no.: 10
Student Name & ID: Bryan Low Keng Seong 20ACB01314	
Supervisor: Prof. Dr. Leung Kar Hang	
Project Title: Video Surveillance: Front-yard Monitoring	

## 1. WORK DONE

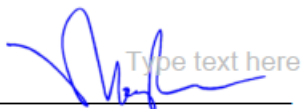
- make the warning message more readable/understandable
- avoid black screen, show the text at all times, find a way to show intensity info better
- show the percentage bar, indicate the intensity of motion
- display a warning when motion exceeds the set threshold

## 2. WORK TO BE DONE

- tweak the values of threshold, and values on how much motion bar increases to better fit more videos
- add yolo bounding box to increase confidence of violence, when human  $> 2$  , and are near each other/overlapping
- report writing

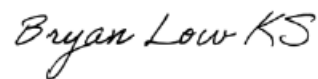
## 3. PROBLEMS ENCOUNTERED

## 4. SELF-EVALUATION OF THE PROGRESS

  
Type text here

Supervisor's signature

8 Apr 2023



Student's signature

## POSTER



Student: Bryan Low Keng Seong  
Supervisor: Prof. Dr. Leung Kar Hang  
Faculty of Information and  
Communication Technology

# Video surveillance : Front-yard monitoring



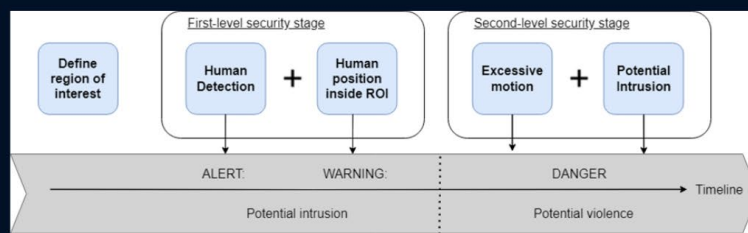
## INTRODUCTION

House break-ins and theft cause disruption in neighbourhood peace and safety. The front yard is the outmost layer of a home, and is vulnerable to criminal activities. An automated smart surveillance system can help detect crime within the front yard.

## OBJECTIVES

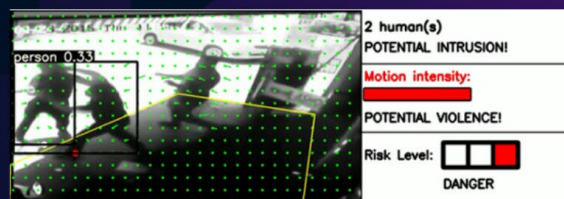
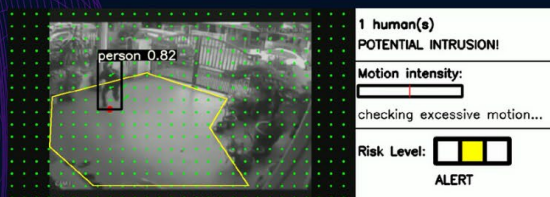
1. To develop a reliable automated front-yard monitoring system
2. To implement a two-level intruder detection feature
  - i. Detect potential intrusion within the front yard.
  - ii. Detect violence/ excessive actions (eg: fighting)

## METHODOLOGY



## RESULTS

Test results of intrusion and violence detection



## PLAGIARISM CHECK RESULT

### ORIGINALITY REPORT

**14%**

SIMILARITY INDEX

**11%**

INTERNET SOURCES

**11%**

PUBLICATIONS

**3%**

STUDENT PAPERS

### PRIMARY SOURCES

**1**

[eprints.utar.edu.my](http://eprints.utar.edu.my)

Internet Source

**2%**

**2**

[cyberleninka.org](http://cyberleninka.org)

Internet Source

**2%**

**3**

[www.robeseafe.uah.es](http://www.robeseafe.uah.es)

Internet Source

**2%**

**4**

[mafiadoc.com](http://mafiadoc.com)

Internet Source

**1%**

**5**

[web.archive.org](http://web.archive.org)

Internet Source

**1%**

**6**

Shubham Shinde, Ashwin Kothari, Vikram Gupta. "YOLO based Human Action Recognition and Localization", Procedia Computer Science, 2018

Publication

**1%**

**7**

[www.groundai.com](http://www.groundai.com)

Internet Source

**1%**

**8**

Submitted to Midlands State University

Student Paper

**<1%**

9	Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, Sung Wook Baik. "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features", IEEE Access, 2018 Publication	<1 %
10	J. Ren, N. H. Reyes, A. L.C. Barczak, C. Scogings, M. Liu. "Towards 3D Human Action Recognition Using a Distilled CNN Model", 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP), 2018 Publication	<1 %
11	Submitted to University of Carthage Student Paper	<1 %
12	Mengyuan Liu, Hong Liu, Chen Chen. "Enhanced skeleton visualization for view invariant human action recognition", Pattern Recognition, 2017 Publication	<1 %
13	Roberto Arroyo, J. Javier Yebes, Luis M. Bergasa, Iván G. Daza, Javier Almazán. "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls", Expert Systems with Applications, 2015 Publication	<1 %
14	Wang, Tian, and Hichem Snoussi. "Detection of Abnormal Events via Optical Flow Feature	<1 %



<b>Universiti Tunku Abdul Rahman</b>			
<b>Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)</b>			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1



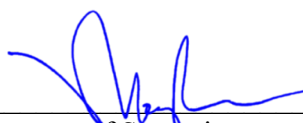
**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

<b>Full Name(s) of Candidate(s)</b>	Bryan Low Keng Seong
<b>ID Number(s)</b>	20ACB01314
<b>Programme / Course</b>	Bachelor of Computer Science
<b>Title of Final Year Project</b>	Video Surveillance: Front-yard monitoring

<b>Similarity</b>	<b>Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)</b>
<b>Overall similarity index: <u>14</u> %</b>  <b>Similarity by source</b> Internet Sources: <u>11</u> % Publications: <u>11</u> % Student Papers: <u>3</u> %	
<b>Number of individual sources listed of more than 3% similarity: <u>0</u></b>	
<b>Parameters of originality required and limits approved by UTAR are as Follows:</b> (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

***Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.***

  
 \_\_\_\_\_  
 Signature of Supervisor

Name: Leung Kar Hang

Date: 23 April 2023

\_\_\_\_\_  
 Signature of Co-Supervisor

Name: \_\_\_\_\_

Date: \_\_\_\_\_





**UNIVERSITI TUNKU ABDUL RAHMAN**

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY  
(KAMPAR CAMPUS)

**CHECKLIST FOR FYP2 THESIS SUBMISSION**

Student Id	20ACB01314
Student Name	Bryan Low Keng Seong
Supervisor Name	Prof. Dr. Leung Kar Hang

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
-	Front Plastic Cover (for hardcopy)
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
-	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
-	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

\*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)

Date: 21/04/23