

**Front Yard Surveillance System: Robbery Scene Detection**

BY

Tang Jia Le

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONS)

INFORMATION SYSTEMS ENGINEERING

Faculty of Information and Communication Technology  
(Kampar Campus)

JAN 2020

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

**Title:** FRONT YARD SURVEILLANCE SYSTEM: ROBBERY SCENE DETECTION

**Academic Session:** JAN 2020

I, TANG JIA LE

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:**

123, JALAN INTAN MAS,  
TAMAN INTAN MAS,  
TELUK INTAN,  
36000, PERAK

Leung Kar Hang

Supervisor's name

**Date:** 24/04/2020

**Date:** 24 April 2020

**Front Yard Surveillance System: Robbery Scene Detection**

BY

Tang Jia Le

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONS)


INFORMATION SYSTEMS ENGINEERING

Faculty of Information and Communication Technology  
(Kampar Campus)

JAN 2020

## DECLARATION OF ORIGINALITY

I declare that this report entitled “**Front Yard Surveillance System: Robbery Scene Detection**” 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.

Signature :   
\_\_\_\_\_

Name : Tang Jia Le

Date : 24/04/2020

## **ACKNOWLEDGEMENT**

I would like to express my deepest appreciation and gratitude to my supervisor, Prof. Maylor Leung Kar Hang, who gave me the opportunity to involve in this computer vision project. He has patiently provided feedback to enhance this project. Without his guidance and promptly help this project would have not been possible. I can no other answer make but thanks, and thanks and ever thanks to Prof. Maylor Leung Kar Hang.

Besides, I want to thank my whole family who support and encourage me to throughout this project. They always comfort me whenever I am facing with difficulties and challenges. Lastly, I would like to thanks my friends who have given me advices and suggestions on this project.

## **ABSTRACT**

This project aims to develop an intelligent surveillance system that can recognize the robbery activities occurring in the front yard of a landed house using a security camera positioned at the place of interest. The focus of this project is to study and understand the pattern of human movement in robbery in order to implement a suitable computer vision algorithm for real-time robbery detection in the front yard of a landed house.

The project is developed with the following technique: YOLO object detection, contour tracking based on temporal subtraction, and motion analysis. First YOLO will start to detect human and car present in the video frame and store its location. When a car is detected, the car stationary time starts to measure. After the initial human position is stored, the subsequent human position is then detected with simple temporal subtraction to reduce the computational resources. The process of tracking the contour is continue until no human motion is detected for the next 10 frames. All of the detected human movement will be drawn as an arrow to indicate the direction of human movement. Analysis is done by calculate the car stationary time and the number of human movements in entrance and car region. Warning is flagged is the probability of human movement in entrance and car region is high and the car stationary time exceed 25 seconds.

## TABLE OF CONTENTS

<b>TITLE PAGE</b>	<b>i</b>
<b>DECLARATION OF ORIGINALITY</b>	<b>ii</b>
<b>ACKNOWLEDGEMENT</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>iv</b>
<b>TABLE OF CONTENTS</b>	<b>v</b>
<b>LIST OF TABLES</b>	<b>viii</b>
<b>LIST OF FIGURES</b>	<b>viii</b>
<b>LIST OF ABBREVIATION</b>	<b>x</b>
<b>CHAPTER 1 INTRODUCTION</b>	<b>1</b>
1.1 Problem Statement	1
1.2 Motivation	1
1.3 Project Scope	2
1.4 Project Objectives	2
1.5 Impact, Significance, and Contribution	3
1.6 Background Information	3
1.7 Proposed Approach	4
1.8 Report Organization	5
<b>CHAPTER 2 LITERATURE REVIEW</b>	<b>7</b>
2.1 Overview of Activity Recognition	7
2.2 Global Representation	9
2.2.1 Background Subtraction	9
2.2.2 Temporal Differencing	10
2.2.3 Optical Flow	10
2.2.4 2D Silhouettes and Shape	11
2.2.5 3D Space-Time Volumes (STVs)	11
2.3 Local Representation	12
2.3.1 Interest Point Detector	12
2.3.2 Local Descriptor	13

2.4	Activity Classification	13
2.4.1	Template-based Approaches	13
2.4.1.1	Dynamic Time Warping	14
2.4.2	Generative Models	14
2.4.2.1	Hidden Markov Model Approach	14
2.4.2.2	Dynamic Bayesian Networks	15
2.4.3	Discriminative Models	16
2.4.3.1	Support Vector Machine	16
2.4.3.2	Conditional Random Field	17
2.4.3.3	Deep Learning Architectures	17
<b>CHAPTER 3 System Design</b>		<b>18</b>
3.1	Real Case Scenario	18
3.1.1	Case A	18
3.1.2	Case B	19
3.1.3	Case C	20
3.1.4	Case Discussion	21
3.2	Design Specification	21
3.2.1	Methodologies	21
3.2.2	Tools to use	22
3.3	System Flow Diagram	23
3.4	Assumptions	23
3.5	Implementation Issues and Challenges	24
3.5.1	Real World Complexity	24
3.5.2	Human Motion	24
<b>CHAPTER 4 System Implementation</b>		<b>25</b>
4.1	Pre-process Video Frame	25
4.2	System Initialization	26
4.3	Information Extraction	27
4.4	Information Tracking	28
4.5	Information Analysis	29



4.6 Decision Making	31
<b>CHAPTER 5 Experimental Results and Discussion</b>	<b>32</b>
5.1 Case A Result	32
5.2 Case B Result	33
5.3 Case C Result	34
<b>CHAPTER 6 Conclusion</b>	<b>36</b>
<b>BIBLIOGRAPHY</b>	<b>37</b>
<b>APPENDICES</b>	<b>A-1</b>
A-1 POSTER	A-1
A-2 PLAGIARISM CHECK RESULT	A-2

## LIST OF TABLES

<b>Table Number</b>	<b>Title</b>	<b>Page</b>
Table 3.1	Tools to use	24

## LIST OF FIGURES

<b>Figure Number</b>	<b>Title</b>	<b>Page</b>
Figure 1.1	System Overview	5
Figure 2.1	Hierarchical Level of Human Activity	7
Figure 2.2	Activity Recognition Approaches	8
Figure 2.3	Global Representation Overview	9
Figure 2.4	Dynamic Bayesian Network (Luo et al, 2003)	16
Figure 3.1	Video Sequence of Case A	19
Figure 3.2	Video Sequence of Case B	20
Figure 3.3	Video Sequence of Case C	21
Figure 3.4	System Overview	22
Figure 3.5	System Flow Diagram	24
Figure 4.1	Pre-process of Frame	26
Figure 4.2	System Initialization	27
Figure 4.3	Information Extraction	28
Figure 4.4	Information Tracking	29
Figure 4.5	Coordinate System	30
Figure 4.6	Direction of Human Movement	31
Figure 4.7	Decision Making	32
Figure 5.1.1	Case A Result 1	33
Figure 5.1.2	Case A Result 2	33
Figure 5.1.3	Case A Result 3	33
Figure 5.2.1	Case B Result 1	34
Figure 5.2.2	Case B Result 2	34
Figure 5.2.3	Case B Result 3	35
Figure 5.3.1	Case C Result 1	35
Figure 5.3.2	Case C Result 2	36



## LIST OF ABBREVIATIONS

<i>CCTV</i>	Closed-circuit Television
<i>HAR</i>	Human Activity Recognition
<i>ROI</i>	Region of Interest
<i>GMM</i>	Gaussian Mixture Model
<i>EM</i>	Expectation-Maximization Algorithm
<i>LKT</i>	Lucas-Kanade-Tomasi
<i>MEI</i>	Motion Energy Image
<i>MHI</i>	Motion History Image
<i>STV</i>	Space-time Volume
<i>STIP</i>	Space-Time Interest Points
<i>SIFT</i>	Scale-Invariant Feature Transform
<i>SURF</i>	Speeded-Up Robust Features
<i>DTW</i>	Dynamic Time Warping
<i>HMM</i>	Hidden Markov Model
<i>DBN</i>	Dynamic Bayesian Network
<i>CHMM</i>	Coupled Hidden Markov Model
<i>HSMM</i>	Hidden Semi-Markov Model
<i>VT-HMM</i>	Variable Transition Hidden Markov Model
<i>S-HSMM</i>	Switching Hidden Semi-Markov Model
<i>SVM</i>	Support Vector Machine
<i>CRF</i>	Conditional Random Field
<i>CNN/ConvNets</i>	Convolutional Neural Network
<i>DNN</i>	Deep Neural Network
<i>RNN</i>	Recurrent Neural Network
<i>LSTM</i>	Long Short-Term Memory
<i>R-CNN</i>	Region-based Convolutional Neural Network
<i>RPN</i>	Region Proposal Network
<i>HOMO</i>	Histogram of Optical Flow Magnitude and Orientation

## **Chapter 1: Introduction**

### **1.1 Problem Statement**

In present day, the closed-circuit television (CCTV) system is commonly adopted by most of the household to monitor activities in the front yard as a crime prevention approach. Robbery is one of the crimes that threaten the safety of the household in residential areas. A surveillance system, equipped with artificial intelligence that can detect robbery scene, will further ensure the safety in housing areas and reduce the occurrence of these crimes.

However, traditional CCTVs are not embedded with artificial intelligence and they are only used to record activities happening in the real-time. In other words, it lacks the ability to analyze the activities captured in the frame to detect suspicious activities such as robbery. As such, traditional CCTVs are generally used for forensic purpose after a certain incident has taken place.

Other approaches such as motion sensor is susceptible to high false alarm rate as any kind of motion detected will trigger the alarm. This kind of sensor failed to distinguish pet activity from the human movement. Unlike smart CCTVs that can tell the difference between pet and human movement. House owner also required to spend a lot of money on purchasing multiple of sensors of every corner of the front yard to detect the motion as opposed to CCTVs approach that required only one camera that facing the gate of the front yard.

### **1.2 Motivation**

Traditional video surveillance is introduced to solve the security issues in the residential areas. However, it lacks the ability to recognize the crime activities in the real-time, to detect the human movement, to analyze the human movement as well as to flag a warning when robbery happens in the premises.

In traditional manual monitoring of CCTV, the security personnel are obliged to concurrently watch over numerous monitoring screens in real time. Therefore, the chances of missing out the crime activities is quite high as these kinds of events normally occur within split second. With the emergence of human action recognition (HAR) for video surveillance, automatic real-time robbery detector can assist the

security personnel to monitor over the huge number of surveillance cameras without missing out the sign of robbery by flagging a warning to the critical region instantly when a potential robbery happened. Thus, an intelligent surveillance system will address the problem of impracticable ratio of cameras to human supervisors.

In general, the frequency of occurrence of anomalous events such robbery is much lower relative to daily activities performed by the household. As such, the surplus of security personnel in monitoring surveillance cameras will result in wasting of money and time resources. Thus, an intelligent surveillance system is in need to curb such wastage.

Moreover, inefficiency of motion-based sensor in detecting suspicious human movement drives the invention of intelligent CCTVs that able to analyze the pattern of human motion in the video in order to detect robbery activities that happen in the front yard.

### **1.3 Project Scope**

This project aims to develop an intelligent surveillance system that can recognize the characteristics of robbery occurring in the front yard of a landed house using a security camera positioned at the place of interest. Robbery are high-level concept and diverse events in the real world, so this project makes some assumptions on the following:

- Robbery detection are based on analysis of human movement in the front yard area
- Pattern of human movement are learned from the studies on footage of robbery activities
- A warning is flagged when suspicious human movement is detected in the front yard

The focus of this project is to study and understand the pattern of human movement in robbery in order to implement a suitable computer vision algorithm for real-time robbery detection in the front yard of a landed house.

### **1.4 Project Objectives**

1. To identify the characteristics of robbery from the footage

- The characteristics will reflect the accuracy of detecting the robbery
  - The characteristics refer to the attributes that distinguish between normal activities and potential robbery
  - Characteristics of robbery are referred to home break-ins and repeating human movement from gate to car or vice versa
2. To develop an automated robbery detection surveillance system
    - The system will be able to detect robbery in the front yard automatically
  3. To develop a real-time surveillance system
    - The system aims to detect the robbery activities in real-time based on the characteristics of the crime activities

### **1.5 Impact, Significance, and Contribution**

In this project, an intelligent surveillance system is developed to automatically recognize robbery crimes that happen in the front yard area of a landed house. The system developed will help the victims of the crime to get help from authorities instantly by raising a warning when these activities are detected. As such, it reduces the fatality or injury rate of the victims and increases the chance of arresting the robbers.

Besides, this system can analyze the motion pattern of the robber and detect the sign of robber activities. The behaviour of the robbers such as repeated movement between gate and car and home breaking-in are the early signs of these crimes. By recognizing these signs, the system can flag a warning to the authorities when robbery takes place.

With the intelligent surveillance system, the security personnel can direct their focus to the critical region which the robbery might take place. Therefore, this will reduce the chance of missing out the signs of crime and allows the authorities to act promptly when these crimes happened. Besides, it will also help to reduce the human supervision on real-time security camera and consequently avoid the wastage of human resources.

### **1.6 Background Information**



Video surveillance system (CCTV) was introduced to monitor the activities that happen in the place of interest and send the signal to monitoring screen. Generally, a few human operators will keep track of all the CCTV screens simultaneously in the real time to detect criminal activities. However, with the growing amount CCTV, human operators will miss out the crime activities due to the drop of attention over the long hours of monitoring the same scene. Furthermore, detecting crime activities are challenging due to their rarity. (Amira & Ezzeddine 2017)

With the emergence of human action recognition for video surveillance, automatic detection of robbery is made possible. Human action recognition (HAR) is a way of deriving information from the action of interest subjects and the scene to recognize the activities captured by the camera. HAR can be done with the advances of image representation approaches and classification methods. Image representation starts from global representation to local representation and more recently on depth-based approaches. For classification methods, there are template-based approaches, generative models and discriminative models. Recent studies focus more on artificial neural network and deep learning.

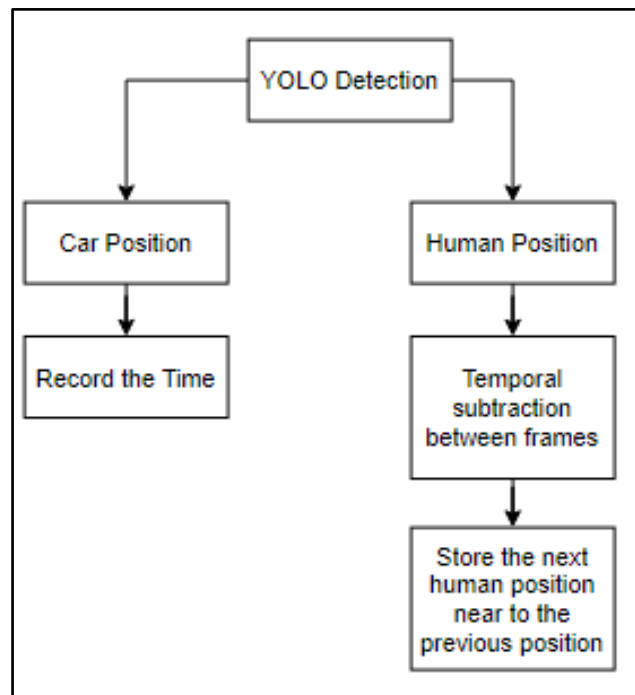
Despite the attention on this topic, most of the previous works emphasized on high-level concept of general events such as violent and abnormal activities without defining a specific area of those activities. In this paper, the scope is narrowed to only detecting robbery in the front yard area of the house as personal safety in a residential area is concerned by most of the urban populations. Since robbery are considered as rare events that does not occur on a daily basis, it can be difficult to detect robbery scene through the manual monitoring of traditional CCTV. Therefore, an intelligent video surveillance system is more reliable in detecting these crime activities in the front yard of the house.

### **1.7 Proposed Approach**

There are two important aspects needed in order to detect robbery:

- Car Stationary Time: Duration of car stop at the front yard
- Human Movement: Spatial and temporal information regards to the human movement

Both aspects will be combined to analyse the whether a potential robbery has happened in the surveillance video.



**Figure 1.1 System Overview**

In this section, only an overview on how the to extract the aspects from the video is discussed. Chapter 3 and 4 will further discuss the design and implementation of the system. The system starts with the YOLO detection and store the pixel position of the car and the first appearance of the human in a particular frame of a video input. When a car is detected in a frame of the video, the system will start to record time. The car stationary time increase when the car is detected again in the future frame of the video. If there is a human detected in the frame, then the system will store human position as initial position and do the temporal subtraction on sobel edge of current and previous frame. The resultant of the subtraction will get dilated to combine the nearby contour of the image in order to reduce the insignificant motion between the previous and current frame. The next human position is stored if the position is close to the initial human position.

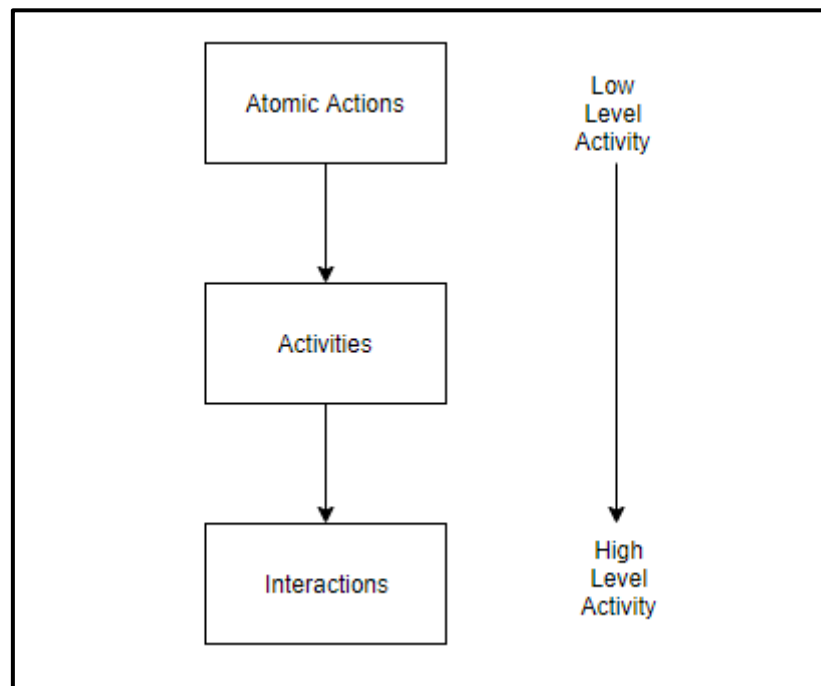
## **1.8 Report Organization**

## Chapter 1: Introduction

This report consists of 6 chapters that starts with the introduction that includes the problem statement, motivation, project scope and objectives, background information and proposed approach. Chapter 2 is presented with literature review. For chapter 3, the design of the system is laid out to describe how to develop this project. Diagram for top-down design of the system will be provided in this chapter. Chapter 4 will provide the implementation of the system and for the subsequent chapter will discuss about the experimental result of the system. Last chapter will give a conclusion for the whole project.

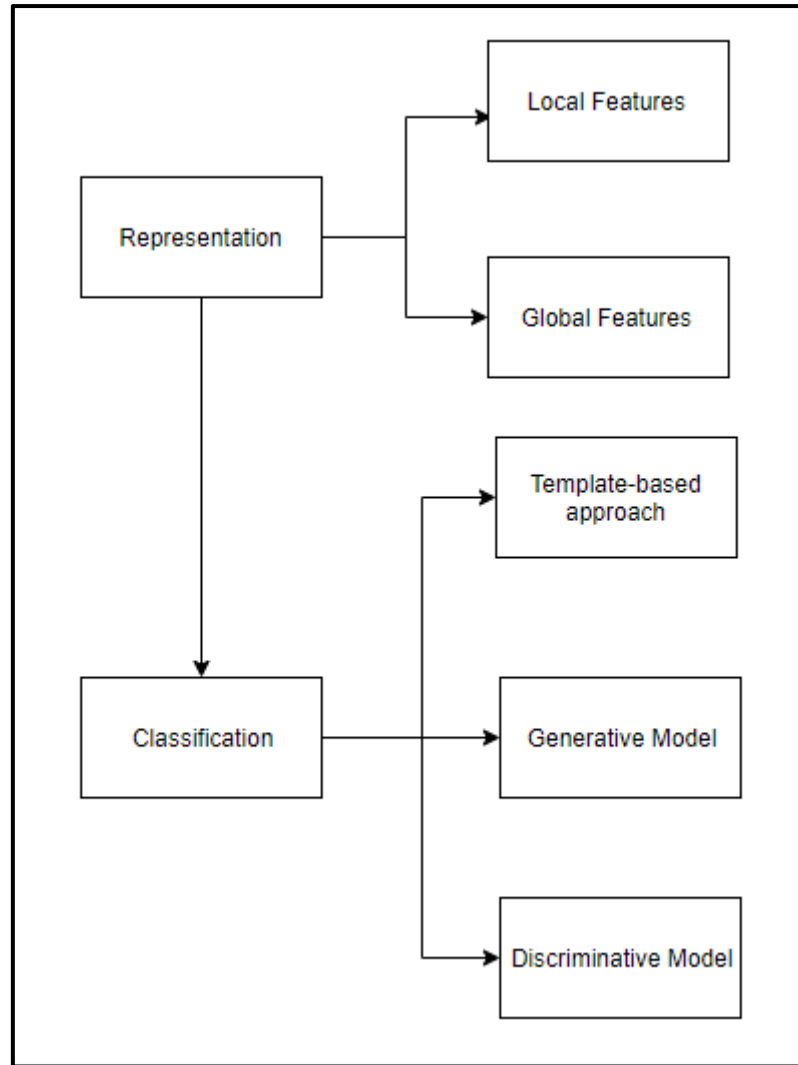
## **Chapter 2: Literature Review**

### **2.1 Overview of Activity Recognition**



**Figure 2.1 Hierarchical Level of Human Activity**

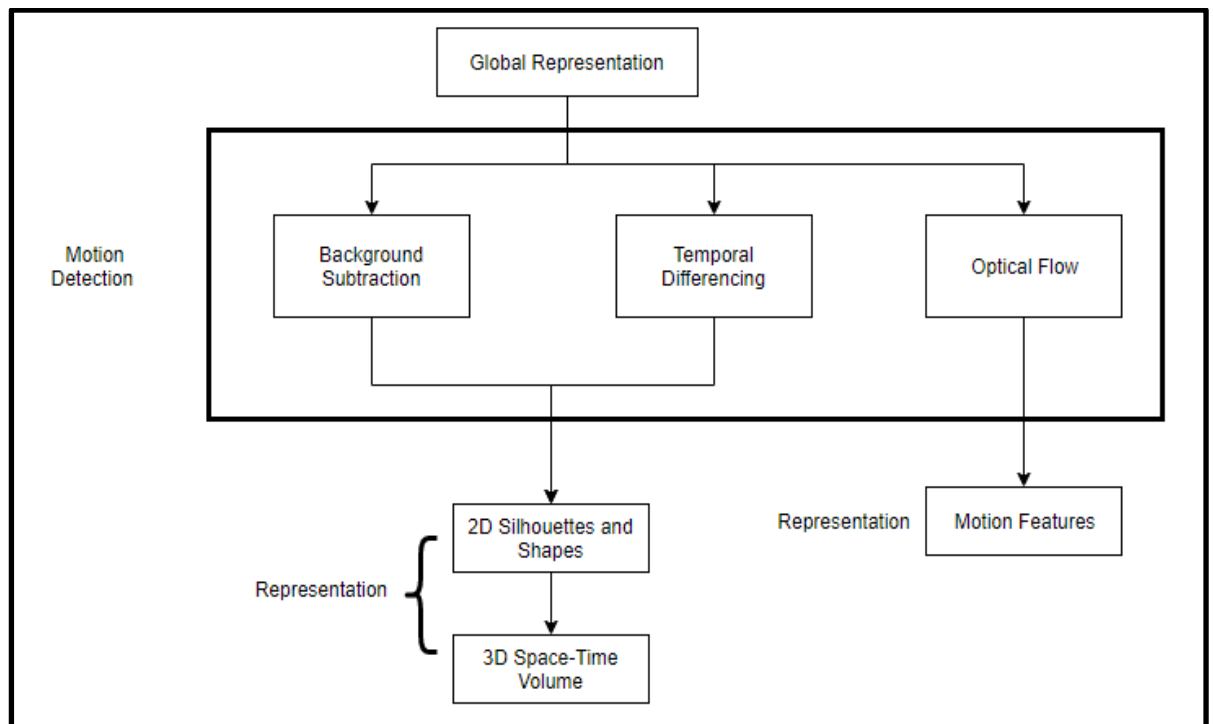
In Zhang et al. (2017) paper, human activities are classified into 3 hierarchical levels which are atomic action, activities, and interactions. Atomic action refers to the movement of specific human body part such as raising hand whereas activities means whole-body movement that composed of multiple atomic actions performed by a single person only in chronological order. Jogging, walking, jumping are example of human activities. The highest level is human interaction. Human interaction can be defined as two or more person/object performing activities. Human-human interaction and human-object interaction are the category of human interaction. For example, interaction can refer to the fight between two person or a person who is sweeping the floor.



**Figure 2.2 Activity Recognition Approaches**

Human activity recognition in video are divided into representation and classification phase. Representation phase are needed to sort out the region of interest from the rest of the image to save the computational cost. Region of interest can be extracted globally or locally. The global and local features are then classified into different activities based on template-based approaches, generative models and discriminative models to be described below.

## 2.2 Global Representation



**Figure 2.3 Global Representation Overview**

Global representation describes the image/video as a whole feature using the global descriptor extracted from the image/video. In this representation, moving objects are detected using background subtraction, statistical method or temporal differencing to form 2D silhouettes and 3D Space-Time Volume for global features. Other global approach such as optical flow is also used to extract and describe the motion features of the silhouettes.

### 2.2.1 Background Subtraction

The idea behind the background subtraction is to extract the foreground from the background before proceeding to recognize human activities in video. In this context, the foreground refers to the region of interest that can be used to represent the global information of the image. This approach is commonly used to detect foreground with static background by calculate the pixel difference between the current image with reference background image. A background model is needed to extract the silhouettes (ROI).

In order to calculate the background model, Wren et al. (1997) introduce Gaussian distribution that calculate the average and standard deviation of each pixel

value for last n frame. A gaussian probability density function will then use to determine which pixel that fall into range of background. Using this approach, some extracted foreground values are unnecessary because the average calculation of foreground value includes the previous foreground pixel. Koller et al. (1994) proposed a selective background update strategy to overcome this problem.

Gaussian mixture model (GMM) is proposed by Stauffer and Grimson (1999) to use the mixture of Gaussians to model the value of each pixel with the expectation maximization (EM) algorithm. Pixel value with high probability in GMM indicate the pixel belong to background. Due to high computational cost of EM algorithm, k-mean clustering algorithm is adopted with slightly sacrifice on accuracy.

Statistical methods use the statistical approach on individual or a group of pixels to construct more advanced background model. Background statistics is used to determine the foreground by matching every pixel of current image with the statistics of background model. This method is more robust to noise, shadow and lighting conditions. (Stauffer & Grimson, 1999).

### **2.2.2 Temporal Differencing**

Temporal differencing approach is suitable to extract moving foreground from dynamic background as it calculates the difference between two or three consecutive frames in pixel-by-pixel manner. Pixel changes due to noise will affect the accuracy of foreground extraction. Due to this, a threshold is applied on the difference of image. (Sarvesh & Agrawal, 2012) However, this approach performs badly in extracting all relevant pixels. (Ko, 2011)

### **2.1.3 Optical Flow**

The notion of point between images is detected and described by optical flow. The optical flow is commonly obtained using the Lucas-Kanade-Tomasi (LKT) feature tracker. Lu, et al (2004) proposed a method that tracks the human joints in key frames as well as actual frames using LKT feature tracker to encode a specific posture. Every key posture is encoded in key frame, and a key posture sequence refers to an activity. Unique posture recorded in actual frame can be detected by identifying the posture similarity between key and actual frame. Subsequently, the posture sequence in actual frame is mapped with key posture sequence by matching the body locations and confirm the activity in actual frame.

#### **2.2.4 2D Silhouettes and Shape**

The silhouettes of the detected objects are stacked to construct a binary motion-energy image (MEI) that show the area of motion in a sequence of images and a motion-history image (MHI) that tell the temporal history of motion at the location. Bobick and Davis (2001) use these temporal templates to represent the human movement and compare temporal templates against predefined instances to recognize the human activity.

#### **2.2.5 3D Space-Time Volumes (STVs)**

A sequence of images forms an activity in a video. By stacking silhouettes of an image along the time axis, a three dimensions shape (space-time volume) is formed. 3D space-time volume consists of two spatial dimension which are X and Y pixel location as well as a temporal dimension T. Blank, et al., (2005) first extend the MEI templates over the time axis in order to produce a 3D shape to represent an action; however, the resulting 3D space-time shape cannot be analysed due to the its nonrigidity shape and contrast between space and time dimensions. To overcome this shortness, Ke, et al. (2007) use mean shift clustering approach to segment the input video into space-time volume. The authors use their proposed shape-matching technique on the over-segmented area (super-voxels), so they do not need to depend on background model for representing shape.



## **2.3 Local Representation**

### **2.3.1 Interest Point Detector**

Local representation refers to the process of finding interest points that contain rich motion and temporal information. These interest points are more robust to noise and occlusions when compared with global representation. In 2005, Laptev build a space-time interest points detector (STIP) from Harris corner detector (Harris and Stephens, 1988) by taking temporal information into consideration. Harris corner detector is introduced to find the features in an image with large variation in intensity when the sliding window moves slightly in any direction. The STIP or know as 3D-Harris corner detector; however, need spatial-temporal information to identify interest point that yield significance spatial changes and non-constant motion.

Lowe (1999) proposed the scale-invariant feature transform (SIFT) to find the interest points that are invariant to rotation and scale. SIFT detector will detect the scale-space extrema by search over several scale and image locations. Convolution of a Gaussian kernel at different scales are apply on the input image to find the scale space of the image function  $L(x,y,\sigma)$ . Maxima and minima of difference of Gaussian in scale space is determined to find the key points candidates. Then, a detailed fit is performing to surrounding data to get the location and scale. The interest points are selected based on its stability.

Willems et al (2008) suggested an efficient way to detect scale-invariant interest points by extending 2D Hessian detector to 3D interest point detector. Determinant of the 3D Hessian matrix is used to select the scale and localize the point of the image. In this way, iterative scheme is eliminated to save the computation time. An improvement on efficiency is done using speeded-up robust features (SURF) to detect the interest points invariant to scale and rotation. Firstly, input image is analysed at various scale to ensure robustness to the changing scale. Then, Hessian-matrix approximation use the integral image for image convolutions to detect the interest points.

Dollar et al (2005) pointed out that 3D detectors face difficulty to detect motion lack of true spatiotemporal corners despite the motion is significant and happening. The authors (Dollar et al, 2005) proposed an alternative spatiotemporal feature detector to address this problem. The idea of the detector is to use response function with two separate filters in spatial and temporal which is a 2D Gaussian smoothing kernel for

spatial dimensions as well as 1D Gabor filters for temporal dimensions. Generally, the response function gives a positive response to any area of complex motion that have the unique spatial characteristics.

### **2.3.2 Local Descriptor**

Local descriptor is used to describe the patches around or at the interest points. Laptev (2005) calculate local spatiotemporal N-jets which is descriptor for 3D Harris corner detector. a normalized spatiotemporal Laplacian operator is fully maximized by the descriptor on spatial and temporal scale to determine the spatiotemporal extends of detected events. The descriptor is said to be robust to cluttered background and occlusion.

In 2007, Scovanner, et al. proposed a 3D SIFT descriptor by extending the 2D gradient magnitude and orientation to 3D, resulting the sub-histograms that encoded the 3D SIFT descriptor. This descriptor uses spatiotemporal words to describe the video. Besides, Willem et al, (2008) applied Haar-wavelet responses for both the 2D and 3D SURF descriptor; however, 3D SURF descriptor stores the vector of the 3 axis response and avoid the inclusion of the total sums over the absolute value as it will increase the size of the descriptor to double.

## **2.4 Activity Classification**

### **2.4.1 Template-based Approaches**

One of the early significant works on action recognition is temporal template introduced by Bobick and Davis (2001). Temporal motion information in an image is represented by temporal template. Temporal template consists of motion energy image and motion history image. Motion energy image (MEI) is a binary cumulative motion image that shows the area of motion in an image while motion history image (MHI) describes the motion in the image as it progresses. MHI template will shows the more recent motion with respect to higher intensities.

In 2007, Shechtman and Irani come up with the similar idea that compare two different video segments of different space-time intensity to detect human behaviour in video sequences. Firstly, intensity information of different activities is built from small video clip. The intensity information is encoded in the form of 3D space-time intensity template. Entire video sequences will then compare with the template in all 3

dimensions. The similarity comparison is done by splitting up both video and template into tinier patch units from the full video segments. The local consistency in between the small patches is then calculated. This approach is able to detect few different activities simultaneously and robust to slightly changes in scale and orientation.

#### **2.4.1.1 Dynamic Time Warping**

Human activities are formed through a series of key frames. One of the dynamic programming algorithms known as dynamic time warping (DTW) can be used to match the sequence of frames with variance. Darrell and Pentland (1993) use a set of view model to represent gestures. The authors then apply the dynamic time warping that calculate the means as well as the variations of correlation scores of the images frames and view models to match the gesture template.

Veeraraghavan et al (2005) proposed a nonparametric model that based on dynamic time warping algorithm (DTW) for recognition of human gait pattern. The authors adjust the DTW algorithm by including the non-Euclidean space where the deformation of shape occurs. The result outcome shows the modified algorithm is more efficiency for the human gait recognition. DTW algorithm requires only a small number of training samples; however, the complexity in computation elevate significantly when activity types increase.

#### **2.4.2 Generative Models**

Generative model makes use of statistics and probability to create the joint probability  $P(X,Y)$  where  $X$  is inputs and  $Y$  is label. This model will calculate the  $P(Y | X)$  using Bayes rules to find out the most likelihood label  $Y$  given  $X$  inputs. The most common algorithms for generative model are Hidden Markov Model (HMM) and Dynamic Bayesian Networks (DBN).

##### **2.4.2.1 Hidden Markov Model Approach**

Hidden Markov Models were originally introduced to overcome the speech recognition problem. In 1992, Yamato et al proposed to use HMM for human action recognition. A set of time-sequential image frames is converted into feature vector sequence before further transform into symbol sequence using vector quantization. Then, the parameters of HMMs is trained to recognize the human activity categories

from the training sequence. The input sequence of human activity is categorized using the HMM model that best suits the sequence.

Since HMM only has single state variable, it faces the difficulties to structure human interactions. Oliver et al introduce coupled HMM (CHMM) to model interaction between people by building two HMMs for two different agents and specifies the probabilities among the hidden states. Besides, the time that consumed on the state is unknown in the HMM duration model. This means that the longer the length of time interval, the less likely a state will be observed for the time interval.

Hidden semi-Markov model (HSMM) as well as the variable transition HMMs (VT-HMM) were proposed to solve the decay duration for state problem. The HSMM equipped with explicit duration model that have specific distribution makes it an ideal choice to solve the mentioned problem. Duong et al (2005) proposed switching hidden semi-Markov model (S-HSMM) that come with the properties of inherent hierarchical structure as well as explicit duration model. High-level and low-level activities are represented individually in two layers of S-HSMM.

#### **2.4.2.2 Dynamic Bayesian Networks**

A dynamic Bayesian network is a Bayesian network that extended with the capability of modelling influences over time axis. State space of DBN is not limited to one random variable unlike HMM. As such, HMM is a simplified version of DBN that comes with restricted number of random variables as well as predetermined graph structure.

In 2010, Suk et al. introduced this dynamic Bayesian network for the recognition of two hands gesture. This DBN represent movements of two hands as well as their spatial information with three hidden variables denoted as square nodes ( $X^1, X^2, X^3$ ) whereas there are five observable variables represented as circle nodes. The five observation included the motion of two hands, position of each hand in relative to face and the spatial relation between hands. Then, the authors make use of first-order Markov assumptions to simplify the DBN structure.

Park and Aggarwal (2004) work on recognizing two-person interactions using hierarchical Bayesian network. Firstly, the body parts are extracted from segmentation system and body part pose is estimated at low level of Bayesian network. Next, all the individual Bayesian network arrange hierarchically to predict or estimate a person's

entire body poses. Finally, a sequence with DBN is formed by combining post estimation results of two-person interactions.

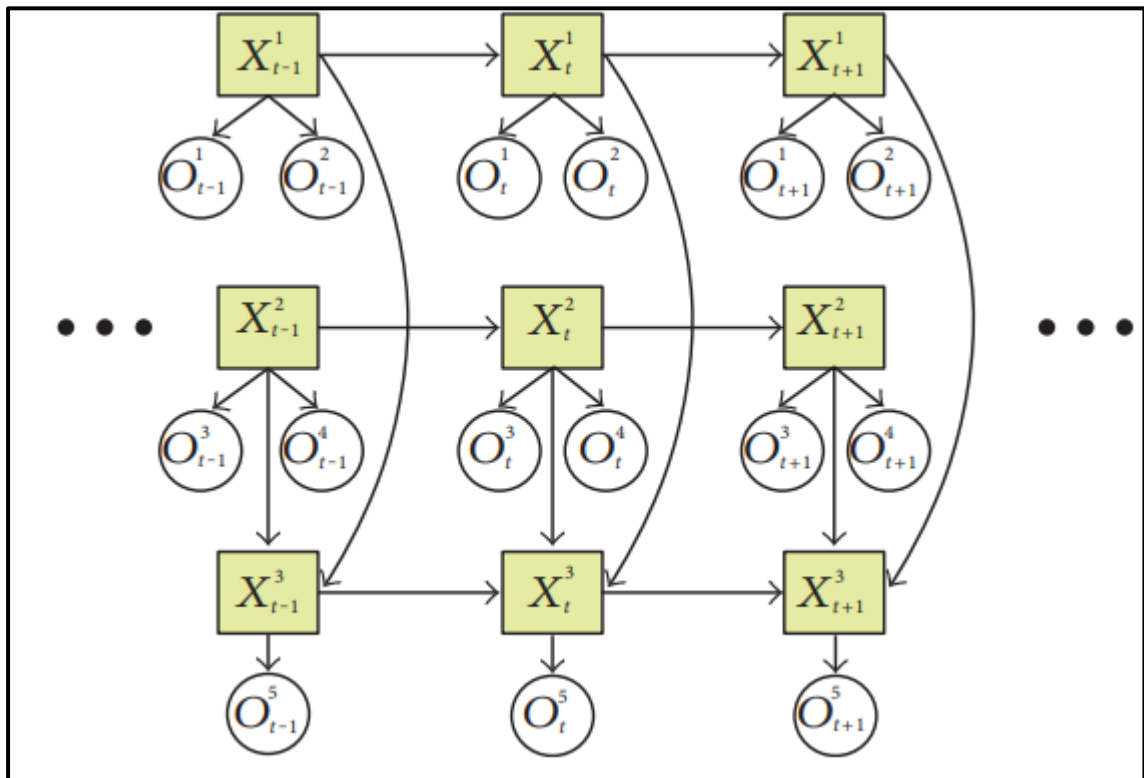


Figure 2.4 Dynamic Bayesian Network (Luo, et al, 2003)

### 2.4.3 Discriminative Models

#### 2.4.3.1 Support Vector Machine

Support Vector Machine (SVMs) are one of the discriminative classifiers used popularly in human activity recognition (HAR). SVM was originally introduced by Vapnik et al (1997) to classify instances into different classes. This is done by determining the hyperplane that has maximum distance to data points of any classes. A local space-time features was combined with SVM for human activity recognition (HAR). (Schuldt et al., 2004). The authors record a video dataset (KTH dataset) for the benchmark of human activity recognition (HAR). Laptev et al (2008) applied multi-channel nonlinear SVMs on KTH dataset and achieve accuracy of 91.8%.

### **2.4.3.2 Conditional Random Field**

Conditional random field (CRF) is a probabilistic model that define a conditional probability for a label sequence with a particular observation sequence. Vail et al (2007) noticed discriminative CRF has a higher performance than HMM despite the model feature follow the HMM independence assumptions. The authors figured out HMM model will be invalidated when the observations are not independent given their labels whereas CRF model put this independence assumption aside and conditions on whole observation.

### **2.4.3.3 Deep Learning Architectures**

Deep learning architectures can be grouped into convolutional neural network (CNNs or ConvNets), deep neural network (DNNs) and recurrent neural networks (RNNs). The most popular approach is ConvNet. In 2012, Krizhevsky et al train a deep ConvNet with a huge dataset that consists of 15 million of labelled images with over 22,000 categories. The outstanding performance of the ConvNet increase the popularity of its usage in pattern recognition fields. Conventional machine learning approaches extract features manually whereas ConvNets learn the features automatically. Mo et al (2016) make use of ConvNets to extract the feature automatically and classify it with multilayer perceptron.

On the other hand, deep neural network (DNNs) use hand-crafted features rather than learn it automatically from the dataset. Harris corner interest points together with histogram-based features has been applied by Berlin and John (2016) as feature vector. They then proposed to embed a stacked auto encoder into deep neural network for classification of human-human interaction. Human activity generally happened in sequential. Due to this property, RNNs are suitable for recognizing human activity. The most widely used RNNs architectures is long short-term memory (LSTM) because it can keep track the observation and store it in memory. Baccouche et al (2010) train a LSTM-RNN to classify the action in soccer video. The experimental result has a 92% of accuracy when combining the two features of Bag-of-Words and dominant motion.

**Chapter 3: System Design**

**3.1 Real Case Scenarios**

**3.1.1 Case A**



**Figure 3.1 Video Sequence of Case A**

- a) A man appeared in the front yard of the house and walking towards the gate of the house.
- b) The last seen position of the man before entering the house through the gate.
- c) A man walks out from the gate and move the things to the car.
- d) A man walks from the car to the gate to take item.

**3.1.2 Case B**



**Figure 3.2 Video Sequence of Case B**

- a) A man appeared in the front yard of the house after climb over the gate and walking toward the front door.
- b) The last seen position of the man who tries to break in the house through the front door.
- c) A man come out from the car and walking toward the front door.
- d) A man carries an item and move from the front door to the gate.



**3.1.3 Case C**



**Figure 3.3 Video Sequence of Case C**

- a) A man walks from the car and heading toward the door of the front yard.
- b) The last seen position of the man before entering the door.
- c) The man walks out form the door.
- d) The man walks toward the car.

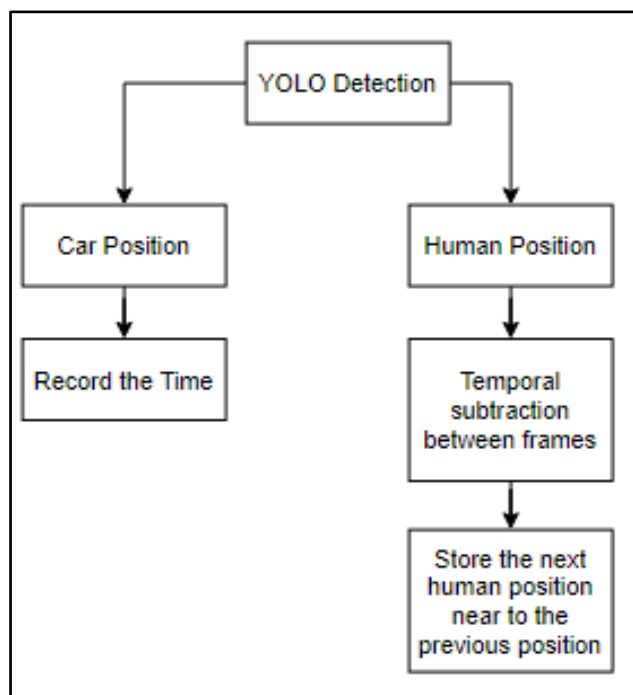
### **3.1.4 Case Discussion**

From above three scenarios:

- A person is first appeared in the front yard and walking towards the entrance before disappear from the video frame.
- A person made repeating movement from the door entrance to the car and from the car to the door entrance.
- By generalizing the mentioned scenario, robbery activities include a person entering the house and made repeating movement from the vehicle to the door entrance and from the door entrance to the vehicle. The long stationary time of the vehicle in the front yard combine with lot of human repeating movement from the vehicle and door entrance and vice versa.
- The alarm should be triggered when such suspicious activities occurs in the front yard of the house.

### **3.2 Design Specification**

#### **3.2.1 Methodologies**



**Figure 3.4 System Overview**

1. You-Only-Look-Once (YOLO) is one of the state-of-art neural network object detectors with high accuracy. YOLO algorithm first divide the input frame into  $S * S$

S grid and tries to find the bounding box of the object that is located in the grid cell. Each of the bounding boxes then assign with box confidence score before a conditional class probability is calculated for the bounding box. Thus, YOLO algorithm required high computational resources.

2. Run YOLO on every frame of the video will take high computational resources, consequently the surveillance system will take long time to extract the information from the video frame. The idea is to reduce the number of YOLO detection to make it possible to extract useful information in the real time.
3. YOLO detection is set to run in every 90 frames (roughly 3 seconds) to detect the object present in the frame. The number of detections can be set to increase the accuracy. For every 90 frames, the system will run the YOLO detection one time to store the position interested object present in the frame.
4. When a car is detected in a frame, the stationary time will start to record. Since the YOLO detection only run for every 90 frames, the stationary time can be calculated based on the consecutive number of times a car is detected in the video. If a car is detected at 90<sup>th</sup> and 180<sup>th</sup> frame then the stationary time is recorded as 90 divided by number video frame per second.
5. When a person is detected in a frame, its pixel position will store in a list. Temporal subtraction is done between the previous and current frame to get the human-like movement and store it in the list. This will continue until no human movement is detected for the next 10 frames. The list will then transfer to another list that keep track of all human movement before it gets emptied.

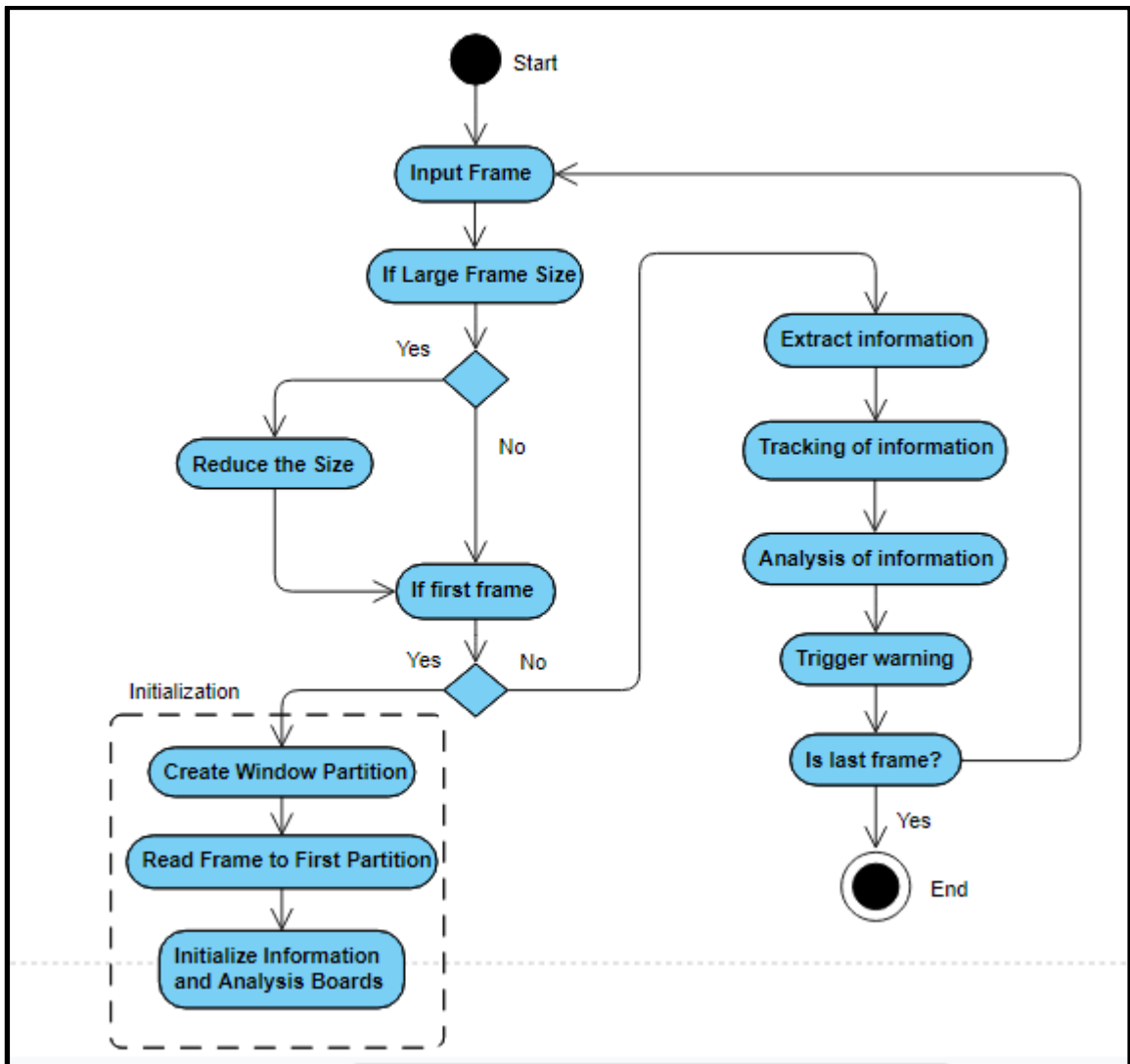
### **3.2.2 Tools to use**

Hardware	Software
Laptop: Dell Inspiron 14	Visual Code
Processor: Intel® Core™ i7-6500U Processor	C++
RAM: 12 GB	Open Source Computer Vision (OpenCV)
Operating System: Window 10	GNU C++ Compiler

**Table 3.1 Tool to use**

### 3.3 System Flow Diagram

The overview process of the system is shown in Figure 3.5. The proposed system consists of few modules such as extract information, tracking of information, analysis of information and decision making to trigger warning. In the next chapter, each of the modules will be described in details.



**Figure 3.5 System Flow Diagram**

### 3.4 Assumptions

1. Direction of the human movement point toward the door and the vehicle most of the time
2. Entrance of the front yard is based on the disappearance of human movement.

3. Long stationary time of vehicle with lot of human movement between the entrance and the car indicate robbery activities.

### **3.5 Implementation Issues and Challenges**

#### **3.5.1 Real World Complexity**

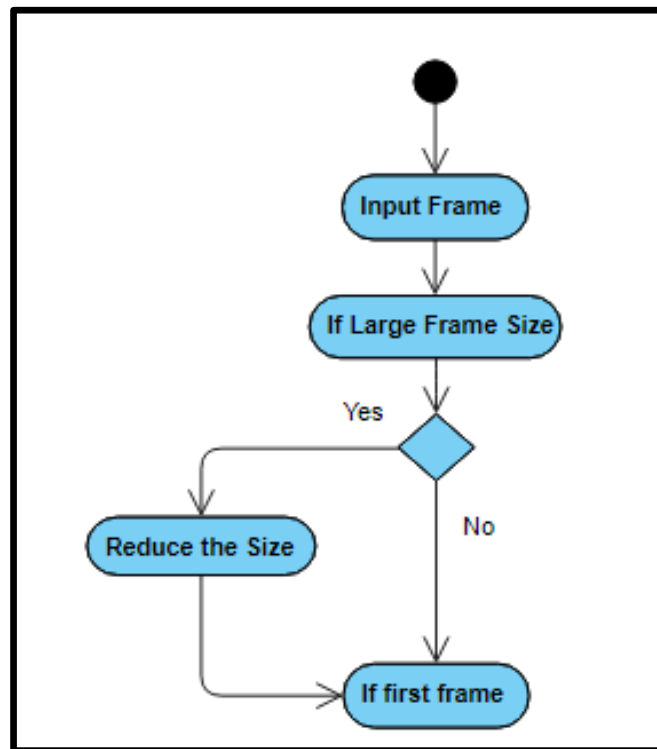
All surveillance videos have different background and are capture at different time. As such, real world videos are prone with variation of brightness, occlusions, different viewpoint of subjects and scale variation. The image processing and recognition tasks become complex when deal with various conditions from unrestricted real world.

#### **3.5.2 Human Motion**

Generalizing human motion from a 3-dimensional video to 2 dimensions will reduce the spatial information present in the video frame. Human motion can be tricky to track when information lost due to the mapping of 3 dimensions to 2 dimensions frame. Moreover, human motion is not always predicted as real world scenarios can varies between cases.

## **Chapter 4: System Implementation**

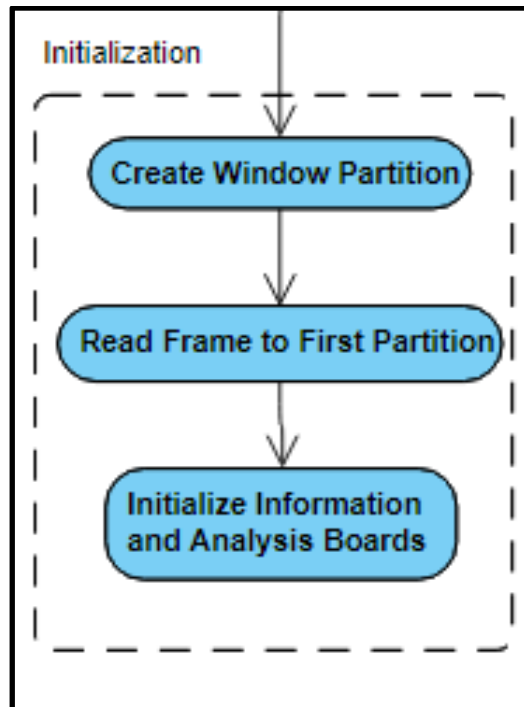
### **4.1 Pre-process of Video Frame**



**Figure 4.1 Pre-process of Frame**

The system starts with reading input frame from the surveillance camera. The system is designed to work with low resolution ranging from 180 x 320 to 240 x 420. All the video frame with the height larger than or equal to 720 will get scale down twice to reduce its size, on the other hand video frame with height larger than or equal to 480 will get scale down once to reduce its size. The frame size reducing is done by scaling down the image frame instead of resizing the image frame to a certain width or height as this will causes image distortion and possible information lost. Pre-processing the image frame is important to ensure the system works correctly.

## **4.2 System Initialization**



**Figure 4.2 System Initialization**

The system is initialized when after the first frame is pre-processed. In this initialization process, a window partition is created based on the pre-processed frame's width and height as well as other properties such as number of colour channels. After that, the frame will display to the first partition of the window and subsequently information and analysis board is initialized to display.

### 4.3 Information Extraction

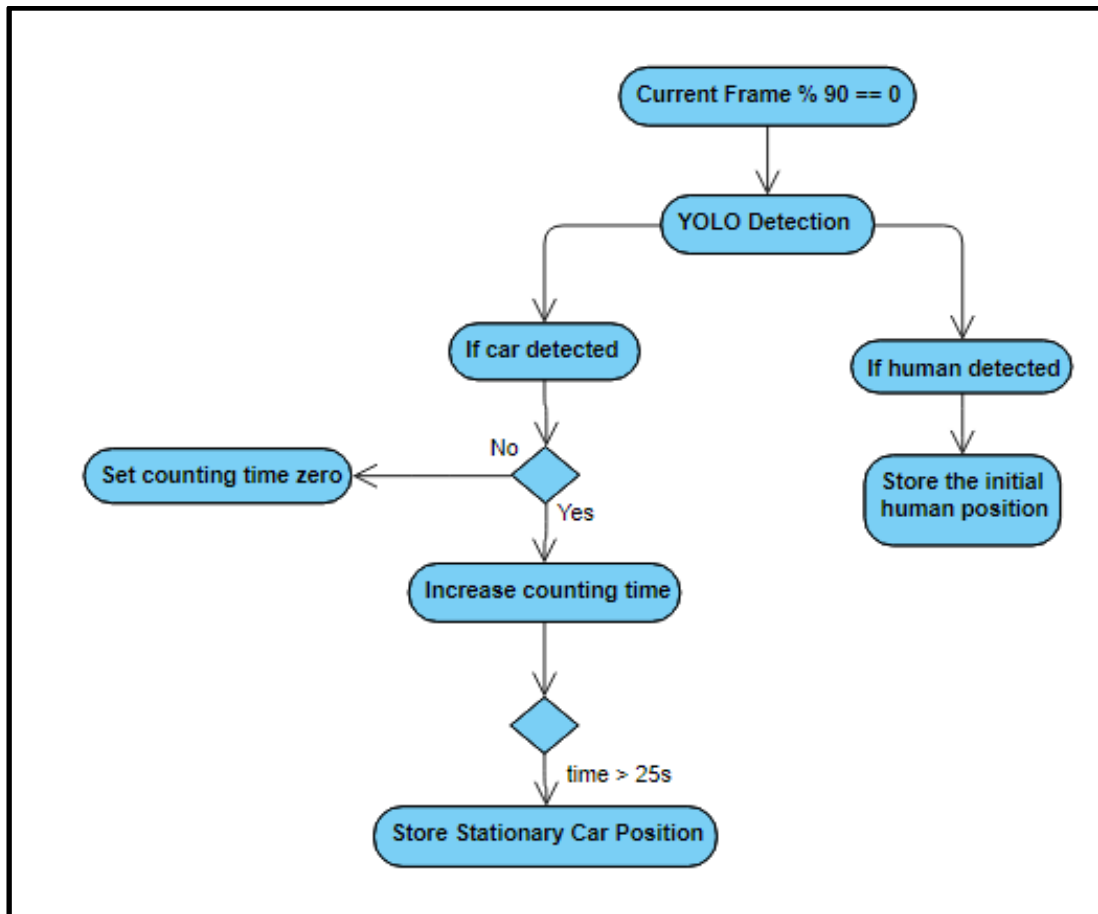


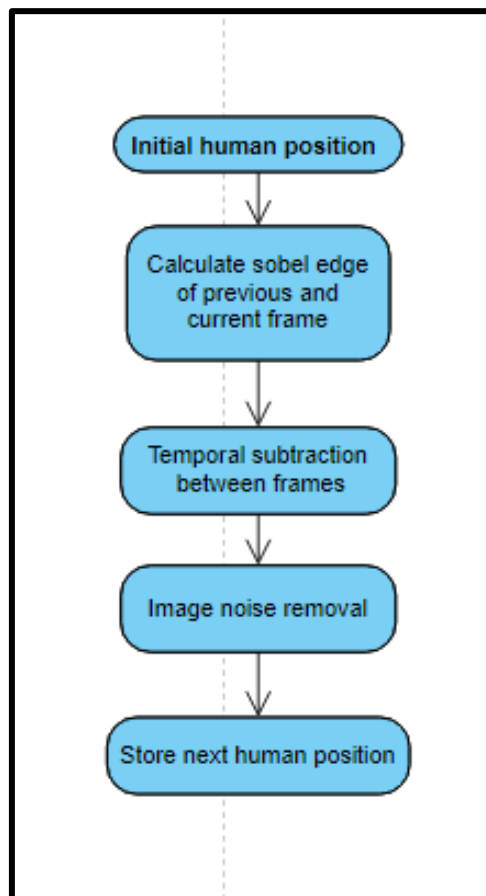
Figure 4.3 Information Extraction

- Information extraction from a video frame usually required high computational resources. You-Only-Look-Once (YOLO) algorithm is chosen to perform object detection in the video frame as it required lower computational resource in relative to other neural network object detectors. However, running this algorithm for every frame in a video is not feasible and wasting the computational resources.
- The idea is to reduce the number of detections performed on the video frame. So, this YOLO algorithm is run once for every 90 frames to detect the interest objects. (current frame % 90 == 0)
- If a car is detected on the particular frame, then the counting time will increase by one. The car stationary time is calculated based on the counting time. Since YOLO algorithm is runs for every 90 frames, then one counting time is roughly equal to 3 seconds. (90 frames divided by video frame captured per seconds).



- If there is no car detected when the next detection is performed, then the stationary time is reset to zero as the car may be passing by in the area. If the stationary time exceed 25 seconds, the system will mark the car as stationary and store its position.
- If a human is detected on the particular frame, then the human position will be stored in a list as initial human position. This initial human position will be treated as starting point of the human in motion.

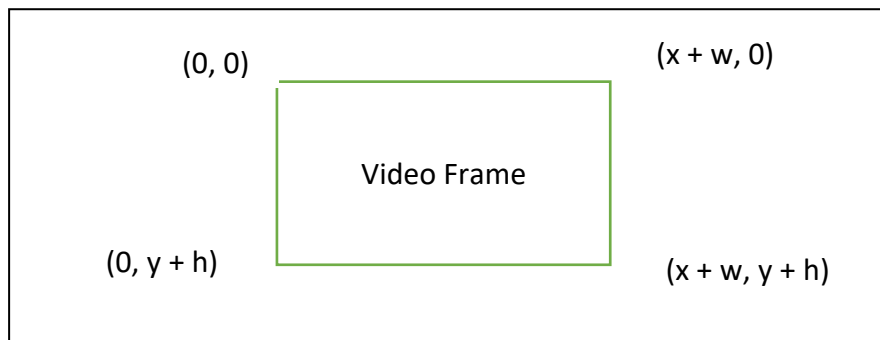
#### **4.4 Information Tracking**



**Figure 4.4 Information Tracking**

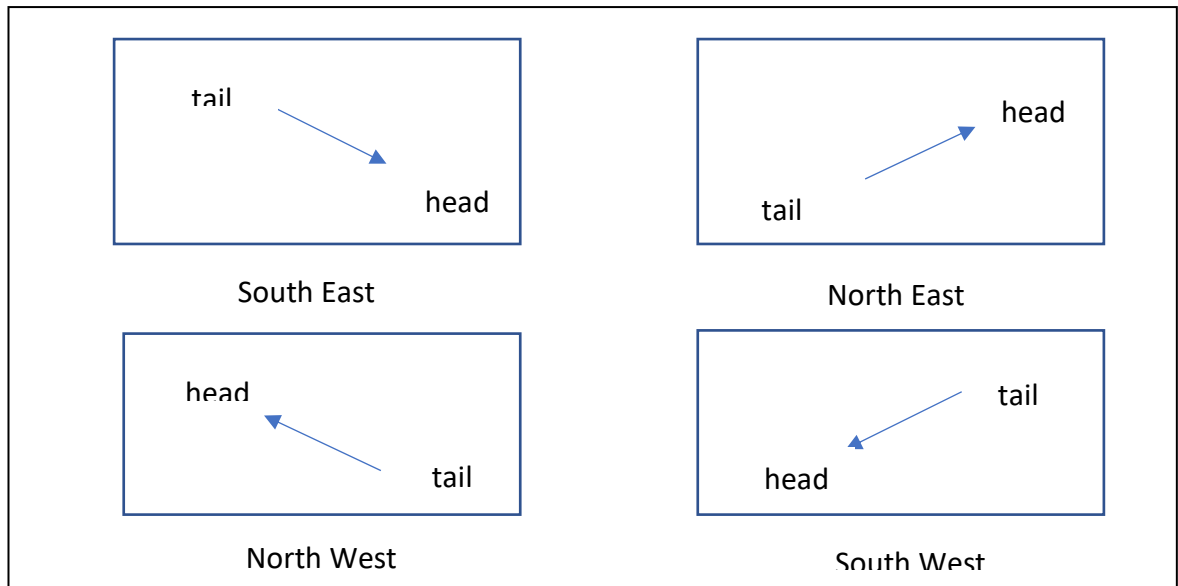
After the initial human position is stored, the system will track for the subsequent human position present in the video frame. This is done by calculating the sobel edge of previous and current frame. Temporal subtraction between the sobel edge of the previous and current frames is dilated to combine the contour that are close to each other's in order to remove the noise such as insignificant motion that present in the video. A bounding box is then used to locate all the possible human contour in the frame. Contour with height larger than its width by 1.5 times will marked as possible human contour. Position of human contour that close to the initial human position/previous human position will push back to the list of human movement. This process will continue until there is no human motion detected for the next 10 frames, then an arrow is drawn and human movement is recorded.

#### **4.5 Information Analysis**



**Figure 4.5 Coordinate System**

Let  $x$  = horizontal value,  $y$  = vertical value,  $w$  = image width and  $h$  = image height. Two-dimensional image is presented by matrix of pixels. The pixel of the image starts at the top left of the image with coordinate of  $(0, 0)$ . The  $x$  value increase when the pixel is moving to the right and the  $y$  value increase when the pixel is moving to the bottom. The  $x$  value of top right pixel is the sum of  $x + w$  and the  $y$  value are zero. The  $x$  value of the bottom right pixel is zero and the  $y$  value is the sum of  $y + h$ . The bottom right coordinate is  $(x + w, y + h)$ .



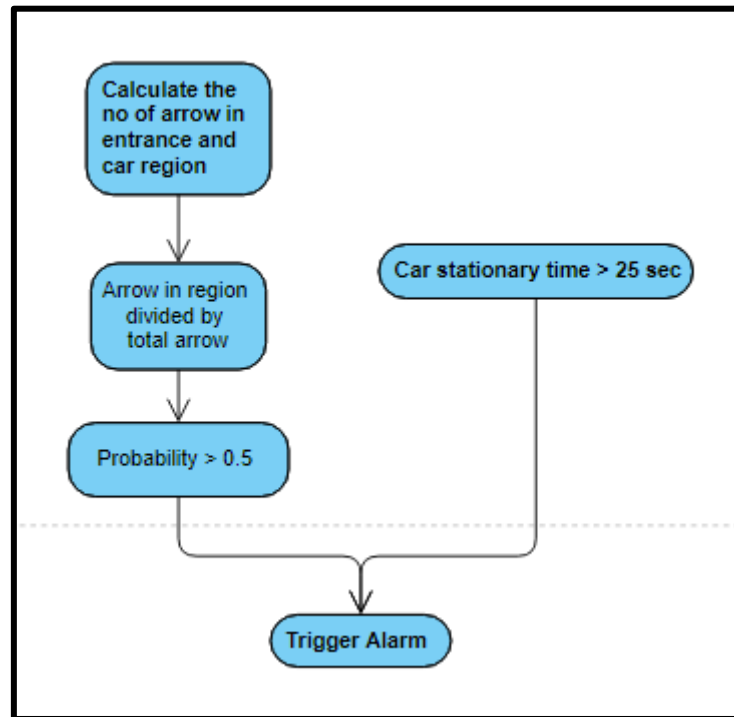
**Figure 4.6 Direction of human movement**

All human movement is stored in a list and draw by using the arrow. The tail of the arrow indicates the first appearance of human and the head of the arrow tell the last position of human. South east direction arrow mean that a human is walking toward south east direction. The direction of arrow is calculated by using the head pixel position  $(x_2, y_2)$  to subtract with the tail pixel position  $(x_1, y_1)$ .

- i. South East Direction:  $x_2 > x_1$  and  $y_2 > y_1$
- ii. North East Direction:  $x_2 > x_1$  and  $y_2 < y_1$
- iii. North West Direction:  $x_2 < x_1$  and  $y_2 < y_1$
- iv. South West Direction:  $x_2 < x_1$  and  $y_2 > y_1$

Entrance of the door is determined by the number of arrows pointed to one of the directions. For instance, if there are three arrows pointing toward south east direction, then the average pixel position of these three arrows' head is calculated. Subsequently, a region with  $25 * 25$  is drawn using the average pixel position.

### 4.6 Decision Making



**Figure 4.7 Decision Making**

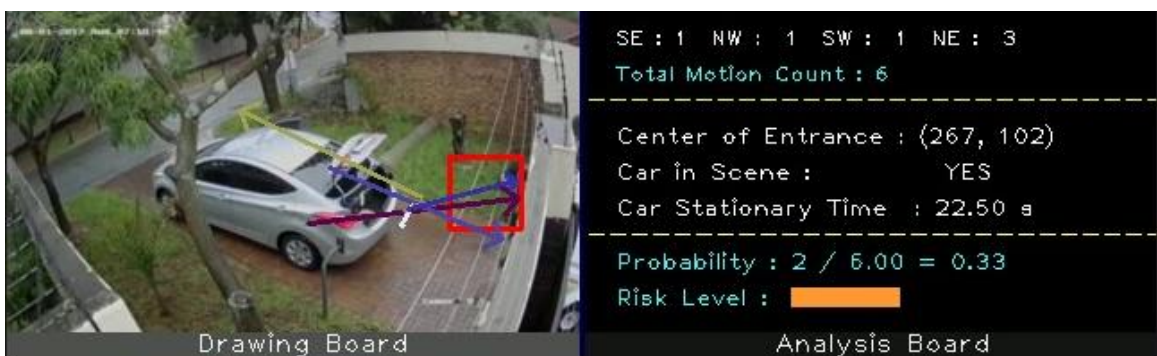
Decision to trigger the warning is based on the number of arrows in entrance and car region as well as car stationary time. If an arrow's head or tail is present in entrance or car region, then the arrow is considered as source or destination of the human movement. The source and destination of human movement will be divided by total movement count exists in the video to get the percentage of suspected movement in a video. If car stationary time exceed 25 seconds and the probability of suspected movement is larger than 0.5 then a high-risk level warning is flagged. If the probability of suspected movement is larger than 0.5 but without any car stationary in the video, then a medium risk level warning is flagged.

**Chapter 5: Experimental Results and Discussion**

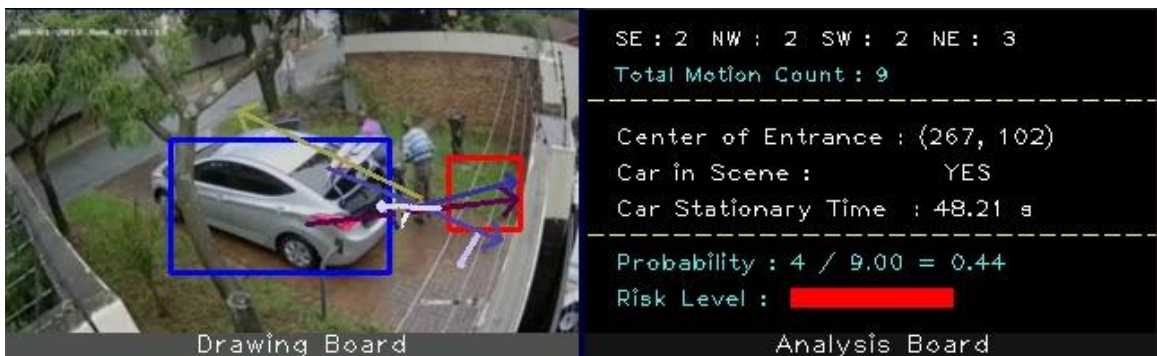
**5.1 Case A Result**



**Figure 5.1.1 Case A Result 1**



**Figure 5.1.2 Case A Result 2**



**Figure 5.1.3 Case A Result 3**

From Figure 3.1 a) and b), a person is walking from the outside of the front-yard and heading toward the entrance of the house. The system will first detect the first appearance of the human position and track the subsequent position until there is no human movement exists in the frame. After that an arrow will be drawn to show the direction of the human movement as shown in Figure 5.1.1. In Figure 5.1.2, an entrance region is drawn on the frame. The region is draw after three arrow point toward the North East direction is recorded. In Figure 5.1.3, a blue box is drawn on the car stationary region when the stationary time exceed 25 seconds. The probability is

calculated by dividing the total human movement counts with the number of arrow's head or tail in the red and blue box. Figure 5.1.3 shows that there are 4 arrows in red and blue box. Hence, the probability is 4 divided by 9 (total movement count).

### 5.2 Case B Result



Figure 5.2.1 Case B Result 1



Figure 5.2.2 Case B Result 2



**Figure 5.2.3 Case B Result 3**

In Figure 5.2.1, an arrow is drawn to indicate that a person in walk from the gate to the front door of the house in the front-yard area (Figure 3.1.2 a) and b)). After the person climb over the gate and enter the front yard area, the robber drives his car into the front-yard area of the house. Figure 5.2.2 shows a car is drawn with a blue box after the car stay stationary for more than 25 seconds. When there are 3 arrows points to the South East direction, the average pixel position of the arrows' head is calculated to draw the red box to indicate the entrance of the house. The risk level is high because more than half of the human movement is located in the blue box (car area) and the car stay stationary for a long time.

**5.3 Case C Result**



**Figure 5.3.1 Case C Result 1**



**Figure 5.3.2 Case C Result 2**



**Figure 5.3.3 Case C Result 3**

Figure 3.1.1 show a drawn arrow that point from the car to the door of the front-yard (Figure 3.1.3 a) and b)). A blue box is drawn as there is car stay stationary for a time longer than 25 seconds as shown in Figure 5.3.2. After the robber successfully open the entrance, he moves from the car to the entrance and his compliance follow him to enter the house. When there are 3 arrows point in North West direction, a red box is drawn to indicate the entrance region of the house as shown in Figure 5.3.3. When the probability exceeds 0.5, the risk level increase to red as a warning signal.



## **Chapter 6: Conclusion**

In this paper, an intelligent surveillance system is developed to automatically detect robbery that happens in the front yard area of a landed house. The surveillance system is made possible by using computer vision and artificial intelligent techniques. The project aims to help the victims of robbery to get attention from authorities in real time by flag an alarm when robbery is detected. As such, it will reduce the fatality rate of robbery victims and increase the chance of arresting robbers.

The project is developed with the following technique: YOLO object detection, contour tracking based on temporal subtraction, and motion analysis. First YOLO will start to detect human and car present in the video frame and store its location. When a car is detected, the car stationary time starts to measure. After the initial human position is stored, the subsequent human position is then detected with simple temporal subtraction to reduce the computational resources. The process of tracking the contour is continue until no human motion is detected for the next 10 frames. All of the detected human movement will be drawn as an arrow to indicate the direction of human movement. Analysis is done by calculate the car stationary time and the number of human movements in entrance and car region. Warning is flagged is the probability of human movement in entrance and car region is high and the car stationary time exceed 25 seconds.

A future improvement of this project can focus on state-of-arts methods in computer vision to improve the performance and accuracy of the system. Next improvement can focus on differentiating between multiple human using color histogram and detecting the item carried by the robbers will increase the accuracy of the system.

**BIBLIOGRAPHY**

Amira, BM, Ezzeddine, Z. 2017, 'Abnormal behavior Recognition for intelligence video surveillance system', *Expert System with Applications*, vol. 91, no. 1, pp. 480-491.

Berlin, S, J, John, M, 2016, 'Human interaction recognition through deep learning network', *IEEE International Carnahan Conference on Security Technology (ICCST)*, pp.1-4/

Blank, M, Gorelick, L, Shechtman, E, Irani, M, Basri, R, 2005, 'Actions as space-time shapes', *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, vol 1, pp. 1395-1402.

Bobick, A, Davis, J, 1996, 'An appearance-based representation of action', *Proceedings of 13<sup>th</sup> International Conference on Pattern Recognition*, pp. 307-312.

Bobick, A, Davis, J, 2001, 'The recognition of human movement using temporal templates', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257-267.

Darrell, T, Pentland, A, 1993, 'Space-time gestures', *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 335-340.

Dollar, P, Rabaud, V, Cottrell, G, Belongie, S, 2005, 'Behavior recognition via sparse spatio-temporal features', *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp.65-72

Duong, T, V, Bui, H, H, Phung, D, Q, Venkatesh, S, 2005, 'Activity Recognition and abnormality decision detection with the switching hidden semi-markov model', *IEEE Computer Society Conference on Computer Vision and Pattern*, pp.838-845

Harris, C, Stephens, M, 1988, 'A combined corner and edge detector' *Alvey Vision Conference*, p.50

Ke, Y, Sukthankar, R, Hebert, M, 2007, 'Spatio-temporal shape and flow correlation for action recognition' *IEEE Conference on Computer Vision and Pattern Recognition*, pp.239-242

Koller, D, Weber, J, Huang, T, 1994, 'Towards robust automatic traffic scene analysis in real-time', *Proceedings of 12<sup>th</sup> International Conference on Pattern Recognition*, vol. 1, pp. 126-131.

Krizhevsky, A, Sutskever, I, Hinton, GE, 2012, 'Imagenet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, pp.1097-1105.

## BIBLIOGRAPHY

Laptev, I, Lindeberg, T, 2003, 'Space-time interest points', *Proceedings Ninth IEEE International Conference on Computer Vision*, vol.1, pp.432-439.

Laptev, I, 2005, 'On space-time interest points', *International Journal of Computer Vision*, vol. 64, pp.107-123.

Lowe, D, G, 2004, 1999, 'Object recognition from local scale-invariant features', *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pp. 1150-1157

Lu, X, Liu, Q, Oe, S., 2004, 'Recognizing non-rigid human actions using joints tracking in space-time', *International Conference on Information Technology: Coding and Computing*, pp. 620-624.

Mo, L, Li, F, Zhu, Y, Huang, A, 'Human physical activity recognition based on computer vision with deep learning model', *IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp.1-6

Oliver, N, M, Rosario, B, Pentland, A, P, 2000, 'A Bayesian computer vision system for modelling human interactions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 831-843.

Park, S, Aggarwal, J, K, 2004, 'A hierarchical Bayesian network for event recognition of human actions and interactions', *Multimedia System*, vol. 10, pp. 164-179

Shechtman, E, Irani, M, 2007, 'Space-time behaviour-based correlation-or-how to tell if two underlying motion fields are similar without computing them?', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 29, pp. 2045-2056.

Schuldt, C, Laptev, I, Caputo, B, 2004, 'Recognizing human actions: a local SVM approach', *Pattern Proceedings of the 17<sup>th</sup> International Conference on Pattern Recognition*, pp.32 -36.

Scovanner, P, Ali, S, Shah, M, 2007, 'A 3-dimensional sift descriptor and its application to action recognition', *Proceedings of the 15<sup>th</sup> ACM International Conference on Multimedia*, pp. 357-360

Stauffer, C, Grimson, W, E, L, 1999, 'Adaptive background mixture models for real-time tracking', *Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 246-252.

Suk, H, I, Sin, B, K, Lee, S, W, 2010, 'Hand gesture recognition based on dynamic Bayesian network framework', *Pattern Recognition*, vol. 10, pp. 994-999

Vail, D, L, Veloso, M, M, Lafferty, 2007, 'Conditional Random Fields for activity recognition', *Proceedings of the 6<sup>th</sup> International Joint Conference on Autonomous Agents and Multiagent Systems*, p.235

## BIBLIOGRAPHY

Vapnik, V, Golowich, S, E, Smola, A, 1996, 'On discriminative vs generative classifiers: a comparison of logistics regression and naïve bayes', *Advances in Neural Information Processing Systems*, vol. 9, pp. 841-848.

Veeraraghavan, A, Roy-Chowdhury, A, K, Chellapa, R, 2005, 'Matching shape sequences in video with applications in human movement analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, pp. 1896-1909.

Willems, G, Tuytelaars, T, Gool, L, V, 2008, 'An efficient dense and scale-invariant spatio-temporal interest point detector', *Computer Vision – European Conference on Computer Vision*, pp. 650-663

Wren, CR., Azarbayejani, A., Darrell, T., Pentland, AP., 1997, 'Pfinder: real-time tracking of the human body', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, pp. 780-785.

Yamato, J, Ohya, J, Ishii, K, 1992, 'Recognizing human action in time-sequential images using hidden markov model', *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 379-385

Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S. and Li, Z. (2017). A Review on Human Activity Recognition Using Vision-Based Method. *Journal of Healthcare Engineering*, 2017, pp.1-31.

## A-1 Poster



Universiti Tunku Abdul Rahman  
Faculty of Information and Communication Technology

Name: Tang Jia Le  
Student ID: 16ACB04284  
Programme: Information System Engineering  
Supervisor: Prof. Maylor Leung Kar Hang

Traditional CCTVs are not embedded with artificial intelligence and they are only used to record activities happening in the real-time. In other words, it lacks the ability to analyze the activities captured in the frame to make early detection of suspicious activities such as robbery.

### System Implementation



## A-2 Plagiarism Check Result

FYP2

---

ORIGINALITY REPORT

---

<b>15%</b>	<b>10%</b>	<b>9%</b>	<b>10%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

---

PRIMARY SOURCES

---

<b>1</b>	<b>Submitted to Heriot-Watt University</b> Student Paper	<b>1%</b>
<b>2</b>	<b>Qingdi Wei, Xiaoqin Zhang, Weiming Hu.</b> <b>"chapter 12 Action Recognition", IGI Global,</b> <b>2010</b> Publication	<b>1%</b>
<b>3</b>	<b>link.springer.com</b> Internet Source	<b>1%</b>
<b>4</b>	<b>www.ms.sapientia.ro</b> Internet Source	<b>1%</b>
<b>5</b>	<b>centaur.reading.ac.uk</b> Internet Source	<b>1%</b>
<b>6</b>	<b>kyutech.repo.nii.ac.jp</b> Internet Source	<b>&lt;1%</b>
<b>7</b>	<b>dspace.lboro.ac.uk</b> Internet Source	<b>&lt;1%</b>
<b>8</b>	<b>ir.lib.ncu.edu.tw:88</b> Internet Source	<b>&lt;1%</b>

---

A-2

9	Md. Uddin, Weria Khaksar, Jim Torresen. "Ambient Sensors for Elderly Care and Independent Living: A Survey", Sensors, 2018 Publication	<1%
10	tel.archives-ouvertes.fr Internet Source	<1%
11	Submitted to iGroup Student Paper	<1%
12	eprints.eemcs.utwente.nl Internet Source	<1%
13	Submitted to University of Bath Student Paper	<1%
14	pt.scribd.com Internet Source	<1%
15	jtiik.ub.ac.id Internet Source	<1%
16	Submitted to Higher Education Commission Pakistan Student Paper	<1%
17	scholarcommons.usf.edu Internet Source	<1%
18	Mahlagha Afrasiabi, Hassan Khotanlou, Theo Gevers. "Spatial-temporal dual-actor CNN for human interaction prediction in video",	<1%

## Multimedia Tools and Applications, 2020

Publication

---

19	<a href="http://www.computing.edu.au">www.computing.edu.au</a> Internet Source	<1%
20	Submitted to Engineers Australia Student Paper	<1%
21	Dong-Jun Park. "Video Event Detection as Matching of Spatiotemporal Projection", Lecture Notes in Computer Science, 2010 Publication	<1%
22	Submitted to Universiti Sains Malaysia Student Paper	<1%
23	Submitted to University of Bristol Student Paper	<1%
24	Ke, Shian-Ru, Hoang Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. "A Review on Video-Based Human Activity Recognition", Computers, 2013. Publication	<1%
25	<a href="http://crcv.ucf.edu">crcv.ucf.edu</a> Internet Source	<1%
26	Submitted to The Hong Kong Polytechnic University Student Paper	<1%
27	<a href="http://clock.uclan.ac.uk">clock.uclan.ac.uk</a> Internet Source	

---



		<1%
28	Sangho Park, J. K. Aggarwal. "A hierarchical Bayesian network for event recognition of human actions and interactions", <i>Multimedia Systems</i> , 2004 Publication	<1%
29	Suk, H.-I.. "Hand gesture recognition based on dynamic Bayesian network framework", <i>Pattern Recognition</i> , 201009 Publication	<1%
30	<a href="http://repository.tudelft.nl">repository.tudelft.nl</a> Internet Source	<1%
31	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
32	Submitted to University of Glasgow Student Paper	<1%
33	Huasong Min, Yunhan Lin, Sijing Wang, Fan Wu, Xia Shen. "Path planning of mobile robot by mixing experience with modified artificial potential field method", <i>Advances in Mechanical Engineering</i> , 2015 Publication	<1%
34	<a href="http://doras.dcu.ie">doras.dcu.ie</a> Internet Source	<1%

35	<a href="http://ies.anthropomatik.kit.edu">ies.anthropomatik.kit.edu</a> Internet Source	<1%
36	Hamm, Jihun, Benjamin Stone, Mikhail Belkin, and Simon Dennis. "Automatic Annotation of Daily Activity from Smartphone-Based Multisensory Streams", Lecture Notes of the Institute for Computer Sciences Social Informatics and Telecommunications Engineering, 2013. Publication	<1%
37	Ni, Bingbing, Shuicheng Yan, and Ashraf Kassim. "HUMAN GROUP ACTIVITIES: DATABASE AND ALGORITHMS", Advanced Topics in Biometrics, 2011. Publication	<1%
38	Submitted to CSU, Chico Student Paper	<1%
39	Submitted to University of Wales Institute, Cardiff Student Paper	<1%
40	<a href="http://www.cs.columbia.edu">www.cs.columbia.edu</a> Internet Source	<1%
41	Submitted to University of Wollongong Student Paper	<1%
42	<a href="http://eprints.soton.ac.uk">eprints.soton.ac.uk</a> Internet Source	<1%

43	<a href="http://era.library.ualberta.ca">era.library.ualberta.ca</a> Internet Source	<1%
44	Alper Yilmaz. "Object tracking", ACM Computing Surveys, 12/25/2006 Publication	<1%
45	Submitted to Cranfield University Student Paper	<1%
46	Lecture Notes in Electrical Engineering, 2014. Publication	<1%
47	Submitted to University of Newcastle Student Paper	<1%
48	Submitted to University of Western Sydney Student Paper	<1%
49	Submitted to University of Leeds Student Paper	<1%
50	<a href="http://www.db-thueringen.de">www.db-thueringen.de</a> Internet Source	<1%
51	<a href="http://en.wikipedia.org">en.wikipedia.org</a> Internet Source	<1%
52	Submitted to Brunel University Student Paper	<1%
53	<a href="http://diglib.eg.org">diglib.eg.org</a> Internet Source	<1%

54	<a href="http://id.scribd.com">id.scribd.com</a> Internet Source	<1%
55	<a href="http://www.frontiersin.org">www.frontiersin.org</a> Internet Source	<1%
56	Liu, X.. "Multi-agent activity recognition using observation decomposed hidden Markov models", Image and Vision Computing, 20060201 Publication	<1%
57	Ahmad, M.. "Human action recognition using shape and CLG-motion flow from multi-view image sequences", Pattern Recognition, 200807 Publication	<1%
58	<a href="http://imagefeatures.org">imagefeatures.org</a> Internet Source	<1%
59	<a href="http://herkules.oulu.fi">herkules.oulu.fi</a> Internet Source	<1%
60	Jones, S.. "Relevance feedback for real-world human action retrieval", Pattern Recognition Letters, 201203 Publication	<1%
61	Engineering Computations, Volume 31, Issue 2 (2014-03-28) Publication	<1%
62	<a href="http://theses.whiterose.ac.uk">theses.whiterose.ac.uk</a> Internet Source	

		<1%
63	Submitted to UT, Dallas Student Paper	<1%
64	Submitted to 87986 Student Paper	<1%
65	Submitted to University of Huddersfield Student Paper	<1%
66	Submitted to KYUNG HEE UNIVERSITY Student Paper	<1%
67	Submitted to University of Technology, Sydney Student Paper	<1%
68	W. Hu, T. Tan, L. Wang, S. Maybank. "A Survey on Visual Surveillance of Object Motion and Behaviors", IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 2004 Publication	<1%
69	Xu, Xin, Jinshan Tang, Xiaolong Zhang, Xiaoming Liu, Hong Zhang, and Yimin Qiu. "Exploring Techniques for Vision Based Human Activity Recognition: Methods, Systems, and Evaluation", Sensors, 2013. Publication	<1%
70	<a href="http://www.lirmm.fr">www.lirmm.fr</a>	

	Internet Source	<1%
71	spod.tarc.edu.my Internet Source	<1%
72	Ziming Zhang, Yiqun Hu, Syin Chan, Liang-Tien Chia. "Chapter 60 Motion Context: A New Representation for Human Action Recognition", Springer Science and Business Media LLC, 2008 Publication	<1%
73	Lecture Notes in Computer Science, 2015. Publication	<1%
74	Industrial Robot: An International Journal, Volume 40, Issue 6 (2013-10-19) Publication	<1%
75	Submitted to GLA University Student Paper	<1%
76	Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, 2006 Publication	<1%
77	"Information Extraction: Algorithms and Prospects in a Retrieval Context", Springer Science and Business Media LLC, 2006 Publication	<1%

**78** "Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence", Springer Science and Business Media LLC, 2012 **<1%**  
Publication

---

**79** Submitted to Universiti Kebangsaan Malaysia **<1%**  
Student Paper

---

**80** Md. Atiqur Rahman Ahad. "Motion History Images for Action Recognition and Understanding", Springer Science and Business Media LLC, 2013 **<1%**  
Publication

---

Exclude quotes  On

Exclude matches  Off

Exclude bibliography  On

<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>	TANG JIA LE
<b>ID Number(s)</b>	1604284
<b>Programme / Course</b>	BIS (Hons) Information Systems Engineering
<b>Title of Final Year Project</b>	Front Yard Robbery Detection Surveillance System

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
<b>Overall similarity index: <u>15</u> %</b>  <b>Similarity by source</b> Internet Sources: 10 % Publications: 9 % Student Papers: 10%	
<b>Number of individual sources listed of more than 3% similarity: None</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</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: 24 April 2020

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

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

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





**UNIVERSITI TUNKU ABDUL RAHMAN**  
**FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR**  
**CAMPUS)**

**CHECKLIST FOR FYP1 THESIS SUBMISSION**

Student Id	1604284
Student Name	Tang Jia Le
Supervisor Name	Prof. Maylor 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.
√	Title Page
√	Signed form of the Declaration of Originality
√	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)
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result – Form Number: FM-IAD-005)

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

<p>I, the author, have checked and confirmed all the items listed in the table are included in my report.</p> <div style="text-align: center; margin-top: 20px;">  </div> <p>_____          (Signature of Student)          Date: 24/04/2020</p>	<p>Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction.</p> <div style="text-align: center; margin-top: 20px;">  </div> <p>_____          (Signature of Supervisor)          Date: 24 April 2020</p>
---	---