Empirical Approaches for Human Behavior Analytics

Jia Lu

A thesis submitted to Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2016

School of Engineering, Computer and Mathematical Sciences

Abstract

Surveillance is ubiquitous in our communities which can be utilized to deal with multiple security issues. However, most of surveillance systems still are not intelligent which are mainly relying on security staff's human labor. Thus, human behavior analysis based on computer vision could tremendously reduce security staff's workload.

To analyze and understand human behaviors in surveillance, the start point is to extract computable features from captured videos based on detected human body, the ultimate goal is to finally recognize human behaviors from motion and event analysis.

This thesis presents comprehensive and in-depth empirical approaches for event recognition in surveillance based on distinct Feature Extraction Techniques (FET), namely: Histograms of Oriented Gradients (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP). Each of the FETs is based on local feature descriptor which depends on adjusting the cell size of the ROI to achieve better performance. In this thesis, we find the cell size will influence not only the computational time, but also the precision rate.

This thesis utilizes the well-known Weizmann video datasets. While both LBP and SILTP features work very well, HOG has shown its superior performance for human behaviour analytics with five selected events (Walking, Running, Skipping, Jumping and Jacking). The simulated results of three classifiers from WEKA (MLP, *k*-NN, decision tree) have reflected rightness of the extracted features.

In this thesis, the empirical approaches for human behaviour analytics in surveillance reduce human labor tremendously. The contributions of this thesis are: (1) The distinct FET makes the best precision of overall human behaviour recognition at the rate above 97.7%. (2) By adjusting the cell size of ROI, the proposed approaches are able to be accelerated, furthermore, computational time could be reduced.

Keywords: Histograms of Oriented Gradients (HOG), Local Binary Pattern (LBP), Scale Invariant Local Ternary Pattern (SILTP), MLP, *k*-NN, Decision Tree.

Table of Contents

Abstract I
Table of Contents II
List of Figures IV
List of Tables V
List of Algorithms VI
Attestation of Authorship VII
Acknowledgment VIII
Chapter 1 Introduction 1
1.1 Background and Motivation 2
1.2 Research Question 4
1.3 Contribution 5
1.4 Objective of This Thesis 6
1.5 Structure of This Thesis 6
Chapter 2 Literature Review 8
2.1 Introduction 9
2.2 The Concept Event 10
2.3 Motion Detection 12
2.3.1 Moving Object Detection 13
2.3.2 Moving Object Tracking 16
2.4 Recognition 18
2.5 Classifications 23
2.5.1 Supervised Learning 24
2.5.2 Unsupervised Learning 28
Chapter 3 Methodology 30
3.1 Research Designing 31
3.1.1 Locating the ROI 32
3.1.2 ROI Segmentation 36
3.1.3 Feature Extraction 38
3.2 Evaluation Methods 45
Chapter 4 Results 47
4.1 Data Collection and Experimental Environment 48

4.2 Human Behavior Recognition and Classification 49
4.2.1 Human Behavior Recognition 50
4.2.2 Human Behavior Classification 59
4.3 Limitations of the Research 64
Chapter 5 Analysis and Discussions 65
5.1 Analysis 66
5.1.1 Feature Extraction 66
5.1.2 Artificial Neural Network 67
5.1.3 Other Classifiers 70
5.2 Discussions 74
Chapter 6 Conclusion and Future Work 76
6.1 Conclusion 77
6.2 Future Work 78
References 79

List of Figures

Figure 2.1 Different approaches to human behavior recognition.....28
Figure 2.2 The structure of an artificial neuron.....34
Figure 2.3 The structure of BP neural network.....36
Figure 3.1 The flowchart of the steps of human behavior recognition.....39
Figure3.2 Locating ROI area in the video frame.....43
Figure3.3 The segmentations of ROI area in a video stream.....44
Figure3.4 Examples of the ROI for five output classes after segmentation.....45
Figure3.5 The flowchart of the HOG feature extraction.....47
Figure3.6 Feature extraction by using HOG for different human behaviors.....47
Figure3.7 The deployment of HOG feature vectors.....48
Figure3.8 HOG feature example with two different parameters of cell size.....50
Figure3.9 The simple mechanism of original LBP operator.....51
Figure3.10 Feature extraction by using LBP for different human behaviors.....51
Figure3.11 The simple mechanism of SILTP operator.....52
Figure3.12 Feature extractions by using SILTP for different human behaviors.....53
Figure 4.1 Video frame example.....57
Figure 4.2 Flowchart of Artificial Neural Network.....58
Figure 4.3 The result of trained neural network by using HOG training dataset.....60
Figure 4.4 The result of trained neural network by using LBP training dataset.....62
Figure 4.5 The result of trained neural network by utilizing SILTP training dataset.....63
Figure 4.6 Matlab result of testing dataset.....64
Figure 4.7 The simulated result of pre-trained neural network by using HOG.....65
Figure 4.8 The simulated result of pre-trained neural network by using LBP.....66
Figure 4.9 The simulated result of pre-trained neural network by using SILTP.....67
Figure 4.10 Results of event recognition and classification.....70
Figure 4.11 Incorrectly classified results of event recognition and classification.....71
Figure 5.1 The general model of decision trees.....78
Figure 5.2 The visualization result of decision trees.....79

List of Tables

Table 2.1 The event description for the previous sentence.....20

Table 2.2 The summary of four different object detection methods.....24

Table 3.1 The sample of the confusion matrix…….…….....54

Table 4.1 Weizmann dataset…….…....56

Table 5.1 The results of different feature extraction techniques…….…….....74

Table 5.2 The precision of different FET with different cell size…….….....76

Table 5.3 The precision of different FET in the classification of training dataset…….....76

Table 5.4 Results of behavior classification and recognition…….…….....77

Table 5.5 The precision of different FET in the classification of testing dataset…….....77

Table 5.6 The results of decision tree C4.5 algorithm……...80

Table 5.7 The results of MLP by utilizing testing dataset……81

Table 5.8 The results of KNN by utilizing testing dataset……82

List of Algorithms

Algorithm 3.1 Finding and Segmentation of Region of Interest.....45

Algorithm 3.2 Histogram of Oriented Gradient Feature Extraction.....50

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature:  Date: 10 December 2016

Acknowledgment

This research work was completed as the part of the Master of Computer and Information Sciences (MCIS) course at the School of Computer and Mathematical Sciences (SCMS) in the Faculty of Design and Creative Technologies (DCT) at the Auckland University of Technology (AUT) in New Zealand. I would like to deeply thank my parents for the financial support they provided during my entire time of academic study in Auckland. I also would like to thank my homestay who supports me in life and language.

My deepest thanks are to my primary supervisor Dr. Wei Qi Yan who has provided me with much appreciated technological guidance and support. I believe that I could not achieve my Master Degree without his invaluable help and supervision. In addition, I would like appreciate my secondary supervisor Dr. Boris Bačić and school administrators for their support and guidance through the MCIS in the past years.

Jia Lu

Auckland, New Zealand

December 2016

Chapter 1
Introduction

The first chapter of this thesis consists of five sections. In the first section, background and motivation of this thesis are introduced, human behavior recognition has not only a wide range of applications that can be utilized to identify particular events, but also has a function that could be applied to find out abnormal human behaviors in order to reduce human labor. Objectives will be discussed in the fourth section. This chapter also covers the details of research questions after the in-depth comprehensive understanding of the relevant literatures and research background. Finally, we will present an overview of the structure of this thesis in section 5.

1.1 Background and Motivation

Nowadays, computing ability of computers are more and more powerful, however, computers as a machine still heavily rely upon human beings to learn and understand. Thus, it is needed that all computers have high intelligence and automatically accomplish the assigned work.

Computational Vision (CV) as an important branch in Artificial Intelligence (AI) and Pattern Recognition (PR) has the abilities to deal with visual information. Like an English idiom says, 'A picture is worth a thousand words'. Computer vision as a research area in information science has been mainly employed to process and analyze image / video data. Its goal is to make computers have the ability to understand our human world by using visual observation via cameras. Moreover, with the widespread use of digital cameras and smartphones, computer vision has become more and more important in our daily life.

Surveillance cameras are ubiquitous in our society, previously it only appears in banks or other highly secure areas. Nowadays, it not only turns up in private places, but also in public premises such as shopping malls, streets, bus and trains, etc. (Petrushin, 2005; Popoola & Wang, 2012). Moreover, with the rapidly increasing number of surveillance cameras used to capture events of interest in the scene, time-consuming for human labor to extract the events of interest from thousands of surveillance videos has become a major problem. Thus, to reduce the time consuming and make it more efficiency should be considered, in addition, we need to find a way to detect events from video streams automatically so as to make the surveillance more intelligent. Despite surveillance videos have been used in decades, object detection and classification have been developed as a particular filed. In recent years, a plenty of researchers paid attention to 'object' and their focus is on object detection and recognition such as pedestrians, vehicles and its trajectories (Krumm, et al., 2000; Siebe & Maybank, 2002). However, object is the basic concept in a video, and it cannot present stories behind the binary stream.

Event in the real world can be defined as occurrence happened in a determinable space and time. In surveillance, videos are generally acquired from cameras. As an active research area in intelligent surveillance, finding the event can improve the accessibility and reusability for a large amount of media collections (Maryam & Reza, 2012). Event mining, which has a wide range of applications, can be applied to mining particular events in transportation (Chen & Zhang, 2006), and as a function to detect abnormal human behaviors (Popoola & Wang, 2012). To analyze and understand human behaviors, the main task is to extract visual features from surviellance videos based on detected human body, analyzing these features leads to finally implement the behavior recognition. As an advanced computer vision processing, human behavior analysis contains digital image processing, pattern recognition, machine learning and other core issues which also has very high theoretical and practical value, especially with a great deal of applications in surveillance.

As we mentioned before, there were two different levels in the applications of surveillance. One is the object motion detection (a low-level process) which contains image pre-processing, foreground segmentation, object detection and tracking etc.

In 2000, morphological change based detection algorithms was implemented for motion detection (Stringa, 2000). For motion detection, there are a majority of traditional methods to be implemented, such as background subtraction, inter-frame difference and optical flow. Background subtraction as the most common method uses the subtraction between current frame and background frame, adjusts the threshold to finally acquire the moving object. It has a good detection effect even with complex background. However, it is very sensitive to outdoor conditions such as weather, lighting and raining, etc. For inter-frame difference method, motions are extracted by the correlation between adjacent frames through modifying the threshold to achieve better performance. But the limitation of inter-frame difference method is that the object is very slow in speed and cannot be effectively detected. Optical flow method uses the changes of the pixels in image sequence in temporal domain and the correlation between adjacent frames.

Pattern recognition as the high-level process contains classification, recognition, etc. In the early research of human behavior analysis and understanding, the focus was on description of human body in the sample geometric models (Aggarwal & Cai, 1997), such as stick figure model (Guo, Xu, & Tsuji, 1994), 2D contour model (Aggarwal, Cai, Liao, & Sabata, 1997) and volumetric model (Rohr, 1994). Afterward, the focus was on the feature extraction which is able to well describe human motion, such as position, direction, trajectory, angles and velocity, etc. to distinguish different behaviors. Moreover, by combining different methods together, this could recognize human behaviors.

Thus, with the theories of image and video processing becoming more mature, in the aid of computing and storing capacity being stronger in computer vision, these make human behavior recognition possible. Due to the in-depth understanding of the relevant literatures and research background, this leads to our research questions of this thesis.

1.2 Research Questions

As we mentioned, this thesis aims at classification for human behavior recognition, and implementations of event based human behavior recognition. Analyzing the methods and techniques helps us to know the basic procedures of implementing an event based human behavior recognition. Therefore, the research questions of this thesis are:

Question:

What computing techniques can be implemented to human behavior recognition?

We attempt to find answers for the following question which was developed from the above main question:

"What are the algorithms for human behavior recognition? which algorithm is the best one for our research?"

The core idea of this thesis is human behavior recognition. Thus, the techniques that we adopted in this research project need to be evaluated and proper techniques need to be

chosen so as to implement the best result of human behavior recognition.

1.3 Contribution

The focus of this thesis is mainly on implementation of event-based human behavior recognition. The similar event based recognition from the videos can be developed step by step in accordance with the provided procedures in this thesis. The proposed human behavior recognition encompasses four parts in this thesis (1) video pre-processing, (2) feature extraction, (3) human behavior recognition, (4) resultant analysis of human behavior recognition. The proposed method and implementation will be illustrated in Chapter 4.

Moreover, the core algorithm and approaches will be analyzed in this thesis which are suitable for human behavior recognition. A comparison of various object detection methods is present in Chapter 2. Moreover, the introduction of various approaches will also be detailed in the next chapter.

Furthermore, this thesis will mainly utilize Artificial Neural Networks (ANNs) for the recognition with different feature extraction techniques. Thus, this thesis provides detailed information on implementing human behavior recognition in MATLAB correspondingly.

Last but not the least, three classification methods will be presented and compared with those existing methods which are implemented by using the platform WEKA.

1.4 Objective of This Thesis

Firstly, this thesis introduces classification techniques related to video/image pre-processing, event based behavior detection and recognition. In addition, the principles of these techniques will also be demonstrated and evaluated in this thesis.

Secondly, in order to achieve the recognition of human behaviors in surveillance, a conceptual framework is also presented for event based human behavior recognition. Therefore, the overall objective of this thesis is divided into four parts which include video pre-processing, object detection, behavior recognition and behavior classification.

Finally, the comparison of algorithms for human behavior recognition will be presented in this thesis. All in all, it is necessary to find the best algorithm that suits human behavior recognition. Moreover, multiple classification methods will be introduced to this thesis.

1.5 Structure of This Thesis

The thesis is structured as follows:

In Chapter 2, literature review will be discussed, such as the previous studies in event detection and recognition in surveillance for the area of human behavior analysis and understanding. Human behavior recognition was studied as a high-level processing in semantic. Thus, Chapter 2 will introduce the fundamentals of video pre-processing and numerous recognition and classification methods which will be applied to human behavior analytics.

In Chapter 3, the explanation of research methodology of this thesis will be stated. In addition, the potential solutions and answers will also be presented. Moreover, the experimental layout and design as well as datasets and implementations with the evaluation methods will be presented.

In Chapter 4, the methodologies and algorithms we presented will be implemented.

Moreover, experimental results and outcomes will be detailed with the support of tables and figures. The limitations of this project will also be addressed as well.

In Chapter 5, the analysis and discussions are depicted based on experimental results and outcomes we acquired in Chapter 4. Finally, the conclusion and future work will be presented in Chapter 6.

Chapter 2
Literature Review

With in-depth analysis of the research questions and rationale reviews of the previous studies, the focus of this thesis is on event based human behavior analytics from videos. For instance, the pre-processing, moving object detection and tracking are applied to human behavior recognition. The state-of-the-art human behavior recognition methods will be summarized in this chapter.

2.1 Introduction

Nowadays, surveillance only has the ability to detect moving objects which are at the level of object. Apparently it will miss a large amount of scene information behind the object. As we mentioned in Chapter 1, to understand and describe the scene, the concept event is possible to represent the stories in a surveillance environment. Therefore, at event level, researchers are paying much attention to the recognition, as a special case, human behavior recognition will be studied based on events in this thesis.

Human behavior recognition as an active research area attracts more and more attention in intelligent surveillance due to the increasing global security issues, as well as soaring requirement of effective monitoring in public places such as shopping malls, airports, arenas, banks, etc. or any other private primises, event-driven is the case where an event triggers the consequential actions.

There are several technologies and applications which have been applied to recognize human behaviors. With advantage of low-cost and convenience, human behavior recognition has become the main mission of this thesis. Surveillance can achieve real-time monitoring and automatic analysis of video footages that were captured from surveillance camera. Moreover, according to the analysis of human behaviors, the suspects, abnormal behaviors or events can be identified. When an exception occurs, surveillance based human behavior analysis could make alarms and alert security staff timely. Intelligent surveillance is able to analyze and understand the tracking objects in order to achieve human behavior recognition in real time. Human behavior recognition is established on the basis of moving object tracking and visual feature extraction in order to analyze the data of human behaviors by utilizing pattern recognition and machine learning. Generally speaking, the goal of machine learning is to recognize human behaviors. Human behaviors could be recognized in multiple ways in this thesis.

2.2 The Concept Event

With rapid development of technologies, there is a mass of sensory data which have been acquired from various devices, such as mobile phones and cameras (Sakaki, Okazaki, & Matsuo, 2010). For all of those images, a huge volume of visual data promoted event detection in the real world. More specifically, for event detection in surveillance, the task is to identify and locate the Region of Interest (ROI) using spatial-temporal patterns.

The definition of event in a dictionary is "Event is something that happens, especially when it is unusual or important". Events can be used to describe all the things that are happening in a particular situation. An event in our real lives can be defined as occurrence happened in a determinable space and time (Xie, Sundaram, & Campbell, 2008; Yan, 2016). A video event also can be viewed as a long-term temporal objects that have thousands of frames (Zelnik-Manor & Irani, 2001). The six elementary aspects of event description includes temporal aspect, spatial aspect, causal aspect, experiential aspect, informational aspect and structural aspect (Westermann & Jain, 2007).

In the early ages, temporal events can be grouped into three classes which encapsulate temporal textures (an uncertain spatial and temporal extent), activities (temporally periodic but the spatial extent is finite), motion events (isolated events that are no repetition in either temporal or spatial scales) (Polana & Nelson, 1994). For a temporal event, all sequences were used to characterize the same temporal scale. In contrast, an object may appear at various spatial scales in all sequences for a spatial event (Zelnik-Manor & Irani, 2001). In 1996, a parameterized model is used to recognize articulated motions (Ju, Black, & Yacoob, 1996). The approaches could be applied to detect periodic activities (Cutler & Davis, 2000; Zhao & Pietikainen, 2007).

Events have the ability to memorize the real-world experiences. Moreover, events can be comprehended as a story that has duration, also with or without participation of human. Typically, an event is a piece of chain which can be the effect of previous occurrence and causes a succeeding occurrence to change the state of a story. Therefore,

duration and location become the most important parameters in the case to describe an event. An event should be unique at its duration and location, the events with different duration and location should be considered as different events (Yan, 2016).

Event-based video surveillance takes the advantages of event detection and recognition. In Chapter 1, the applications of detected events in transportation, or detected abnormal human behaviors were wildly used to public. Thus, event analysis has primarily focused on understanding the predefined events, and the major of event detection in video surveillance is to detect and analyze unusual/abnormal events, for an abnormal event not only has a low frequency, but also attracts lots of attention to human. For example, most of the customers go to a shopping mall that leads the event for customers in shopping mall into three categories (or three normal event actions) which contain walking, standing and sitting, so for video surveillances these are considered as normal events. All the abnormal events should be detected and analyzed; the specific frames that contain abnormal events in surveillance also should be stored as alarms for the security staff or an evidence in dispute court.

In line with literature review on details of video mining (Dai, Zhang, & Li, 2006) and video retrieval (Geetha & Narayanan, 2010), an event consists of six aspects and five major components (Xie, Sundaram, & Campbell, 2008) Moreover, three interrogatives (what, where, who) are used to describe an event (Li & Fei-Fei, 2007). In 2008, an event was described and characterized by six common sense aspects ‘5W1H’ (Xie, Sundaram, & Campbell, 2008). The 5W1H refers to the six interrogatives which include who, when, where, why, what, and how.

‘Who’ means object, it refers to a person or a thing that can lead to occurrence of the event. ‘When’ is equavalent to duration or time that is one of the most important components of an event which contains two different scales (temporal and spatial) and various types (such as ‘after Christmas’ and ‘Monday’); ‘Where’ is the location of the event, and it can have multiple categroies (such as ‘not far from Auckland University of Technology’ and ‘55 Wellesley St.’) as well as duration. ‘Why’ tells us the background or

context of the event which provides the cause of an event. ‘What’ defines the actions behind an event, and it can be grouped into two different classes (normal or abnormal). While ‘How’ reflects the dynamic state of an event which usually is to tell the story happening related to event.

Table 2.1 Event description for the previous sentence

Interrogatives	**Event Description**
Who	Frank
When	Monday and Friday
Where	Classroom
Why	Study
What	Two papers

Therefore, we can describe an event in a concise way. Table 2.1 shows the description for the previous sentence.

Events are described as temporal objects and spatial objects with multiple scales (Zelnik-Manor & Irani, 2006). However, the core idea of object detection is to detect a new or moving object (foreground) in the same background. Video event mining is still playing a key role in the development of next generation of video search (Xie & Yan, 2008). At present, the problem that still remains for video event search is a lack of effective indicators to describe the content of video data. In four phases of event mining (Valera & Velastin, 2005), it is important to extract existing semantic patterns (Maryam & Reza, 2012).

2.3 Motion Detection

Event detection as one of the major components of this thesis is divided into three sections: video pre-processing, moving object detection and tracking. Moreover, moving object detection is the first step of event detection. In this section, the focus is on introduction of video pre-processing, moving object detection and tracking.

2.3.1 Moving Object Detection

As we mentioned, the object is one of the most important elements in an event. Thus, moving object detection will become a major component in event detection, and it is critical task for most of computer vision applications. Moreover, moving object detection is the basic step for further analysis in the streams (Papageorgiou, Oren, & Poggio, 1998; Poppe, 2007; Joshi & Thakore, 2012). In order to detect a moving object with respect to background of a scene, the primary objective of moving object detection in a video is to analyze the image sequences (Elhabian, El-Sayed, M., & Ahmed, 2008). Moreover, the video can be regarded as the combinations of continuous static frames and all the individual frame extracted from video stream should be able to be processed by all kind of image processing operators.

In order to achieve moving object detection, there are four different approaches which comprise background subtraction, inter-frame differencing, temporal differencing and optical flow (Kulchandani & Dangarwala, 2015).

The approach of background subtraction for moving object detection is used by subtraction between current frame and background frame, the approach is usually adopted in the video that from static cameras (Wren, Azarbayejani, Darrell, & Pentland, 1997; Piccardi, 2004). There are eight critical situations that should be handled, which mainly comprehend an image having noise, non-static objects with small movements, illumination changes, shadow regions, multiple objects, undeviating variations of the objects and background movement changes (Rakibe & Patil, 2013). The formula for background subtraction is as eq. (2.1):

$$D_k(x,y) = \begin{cases} 1; \mid I_k(x,y) - B_k(x,y) \mid \geq T \\ 0; \mid I_k(x,y) - B_k(x,y) \mid < T \end{cases} \tag{2.1}$$

where $I_k(x,y)$ is current frame and $B_k(x,y)$ is background frame, T represents predefined threshold and '1' in $D_k(x,y)$ is the binary result.

The inter-frame differencing method identifies moving object by comparing the difference between two adjacent frames, moreover, the moving object will be extracted by adjusting the threshold. The formula for inter-frame differencing is shown as eq. (2.2 & 3):

$$D_m(x,y) = |F_m(x,y) - F_{m-1}(x,y)| \tag{2.2}$$

$$B_m(x,y) = \begin{cases} 1, & D_m(x,y) > T \\ 0 & D_m(x,y) < T \end{cases} \tag{2.3}$$

where $F_m(x, y)$ and $F_{m-1}(x, y)$ are the video frames which at m frame and m-1 frame, $D_m(x, y)$ is the difference between $F_m(x, y)$ and $F_{m-1}(x, y)$, $B_m(x, y)$ is the binary image of $D_m(x, y)$ after adjusting the threshold T. Thus, if the intensity of difference $D_m(x, y)$ is greater than T, then the moving object will be represented in a binary image. The inter-frame differencing method also has disadvantages to be handled which consist of chosen value of the threshold and time difference between the frames, the threshold may influence object segmentation.

Temporal differencing approach detects a moving object by comparing pixel-wise differences between two or three sequential frames. In return, the moving regions are extracted (Hu, Tan, Wang, & Maybank, 2004; Kulchandani & Dangarwala, 2015). The approach is similar with background subtraction; the only difference is that the background image was replaced by the previous one.

Optical flow is a kind of image motion. In temporal domain, motion object can be represented as the changes of different pixels in video frames. Optical flow can be represented by projection of motion vectors from a random filed. Thus, in a video, if there is no moving object, motion vectors related to optical flow should be continuous. Once a moving object is detected, the moving object will generate a relative motion, the motion vectors will be salient. Thus positions of the moving object in a video will be located.

There are other approaches that can be used to extract moving object, which is related to deal with the complex scenario. An algorithm uses a range of grayscale intensity instead of a single grayscale one as background by calculating the minimum and maximum values of pixel and the difference of two adjacent frames constructs a background model (Haritaoglu, Harwood, & Davis, 1998).

Single Gaussian Model for each pixel could achieve background modeling (Wren, Azarbayejani, Darrell, & Pentland, 1997). Gaussian Mixture may correspond to background colors which can be described by multiple Gaussian distributions (Stauffer & Grimson, 1999). A non-parametric method for background modeling based on Kernel Density Estimation fully utilized the information in recent history of frames to update the background model which not only is adapt to the complex pixel distribution density but also overcomes the frequent changes of pixel values in a short period of time (Elgammal, Harwood, & Davis, 2000). A method of background modeling based on Bayesian decision rule classifies background and foreground from the selected visual feature vectors. The background object can be described by colors, and the color co-occurrence feature can be represented by the moving object. Thus, the objects can be classified from the foreground by selecting an appropriate feature vector (Li, Huang, Gu, & Tian, 2003). Table 2.2 summarizes the four methods with their advantages and disadvantages.

Table 2.2 Summary of four methods of object detection

Methods	**Advantages**	**Disadvantages**
Background subtraction	Fast computation; Easy to implement; Fully complete extraction of moving object;	Effected by the complex background and external environment;
Inter-frame differencing	Easy to implement; Low complexity; Not affected by external environment; Better stability;	Relied on the time interval between frames; Cannot extract a complete area of the object;
Temporal differencing	Easy to implement; Not affected by external environment; Lower computation cost;	Cannot deal with the moving object which has slow or fast movements;
Optical flow	Strong adaptability; Easily to detect the moving object;	Computationally intensive; Hard to implement; Low ability to de-noising;

2.3.2 Moving Object Tracking

For more complicated scene, tracking of moving objects needs to be considered. The major mission for object tracking is to allocate the consistent labels for moving objects in a scene. In video analysis, event detection should be thought as the one important step which motion trajectory of a moving object may need to be analyzed in temporal domain.

In 2006, object tracing and its features for object representation have been well studied (Yilmaz, Javed, & Shah, 2006). Moreover, moving object tracking is classified into four groups including feature based tracking (Comaniciu, Ramesh, & Meer, 2003), region based tracking (Wren, Azarbayejani, Darrell, & Pentland, 1997; McKenna, Jabri, Duric, Rosenfeld, & Wechsler, 2000), model based tracking (Hu, Tan, Wang, & Maybank,

2004) and silhouette based tracking (Freedman & Zhang, 2004). Mean Shift as a typical algorithm can be applied to moving object tracking (Comaniciu, Ramesh, & Meer, 2000; Yang, Duraiswami, & Davis, 2005) as well.

Kernel tracking refers to track a moving object with its shape and appearance, which is a appearance model based on template and density (Yilmaz, Javed, & Shah, 2006). The moving object will be labeled and tracked by estimating kernel motion in the continuous frames. Point tracking represents moving object as the point / centroid in the continuous frames with the correlation based on previous state of an object including position and motion of the object. Moreover, deterministic and statistical approaches were employed to point tracking, such as Kalman filtering (Weng, Kuo, & Tu, 2006). Silhouette tracking contains contour evolution and shape matching. The goal is to use the previous frame to generate object model so as to find the object region in each frame which includes object edges or contour of the moving object.

Optical flow can be utilized to moving object detection and tracking such as Horn-Schunck algorithm (Horn & Schunck, 1981) and Lucas-Kanade algorithm (Lucas & Kanade, 1981). An algorithm is developed to update template so as to avoid the drifting inherent in LK algorithm, which has been used to improve the robustness of object tracking (Matthews, Ishikawa, & Baker, 2004).

Motion can be characterized by three-dimensional velocity; two-dimensional velocity can be acquired from the projection of three-dimensional velocity. The changes of optical flow can be seen as the characteristics of moving objects (Klette, 2014).

Lucas-Kanade (LK) algorithm is a differential method which assumes the flow is constant in a local neighborhood of the pixel and could be solved in the least-square which has a less amount of calculations. The metrix is,

$$\begin{bmatrix} I_{x1} & I_{y1} \\ I_{x2} & I_{y2} \\ \vdots & \vdots \\ I_{xn} & I_{yn} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_{t1} \\ I_{t2} \\ \vdots \\ I_{tn} \end{bmatrix} \tag{2.4}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \Sigma I_{x1}^2 & \Sigma I_{x1} I_{y1} \\ \Sigma I_{x1} I_{y1} & \Sigma I_{x1}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\Sigma I_{x1} I_{t1} \\ -\Sigma I_{x1} I_{t1} \end{bmatrix} \tag{2.5}$$

where the n refers to the number of pixels, I_x and I_y represent the position of pixel at (x, y) and I_t refers to the time t.

Motion detection as the low-level process should be finished at the very begining, the video frame pre-processing can find the Region of Interest (ROI), moving object detection based on the pre-processing could find the moving object. Basically, there are four different approaches for moving object tracking, which helps us to tackle the complex sense in numerous video frames. The fundamental processing provides a great deal of support for subsequent high-level processes such as human behavior recognition.

2.4 Recognition

In order to discover the stories behind a scene, we need to detect event, therefore a high-level process is presented. In order to recognize human behaviors, visual feature vectors as the major component should be considered, which is to find the Region of Interest. Nowadays, most of surveillance systems only pay attention to the simple object detection and tracking, but there is no further process to recognize normal and abnormal human behaviors. Event based human behavior recognition is to understand human behaviors automatically in order to reduce human labor of security staff.

Spatial-temporal model (Rui & Anandan, 2000) and periodic model (Cutler & Davis, 2000) are proposed to detect and analyze the periodic motions. Moreover, event recognition can be split into two kind of approaches, which encompass model-based approaches and appearance-based approaches. In model-based approaches, Bayesian

networks typically have been used to recognize the simple events or static postures from video frames (Intille & Bobick, 2001; Hongeng, Nevatia, & Bremond, 2004), meanwhile, Hidden Markov Model (HMM) also has been applied to human behavior recognition (Oliver, Rosario, & Pentland, 2000; Tran & Davis, 2008). Appearance-based approaches are based on salient regions of local variations in both spatial and temporal domains (Laptev, 2005; Niebles, Wang, & Fei-Fei, 2008), Principal Component Analysis (PCA) as one of the examples is able to acquire eigenimages and eigenvectors by utilizing dimensional reduction which was successfully applied to face recognition (Yang, Zhang, Frangi, & Yang, 2004). Boosting is adopted to learn from a cascade of filters for efficient visual event detection (Ke, Sukthankar, & Hebert, 2005).

In addition, grammar-based and statistical-based methods (Naphade & Huang, 2002) could be categorized by using sampling support, characteristics, and mathematical modeling. Moreover, the supportive samples can be a pixel, a region or a frame of the abnormality (Tziakos, Cavallaro, & Xu, 2010). Relevance Feedback (RF) was introduced to retrieve a specific query by enquiring subjective opinion incorporated in the learning process (Su, Zhang, Li, & Ma, 2003). It is found that event description language based on multi-camera surveillance network is possible to be used in event recognition (Velipasalar, Brown, & Hampapur, 2006). An unsupervised model (Xie, Sundaram, & Campbell, 2008), consists of outliner identification and model adaptation.

Generally speaking, there are numerous approaches that can achieve a human behavior recognition, and it can also be categorized into twofold: single-layered approaches and hierarchical approaches (Aggarwal & Ryoo, 2011).

The single-layered approaches are able to recognize human behaviors from a sequence of video frames directly. In this case, the approaches are suitable for recognizing gestures and behaviors. Hierarchical approaches split the behaviors into sub-events, these sub-events are set to describe the motion which suits the advanced and complex behavior analysis and understanding. Moreover, in the latest study, orientation histogram based

approaches have been proposed which are scale invariance and rotation invariance when the gradient direction is unified.

Figure 2.1 shows the structure of human behavior recognition from surveillance videos with various approaches. At the third level, it indicates sub-approaches.

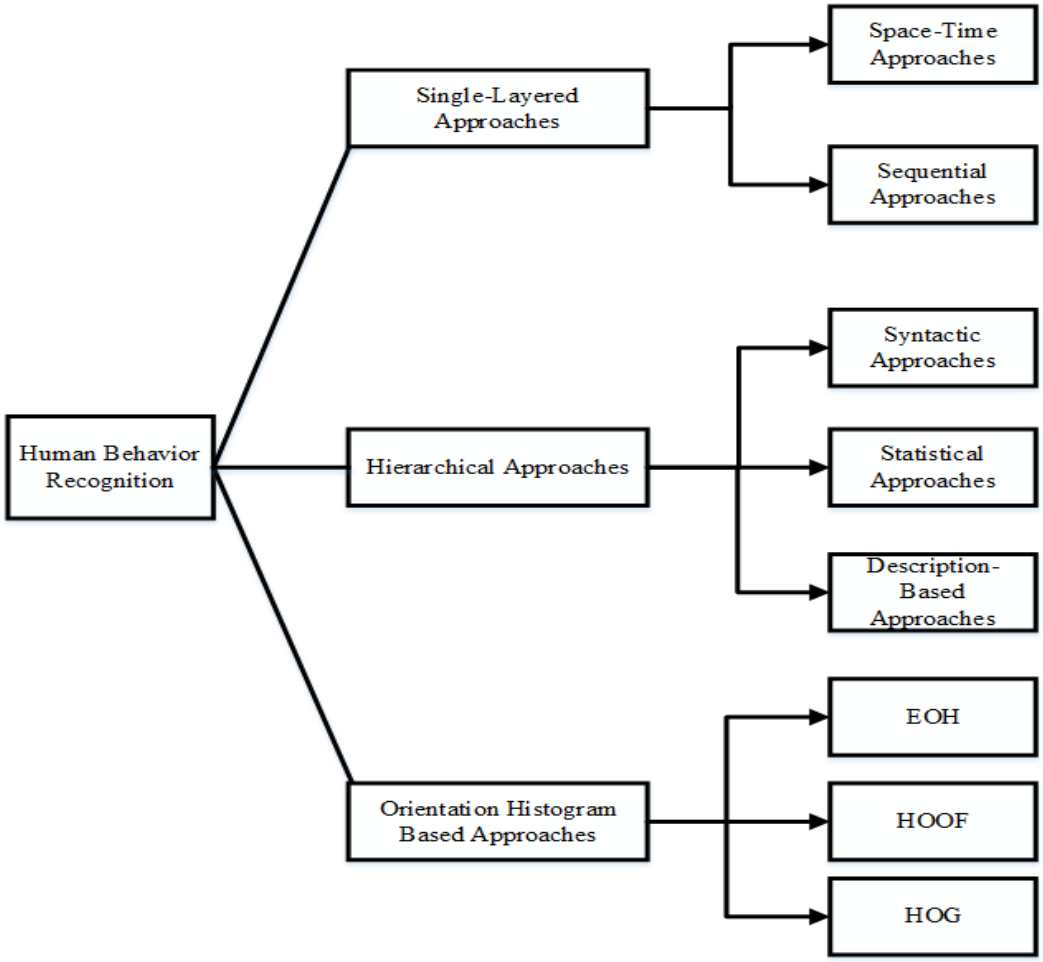


Figure 2.1 Different approaches to human behavior recognition

Single-layered approaches can be divided into two groups which include space-time approaches and sequential approaches. In 2008, a template-based method was proposed to recognize human behavior based on three-dimensional spatio-temporal volume for a given class (Rodriguez, Ahmed, & Shah, 2008).

A Motion-energy Image (MEI) and Motion-history Image (MHI) were proposed to replace the three-dimensional spatio-temporal volume, the combination of these two methodologies with template matching is able to recognize simple human behaviors.

While the MEI determines location of the moving object in spatial domain occurred in the frame sequence, and the MHI reflects the moving object and its motion intensity in different locations (Bobick & Davis, 2001).

Correlated spatio-temporal shapes have been segmented by utilizing the shape and flow-based features in order to achieve human behavior recognition, the mean shift was adopted to cluster the voxel location and color (Ke, Sukthankar, & Hebert, 2007). Three-dimensional spatio-temporal volume is utilized to correlate dynamic behaviors based on video templet which allows recognition of complex behaviors in the video frames (Shechtman & Irani, 2005).

Local spatio-temporal features at a multiple scales are proposed in 2001 (Zelnik-Manor & Irani, 2001). The sparse interest points in spatio-temporal domain are extracted from video frames which could successfully recognize human behavior by using the Harris feature descriptor (Harris & Stephens, 1988) (Laptev & Lindeberg, 2003). A spatio-temporal matching is to measure the structural similarity of features extracted from different video frames achieves a complex non-periodic behavior recognition (Ryoo & Aggarwal, 2009b). A Linear Time Invariant (LTI) converts the sequence of video frames into silhouette sequence, then utilizes vectors of width and Fourier descriptors to recognize human behaviors (Lublinerman, Ozay, Zarpalas, & Camps, 2005). A dynamic time warping (DTW) matching algorithm was developed to recognize human behavior, the approach learned the time warps at the same tine allows the variations in descriptors (Veeraraghavan, Chellappa, & Roy-Chowdhury, 2006). Moreover, Hidden Markov Models (HMMs) (Natarajan & Nevatia, 2007) and Dynamic Bayesian Networks (DBNs) (Park & Aggarwal, 2004) have also been used in the sequential approaches for human behavior recognitions.

Hierarchical approaches mainly have three different divisions including statistical approaches, syntactic approaches and description-based approaches. The statistical approaches recognize human behaviors by utilizing multiple layers of the statistical state-based model. Two layers of hierarchical HMMs recognize human behaviors for the

complex sequential structures (Nguyen, Phung, Venkatesh, & Bui, 2005). In this case, DBNs include multiple levels of hidden states to represent human behaviors hierarchically (Dai, Di, Dong, Tao, & Xu, 2008). The syntactic approaches take human behaviors as a string of symbols, and each of symbols represents at atomic level. In 2000, a hierarchical approach of stochastic context-free grammars (SCFGs) is applied to recognize high-level human behaviors, two layers were adopted in the approach which includes an HMMs layer to recognize the atomic behaver and stochastic parsing as another layer to classify high-level behaviors (Ivanov & Bobick, 2000). The syntactic approaches are as same as the statistical approaches which are able to recognize the sequential behaviors. But it also has a limitation that the approach cannot recognize the sub-events with a complex behavior.

A description-based approach is usually designed for complex behavior recognition. Using Context-free Grammar (CFG) as a language to describe the complex relationship between temporal and spatial among the sub-event, Bayesian networks and HMM were also utilized for recognizing atomic actions from video frames (Ryoo & Aggarwal, 2006). Later, a probabilistic model was extended for human behavior recognition in order to reduce failures of the low-level components (Ryoo & Aggarwal, 2009a). In 2009, a probabilistic analysis was presented by using a description-based approach which aims to recognize human actions at the atomic-level (Gupta, Srinivasan, Shi, & Davis, 2009).

Orientation histogram based approaches are adopted to recognize human behaviors. Edge Orientation Histogram (EOH) was offered to the gradient orientation for feature extraction, and the feature characterizes one true value of the orientation, and the EOH was proposed for face detection (Levi & Weiss, 2004). Later a pedestrian detection framework based on the EOH method was developed (Chen & Chen, 2008). Motion orientation histogram (MOH) was proposed to recognize human motions in video based on Motion History Image (MHI). A gradient based motion pyramid method was also developed to extract the local motion from images (Davis, 2001). Histogram of Oriented Gradient (HOG) was proposed to achieve pedestrian detection (Dalal & Triggs, 2005) and it also is utilized as the descriptor of human behavior, the HOG feature has the scale

invariance. Histogram of Oriented Optical Flow (HOOF) approach was utilized to a Nonlinear Dynamical System (NLDS) with the output of space of histograms (Chaudhry, Ravichandran, Hager, & Vidal, 2009).

There are three different approaches which can extract key features of human behavers. Furthermore, for the space-time approaches utilized for the recognition of periodic actions, the limitation is that the approaches were not suitable for more complex behavers. The sequential approaches rely on sequential relationship of visual features that can be adopted to detect more complex behaviors but may require numerous videos for training. The hierarchical approaches are working on high-level behavior recognitions, the approaches can decompose these high-level behaviors into sub-events so as to make the behavior recognition simple, and it can deal with less training data. However, the limitation is that both statistical and syntactic approaches cannot recognize behaviors with the concurrently organized sub-events, but the description-based approaches can overcome this. The orientation histogram based approaches can get the features from the spatial domain and also can acquire information from the temporal domain. Moreover, the histogram also can help in recognizing the behaviors that have large motion.

2.5 Classifications

We introduced multiple recognition approaches in order to extract a key feature of human behaviors. As a high-level processing, the recognition helps us recognize human behaviors such as Running, Walking, Hand waving, etc. Moreover, machine learning based human behavior classification should be considered in behavior recognition.

Machine Learning (ML) is mainly working for a classifier to classify the computable feature vectors extracted from video frames (Duda, Hart, & Stork, 2001; Turaga, Chellappa, Subrahmanian, & Udrea, 2008). There are multiple classifiers which are basically divided into twofold which include supervised learning and unsupervised learning. This section will introduce different approaches for human behavior recognition.

2.5.1 Supervised Learning

In supervised learning, the training set will be manually assigned to a class number or label, then a classifier will map all the input for the corresponding output. As the result, the simple judgment of an output will reflect the classification output and it will have the ability to classify further coming unknown data. There are numerous supervised learning methods, such as Support Vector Machine (SVM), *k*-Nearest Neighbor (KNN), Artificial Neural Networks (ANNs), etc.

For the recognition on spatial domain, local features were usually combined with SVM for a robust classification. SVM more relies on the number of the support vectors instead of dimensionality of the feature space (Vapnik, 1999). Therefore, it was developed for optimal classification under a linear situation to maximize the margin between different categories.

A multiclass SVM classifier contains three different binary classification methods: 'one-against-all', 'one-against-one' and Directed Acyclic Graph SVM (DAGSVM), the last two methods should be more suitable for practical use (Hsu & Lin, 2002). SVM classification was utilized with the combinations of local features and local histograms to recognize human behaviors (Schuldt, Laptev, & Caputo, 2004).

Dynamic motion features with multi-layer perceptron (MLP) and SVM were taken advantage for implementing an efficient result in human motion recognitions (Kapur, Kapur, Virji-Babul, Tzanetakis, & Driessen, 2005). Local space-time features and space-time pyramids with the multichannel non-linear SVMs were employed to learn the realistic human actions from movies (Laptev, Marszalek, Schmid, & Rozenfeld, 2008). A hierarchical classification with aggregated multiple SVMs was adopted to recognize human behavours.

SVM is trained to choose the proper features and then aggregated together into a hierarchical classification pattern in order to achieve the best performance (Qian, Mao, Xiang, & Wang, 2010). A SVM with Kernel Principal Component Analysis (KPCA) has

the ability to make the consistent kernel transformations over training and testing samples which are possible to solve both binary and multiclass classification problems (Gu & Guo, 2012).

The *k*-Nearest Neighbor (KNN) as a supervised learning method can be utilized for either classification or regression, it is a fuzzy algorithm which was first introduced by Keller et al. (Keller, Gray, & Givens, 1985). For classification, the output of *k*-NN is the class label, the main core of KNN is that the class will be classified by the majority vote of its neighbors and should be assigned to the class which is most commonly appeared in its *k* (integer number) nearest neighbors.

An optical flow based motion descriptor was used in a spatio-temporal domain combined with the nearest neighbor framework (Efros, Berg, Mori, & Malik, 2003). In 2007, the nearest neighbor with Euclidian distance is used to operate the normalized features (based on a space-time cubes), correctly classified behavior was related to the retrieved nearest neighbor (Gorelick, Blank, Shechtman, Irani, & Basri, 2007). An optimal alignment to compute the distances of action-to-action combines with the *k*-NN classifier, the non-modeled actions will be rejected by thresholding the distances (Lin, Jiang, & Davis, 2009). A comparative study of classifying human actions emphasized classifiers including SVM, ANN, KNN, DTW, LSM (least-squares method) and BDM (Bayesian decision making). The study shows *k*-NN method is accurate, and it also requires considerable time for classification (Altun & Barshan, 2010). There are other disadvantages of *k*-NN algorithm such as learning rate and a number of training samples in the different classes. The classification deals with a large amount of calculations that will lead to reduce the efficiency. For unbalanced samples of different classes, the output sample may be classified into the class which has a large amount of samples.

ANNs have been well studied in past decades (Bajpai, Jain, & Jain, 2011). Using ANN, computations could be processed by a network of simple binary neurons (Sondak & Sondak, 1989). Most ANN algorithms are based on supervised learning model, for an example, BP (Back Propagation) algorithm is based on multilayer feedforward network

which also consists of three layers and required training data. The traditional BP algorithm is a learning method which converts the machine learning problem to a non-linear optimization problem by using gradient descent algorithm and iterative computations. BP algorithm is one of the most multilayer networks which have been utilized in supervised learning. Although ANNs consists of different layers, at the same layer, each node has not relationship with others, and different level of layer nodes may present as a decreasing trend (Basheer & Hajmeer, 2000). However, the result may converge to the local minimum and also have slow learning speed associated with computational complexity (Burse, Manoria, & Kirar, 2011).

An experimental was conducted and applied the classifiers to detect pedestrian (Munder & Gavrila, 2006). A stereo-based segmentation algorithm is combined with a neural network in order to achieve object extraction and recognition (Zhao & Thorpe, 2000).

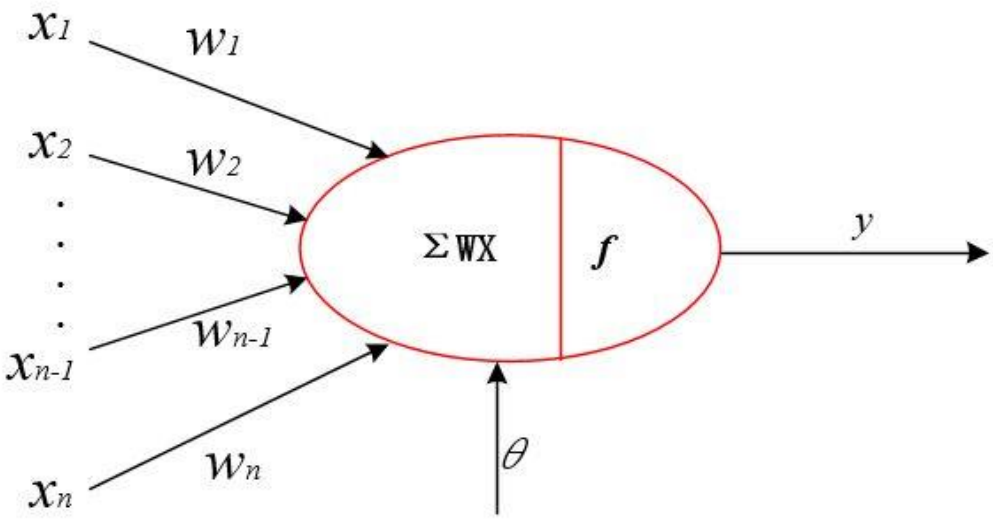


Figure 2.2 The structure of an artificial neuron.

As shown in Figure 2.2, the structure of an artificial neuron is introduced and it has been shown as a Processing Element (PE) for the node with a weighted directed chain connected to each other. Thus, x_i represents the information from PE and the intensity of

interaction with the PE is w_i^{θ} represents the internal threshold of the PE and *f* stands for the activation function which determines the output of ANN neurons.

Generally, ANN is similar to our human brain. Neurons are the basic processing unit in an ANN framework with the aggregation of numbers of neurons which construct a neural network. ANNs are more like a brain because of combinations amongst its layers. However, the ANNs are not a real human brain because of an artificial neuron, the weights as a number show up to represent the synapse, an activation function is required in ANN model to control amplitude of the output (Bajpai, Jain, & Jain, 2011).

More importantly, after the training, it cannot be changed. If any new data is added into the network, the mode needs to be trained again. (Miller & Khan, 2011).

$$E = \frac{1}{2}\sum_{p=1}^{n}\left(d_p - y_p\right)^2 \tag{2.6}$$

The BP algorithm uses a gradient search technique which decreases the error function as shown in eq. (2.6), where the y_p is the output, d_p is the desired output of the input pattern. Moreover, it also presents the input samples as well as the corresponding desired output. The network neurons were operated to calculate the hidden layers until the output data shows each of the output values. Figure 2.3 illustrates a structure of BP neural network.

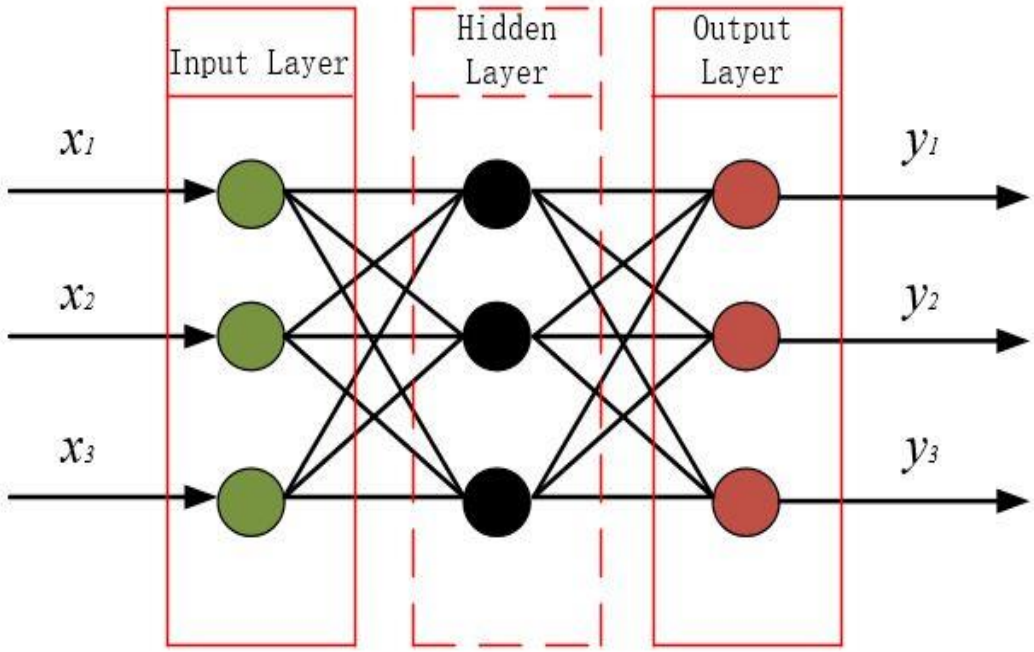


Figure 2.3 Structure of BP neural network.

The steps for ANNs to recognize human behaviors are shown as below.

- Prepare the training dataset (feature extraction) and the target dataset (labeled the different classes).
- Train the neural network with the adjusted parameters.
- Simulate the train neural networks with the testing dataset.

2.5.2 Unsupervised Learning

For unsupervised learning, the major difference with supervised learning method is that it does not require pre-trained samples which means the classes between different samples are unknown.

A multi-observation Hidden Markov Model with spectral clustering is used to achieve an unsupervised training model that can detect abnormal activities (Xiang & Gong, 2005). A Self-organizing Map (SOM) approach was offered to cluster and visualize events so as to achieve the recogniztion of abnormal behaviors (Petrushin, 2005). An unsupervised

learning based on Gaussian mixture model (GMM) is used to recognize various human behaviors (Allen, Ambikairajah, Lovell, & Celler, 2006). Moreover, the algorithm also can localize multiple actions in a complex scenario which contains multiple actions (Niebles, Wang, & Fei-Fei, 2008). Spatial-temporal local motion features combine with an unsupervised generative model to learn different actions (Savarese, DelPozo, Niebles, & Fei-Fei, 2008). The maximization of mutual information (MMI) is with *k*-means in order to learn human behaviors (Liu & Shah, 2008). In 2013, a method was proposed based on multi-dimensional time series using HMM in multiple regression, combined with expectation-maximization (EM) algorithm as an unsupervised framework, which has not annotations for all activities (Trabelsi, Mohammed, Chamroukhi, Oukhellou, & Amirat, 2013).

In this section, we introduced multiple machine learning methods, both supervised learning and unsupervised learning can be utilized to recognization of human behaviors. Supervised learning method requires training samples in order to get an optimal model and uses it to predict the given sample from the well-trained models. Thus, all the classes are known, the ambiguity of the training samples is low. However, unsupervised learning does not require training samples, it studies structural knowledge of training samples without labels. Therefore, all the samples are unknown, the ambiguity of the training samples are high.

Chapter 3
Methodology

The main content of this chapter is to clearly articulate research methods, which satisfy the objectives of this thesis. The chapter mainly covers the details of research methodology for human behavior analytics which will be clearly introduced with the confident and imaginative use of the feature description methods.

3.1 Research Designing

To design and implement human behavior recognition is the main purpose of this thesis. As shown in Figure 3.1, the flowchart of human behavior recognition for each step is clearly pointed out through six stages as the structure of this research.

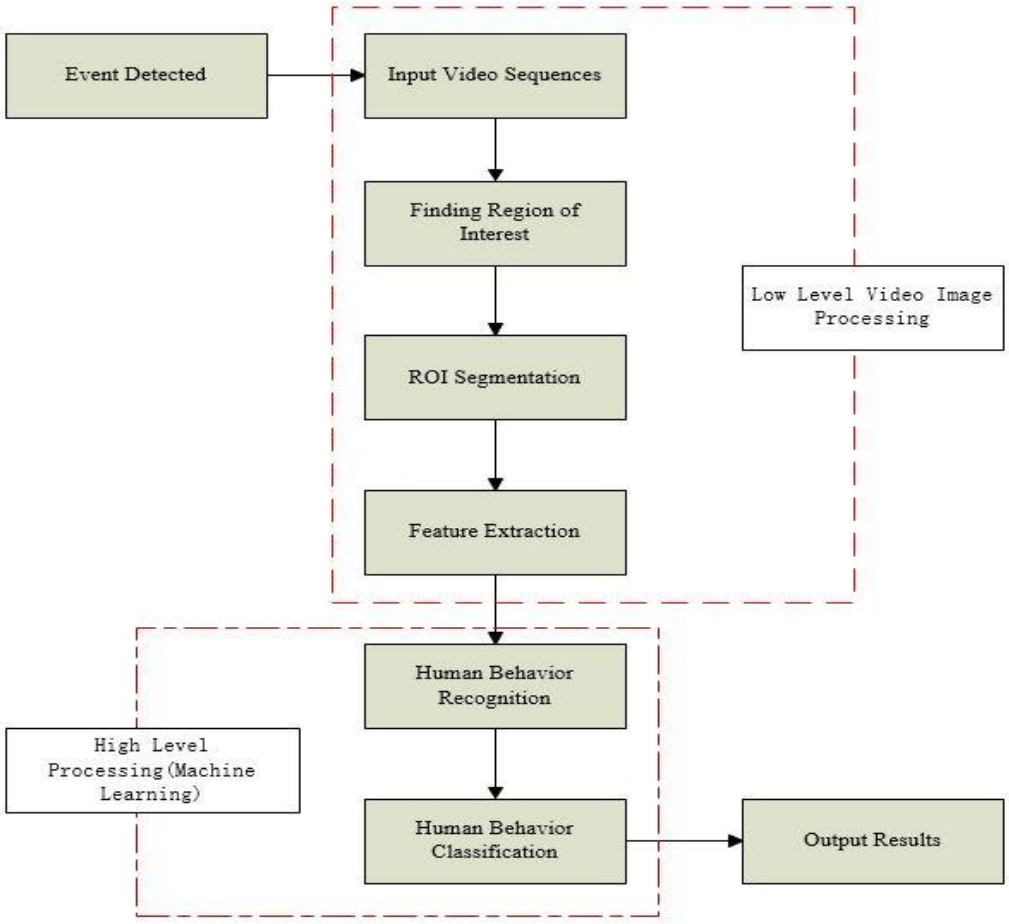


Figure 3.1 The flowchart of the steps of human behavior recognition

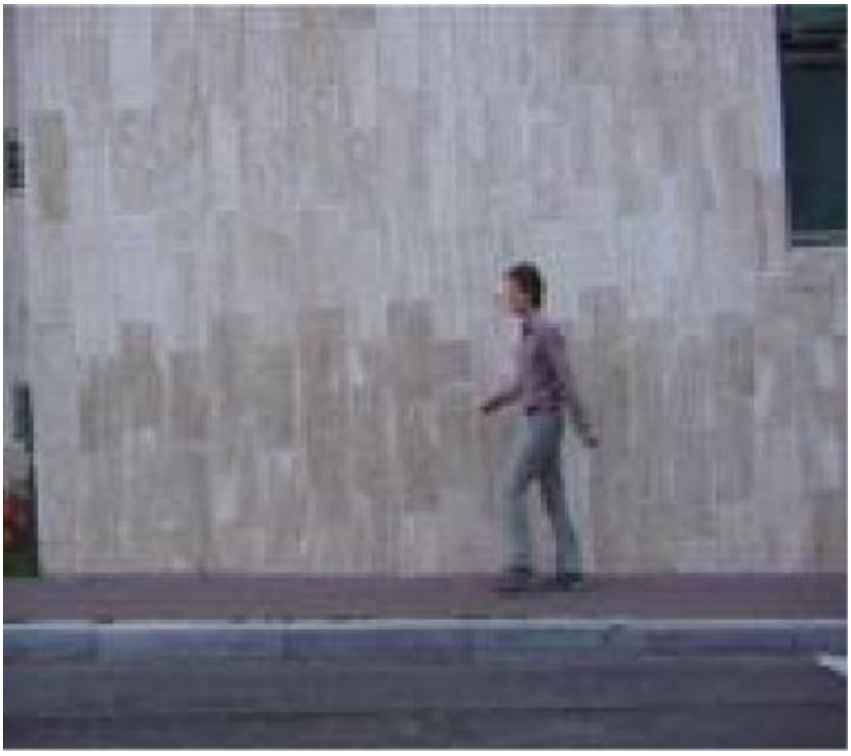
The first three steps are thought as the low-level processing which mainly includes detecting the Region of Interest (ROI), segmenting ROI region and extracting computable features. To achieve human behavior recognition, the high-level processing for semantics and annotation will also need to be taken into account.

3.1.1 Locating the ROI

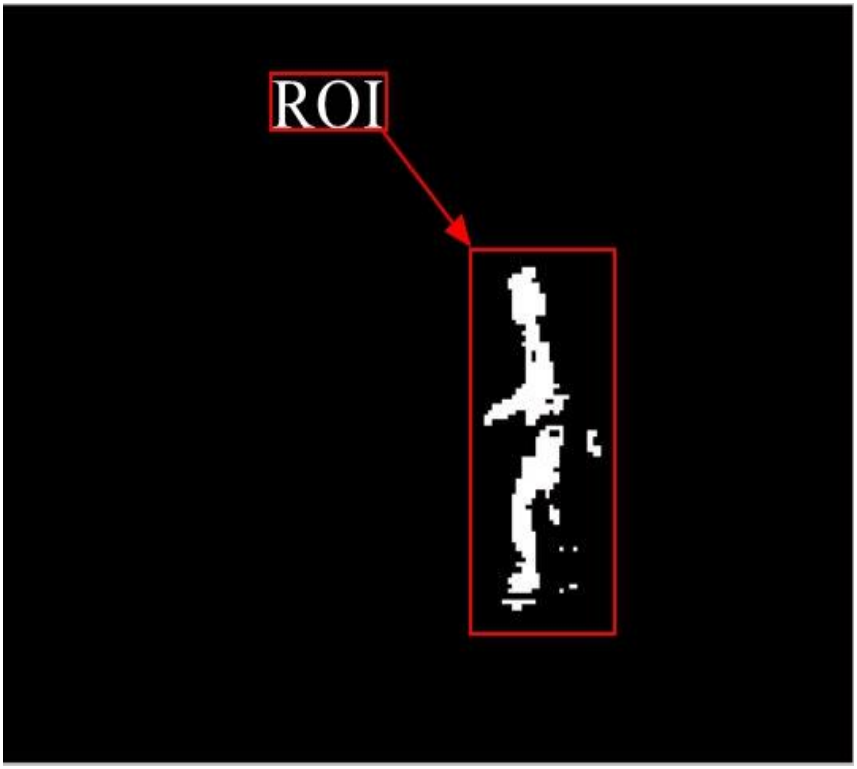
In the most of surveillance applications, object detection is the main step before the further coming processing. In this case, finding the location of an object or ROI needs to be considered at the very beginning. Therefore, locating the ROI is the first work of human behavior recognition in this thesis project.

To recognize human behaviors, we need to distinguish background and foreground of a video. The event of 'Running through a surveillance camera' could be determined by detecting the ROI and further high-level understanding in semantics. In this section, we will introduce the ROI detection which is mainly utilized to human behavior recognition. As we mentioned in Chapter 2, there are different methods to find the ROI, we usually think the image that contains a human body is thought as the ROI which eventually needs to be detected and recognized. The ROI needs to be segmented from the original frames just in order to extract visual features easily and reduce the training time of classification.

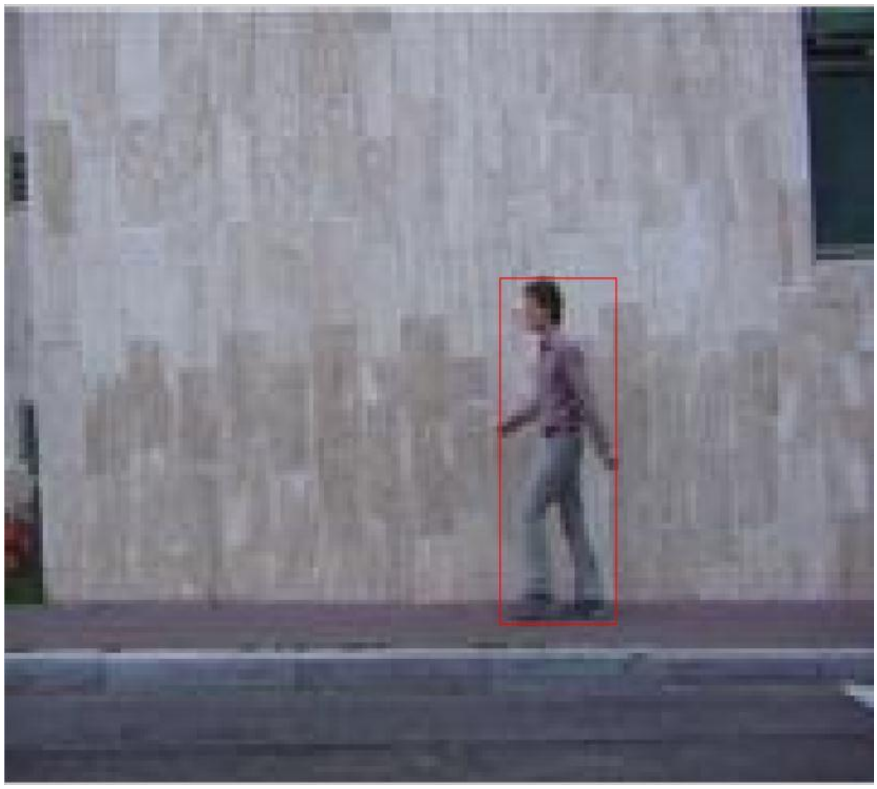
Figure 3.2 illustrates how to locate the ROI, Figure 3.2(a) shows the original image in a video. In Figure 3.2(b), the red rectangle shows the area of ROI which needs to be segmented, Figure 3.3(c) shows the ROI in RGB color space, Figure 3.2(d) and (e) display the grayscale images, Figure 3.2(f) indicates the binary image followed by the direct result of Figure 3.2(e). Figure 3.3 demonstrates an example of the ROI area, which needs to be segmented in the consecutive frames of a video.



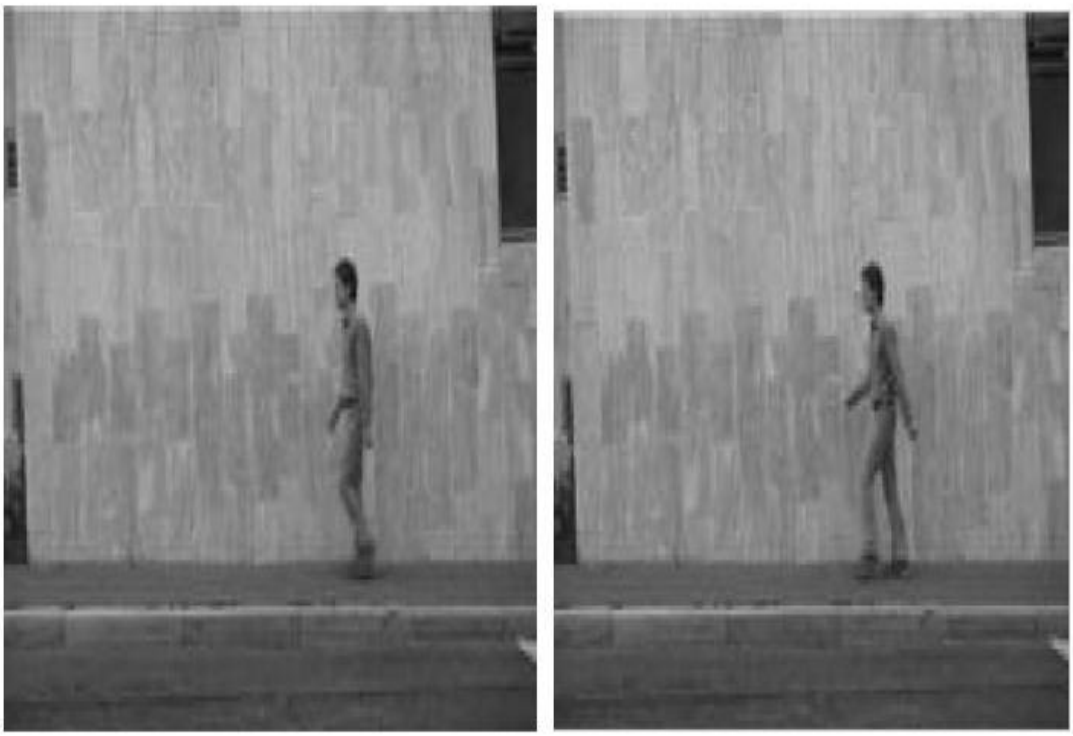
(a) The original image (in Walking frame 23)



(b) The area of ROI that need to be segmented



(c) The area of ROI in RGB color space



(d) The single grayscale image (in Walking frame 22 and 23)

(e) The grayscale image by utilizing Inter-frame differencing

(f) The binary image followed by the directly result of (e)

Figure 3.2 Locating ROI area in the video frame

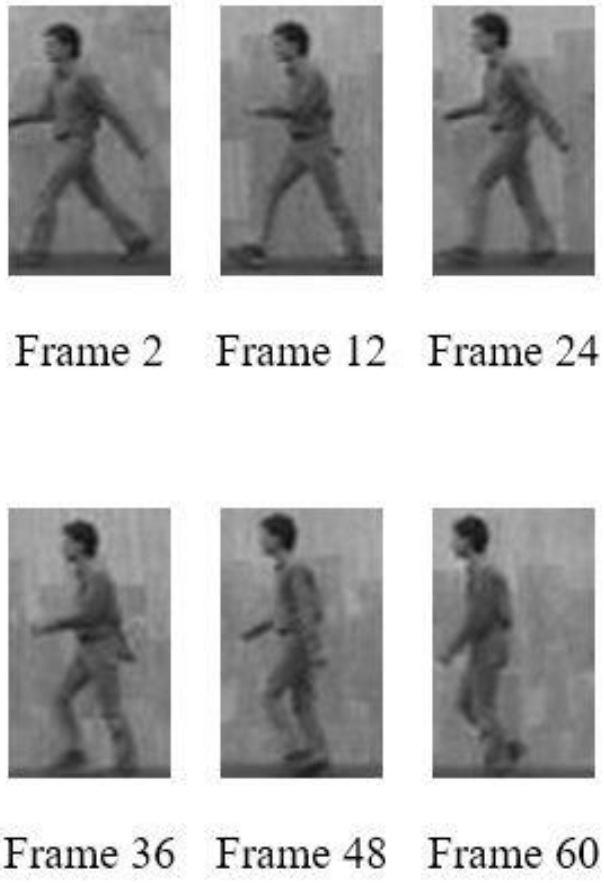


Figure 3.3 The segmentations of ROI area in a video stream

3.1.2 ROI Segmentation

Once the area of ROI was detected, the next step is to segment the ROI from the foreground. As we mentioned in Chapter 2, the majority of this thesis is to recognize the behaviors of a human body, which involves visual understandings at semantic level. Due to the segmentation in each video frame, the size of ROI may be different. To achieve the high-level processing, the size of each ROI image needs to be normalized.

Figure 3.4 shows the five output classes with the ROI in each of greyscale image which is extracted from various video frames. Algorithm 3.1 represents how to locate the ROI as well as how to segment the ROI.

(a) Jack (b) Jump (c) Skip (d) Walk (e) Run

Figure 3.4 Examples of the ROI for five output classes after segmentation.

Algorithm 3.1 Finding and Segmentation of Region of Interest

Input: Video which contained human behavior

Output: The segmented Region of Interest video frames

```
Set wmax=0, hmax=0;
Read image rgb_image from video frame;
Convert rgbFrame to grayFrame;
Convert grayFrame to BW;
[row, col] = size(BW);
For a=1 to row do
    For j=1 to col do
        If BW(a, j) == 1 then
            y1=a;
        End if
    End for
End for
For a=row to -1 to 1 do
    For j=col to -1 to 1 do
        If BW(a,j) == 1 then
            y2=a;
        End if
    End for
End for
For a=1 to col do
    For j=1 to row do
        If BW(j,a) == 1 then
            x1=a;
        End if
    End for
End for
For a=col to -1 to 1 do
```

```
        For j=row to -1 to 1 do
            If BW(j,a) == 1 then
                x2=a;
            End if
        End for
    End for
    If wmax less than (x2-x1) then
        wmax=(x2-x1);
    End if
    If hmax less than (y2-y1) then
        hmax=(y2-y1);
    End if
Set xc= centroid of weight and yc= centroid of hight;
cp=imcrop(grayFrame,[xc-wmax/2,yc-hmax/2,wmax,hmax]);
```

3.1.3 Feature Extraction

Three distinct Feature Extraction Techniques (FET) were investigated and implemented in human behavior recognition in order to achieve the objective of this thesis, which contains Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP).

To achieve video classification of human exercising activities (Walking, Skipping, Running, Jacking, and Jumping), we take into consideration of visual features having spatial and temporal information between adjacent video frames. HOG is the visual feature which is adopted for object recognition in computer vision. As we know HOG descriptors provide better performance comparing to other features for human behavior detection and recognition. Figure 3.5 illustrates the flowchart of HOG feature extraction.

The HOG feature extraction is one of the important steps for pedestrian detection and recognition. The HOG features for human detection have been trained and tested after normalization, gradient computation and spatial organization. The main idea of HOG descriptor is to calculate occurrences of gradient orientation in localized portions of an image.

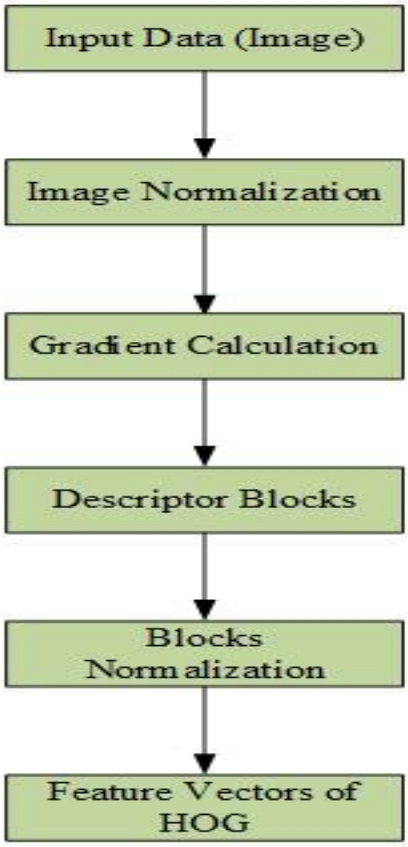


Figure 3.5 The flowchart of the HOG feature extraction.

The core idea of HOG descriptors is to detect local object shapes using light intensity gradient or edge distribution. The implementation of HOG descriptors has been achieved by segmenting the image into small connection regions which called 'cells'. In the HOG experiment, the 8×8 cell size was adopted. Each of cells generates a histogram of oriented gradients, or cell edge direction of the pixel, the combination of these histograms is applied to express a descriptor. Figure 3.6 shows the extracted HOG features of different human behaviors based on Weizmann dataset.

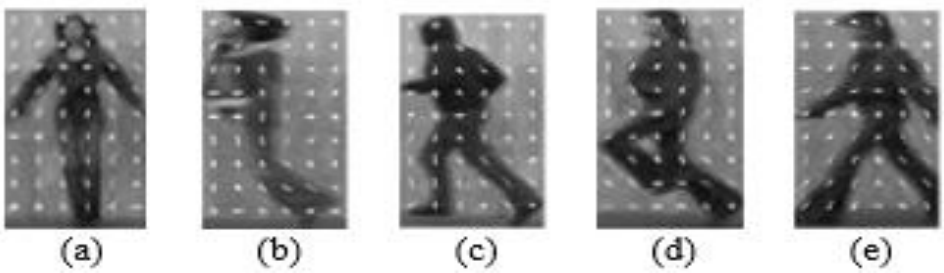


Figure 3.6 Feature extractions by using HOG for different human behaviors: (a) Jacking, (b) Jumping, (c) Running (d) Skipping and (e) Walking.

To improve the performance, local histogram is normalized by calculating the intensity in a large region across the image called 'block', which illustrates a number of cells in a block and the 2×2 block size was adopted in the experiment. Then it is used to normalize all cells within the block. After normalization, illumination and shadow have been greatly changed. Figure 3.3 (e) shows gray scale images from the Weizmann dataset. Histograms of Oriented Gradients (HOG) descriptors have the advantages, which effectively describe local shapes. By changing the number of bins and cell size of the histogram, it is able to capture an image of the local region.

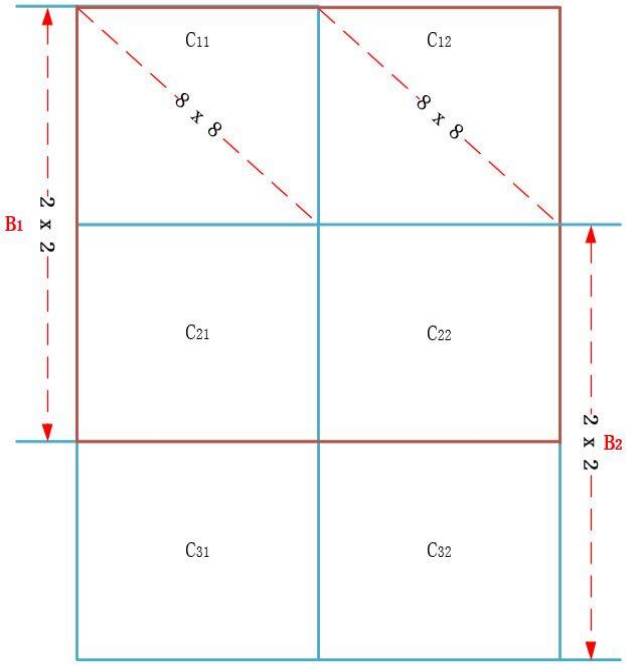
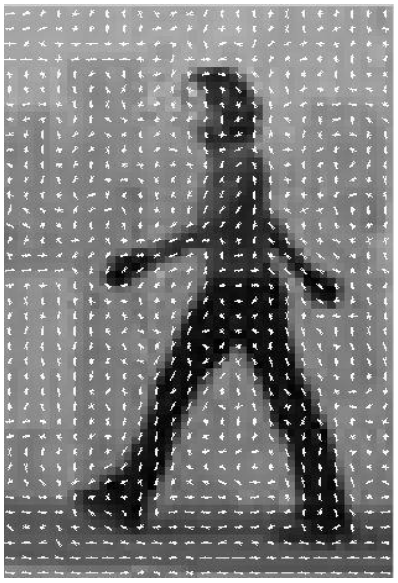
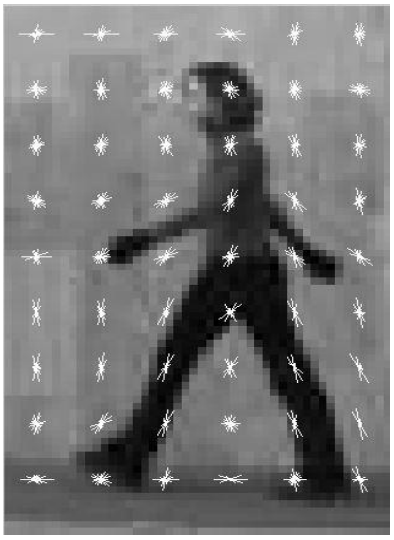


Figure 3.7 The deployment of HOG feature vectors

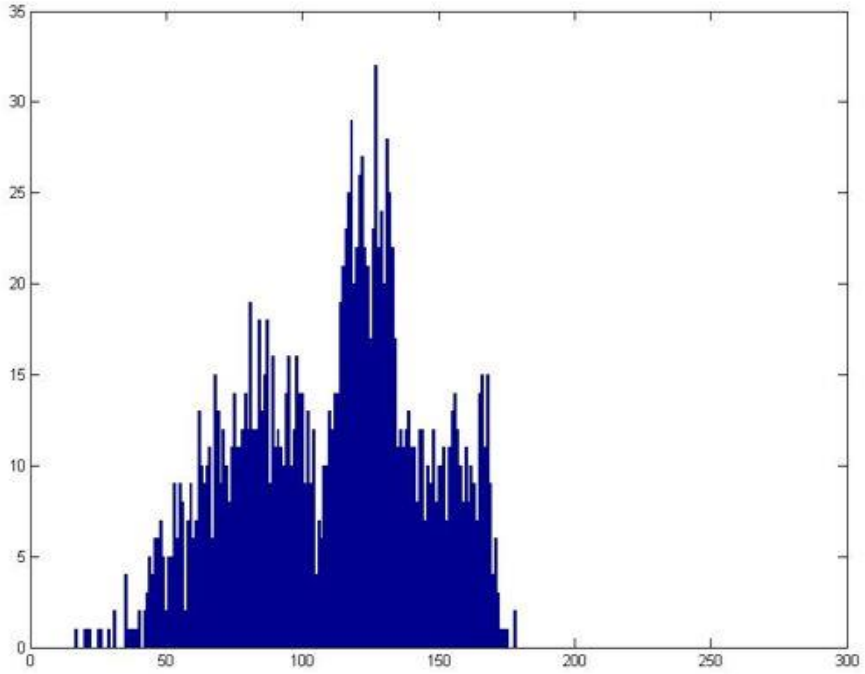
Figure 3.7 illustrates an example of the deployment of HOG feature vectors, one block (B_1) contains four cells (C_{11}, C_{12}, C_{21}, C_{22}). The second block B_2 also contains four cells (C_{21}, C_{22}, C_{31}, C_{32}). Therefore, HOG descriptor in this experiment is organized as 8×8 cell size. The 8×8 cell size does not include too much shape information. However, cell size will increase the dimensionality of the HOG feature vectors which also increase the computational time. The length of HOG feature vector in the experiment is 1440. Figure 3.8 shows HOG features in the single segmented frame.



(a) 2-by-2 cell size of HOG feature



(b) 8-by-8 cell size of HOG feature



(c) Projection result of the histogram in the single segmented frame

Figure 3.8 HOG feature example with two different parameters of cell size

HOG descriptors have the advantages which effectively describe the local shape of the images (Dalal & Triggs, 2005). By changing number of bins and cell size of the histogram, it is able to capture an image of the local region. However, because of the nature of gradient descriptor is quite sensitive to noises, Algorithm 3.2 shows the basic step of HOG feature extraction.

Algorithm 3.2 Histogram of Oriented Gradient Feature Extraction

Input: The segmented Region of Interest video frames

Output: Extracted HOG features

For video =1 to 30 **do**
    For i =3 to 28 **do**
        Read image *I* from the segmented Region of Interest video frame;
        extractHOGFeatures(I,'CellSize',[8 8]);
        save featureVector to datafeature;
    End for
End for
Convert cell to mat file;
Save mat file;

Local Binary Pattern (LBP) as another feature extraction method is also studied in

this thesis. The LBP operator is used to describe local texture feature of the operator which has the rotation invariance and gray scale invariance. It is simple but very effective, it compares each pixel with its nearby pixels and then stores the result as a binary number.

The original LBP operator splits the image into 16×16 cells, and each cell is defined within a fixed 3×3 window, the pixel in the central window will be defined as the threshold (red circle), and the average value of adjacent eight pixels is shown in Figure 3.9. If the value of surrounding pixels is greater than the value of center pixel, then that pixel will be marked as '1'. On the contrary, the pixel will be marked as '0'. Therefore, the marked eight numbers are combined together as LBP feature of the central pixel.

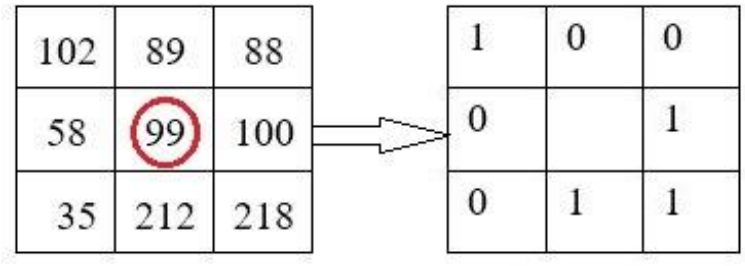


Figure 3.9 The simple mechanism of original LBP operator

After we get LBP value of each center pixel, histogram of each cell (frequency of each number) will be calculated, normalization for the histograms will also be conducted. Thus, all the histogram of cells will represent the LBP feature vector of an image. The length of LBP feature vector in the experiment is 708. Figure 3.10 shows the outcomes of extracted LBP features of different human behaviors based on the Weizmann dataset.

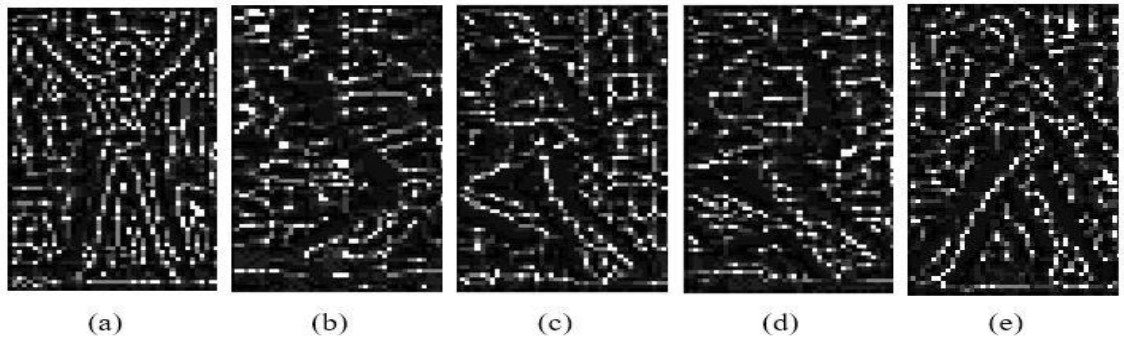


Figure 3.10 Feature extractions by using LBP for different human behaviors: (a) Jacking, (b) Jumping, (c) Running (d) Skipping and (e) Walking.

Like Local Binary Pattern (LBP), Scale Invariant Local Ternary Pattern (SILTP) as another local pattern was studied in 2010 which is effective for illumination variations (Liao, Zhao, Kellokumpu, Pietikainen, & Li, 2010).

SILTP is very similar to LBP, the grayscale intensity of the central pixel will be compared with its neighbour pixels. However, a scale factor was proposed in SILTP to indicate the range and one more comparison will be conducted in SILTP. Figure 3.11 illustrates the simple mechanism of the SILTP operator. Once the grayscale intensity of one pixel is less than the range of the minimum value of the central pixel, the pixel will be marked as '10'. If the intensity value of one pixel is greater than the range of the maxima value of the central pixel, the pixel will be marked as '01'. And if the grayscale value of one pixel is within the range of central pixel ($64\times(1\pm0.1)$), the pixel will be marked as '00'.

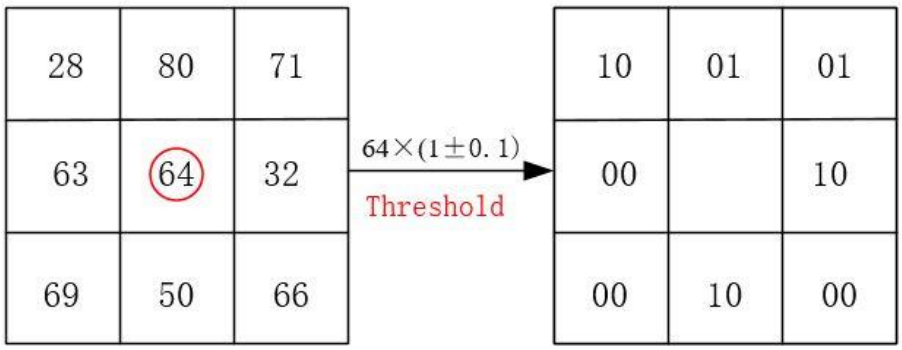


Figure 3.11 The mechanism of SILTP operator

Scale Invariant Local Ternary Pattern (SILTP) has three advantages in the local image including efficiency in computation, robustness to local noises and robustness to illumination changes. The length of a SILTP feature vector in the experiment is 3648. Figure 3.12 shows the extracted output of SILTP features for different human behavior analysis based on the Weizmann dataset.

Figure 3.12 Feature extractions by using SILTP for different human behaviors: (a) Jacking, (b) Jumping, (c) Running (d) Skipping and (e) Walking.

3.2 Evaluation Methods

In this project, the evaluation was performed by using different classifications in order to calculate the classification accuracy. For the confusion matrix, the main purpose is to compare the classification results with the ground truth, and provide a visualized classification accuracy within the confusion matrix. In other words, the matrix shows the relationship between the outcomes and the predicted classes.

Table 3.1 illustrates the basic format of the confusion matrix and the meanings:

- A is the number of correct predictions, but the instance is negative.
- B is the number of incorrect predictions, but the instance is positive.
- C is the number of incorrect predictions which the instance is also negative.
- D is the number of correct predictions which the instance is positive.

There are several standard terms defined in the matrix including accuracy, recall (also called true positive rate), false positive (FP) rate, true negative (TN) rate, false negative (FN) rate, and the precision.

Table 3.1 The confusion matrix

		Predicted	
		Negative	Positive
Actual	Negative	A	B
	Positive	C	D

Accuracy (AC) refers to the proportion of total number of the correctly predicted. For the equation it is presented as eq. (3.1).

$$\text{AC (accuracy)} = \frac{(A+D)}{(A+B+C+D)} \tag{3.1}$$

Recall (R) is called true positive (TP) which is the ratio of the positive cases which are identified correctly as eq. (3.2).

$$\text{TP Rate (Recall)} = \frac{D}{(C+D)} \tag{3.2}$$

False positive (FP) is the proportion of negative cases, which are classified incorrectly as the positive. It is calculated by using eq. (3.3).

$$\text{FP Rate} = \frac{B}{(A+B)} \tag{3.3}$$

True negative (TN) is defined as percentage of the negative cases, which are classified correctly. It is computed by using eq. (3.4).

$$\text{TN Rate} = \frac{A}{(A+B)} \tag{3.4}$$

False negative (FN) is rate of the positive cases, which are classified incorrectly as the negative, the equation is shown as (3.5).

$$\text{FN Rate} = \frac{C}{(C+D)} \tag{3.5}$$

The last is precision, which is a fraction of the predicted positive cases, it is correct and obtained by using eq. (3.6).

$$\text{P (precision)} = \frac{D}{(B+D)} \tag{3.6}$$

Chapter 4
Results

The main content of this chapter is to introduce the schema of method and implementation of human behavior recognition. Each step in human behavior recognition will be detailed; in addition, data collection with the configured experimental environment will be articulated in this chapter, the results of human behavior recognition will be clarified. Moreover, the results and findings will be evaluated as well as the limitations of this thesis will be pointed out at end of this chapter.

4.1 Data Collection and Experimental Environment

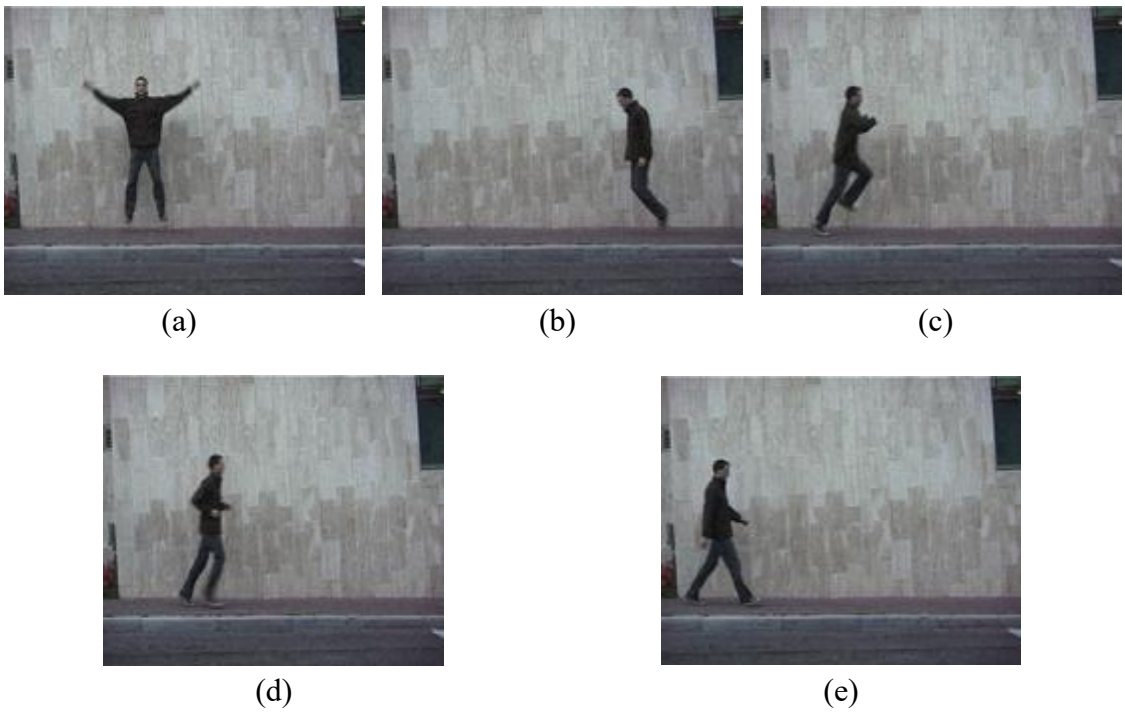
Since focus of this thesis is on implementing a method which is able to detect and analyze moving object in front of a surveillance camera in order to achieve human behavior recognition. Thus, the data needs to be collected in this research project. Moreover, data should be prepared in different ways, it does not only need to contain series of the human behaviors, but also need to contain a series of the participants.

There are numerous video databases, which are well organized for human behavior recognition. Moreover, the behavior of human body is regarded as a special case with two-dimensional of silhouettes motion. There are multiple research groups provided the datasets on their websites for human behavior recognition, such as Weizmann dataset, KTH dataset , and UCF dataset, etc. In this thesis, the Weizmann dataset is chosen for human behavior recognition.

Table 4.1 Weizmann dataset.

Class	**Verb set**	**Class**	**Verb set**
1	Walk	6	One-hand wave
2	Run	7	Two-hands wave
3	Jump	8	Jump in place
4	Gallop sideways	9	Jumping jack
5	Bend	10	Skip

There are ten behavior classes in Weizmann dataset, each of the classes has at least nine video footages captured by a static camera view. Table 4.1 shows ten output classes associated with event labels of the Weizmann dataset which contains 90 video clips with the resolution of 180 × 144. The videos show nine participants in total. Figure 4.1 shows an example of video frames in Weizmann datasets which are adopted for this research project. There are totally five images in this example; each image represents one behavior of a person.



(a) (b) (c)

(d) (e)

Figure 4.1 Video frame example. (a) Jacking, (b) Jumping, (c) Skipping (d) Running and (e) Walking.

The experiment is run on a laptop installed Microsoft Windows 10 Operating System using the Intel Core i7 CPU 2.30GHZ. The project of human behavior recognition is developed and implemented in Matlab R2015b, and simulated by utilizing WEKA 3.

4.2 Human Behavior Recognition and Classification

After feature extraction, the next step is to utilize machine learning for training and testing the extracted features in order to achieve human behavior recognition and classification. As we mentioned in Chapter 2, both supervised learning and unsupervised learning are able to implement for human behavior analytics such as recognition and classification. In line with this view, Artificial Neural Networks (ANNs) as the main method of human behavior recognition is implemented in this thesis. Figure 4.2 shows the basic flowchart of how the ANN works. Firstly, we need to prepare the training dataset, and then we train the neural network by tuning parameters. Once we acquire the pre-trained neural network, we will easily simulate the network by using testing dataset.

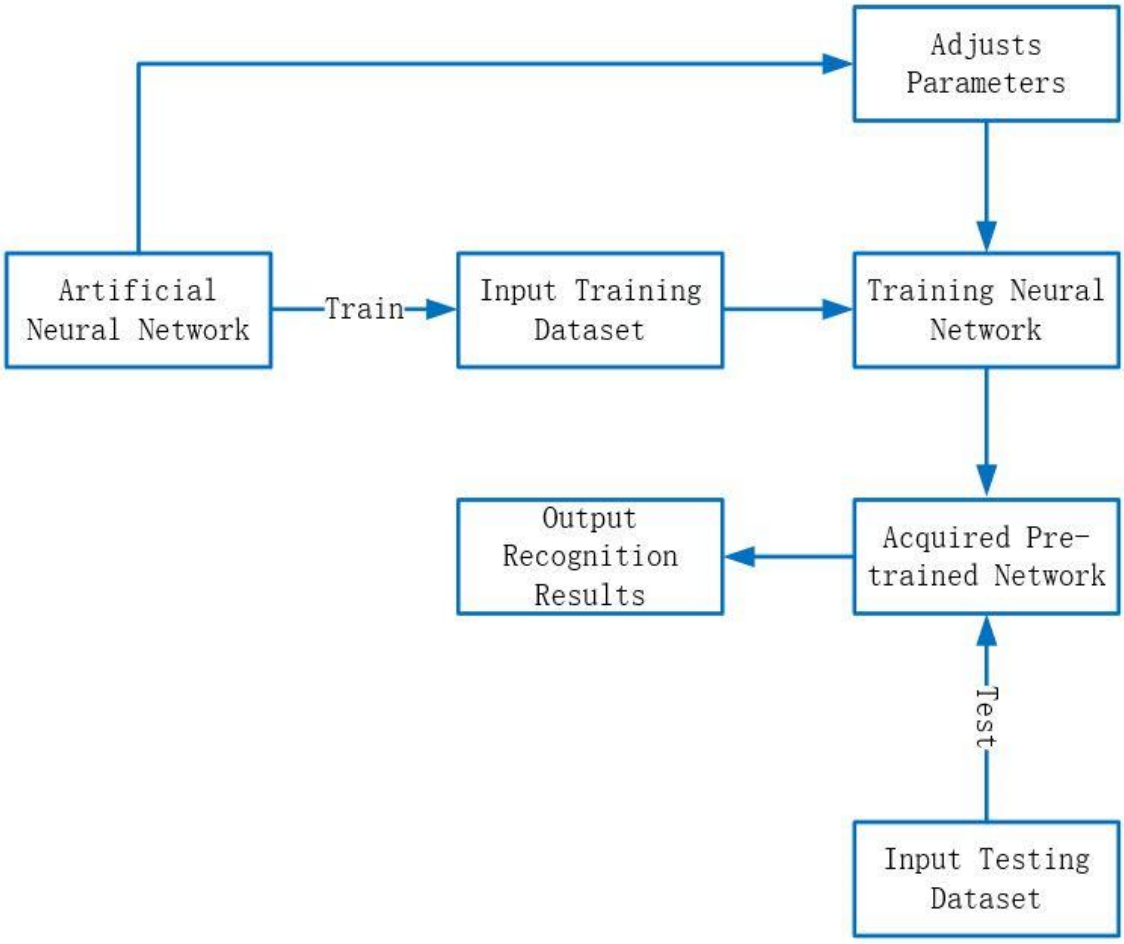
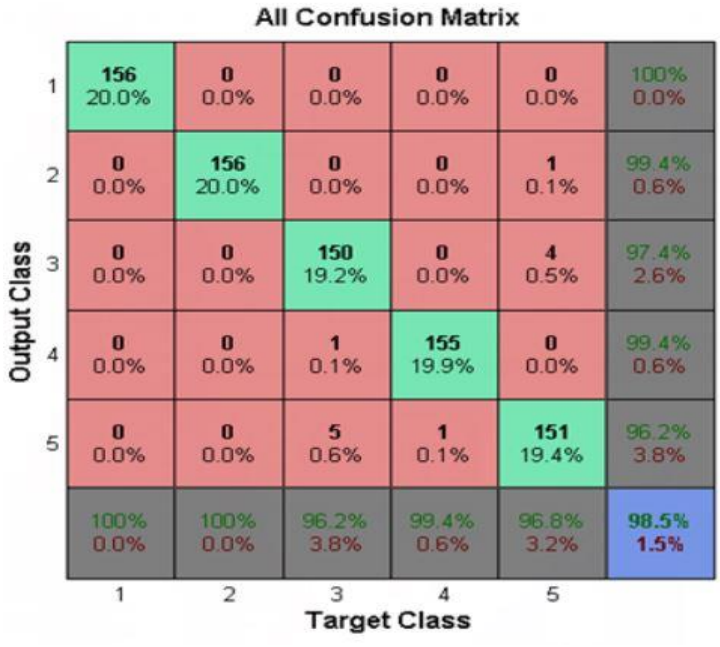


Figure 4.2 Flowchart of Artificial Neural Network

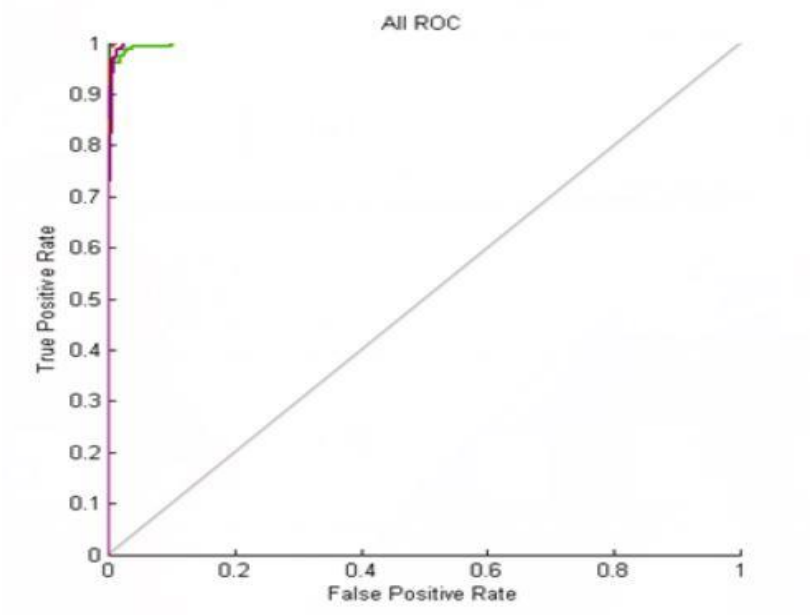
4.2.1 Human Behavior Recognition

ANNs as the main classifier of human behavior recognition were well studied. Moreover, the ANNs should be well trained before the human behavior recognition. The procedure of implementing the ANN is comprised of preparing the training and testing datasets, creating the neural network and simulating the pre-trained neural network.

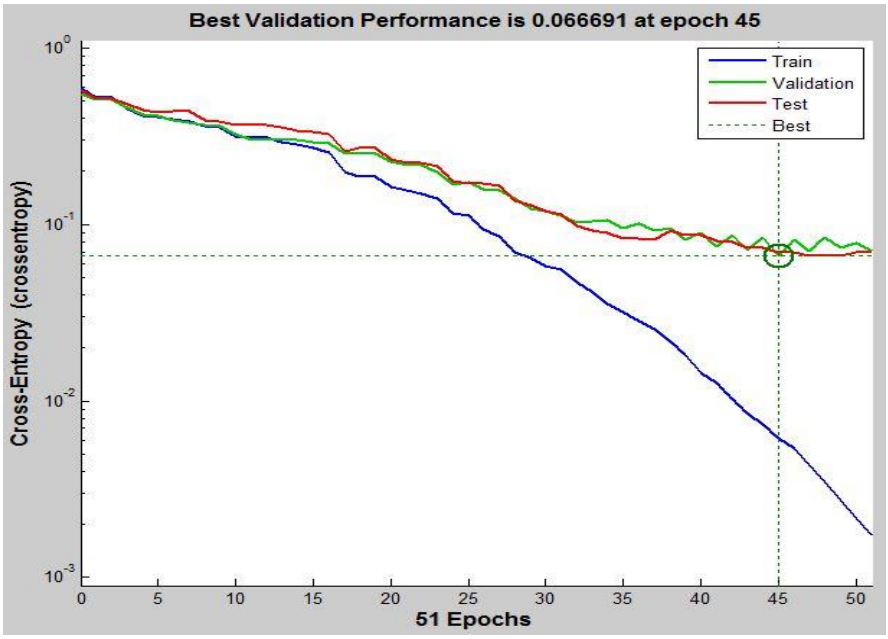
As we mentioned before, normal ANNs contain three layers (input layer, hidden layer, and output layer). In our experiments with the same datasets, a five-layer neural network and a ten-layer neural network both with hidden and output neurons are adopted for recognition by using HOG and LBP feature vectors. For each of the hidden layer, it has ten hidden neurons. For the output layer, it contains five neurons and each neuron represent different behavior.



(a) The confusion matrix of training dataset



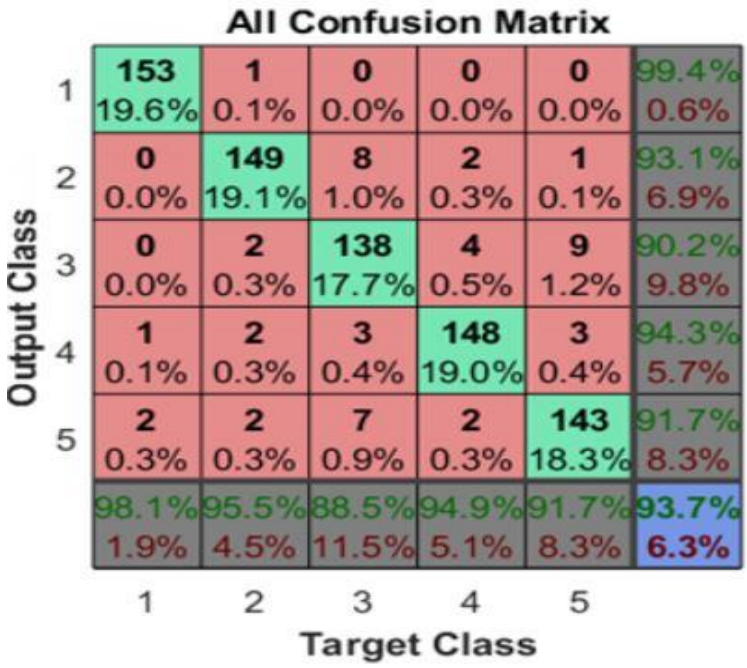
(b) The ROC result for training dataset



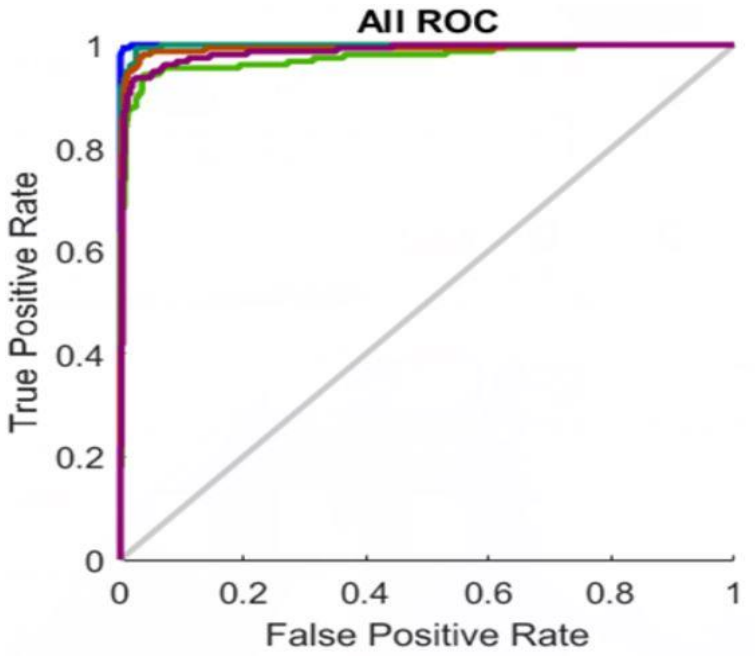
(c) The performance of training dataset

Figure 4.3 The result of trained neural network by using HOG training dataset

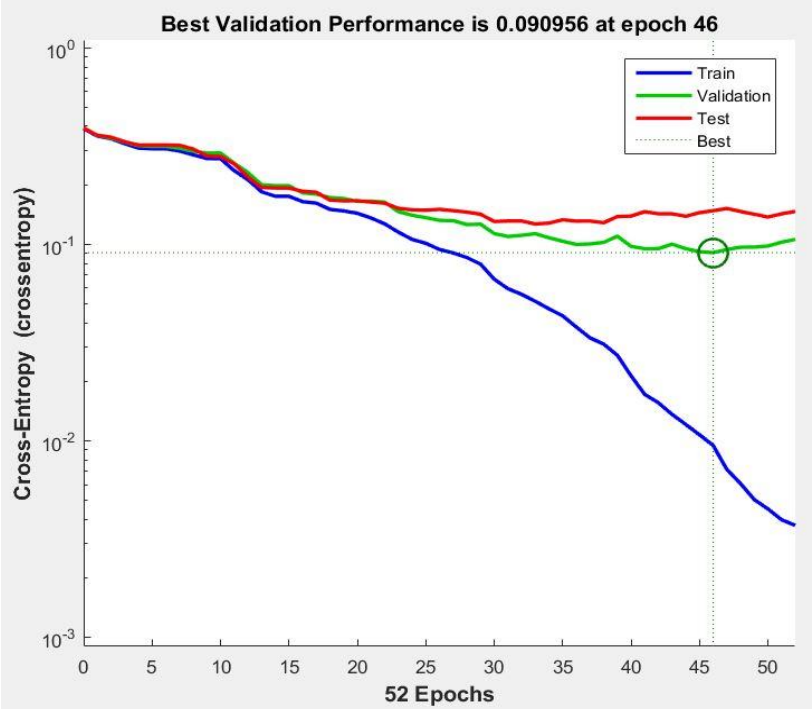
In HOG feature, the length of visual feature vector is 1440. Thus, it makes the number of neurons in input layer as 1440. For LBP feature extraction, the length of LBP feature vector is 708. In the target matrix, we defined the rows as the classes of training dataset, therefore, the training dataset contains five different behaviors; For each behavior, we choose 30 videos in total for training. Each video has 26 frames (samples) and each behavior has 156 frames (samples), thus we totally have 780 frames (samples) in the training dataset. For the purpose of cross-validation, the dataset was split into 70% of training ratio (546 samples) and the remained is equally divided into half, 15% is for validation and another 15% is for testing data (117 samples each). Figure 4.3 and Figure 4.4 show the result of trained neural network by using HOG training dataset and LBP training dataset. Figure 4.5 illustrated the result of trained neural network by utilizing the SILTP training dataset.



(a) The confusion matrix of training dataset

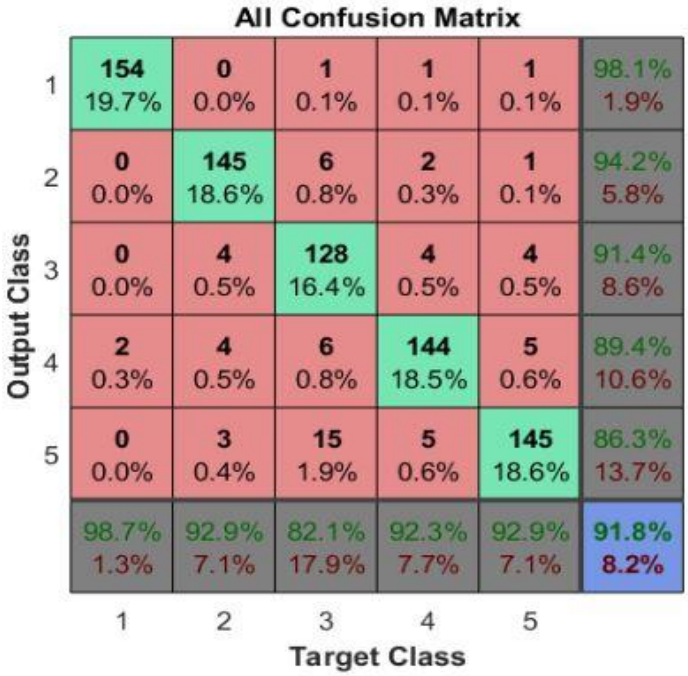


(b) The ROC result for training dataset

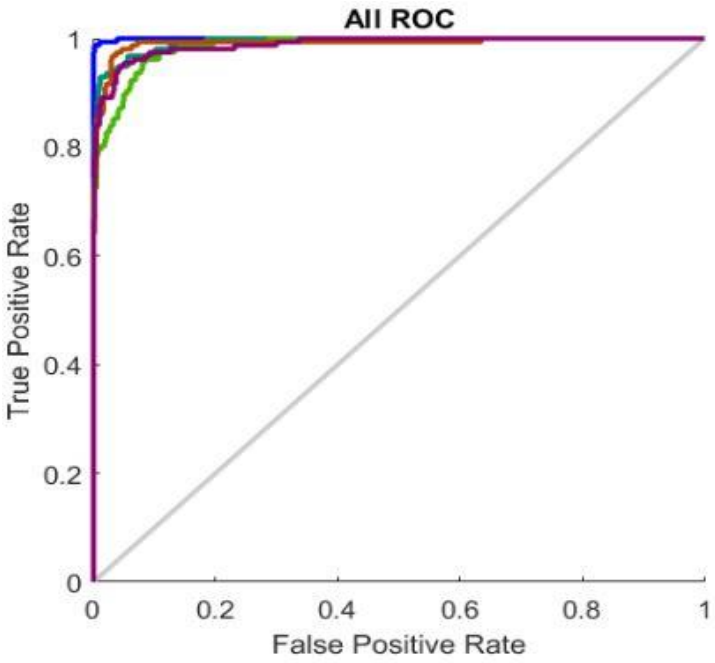


(c) The performance of training datasest

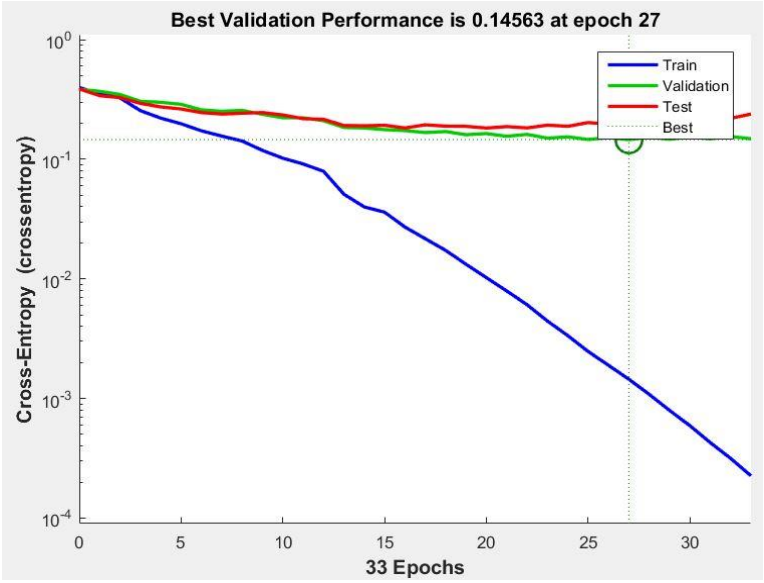
Figure 4.4 The result of trained neural network by using LBP training dataset



(a) The confusion matrix of training dataset



(b) The ROC result for training dataset



(c) The performance of training datasest

Figure 4.5 The result of trained neural network by utilizing SILTP training dataset

After the training work is completed, the second step is to test the neural network with the testing dataset. The testing dataset also includes five different behaviors and we select ten videos in the total testing dataset for each class. For the test target matrix, rows represent pre-defined human behaviors. Each video also has 26 frames (samples) and each behavior has 52 frames (samples) therefore totally we have 260 frames (samples) in the testing dataset. Figure 4.6 shows Matlab results where each column represents one

sample, the values between 0 to 1 show the probabilities of the classification. Each row represents one class which contains Jacking, Jumping, Running, Skipping and Walking. Figure 4.7 and Figure 4.8 show the simulated results of the pre-trained neural network by using HOG testing dataset and LBP testing dataset. Figure 4.9 illustrates the simulated result of pre-trained neural network by using SILTP testing dataset.

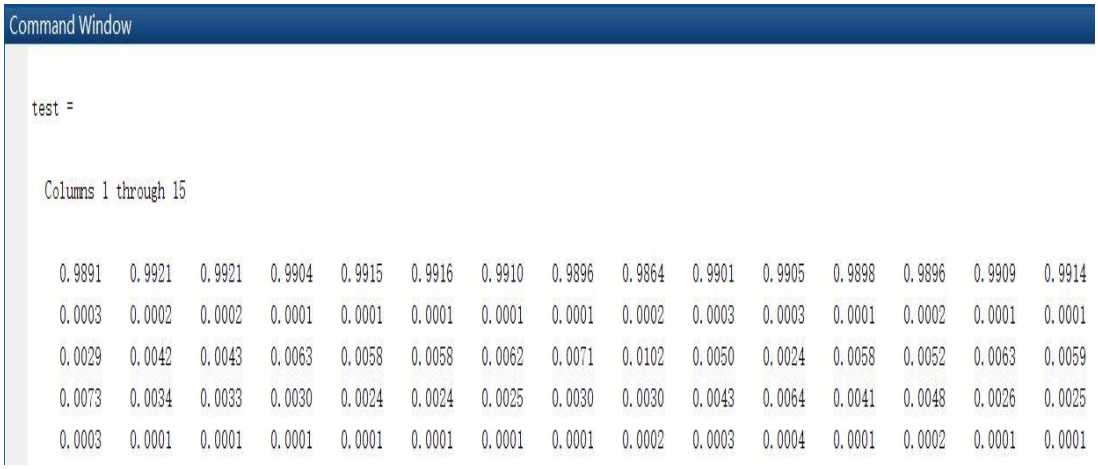
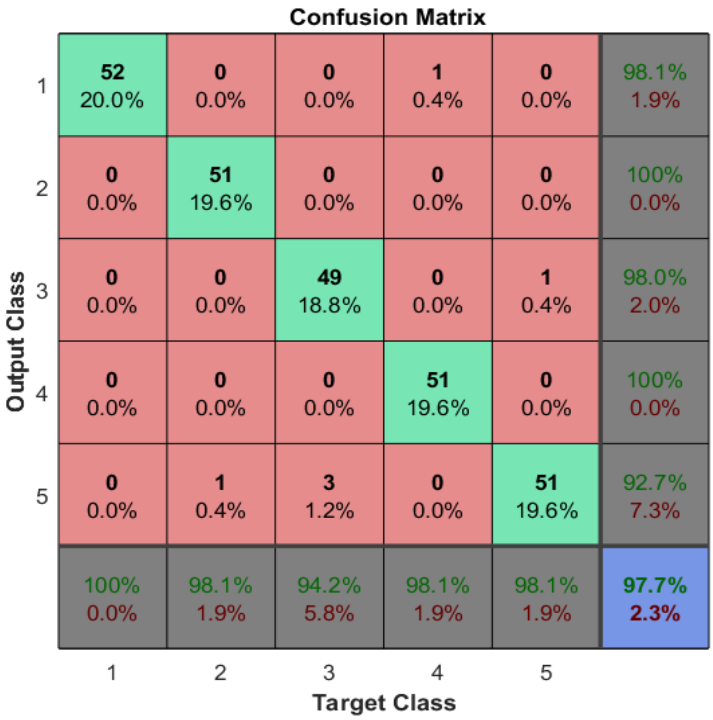
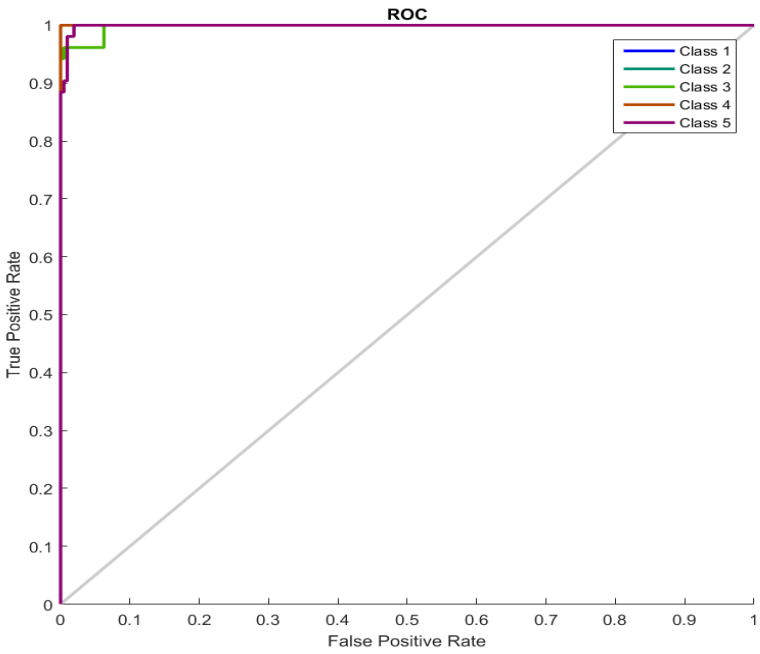


Figure 4.6 Matlab result of testing dataset

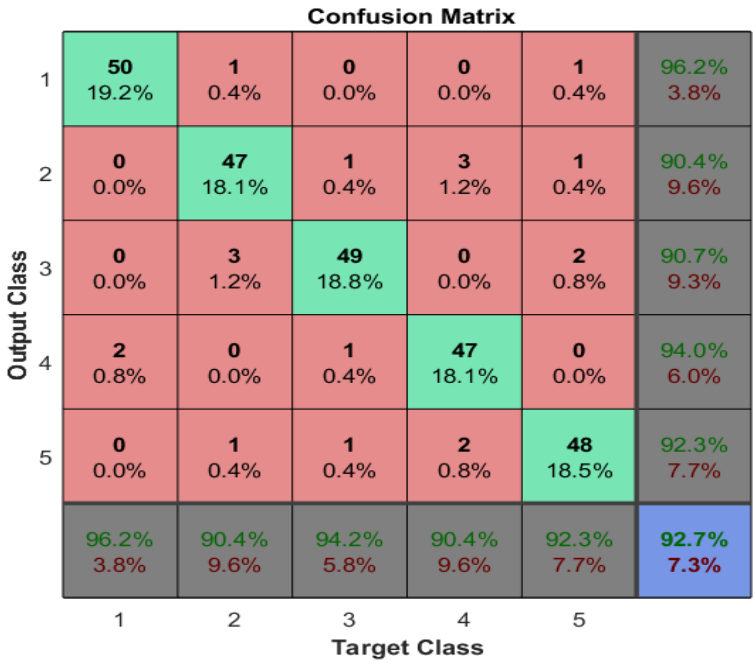


(a) The confusion matrix of testing dataset

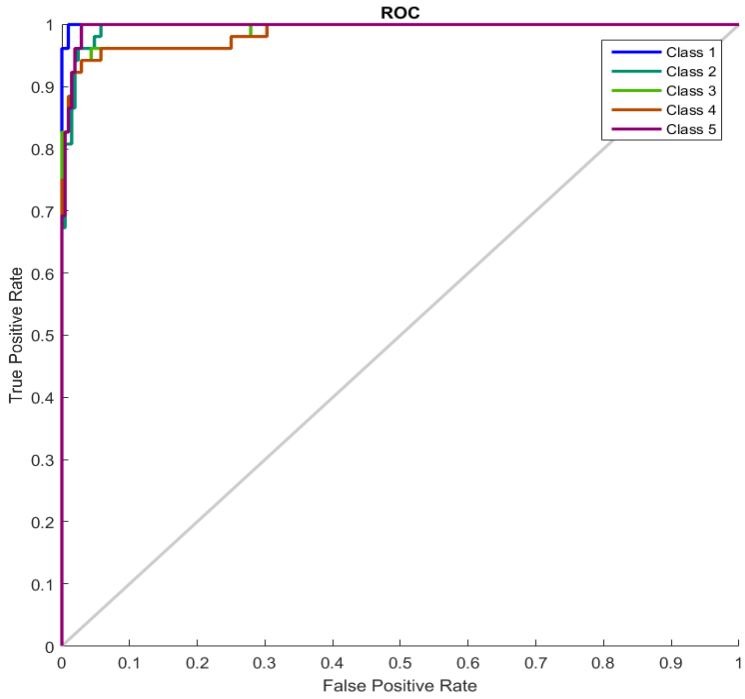


(b) The ROC result for testing dataset

Figure 4.7 The simulated result of pre-trained neural network by using HOG testing dataset

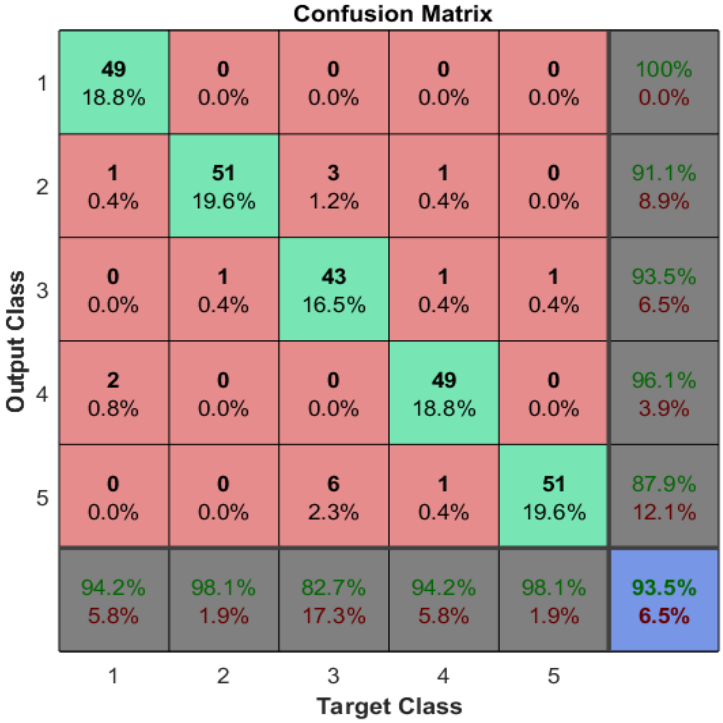


(a) The confusion matrix of testing dataset

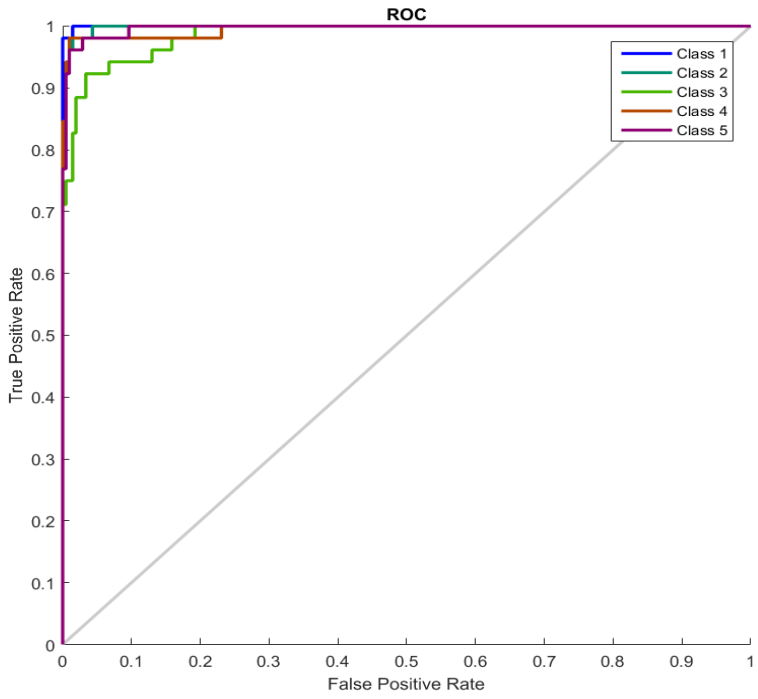


(b) The ROC result for testing dataset

Figure 4.8 The simulated result of pre-trained neural network by using LBP testing dataset



(a) The confusion matrix of testing dataset

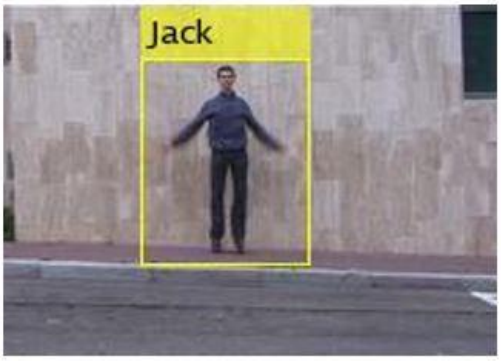


(b) The ROC result for testing dataset

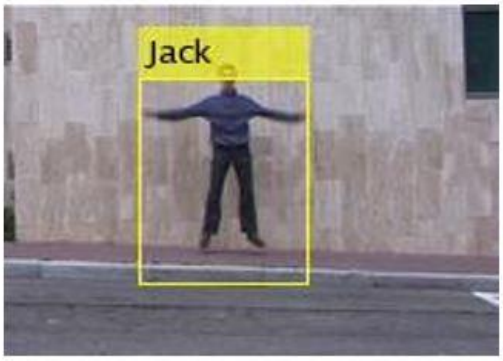
Figure 4.9 The simulated result of pre-trained neural network by using SILTP testing dataset

4.2.2 Human Behavior Classification

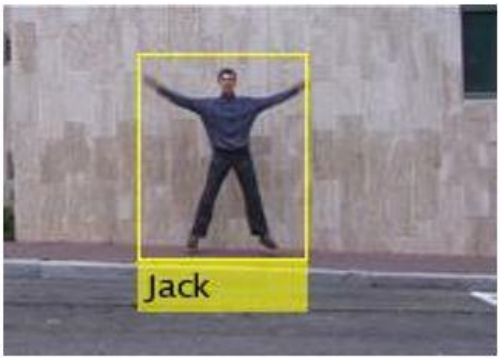
The final step of human behavior recognition is to conduct pattern classification using testing dataset. The classification is the most important step, it directly influences the results of the classification. In this thesis, we investigate five different human behaviors including Jacking, Skipping, Jumping, Walking and Running. Our vision is that when a person passes through the camera, our computer will automatically detect the Region of Interest (ROI) and recognize one of these five different behaviors or more. Moreover, a comparison of performance amongst these classifiers by utilizing WEKA will be presented in the following chapters.



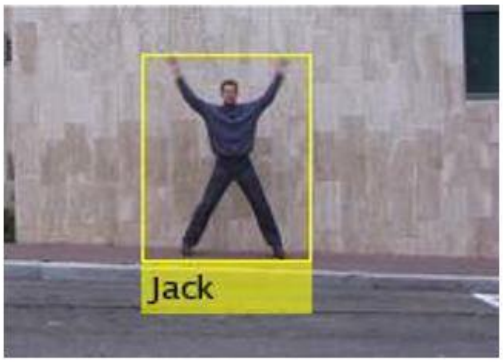
Frame 2



Frame 4



Frame 6



Frame 8

(a) Testing result in Jacking

Frame 3

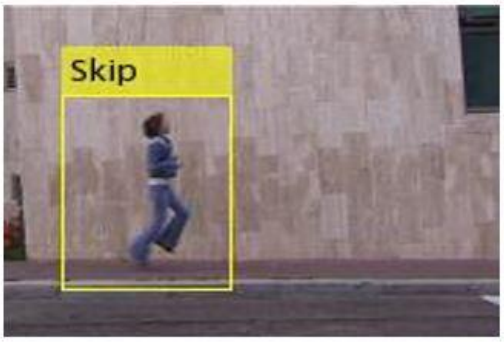
Frame 5

Frame 7

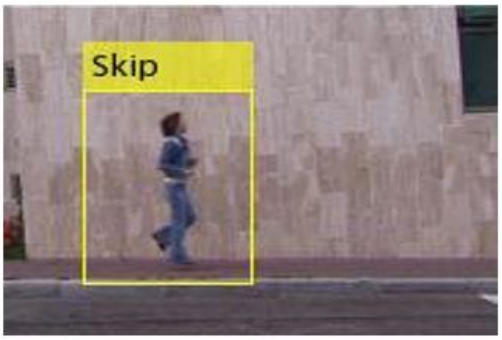
Frame 9

(b) Testing result in Jumping

Frame 2



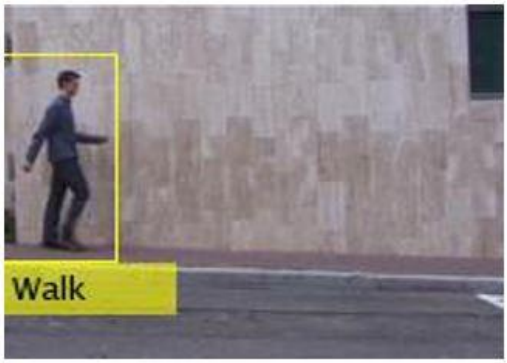
Frame 4



Frame 6

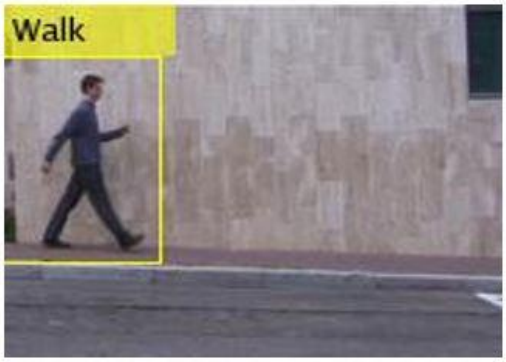
Frame 8

(c) Testing result in Skipping

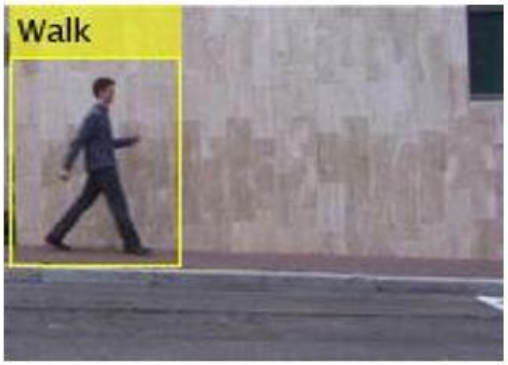


Frame 1

Frame 3

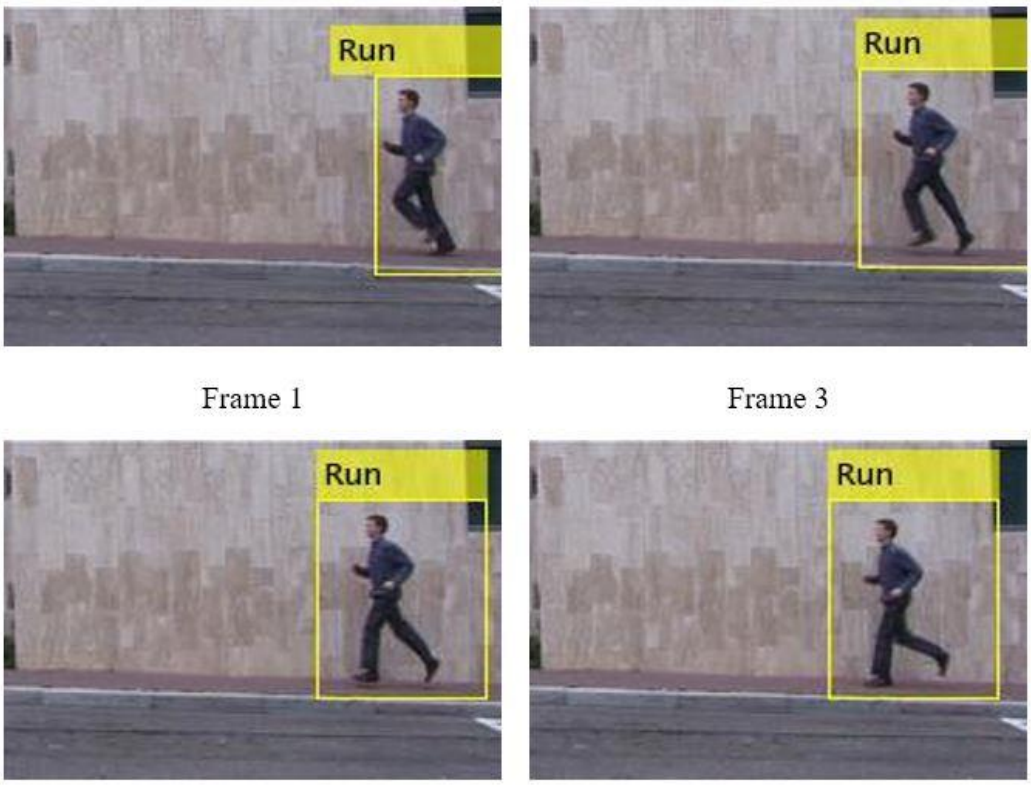


Frame 5



Frame 7

(d) Testing result in Walking



(e) Testing result in Running

Figure 4.10 Results of event recognition and classification: (a) Jacking, (b) Jumping, (c) Skipping, (d) Walking and (e) Running

Figure 4.10 shows the results of event recognition and classification for human behaviors: (a) Jacking, (b) Jumping, (c) Skipping, (d) Walking and (e) Running.

On the contrary, Figure 4.11 illustrates the examples of event recognition for various human behaviors which are classified incorrectly. Figure 4.11(a) is based on HOG feature extraction. In the Skipping videos, there are three frames which are incorrectly classified into Running, Walking and Jump. In the Running test video, two frames are incorrectly classified into Skipping. Figure 4.11(b) is based on LBP feature extraction. There are four different frames, which are placed incorrectly into the Skipping class and also two frames are categoried incorrectly into Jumping class.

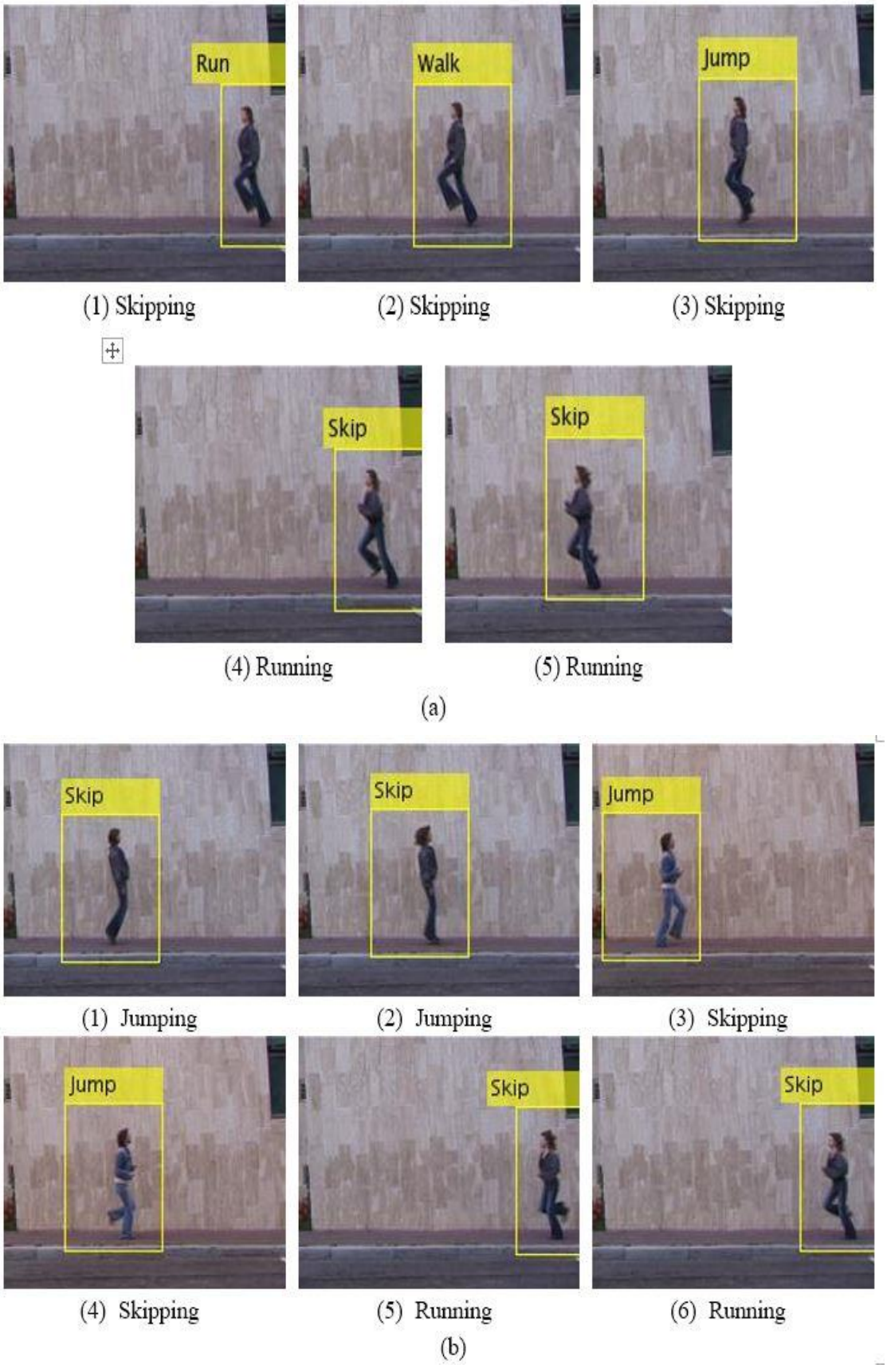


Figure 4.11 Incorrectly classified results of event recognition and classification based on (a) HOG feature extraction, (b) LBP feature extraction

4.3 Limitations of the Research

The proposed algorithms have been implemented successfully in this thesis for human behaviour analytics. However, there are still some limitations that should be improved in future. The limitations may include:

(1) Because the test videos only have single participant in each video which is only suitable for private space.

(2) Surveillance videos were acquired from a static camera within a static representation of viewed outdoor environment.

(3) Only five of human behaviours were taken into account for behaviour recognition in this project, in future we will take more natural human behaviours into consideration.

Chapter 5
Analysis and Discussions

In this chapter, the discussion and result analysis with respect to outcomes of the experiments are clearly demonstrated and presented. More specifically, comparisons with regarding to performance of various classifiers will be discussed in this chapter. The clear demonstration of outcomes from human behavior recognition will be addressed. Finally, the significance will be also stated through analyzing the outcomes.

5.1 Analysis

We have introduced the direct results of human behavior analysis by utilizing MATLAB as stated in the previous chapter. In this chapter, the result analysis will be detailed with our feature extraction techniques. Moreover, three classifiers will be elucidated with the result analysis in this chapter.

5.1.1 Feature Extraction

Histograms of Oriented Gradient (HOG) as one of the Feature Extraction Techniques (FET) is proposed in this thesis and depicted in the previous chapters. In this thesis, we investigated the influence of HOG parameters on human behavior recognition. The comparisons between three feature extraction techniques will be explained in this section.

In the previous chapters, we pointed out that HOG feature adopts various parameters and cell sizes that directly affect the HOG descriptor. If the cell size is too small, it will increase the computational time of feature extraction. On the contrary, increasing the cell size could not include too much shape information which may affect result of the recognition.

Table 5.1 The results of feature extraction

Features	Samples	Cell Size	Feature Extraction Time	Feature Length
HOG	780	2×2	90.178s	30636
HOG	780	8×8	15.001s	1440
LBP	780	8×8	4.703s	3186
LBP	780	16×16	4.534s	708

Table 5.1 shows the results from multiple feature extraction techniques. As it is shown this table, when the number of samples are same, the length of feature vectors will affect the computational time. Moreover, LBP shows more efficient than HOG during the feature extraction because of the less time consuming.

5.1.2 Artificial Neural Network

In previous chapters, Artificial Neural Network (ANN) as the main machine learning method was taken in account this thesis, which is utilized to recognize and classify the pre-defined human behaviors. Three Feature Extraction Techniques (FET) were adopted in this thesis, which include Histograms of Oriented Gradient (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP).

We investigated the three feature extraction techniques which are able to implement the event-based human behavior recognition. The HOG feature with the 8-by-8 cell size and 16-by-16 cell size of LBP feature is adopted in this thesis. In the previous section, we introduced different parameters of HOG and LBP. For the remained feature extraction techniques, the results are much lower than the adopted features. Table 5.2 shows the results of recognition precisions with different cell sizes. As shown in the table, the adopted cell size has a better precision rate in event based human behavior recognition which is up to 90.0%, in the case of using 8×8 cell size, it even reaches a 97.7% of precision.

Table 5.2 The precisions of different FET with different cell size

	Cell Size	Precision
HOG	2×2	89.6%
HOG	8×8	97.7%
LBP	8×8	87.3%
LBP	16×16	92.7%

In the training stage, HOG feature achieved a 98.5% of precision in total; LBP feature only has a 93.7% of precision. Table 5.3 shows the results of training dataset with the precision of human behavior recognition, it has a 100% of precision in classification of Jacking and Jumping, the lowest 96.2% of precision in the Running behavior is classified by utilizing HOG feature. However, for LBP feature, 98.1% of precision in Jacking shows the best classification output, and the lowest precision is 88.5% in Skipping. If using SILTP feature, classification of Jacking class achieves a 98.7% of precision. However, Skipping class is up to only 82.1% of precision.

Table 5.3 The precisions of FETs in the classification of training dataset

	Jacking	Jumping	Skipping	Walking	Running	Total
HOG (8-by-8)	100%	100%	96.2%	99.4%	96.8%	98.5%
LBP (16-by-16)	98.1%	95.5%	88.5%	94.9%	91.7%	93.7%
SILTP	98.7%	92.9%	82.1%	92.3%	92.9%	91.8%

In the testing stage, HOG feature reaches a percentage of 97.7% of precision and a rate of 92.7% for the classification by utilizing LBP feature. Table 5.4 presents the results of human behavior classification. There are total 260 frames (samples) and each behavior has 52 frames (samples). As shown in Table 5.4, both HOG and LBP features have a good

result of classification in Jacking, all the frames are classified correctly. However, HOG features have a low rate of classification in Skipping where only 49 frames are classified correctly. Both Jumping and Walking using the LBP feature also have a low precision of classification which there are only 47 frames classified correctly.

Table 5.4 Results of behavior classification and recognition

Features	Frame Numbers	Correct Number				
		Jacking	Jumping	Skipping	Walking	Running
HOG (8-by-8)	260	52	51	49	51	51
LBP (16-by-16)	260	50	47	49	47	48
SILTP	260	49	51	43	49	51

Table 5.5 shows results of the testing dataset with the precision of each human behavior recognition. In the dataset, HOG feature achieves a 100% of precision in Jacking behavior recognition and the lowest precision rate of the human behavior Skipping only has 94.2%. The LBP feature shows positive result in testing; the best precision is recognition rate of human behavior Jacking which achieves the precision at the rate of 96.2% and the lowest is 90.4%. If we use SILTP features, both Jumping and Running behaviors show 98.1% of precision rate and the lowest precision rate is human Skipping behavior.

Table 5.5 The precisions of FETs in the classification of testing dataset

	Jacking	Jumping	Skipping	Walking	Running	Total
HOG (8-by-8)	100%	98.1%	94.2%	98.1%	98.1%	97.7%
LBP (16-by-16)	96.2%	90.4%	94.2%	90.4%	92.3%	92.7%
SILTP	94.2%	98.1%	82.7%	94.2%	98.1%	93.5%

5.1.3 Other Classifiers

In this thesis, we implemented an event-based human behavior recognition by utilizing ANN. In this section, three classifiers are presented to compare the differences. Our classifiers contain Decision Tree (DT), *k*-Nearest Neighbor (*k*-NN) and Multi-Layer Perceptron (MLP), that are all based on supervised learning. The classifications are conducted by utilizing software WEKA.

Decision Trees (DT) takes an object described by a set of properties as the input, the output is a 'yes' or 'no' decision. The Decision Trees (DT) as a rapid and effective method is defined in the form of a tree structure, each node may represent as a leaf or a decision node in different cases. It will specify the tests, a branch and sub-tree for each of the possible outcomes of the test. Figure 5.1 shows the general model of Decision Tree. More specifically, the decision reaches to one node, the target attribute of data should also be carried out. For instance, Figure 5.1 shows the process starts at root *X*, and then follows the path (Y or Z) which solve the problem until the sub-tree is reached.

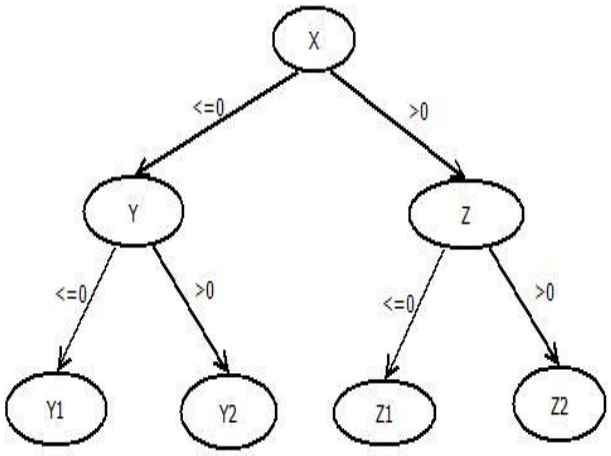


Figure 5.1 The general model of decision trees

DT is the popular and powerful tool for classification and prediction. To compare with other models, the advantage of DT is to represent the fact of the rules which are easy to be expressed.

Due to advantages of the DT classification with a readable structure, decision trees

are able to provide a better explanation of the dataset associated with each sub-tree. Figure 5.2 shows visualized tree structure of the testing dataset. The DT family consists of classical algorithms which include ID3, C4.5 and CART. In this thesis, the C4.5 is adopted to implement classification (in WEKA program, the C4.5 algorithm which also called J48). The C4.5 algorithm addresses the shortcomings of ID3 avoiding the overfitting, processing the training dataset with the missing values and improving the efficiency of classifications.

Figure 5.2 The visualization of decision trees

Table 5.6 shows the results of classifications by utilizing Decision Tree C4.5 algorithm. Moreover, there are 258 frames correctly classified by utilizing HOG feature, for both Jacking and Running behaviors, the classifications achieve the best precision at the rate of 100% and the lowest precision is Walking, which only has 96.15%. For LBP feature, only 252 frames are correctly classified, the class Jacking obtained a precision

rate at 98.08% and the lowest precision is the class of Skipping behavior which only has 86.67%. There are 257 frames which are correctly classified bu using SILTP feature, Jacking class gets a 100% precision and Skipping as the lowest recognition behavior only obtains a precision rate at 96.30%.

Table 5.6 The results of decision tree C4.5 algorithm

		Jacking	**Jumping**	**Skipping**	**Walking**	**Running**
HOG	**Frame Number**	52	53	53	50	52
	Precision	100%	98.11%	98.11%	96.15%	100%
LBP	**Frame Number**	51	48	60	50	51
	Precision	98.08%	92.31%	86.67%	96.15%	98.08%
SILTP	**Frame Number**	52	50	54	53	51
	Precision	100 %	96.15%	96.30%	98.11%	98.08%

The Multilayer Perceptron (MLP) converts I/O problems into nonlinear optimization ones which consist of different layers. Generally, there are three layers in MLP: the input layer which receives the external data (one for each attribute), the hidden layer and output layer which generates the results of classification (one for each class), respectively. Although the MLP consists of multiple layers, but at the same level, each node is not related to others, and the layer nodes maybe presented as a decreasing trend. For each neuron, the input layer and hidden layers (20 hidden layers) were connected to all others in the next layer and controlled by weights.

Table 5.7 shows the results of the testing dataset by utilizing Multilayer Perceptron (MLP). For HOG feature, there are 256 frames which are correctly classified. Human Jumping behavior shows 100 % of precision rate and the Running behavior only has 92.31% of precision rate. For LBP feature, both Jumping and Running behaviors get a 100% percentage of precision, Jacking behavior is only at the lowest 96.30% rate. For SILTP

feature, there are 252 frames which are correctly classified, both Jacking and Running behaviors are at a rate 98.08% of precision and Walking class acquires a 100% precision rate, but for the Jumping behavior, it only meets the lowest 86.67% of precision rate.

Table 5.7 The results of MLP by utilizing testing dataset

		Jacking	**Jumping**	**Skipping**	**Walking**	**Running**
HOG	**Frame Number**	53	52	55	52	48
	Precision	98.11%	100%	94.55%	100%	92.31%
LBP	**Frame Number**	54	52	51	51	52
	Precision	96.30%	100%	98.08%	98.08%	100%
SILTP	**Frame Number**	51	60	46	52	51
	Precision	98.08%	86.67%	88.64%	100 %	98.08%

In WEKA software, *k*-NN algorithm is also called the IBK, which is a lazy learner. The *k*-NN classifier only is working for one part of the dataset and the computation will be deferred until the classification is completed, and one object is classified by using its major neighbors (to choose the major class amongst several of neighbors). For instance, if a sample in the feature space is most similar (adjacent) to the sample which closes to a category, then the sample also belongs to this category. Therefore, the principle of the *k*-NN is that all the testing values will be memorized and the classification will only be conducted when the object attributes match the training values. For *k*-NN approach, the model performance will get affected by the *k* value. Once the value is too small, the resulting model will be over fitted by learning from noisy samples and is not likely to generalize the future data well. On the contrary, when the *k* value is too large, the neighbour may include lots of data points from other classes, and the model will generate suboptimal results.

The value of *k* needs to be chosen particularly. In the experiments the *k* value is

determined by the cross-validation and should be less than the square root of the training set. The results of *k*-NN classification were presented in Table 5.8 with the *k* value of 1. For HOG feature, there are 258 frames which are correctly classified. The class Jumping obtained 100% of precision, both Skipping and Walking behaviors have the lowest precision rate of 98.08%. There are 258 frames to be correctly classified in LBP feature. For the classes Jacking and Jumping behaviors, they earn 100% of precision and Skipping behavior only has achieved the lowest 96.15% of precision rate. For SILTP feature, there are 258 frames which are correctly classified, both Skipping and Running classes get a 100% precision rate, but the class of Jumping behavior only has the lowest 96.30% of precision rate.

Table 5.8 The results of KNN by utilizing testing dataset

		Jacking	**Jumping**	**Skipping**	**Walking**	**Running**
HOG	**Frame Number**	53	52	51	51	53
	Precision	98.11%	100%	98.08%	98.08%	98.11%
LBP	**Frame Number**	52	52	50	53	53
	Precision	100%	100%	96.15%	98.11%	98.11%
SILTP	**Frame Number**	51	54	52	51	52
	Precision	98.08%	96.30%	100%	98.08%	100%

5.2 Discussions

Experimental results were detailed and demonstrated in previously chapters. Firstly, the focus of these experiments is on the event detection in order to detect the Region of Interest (ROI), which was verified in Section 3.3.1. After found the ROI, three different Feature Extraction Techniques (FET): Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP) are detailed and

compared in Section 3.3.3 and Subsection 5.1.1 which show that the different cell size will cause the variety of the feature vector length and also effect comoutational time of feature extraction.

In this thesis, the experiments are detailed in Section 4.2 and the discussions will be followed in Subsection 5.1.2. In Section 4.2, we demonstrate the direct result and most of classes of human Skipping behavior were incorrectly classified into others. In the Subsection 5.1.2, the results of human behavior recognition by utilizing different FET are descripted. An 8×8 cell size Histogram of Oriented Gradient (HOG) feature has obtained the best precision rate of 97.7% overall, LBP feature shows a 92.7% precision rate in total, the SILTP feature shows a 93.5% precision rate for the testing dataset. The FET is able to get above 90% of precision rate. Even in the worst case, the 8×8 cell size of LBP feature is able to earn 87.3% of precision rate. Finally, in the Subsection 5.1.3, three different classifiers are discussed with the results by using WEKA and most of the results are able to get 90% of precision rate. Thus, the result shows that three FETs are able to be implemented into human behavior recognition. Moreover, from the previous tables we see that the best precision for recognizing human behaviours is Jacking and Jumping. However, the classes Running and Skipping have obtained very lower precision. The reason why we earned these lower results than the expected is that the behaviours amongst these frames were similar to those associated with human Skipping behaviour.

Chapter 6
Conclusion and Future Work

In this thesis, in-depth articulation of the techniques was discussed which can be utilized to analyze the pre-defined human behavior. The corresponding approaches for each step have been implemented as the results of this thesis. In this chapter, we will present this thesis at a scholarly level, also highly organize and integrate the conclusion into the context, meanwhile the future work will be pointed out by the end of this thesis.

6.1 Conclusion

The objective of this thesis is to develop those event recognition methods from surveillance videos, particularly in the domain of human behaviours analytics. We proposed the empirical approaches for human behaviour recognition utilizing three distinct Feature Extraction Techniques combined with multiple classifiers. In this thesis, we demonstrated these visual features are able to be employed for human behaviour classification. We have verified the features could achieve satisfactory results by adjusting the cell size of FET. After completed the procedures of human behaviour analytics, the main contributions are summarized as below.

By adjusting the cell size, the proposed empirical approaches are able to reduce the computational time of feature extraction. LBP feature which has the fastest extraction time is carried out within 4.5 seconds. Moreover, due to changes of various cell sizes this makes the computational time greatly reduced.

In feature extraction, HOG generates better results than those of LBP and SILTP which have greater potential in surveillance. The best precision of overall recognition for human behaviour achieves 97.7%, and both behaviours of Jacking and Jumping are recognized very well. Both features of HOG and LPB are relatively simple to implement in low consuming so as to eliminate redundancy.

6.2 Future Work

Our future work includes,

(1) We will work for human behaver analysis from multiple participants. In addition, more complex human behaviours such as fighting, robbery, etc. should be added in this project.

(2) Since the testing surveillance videos were acquired from static cameras. In future, the high-speed camera could be taken into consideration in order to find more details of human behaviours, meanwhile more complex background should be taken into account. Furthermore, lighting conditions also could be improved which makes the method more operatable.

(3) In future, other Feature Extraction Techniques (FET) should be taken into account to investigate human behaviour recognition. This is because the patterns of different behaviours need accurate and precise description.

(4) The future work could be extended to develop end-to-end classifiers using deep learning such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN).

References

Aggarwal, J. K., & Cai, Q. (1997). Human motion analysis: A review. *Nonrigid and Articulated Motion Workshop*, pp. 90-102.

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR), 43*(3), pp. 1-47.

Aggarwal, J. K., Cai, Q., Liao, W., & Sabata, B. (1997). Nonrigid motion analysis: Articulated and elastic motion. *Computer Vision and Image Understanding, 70*(2), pp. 142-156.

Allen, F. R., Ambikairajah, E., Lovell, N. H., & Celler, B. G. (2006). Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models. *Physiological Measurement, 27*(10), pp. 935-951.

Altun, K., & Barshan, B. (2010). Human activity recognition using Inertial/Magnetic Sensor Units. *International Workshop on Human Behavior Understanding*, pp. 38-51.

Bajpai, S., Jain, K., & Jain, N. (2011). Artificial neural networks. *International Journal of Soft Computing and Engineering*, pp. 27-31.

Basheer, I., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods Neural Computting in Micrbiology, 43*(1), pp. 3–31.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(3), pp. 257-267.

Burse, K., Manoria, M., & Kirar, V. P. (2011). Improved backpropagation algorithm to avoid local minima in multiplicative Neuron Model. *Information Technology and Mobile Communication, 147*, pp. 67-73.

Chaudhry, R., Ravichandran, A., Hager, G., & Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1932 - 1939.

Chen, X., & Zhang, C. (2006). An interactive semantic video mining and retrieval platform - Application in Transportation Surveillance Video for Incident Detection. *International Conference on Data Mining (ICDM'06)*, pp. 129-138.

Chen, Y. T., & Chen, C. S. (2008). Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Transactions on Image Processing, 17*(8), pp. 1452-1464.

Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 142 - 149.

Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(5), pp. 564-577.

Cutler, R., & Davis, L. S. (2000). Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), pp. 781-796.

Dai, K., Zhang, J., & Li, G. (2006). Video mining: concepts, approaches and applications. *12 th InternationalMulti-Media Modelling Conference Proceedings*, pp. 477-480.

Dai, P., Di, H., Dong, L., Tao, L., & Xu, G. (2008). Group interaction analysis in dynamic context. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38*(1), pp. 275-282.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, pp. 886-893.

Davis, J. W. (2001). Hierarchical motion history images for recognizing human motion. *IEEE Workshop on Detection and Recognition of Events in Video*, (pp. 39-46).

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification.* New York: Wiley.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. *IEEE International Conference on Computer Vision*, *2*, pp. 726-733.

Elgammal, A., Harwood, D., & Davis, L. (2000). Non-parametric model for background subtraction. *European Conference on Computer Vision*, (pp. 751-767).

Elhabian, S. Y., El-Sayed, M., K., & Ahmed, S. H. (2008). Moving object detection in spatial domain using background removal techniques-state-of-art. *Recent patents on computer science, 1*(1), pp. 32-54.

Freedman, D., & Zhang, T. (2004). Active contours for tracking distributions. *IEEE Transactions on Image Processing, 13*(4), pp. 518 - 526.

Geetha, P., & Narayanan, V. (2010). A survey of content-based video retrieval. *Journal of Computer Science, 4*(6), 734 - 734.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(12), pp. 2247-2253.

Gu, S., & Guo, Y. (2012). Learning SVM classifiers with indefinite kernels. *AAAI Conference on Artificial Intelligence*, pp. 942 - 948.

Guo, Y., Xu, G., & Tsuji, S. (1994). Tracking human body motion based on a stick figure model. *Journal of Visual Communication and Image Representation, 5*(1), pp. 1-9.

Gupta, A., Srinivasan, P., Shi, J., & Davis, L. S. (2009). Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2012

- 2019.

Haritaoglu, I., Harwood, D., & Davis, L. S. (1998). W4S: A real-time system for detecting and tracking people in 2 1/2D. *European Conference on Computer Vision*, *1*, pp. 877-892.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. *Alvey Vision Conference*, *15*, pp. 147-151.

Hongeng, S., Nevatia, R., & Bremond, F. (2004). Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding, 96*(2), pp. 129-162.

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*(1-3), pp. 185-203.

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks, 13*(2), pp. 415 - 425.

Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34*(3), pp. 334-352.

Intille, S. S., & Bobick, A. F. (2001). Recognizing planned, multiperson action. *Computer Vision and Image Understanding, 81*(3), pp. 414-445.

Ivanov, Y. A., & Bobick, A. F. (2000). Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), pp. 852-872.

Joshi, K. A., & Thakore, D. G. (2012). A survey on moving object detection and tracking in video surveillance system. *International Journal of Soft Computing and Engineering (IJSCE), 2*(3), pp. 44-48.

Ju, S. X., Black, M. J., & Yacoob, Y. (1996). Cardboard people: A parametrized model of

articulated image motion. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 38 - 44.

Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., & Driessen, P. F. (2005). Gesture-based affective computing on motion capture data. *International Conference on Affective Computing and Intelligent Interaction*, (pp. 1-7).

Ke, Y., Sukthankar, R., & Hebert, M. (2005). Efficient visual event detection using volumetric features. *IEEE International Conference on Computer Vision (ICCV'05)*, *1*, pp. 166-173.

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. 1-8.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, 15*(4), pp. 580 - 585.

Klette, R. (2014). *Concise computer vision.* Springer London. doi:10.1007/978-1-4471-6320-6

Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., & Shafer, S. (2000). Multi-camera multi-person tracking for EasyLiving. *IEEE International Workshop on Visual Surveillance*, pp. 3-10.

Kulchandani, J. S., & Dangarwala, K. J. (2015). Moving object detection: review of recent research trends. *International Conference on Pervasive Computing (ICPC)*, pp. 1-5.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision,* 64(2-3), pp. 107-123.

Laptev, I., & Lindeberg, T. (2003). Space-time interest points. *IEEE International Conference on Computer Vision (ICCV)*, pp. 432-439.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

Levi, K., & Weiss, Y. (2004). Learning object detection from a small number of examples: The importance of good features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, pp. II-53 - II-60.

Li, L. J., & Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. *International Conference on Computer Vision*, pp. 1-8.

Li, L., Huang, W., Gu, I. Y., & Tian, Q. (2003). Foreground object detection from videos containing complex background. *Proceedings of the eleventh ACM International Conference on Multimedia*, pp. 2-10.

Liao, S., Zhao, G., Kellokumpu, V., Pietikainen, M., & Li, S. (2010). Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *25*, pp. 1301-1306.

Lin, Z., Jiang, Z., & Davis, L. S. (2009). Recognizing actions by shape-motion prototype trees. *IEEE International Conference on Computer Vision*, pp. 444-451.

Liu, J., & Shah, M. (2008). Learning human actions via information maximization. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

Lublinerman, R., Ozay, N., Zarpalas, D., & Camps, O. (2005). Activity recognition from silhouettes using linear systems and model (In)validation techniques. *18th International Conference on Pattern Recognition (ICPR'06)*, *1*, pp. 347-350.

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence, 81*(1), pp. 674-679.

Maryam, K., & Reza, K. M. (2012). An analytical framework for event mining in video data. *Artificial Intelligence Review, 41*(3), pp. 401-413.

Matthews, I., Ishikawa, T., & Baker, S. (2004). The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(6), pp. 810-815.

McKenna, S. J., Jabri, S., Duric, Z., Rosenfeld, A., & Wechsler, H. (2000). Tracking groups of people. *Computer Vision and Image Understanding, 80*(1), pp. 42-56.

Miller, J. F., & Khan, G. M. (2011). Where is the brain inside the brain? On why artificial neural networks should be developmental. *Memetic Comp.*, pp. 217-228.

Munder, S., & Gavrila, D. M. (2006). An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(11), pp. 1863 - 1868.

Naphade, M. R., & Huang, T. S. (2002). Discovering recurrent events in video using unsupervised methods. *International Conference on Image Processing*, *2*, pp. II-13.

Natarajan, P., & Nevatia, R. (2007). Coupled hidden semi Markov models for activity recognition. *IEEE Workshop on Motion and Video Computing,* p. 10.

Nguyen, N. T., Phung, D. Q., Venkatesh, S., & Bui, H. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05), 2*, pp. 955-960.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision, 79*(3), pp. 299-318.

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, 22(8), pp. 831-843.

Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. *sixth International Conference on Computer Vision*, (pp. 555-562).

Park, S., & Aggarwal, J. K. (2004). A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Systems, 10*(2), pp. 164-179.

Petrushin, V. A. (2005, 8). Mining rare and frequent events in multi-camera surveillance video using self-organizing maps. *ACM SIGKDD conference on Knowledge Discovery in Data Mining*, pp. 794-800.

Piccardi, M. (2004). Background subtraction techniques: a review. *IEEE International Conference on Systems, Man and Cybernetics*, *4*, pp. 3099-3104.

Polana, R., & Nelson, R. (1994). Detecting activities. *Journal of Visual Communication and Image Representation, 5*(2), pp. 172-180.

Popoola, O. P., & Wang, K. (2012, 11). Video-based abnormal human behavior recognition—A Review. *IEEE Transactions on Systems, Man & Cybernetics: Part C - Applications & Reviews, 42*(6), pp. 865-878.

Poppe, R. (2007). Vision-based human motion analysis: an overview. *Computer Vision and Image Understanding, 108*(1-2), pp. 4–18.

Qian, H., Mao, Y., Xiang, W., & Wang, Z. (2010). Recognition of human activities using SVM multi-class classifier. *Pattern Recognition Letters, 31*(2), pp. 100–111.

Rakibe, R. S., & Patil, B. D. (2013). Background subtraction algorithm based human motion detection. *International Journal of Scientific and Research Publications (IJSRP), 3*(5), 1-4.

Rodriguez, M., Ahmed, J., & Shah, M. (2008). Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2008.*, pp. 1-8.

Rohr, K. (1994). Towards model—based recognition of human movements in image sequences. *CVGIP: Image understanding,* pp. 94-115.

Rui, Y., & Anandan, P. (2000). Segmenting visual actions based on spatio-temporal motion patterns. *IEEE Conference on Computer Vision and Pattern Recognition*, *1*, pp. 111-118.

Ryoo, M. S., & Aggarwal, J. K. (2006). Recognition of composite human activities through context-free grammar based representation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, pp. 1709-1718.

Ryoo, M. S., & Aggarwal, J. K. (2009a). Semantic representation and recognition of continued and recursive human activities. *International Journal of Computer Vision, 82*(1), pp. 1-24.

Ryoo, M. S., & Aggarwal, J. K. (2009b). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *IEEE International Conference on Computer Vision*, pp. 1593-1600.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *International Conference on World Wide Web*, pp. 851-860.

Savarese, S., DelPozo, A., Niebles, J. C., & Fei-Fei, L. (2008). Spatial-temporal correlatons for unsupervised action classification. *IEEE Workshop on Motion and video Computing*, pp. 1-8.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. *International Conference on Pattern Recognition*, *3*, pp. 32-36.

Shechtman, E., & Irani, M. (2005). Space-time behavior based correlation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, pp. 405-412.

Siebe, N., & Maybank, S. (2002). Fusion of multiple tracking algorithms for robust prople tracking. *European Conference on Computer Vision*, *4*, pp. 373-387.

Sondak, N. E., & Sondak, V. K. (1989). Neural networks and artificial intelligence. *ACM SIGCSE Bulletin, 21*(1), pp. 241-245.

Stauffer, C., & Grimson, W. E. (1999). Adaptive background mixture models for real-time tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, *2.*

Stringa, E. (2000). Morphological change detection algorithms for surveillance applications. *British Machine Vision Conference*, pp. 1-10.

Su, Z., Zhang, H., Li, S., & Ma, S. (2003). Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning. *IEEE Transactions on Image Processing, 12*(8), pp. 924-937.

Trabelsi, D., Mohammed, S., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2013). An unsupervised approach for automatic activity recognition based on hidden Markov model regression. *IEEE Transactions on Automation Science and Engineering, 10*(3), pp. 829-835.

Tran, S. D., & Davis, L. S. (2008). Event modeling and recognition using markov logic networks. *European Conference on Computer Vision*, pp. 610-623.

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: a survey. *IEEE Transactions on Circuits and Systems for Video Technology, 18*(11), pp. 1473-1488.

Tziakos, I., Cavallaro, A., & Xu, L. Q. (2010). Event monitoring via local motion abnormality detection in non-linear subspace. *Neurocomputing, 73*(10), pp. 1881-1891.

Valera, M., & Velastin, S. A. (2005). Intelligent distributed surveillance systems: a review. *IEE Proceedings - Vision, Image and Signal Processing, 152*(2), pp. 192-204.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), pp. 988 - 999.

Veeraraghavan, A., Chellappa, R., & Roy-Chowdhury, A. K. (2006). The function space of an activity. *IEEE Conference on Computer Vision and Pattern Recognition*, *1*, pp. 959 - 968.

Velipasalar, S., Brown, L. M., & Hampapur, A. (2006). Specifying, interpreting and detecting high-level, spatio-temporal composite events in single and multi-camera systems. *Computer Vision and Pattern Recognition Workshop*, pp. 110-110.

Weng, S. K., Kuo, C. M., & Tu, S. K. (2006). Video object tracking using adaptive Kalman filter. *Journal of Visual Communication and Image Representation, 17*(6), pp. 1190-1208.

Westermann, U., & Jain, R. (2007). Toward a common event model for multimedia applications. *IEEE MultiMedia, 14*(1), pp. 19-29.

Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. P. (1997). Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), pp. 780-785.

Xiang, T., & Gong, S. (2005). Video behaviour profiling and abnormality detection without manual labelling. *IEEE International Conference on Computer Vision*, *1*, pp. 1238-1245.

Xie, L., & Yan, R. (2008). Extracting semantics from multimedia content: challenges and solutions. *Multimedia Content Analysis*, pp. 1-31.

Xie, L., Sundaram, H., & Campbell, M. (2008). Event mining in multimedia streams. *Proceedings of the IEEE, 96*(4), pp. 623-647.

Yan, W. Q. (2016). *Introduction to Intelligent Surveillance.* Springer. doi:10.1007/978-3-319-28515-3

Yang, C., Duraiswami, R., & Davis, L. (2005). Efficient mean-shift tracking via a new similarity measure. *IEEE Conference on Computer Vision and Pattern Recognition*, *1*, pp. 176-183.

Yang, J., Zhang, D., Frangi, A. F., & Yang, J. Y. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(1), pp. 131-137.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: a survey. *ACM Computing Surveys, 38*(4), 13.

Zelnik-Manor, L., & Irani, M. (2001). Event-based analysis of video. *IEEE Conference on Computer Vision and Pattern Recognition*, *2*, pp. II-123.

Zelnik-Manor, L., & Irani, M. (2006). Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(9), pp. 1530-1535.

Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), pp. 915 - 928.

Zhao, L., & Thorpe, C. E. (2000). Stereo- and neural network-based pedestrian. *IEEE Transactions on Intelligent Transportation Systems, 1*(3), pp. 148-154.