UNIVERSITY OF SOUTHAMPTON

Biologically-Inspired Motion Detection and Classification: Human and Machine Perception

by

Vijay Laxmi

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering and Applied Science
Department of Electronics and Computer Science

March 2003

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Vijay Laxmi

Humans are good at the perception of biological motion, i.e. the motion of living things. The human perceptual system can tolerate not only variations in lighting conditions, distance, etc., but it can also perceive such motion and categorise it as walking, running, jumping etc. from minimal information systems such as moving light displays (MLDs). In these displays only specific points (e.g. joints in the case of a human being) are visible. Although a static display looks like a random configuration of dots, a dynamic display is perceptually organised into a moving figure. Some kind of temporal integration of the spatial contents seems to be a part of the perception mechanism; as manifested from the minimum presentation time required for biological motion to become apparent. One possible way to understand human perception may be to build an equivalent machine model. An analysis of the workings of this machine may lend us an insight into human perception. In this work, we considered a closed set of 12 different categories of MLD sequences. These sequences were shown to 93 participants and their responses are used as the basis of comparison of human and machine perception. Human responses were compared with the performance of k-nearest neighbour and neural network detectors. Machine perception is found to differ from human perception in some important respects. We also examined the related aspect of person identification on the basis of gait. This has important applications in the fields of surveillance and biometrics. In recent years, gait has been investigated as a potential biometric; as this may be the only information available to identify a distant and/or otherwise masked person. Humans can learn to recognise different subjects in MLDs. In our experiments with a dataset of 21 subjects, an accuracy of nearly 90% and 100% was achieved with neural network and support vector machine classifiers respectively. Also the machines were able to make this recognition in a fraction of a gait cycle.

Contents

List of Figures

List of Tables

Nomenclature

MLD	Moving light display.
Detection	Detection of human motion.
Classification	Person identification.

Absolute mode	Translatory component present.
Spot mode	Translatory component removed.

k-NN	k-nearest neighbour.
ANN	Artificial neural network.
SVM	Support vector machine.

F	Frames per sequence.
M	Dots per frame.
N	Frames per data-point.

NOR	Fronto-parallel view.
DIR	Change of direction.
WBK	Walking backward.
INV	Upside-down or inverted view.
TOP	View from top.
OBQ	View from top at an angle.
PER	Permuted sequence.
SPT	Dots are slightly perturbed.
LPT	Dots are largely perturbed.
SSR	Spatially scrambled dots.
PSR	Inter-dot phase scrambled.
RAN	Random configuration of dots.

Acknowledgements

During the period of my PhD, I have had help from many quarters. I wish to thank all who have helped me through this. I am also grateful for the constant support and encouragement from my supervisors, Dr. JN Carter and Dr. RI Damper. I am also thankful to all my friends and colleagues in the ISIS group. I gratefully acknowledge partial support by the European Research Office of the US Army under contract N68171-01-C-9002.

I wish to thank my parents who helped taking care of little Viraj. I would like to thank Manoj, my beloved husband, for his patient understanding and loving support. Last but not the least, little Viraj who has brought moments of happiness with his cheerful nature during the times of stress.

To my father-in-law . . .

Chapter 1

Introduction

It is a well-known fact that we, human beings, are capable of recognising biological motion i.e. the motion of living things, be it human or animal. This capability remains unaffected by distance variations or poor visibility conditions. Even in poor quality videos or blurred images, human beings not only perceive the motion but also the kind of motion, e.g. jumping, dancing, hopping, running or walking (Cédras and Shah 1995; Bobick and Davis 1996a). Often we can recognise a friend walking at a distance. Any familiarity clues such as clothes, appearance, hair-style are obliterated at large distances. So it must be the motion which is responsible for this identification.

The term "biological motion" encompasses the motion of all living beings. In this thesis, however, we will use this term synonymously with human gait. Human walking is a complex motion and comprises of many simpler motions which are translatory and/or rotational, more specifically pendular, in nature. A good understanding of gait perception may give an insight into the visual perception mechanism. This has led psychologists to use human motion as a basis for perception-related studies. Johansson (1973) was the first to apply moving light displays to such studies. These display only a set of moving dots located on the body with no other pointers to the shape of the moving object. With the help of such studies it was concluded that motion perception is possible even from moving light displays. As these displays are found to be useful tools in human perception, they remain central to this thesis. A brief discussion about these displays is presented next.

1.1 Moving Light Displays

Moving light displays, also known as MLDs, are obtained by affixing small lights to the specific points of the object under consideration and filming it in nearly dark conditions such that the resulting displays do not carry explicit information about the shape, structure or contours of the object. Another alternative is to attach reflector patches to the

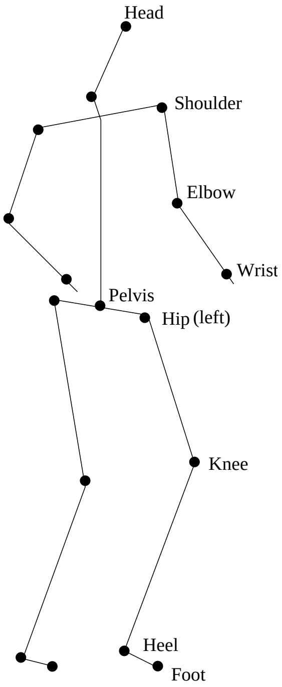


FIGURE 1.1: Major joints of a human body. Only the joints referenced in this study have been shown.

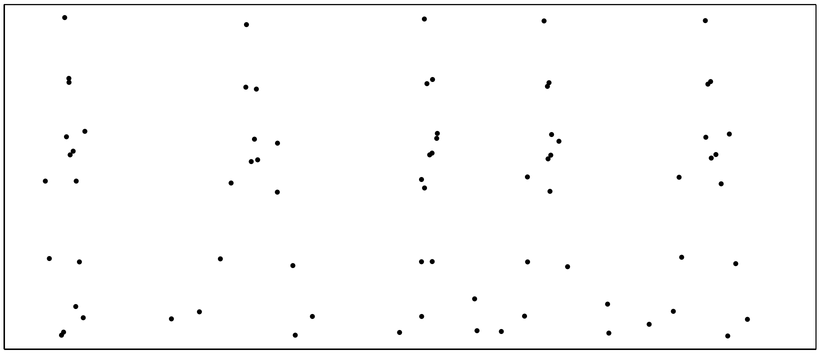


FIGURE 1.2: Snap-shots from a moving light display sequence of a human subject walking from right to left.

object and filming it in minimal lighting conditions. In either case, the recorded film displays only the points. In perception studies related to human motion, these markers are attached to the major joints e.g. head, shoulder, elbow, hip, knee, ankle, etc., as shown in Figure 1.1. MLDs are used as they do not provide explicit information about the human shape; the motion and relative location of dots are the only available cues. The only information available is a set of moving points or dots. This mimics the minimal data input sequence for the motion perception experiments. In spite of this, the perception of the human motion displayed remains vivid. Ever since Johansson (1973) used such displays to demonstrate the capability of the human perceptual systems, researchers have extensively used them, or their computerised adaptations, to determine the processes or mechanisms underlying visual motion perception. Figure 1.2 shows some snapshots from a moving light display for a human walking motion.

1.2 Motivation

The work presented in this thesis is inspired by the ease with which human beings can detect biological motion even in MLDs. Although person identification from MLDs appears to be much harder task (Cutting and Kozlowski 1977), humans can, nevertheless, learn to achieve 100% accuracy for a set of 6 walkers (Stevenage, Nixon, and Vince 1999). This suggests that MLD data are rich in information for perception. If humans can detect a moving subject, probably machines can detect it as well.

1.2.1 Primary Motivation

A biologically-inspired system, which is based on how humans deal with the visual input, is likely to be more robust, noise-tolerant and view invariant, as the human perceptual system also possesses all these characteristics. Development of such a system, however, needs a better understanding of the perception mechanism. It needs to be ascertained what factor(s) are responsible for the human perception. From the point of view of machine learning, an insight into how humans perceive biological motion may greatly aid machine vision especially in the fields of surveillance, tracking and security applications. This is the primary motivation for the work presented in this thesis.

1.2.2 Secondary Motivation

Another motivating factor is the need for personal identification systems and the significant role a biometric can play in such systems (Jain, Bolle, and Pankanti 1999). Personal identification or determination of the identity of a person is critical in today's complex, geographically mobile and vastly interconnected information society. The need is further augmented with the rise in global terrorism. The task of ascertaining the correct identity of an individual trying to access a system or enter a country becomes critically important especially for security and military applications. Ever-increasing use of machines and availability of massive volumes of data deemed it necessary to implement automated personal identification systems. Another application where these systems can play a significant role would be the prevention of crime.

In this thesis, the term *detection* will be used for biological motion detection i.e. recognition of human-like motion in an MLD sequence. The term *classification* will be used synonymously with person identification.

1.3 Human Motion Studies – An Overview

Human motion studies can be broadly divided into three categories. The first category is concerned with detection of the presence of human motion in an image sequence. Major areas for the first category would be tracking applications and detection of pedestrians in an automated traffic system. The second category involves the detection and categorisation of various human activities in an image sequence. The third category is a classification problem and is solely concerned with person identification.

The approaches to human motion detection are based on determining a generic model or template of a moving human shape. Baumberg and Hogg (1993) described a flexible shape vector derived from the cubic B-spline description of the body contours. Niyogi and Adelson (1994a) used the braided pattern in XT-slice of an image sequence to signal the presence of human motion; an active contour of the body was then used to determine the identity of the walker. Niyogi and Adelson (1994b) also described another method based on spatiotemporal surfaces, the surface being a combination of standard parametrised surface – the canonical walk – and a deviation surface specific to the individual walk. Yacoob and Davis (1998) used learnt parameterised temporal-flow models with spatial constraints, models being sets of orthogonal principal components, to track rigid and articulated motions. Bregler and Malik (1998) described a mathematical tracking technique, the product of exponential maps and twist motions.

Bobick and Davis (1996a, 1996b) and Davis and Bobick (1997) employed an optical flow based technique to categorise the various human motions. Advantages of the technique include no adverse effect of blurred image, view-based categorisation and localisation of areas where motion takes place. Yacoob and Black (1998) described parameterised modelling and recognition of activities using principal components of motion parameters. Bregler (1997) used a hierarchical approach to human motion categorisation. The image sequence was first segmented into blobs and motion parameters of these blobs were estimated using an expectation maximisation algorithm. These blob parameters were used to determine the dynamical category of the image sequence. Complex gestures of the image were then determined using learnt hidden Markov models. Takahashi and Ohya (1999) discussed application of neural networks to human motion recognition and categorisation.

1.4 Methodology

The previous section summarised computational approaches previously reported in the literature for motion detection. However, work presented in this thesis differs in the sense that here emphasis is not on the detection of the moving subject but on identifying the factors which allow this. The main objective is to gain understanding into how humans are able to do this detection. A better understanding of human perception may be achieved by analysing perception of an equivalent machine. The work presented in this thesis investigates if it is possible for a machine to detect biological motion in a similar manner. And if the answer is 'yes', does this lead us any closer to the understanding of human perception? As humans can detect motion even from MLDs, where only information of dot positions is used as the input, the machine is also fed with this information only. The underlying assumption is that MLD data has sufficient information for equivalent perception. It needs to be seen if the machine can use this information in an equivalent manner or not. As only a dynamic sequence invokes the perception, some kind of mechanism needs to be established to impart temporal information to the machine. Human responses were collected through an experimental design. Framework and responses of which appear in Chapter 5. These responses formed the basis of comparison of human and machine perception.

This thesis also investigates the applicability of this approach to identification of people on the basis of their walking. As human beings can perform this task even from MLDs that lack any information about the shape of the object, labelled data that contains information only about the coordinates of body joints are used as the input for the classification problem.

For both detection and classification, three machine learning models – k nearest neighbour (k-NN), artificial neural network (ANN) and support vector machines (SVM) – have been considered. Neural networks were used as they are good exploratory tools to investigate the feasibility of our problem. For the simplification of analysis of machine learning, only simple feed-forward neural architectures with back-propagation learning were employed. Initially a Python (van Rossum 1999) program was developed. Python was used as it is an interpreted, object-oriented, dynamic language suitable for rapid application development. Later a more efficient C++ version of the same algorithm was used. Both programs were tested on parity problem and benchmarked Iris data (Fisher 1936) to ensure a good implementation. However, neural architectures suffer from problems of non-convergence and local minima traps. SVMs, on the other hand, can produce stable results and are good for two-class problems. For the purpose of detection, this is justified as the problem is inherently two class. A practical limitation of the inability of SVM implementations (Gunn 1997; Joachims 1999; Rifkin 2000) to handle large amount of data, however, meant that these could not be used for biological motion detection as intended.

For classification problems, the SVM approach needs to be modified. Two different multi-class implementations based on one-against-one and one-against-rest criteria were developed to extend the two-class MATLAB toolbox by Gunn (1997). Both of these implementations were tested on Iris data (Fisher 1936). A k-NN approach was used as it is simpler to analyse and more informative for inherent clustering, if any, in data. It also provides a bound on what is the best a machine learning system can achieve. In theory, as the number of points in feature space $n \to \infty$, the error rate of k-NN is never twice that of the Bayesian classifier and monotonically approaches the Bayesian error as k increases (Devijver and Kittler 1982; Chen, Damper, and Nixon 1997). All these learning models are detailed in Appendix C.

1.5 Contributions of this Thesis

A biologically-inspired approach to motion detection and classification is presented in this work. We are motivated by the fact that humans can recognise a moving subject in an MLD within 0.2 s from a small number of lights attached to the body (Johansson 1973). Humans can also infer something about the personal characteristics or identity of a walker from MLDs (Cutting and Kozlowski 1977; Stevenage, Nixon, and Vince 1999). Motivated by these findings, we have trained computational models to discriminate natural human motion sequences from “near miss” sequences (e.g. random, scrambled, perturbed, etc.) using information about the dots only. Initially only three different categories of sequences were considered (Laxmi, Carter, and Damper 2002b). Later the total number of categories was increased to 12. Extensive psychophysical experiments were undertaken in which 93 participants were shown various MLD sequences and were asked to indicate the presence/absence of natural human motion. The results act as the training and validation data for the machine model, which uses the same categories as inputs (Laxmi, Carter, and Damper 2003). Since humans can infer identity from MLDs, we have also tested ANN and SVM machine models for MLD based classification. A correct identification rate of 95–100% for a set of 4 manually labelled subjects (Laxmi, Carter, and Damper 2002a; Laxmi, Carter, and Damper 2002c), and 85–90% for a set of precisely labelled dataset of 21 subjects (Laxmi, Carter, and Damper 2003) were achieved.

An underlying assumption for the work presented is that the MLD data have enough information for the purposes of both detection and classification. Unlike other related works, which use image-processing techniques, the work presented here is motivated by data and not by any specific methodology. Hence, no pre-processing was done for the input data. The main aim is to test if a machine can learn discriminatory features of the data as humans do. To the best of our knowledge, this is the first attempt to evaluate machine perception in context of human perception.

1.6 Thesis Organisation

The thesis is structured as follows. Chapter 2 is a summary of abilities of human perceptual system in context of MLDs. It also summarises various theories of visual perception, which form the basis of the experiments conducted to evaluate human responses for different kinds of MLD sequences. Chapter 3 is a detailed description of the methodology adopted for machine perception. What constitutes a human or non-human motion in the context of this thesis is presented in this chapter. Chapter 4 reports results of an investigative study to test if machines can detect biological motion or not. As the present work uses some MLDs with no reported literature on the human responses they elicit, an experiment was performed to collect such responses. The experimental framework and the human responses are discussed in Chapter 5. Chapter 6 is a summary of results of biological motion detection by machines – k-NN and ANN. This chapter attempts to determine factors responsible for the machine perception. A comparison of the machine perception with the human perception is also presented. Chapter 7 discusses suitability of gait as a biometric; the number and complexity of motions constituting human gait lends credibility for using this as a biometric. This chapter also presents results obtained by different classifiers with reference to the gait based person identification. Chapter 8 concludes the work presented in this thesis with some pointers to future work.

Chapter 2

Human Perception of Motion

Our primary aim is to gain knowledge of the human perception of biological motion. To this end, it is necessary to know the capabilities of the human perceptual system. This chapter summarises outcomes of the perception related studies. It is structured as follows. Sections 2.1 and 2.2 present structure-based hierarchical theories of perception. A coding theory of perception is presented in Section 2.3. Section 2.4 reviews the perception of upside-down or inverted displays. Processing constraints which may be responsible for a stable perception of otherwise mathematically ambiguous displays are discussed in Section 2.5. A review of possible minimal sub-configurations sufficient for biological motion perception is presented in Section 2.6. Human performance in the context of person identification from MLDs is discussed in Section 2.7. A summary of the chapter and its implications for this thesis is discussed in Section 2.8.

2.1 Johansson's Model of Visual Perception

Johansson (1973, 1975, 1976) was the first researcher to use MLDs for studies on biological motion perception. In his experimental setup, he showed MLD sequences displaying humans carrying out the various activities like walking, running, hopping, dancing, cycling, etc. In his studies, 12 light elements, as illustrated in Figure 2.1, were used. The following responses were the outcome of these studies:

- Observers were able to recognise a human figure and categorise the types of motion. Spontaneous and correct identification of various activities, e.g. running in different directions, cycling, climbing, dancing in couples, etc. was made without exception.
- Recognition was possible only in a dynamic display of lights. Static configuration of lights was not recognised.

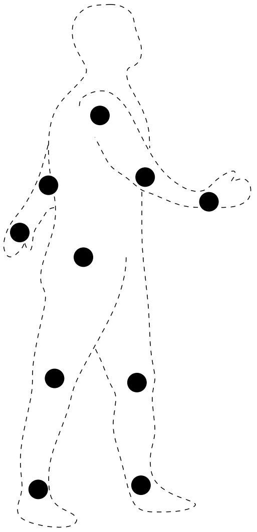


FIGURE 2.1: Johansson used 12 light elements, as depicted here, in his displays. The actual display showed no body contour. It has been drawn to clarify the actual joints.

- For MLDs containing a human walk, a display of 0.2 s, about five frames of Johansson display, was enough for the perceptual organisation of the pattern to a meaningful unit.
- Direction of the subject relative to camera (45°–80°) did not affect recognition.
- Human walking was still recognised when the number of recorded elements in the MLD was reduced. Five elements representing hip-and-legs part were sufficient, though some subjects described these patterns as moving arms of a walking person.
- Subtraction of semi-translational motion at hips did not affect recognition. Adding common components to the element motions, too, did not have an adverse effect on the identification of walking patterns.
- Observers were able to identify walking left, walking right, walking backward right, walking backward left, running right, running left, puppet walking left and puppet walking right from a series of display sequences (Johansson 1976). An exposure time of 0.4 s was required to correctly identify all patterns, "artificial" patterns with puppet-like motions were found to be the hardest to recognise.

2.1.1 Perceptual Units

Johansson concluded that the definite grouping of light points was determined by some general principles but the vividness of the percept was due to prior learning. These general principles, as described below, determined the perceptually connected units:

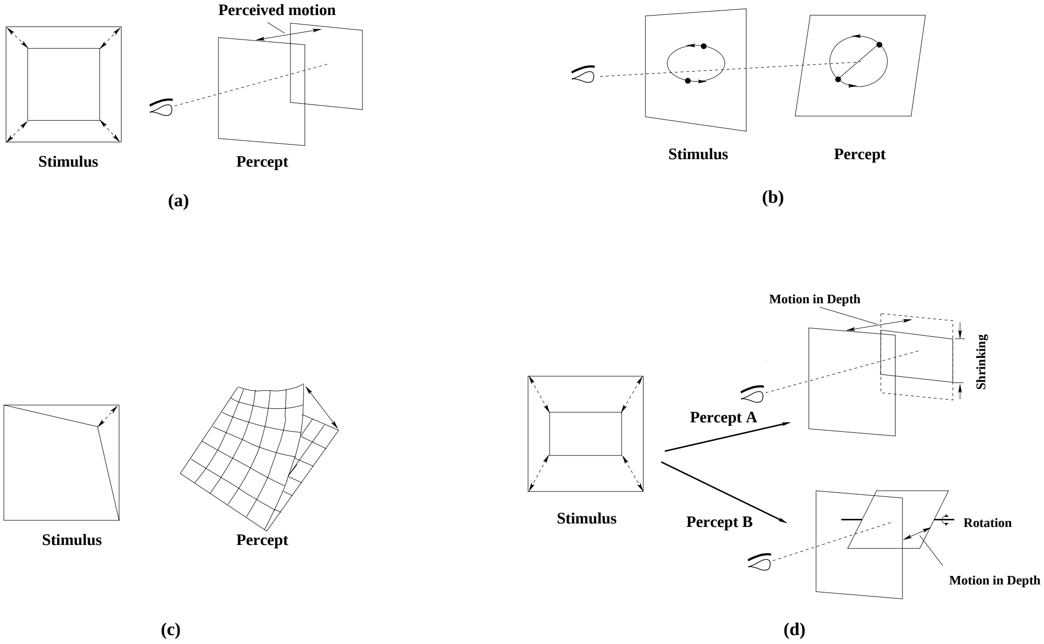


FIGURE 2.2: (a) Four corner points of a shrinking/expanding square and moving diagonally are perceived as an advancing/retreating square. (b) Dots traversing an ellipse are perceived as a rigid rod rotating in depth. (c) Four corner points of a shrinking/expanding square and moving diagonally are perceived as an advancing/retreating square. (d) Four points of a square changing to rectangle and then back to square has two possible perceptions.

- Dots in an MLD, as projected on the picture plane of the eye, are always perceptually related to each other.
- Equal and simultaneous motions in a series of proximal dots automatically connect these elements to rigid perceptual units.
- When equal simultaneous motion vectors can be mathematically abstracted from the motions of a set of proximal dots according to some simple rules, these dots are perceptually isolated and perceived as one unitary motion.

The term *equal* used here includes all motions that follow tracks converging to a common point at infinity on the picture plane and whose velocities are mutually proportional relative to this point. For fronto-parallel projections, this reduces to equal motion directions and velocities.

2.1.2 Visual Vector Analysis

To determine the principles governing the visual perception, Johansson (1975) did some simple experiments (see Figure 2.2). Four points of a contracting/expanding square were perceived as advancing or retreating square. Two points traversing an ellipse were perceived as a rod rotating in a slanted plane. Four points of a square undergoing transformation to a rectangle resulted in two different perceptions. In one percept, the square was perceived as advancing/retreating and simultaneously undergoing shrinking/expanding and in the other percept, these points were perceived as a rotating square. Four points of a square in which only one corner point was moving along the diagonal path evoked the perception of a surface undergoing bending.

On the basis of these observations, Johansson concluded that the visual perception system follows the principles of central perspective and not Euclidean space and prefers maintaining figural constancy. Any change in the figure/shape is perceived as a central perspective transformation, rather than the actual change happening in Euclidean space. Thus the visual system prefers invariance of figure size by inferring motion in 3D space. He formulated a geometric framework, termed visual vector analysis, to explain the perception mechanism. This analysis is based upon the following principles:

1. Sets of spatially invariant relations in the stimulus pattern are spontaneously isolated.

2. For the moving elements of the stimulus pattern, the perceptually lower speed generally forms a reference frame for the higher one.

3. The eye fixation can play an important role for the perceived structure. What is fixated tends to act as a primary reference for other motion components. Therefore, principle (2) is fully valid only under fixation on the background.

For complex motion patterns such as biological motion pattern, the visual system abstracts a hierarchical series of moving frames of reference and motions relative to each of them. *Equal* vectors or vector components form a perceptual unit that acts as a moving frame of reference in relation to which secondary components seem to move. As the hip-shoulder unit has the lowest speed relative to the background and it connects the four pendulum subsystems to a unified system of higher order, the hip element is taken as a starting point in the stimulus-percept analysis.

In visual vector analysis, the motion vectors of the hip in any fixed frame of reference, define the frame of reference in which motion of the knee is abstracted. The motion vectors of the knee, then, define the frame of reference in which the ankle motion is analysed. The motion vectors of any joint are, in effect, vector differentials. Thus motion abstraction is akin to spatio-temporal differentiation. For the grouping of these perceptual elements, the visual memory needs to perform a continuous integration of these differentials. So, mathematically speaking, the visual perception is a process involving an integration of spatio-temporal differentials (Johansson 1976), which can be abstracted as visual vector differentials.

2.2 Centre of Moment and Motion Perception

Further studies (Cutting and Kozlowski 1977; Kozlowski and Cutting 1977; Barclay, Cutting, and Kozlowski 1978) were directed toward the identification of the sex of a walker. These studies revealed some interesting points about the visual perception of motion:

- Gender determination rate was very high when the observer was maximally confident, but no better than the chance rate when minimally confident.
- Arm-swing and walking speed are correlated in natural gait, the faster the gait, the more is the arm-swing.
- Men have larger shoulder swing and women have larger hip swing. In general, women swing arms more, walk faster but take smaller steps than men. But none of these features is necessary for identification of the gender as increasing arm-swing did not make walkers appear more feminine; nor did decreasing arm-swing make them appear more masculine.
- Any deviation from normal arm-swing or change in subject's speed adversely affected the gender determination.
- Upper-body and full-body conditions were significantly more identifiable than the lower body conditions. Arm-swing information appeared to be helpful in recognising the sex of a walker.
- Any joint is sufficient, and no joint is necessary, for identification. Ankles alone resulted in a better determination rate than chance. Although the hip, by itself, did not seem particularly important, but in conjunction with information from the upper body it enhanced perception.
- Gender determination required about 2.7 s or roughly two step cycles. This is significantly larger than the time required for recognition of a human shape.
- Inversion of the stimulus display produced the unexpected effect of reversing the apparent sex of most walkers. When presented upside down, the male walkers appeared female and female walkers appeared male.
- Although participants did not depend on particular joints for accurate judgement of gender, some aspects of joints might have been a cue. Blurring of the point-lights attached to the joints lowered the perceived accuracy to the chance rate.
- When faced with the upside down displays, viewers were significantly more confident when incorrect.
- Although on-joint displays have better detectability, off-joint displays were also found adequate for representing human motion (Cutting 1981).

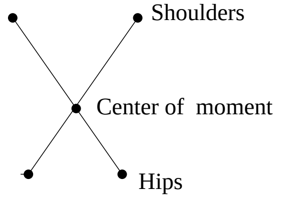


FIGURE 2.3: Analytic way to determine the approximate location of centre of moment as suggested by Cutting, Proffitt, and Kozlowski (1978).

On the basis of the above observations, Cutting, Proffitt, and Kozlowski (1978) concluded that some structural and transformational invariants play a significant role in the gender perception in motion. According to them, centre of moment (Figure 2.3) has the requisite characteristics of a structural invariant. Each body has two centres. The first is the centre of moment about which all the movements occur and the second is the centre of gravity about which the mass is distributed. Since arms and legs are pendular, they rotate about their pivots namely the shoulders and hips. As these pivots undergo torsion, these have a higher order of moment. The centre of moment is slightly lower for males than females due to wider shoulders and narrower hips. Also the centre of gravity is lower for females due to heavy thighs. Cutting, Proffitt, and Kozlowski (1978) suggested that these differences might account for the gender determination with the centre of moment acting as a structural invariant and the pendular and torsional moments acting as transformational invariants. For accurate gait perception, both of these invariants need to be considered.

Cutting and Proffitt (1981) suggested that, while perceiving biological motion from the MLD displays, we extract information in logical steps and perceive the parts of the body as a system of dynamic nested dependencies. The motions and locations of the hip and shoulder are extracted first as they are close to the body's deepest centre of moment within the torso. These points now serve as the static centres of moment for the analysis of the lower and upper body respectively. For example, for the lower body, the hip acts as the centre of moment for the knee, which moves in pendulum fashion about it. One perceives the knee motion only by its motion relative to the hip. Therefore, "the motions of a whole are perceived as the movements of parts about a point that is analytic to the whole, whereas the observer-relative displacement of the whole is perceived as the dynamics of that point".

Minimum Principle: Given a choice between two perceivable patterns of an ambiguous stimulus, the perceptual system seems to choose the simplest (Cutting and Proffitt 1982). This simplicity heuristic in perception is called the minimum principle. Motion of a moving body consists of:

- **Absolute motion** of each element in the display.
- **Common motion** of the whole configuration relative to the observer.

- **Relative motion** of each element to other elements within the configuration.

These three components are related through the following equation for all elements perceived as a single unit.

$$\text{common motion} + \text{relative motion} \ = \text{absolute motion} \tag{2.1}$$

Only relative and common motions are usually perceived when viewing an event. According to Cutting and Proffitt (1982), the minimum principle is operative in two processes, one involving common motions and the other relative motions. Thus, common and relative motion abstraction processes are concurrent. If one of the processes reaches the solution first, the solution for the other process is determined residually. If the relative motions are first to be minimised, the common motions fall out as residuals and vice versa.

Cutting and Proffitt (1982) applied this principle to biological motion perception also. In terms of the minimum principle, the perception within the torso is determined first by minimising the motions of the shoulders and the hips along the stress lines of the twisting torso. The component structures of the upper and lower body are perceived through a minimisation of the common motions. Once the shoulder motion is extracted, this point becomes a static pivot for obtaining the pendular movement of the elbow. Similarly, the elbow becomes the static pivot for the wrist movement. In perceiving the nested component structures of the upper and lower body, each step in information processing causes one point to become static. The common motion within the subsystem is achieved relative to this point. The common motion of the whole body is seen in the moment within the torso. Thus "motion of a whole is perceived as the movement of parts, motion and location of hip and shoulders being extracted first, as these are the deepest of moments within the torso".

2.3 Coding Theory for Motion Perception

Restle (1979) proposed a coding theory for motion perception. According to this theory, of all possible interpretations for a moving light display, the perceptual system chooses the one having minimum information load. If two or more interpretations have the same information load, the perception is not stable. Some participants would perceive one interpretation while the others would perceive another. Information load is the set of parameters necessary to specify the interpretation. The set of parameters can be classified as follows:

- **Structural:** Describe the organisation of dots. This parameter may come into existence only when the dots are perceptually grouped and are regularly placed.

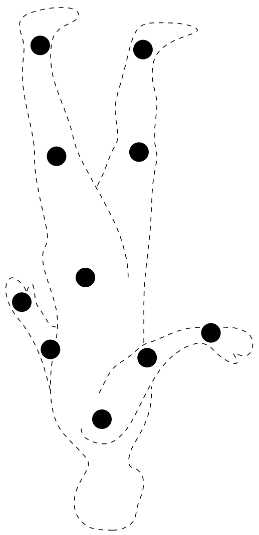


FIGURE 2.4: Static illustration of inverted Johansson display.

- **Positional:** Describe the location of dots. One location parameter per perceptual group is sufficient.
- **Motion:** Describe the motion of dots. Pendulum-like oscillatory and elliptical motions, encountered in biological motions, are projections of the circular motions. Five motion parameters – amplitude (α), phase (ϕ), wavelength (λ) and two angle parameters (β, τ) to describe the orientation of motion plane with respect to image plane – are sufficient for the purpose.

Restle's model is restricted to the motions generated on a circle. Parameters used by him – phase, amplitude, rate, axis of movement and tilt – could apply both to the circular and pendular motion but there is no mechanism in the model to distinguish between these two. Also there is no simple way to code non-repeating movements like translation. Cutting (1981) overcame these limitations. By application of the coding theory adapted to moving dots display obtained by on-joint and off-joint stimuli for a canonical walker and on-joint stimulus for a spatially anomalous walker, he suggested that both movements and the spatial relations are necessary for an event perception.

The major limitation of the coding theory is that it does not explain how to obtain a perceptual group. This theory can be used to explain the perception of an event only after the perceptual grouping is known.

2.4 Perception of Inverted Moving Light Displays

Sumi (1984) found that when the Johansson displays were inverted, and run backward, some sort of human movement could still be perceived. However, it was perceived

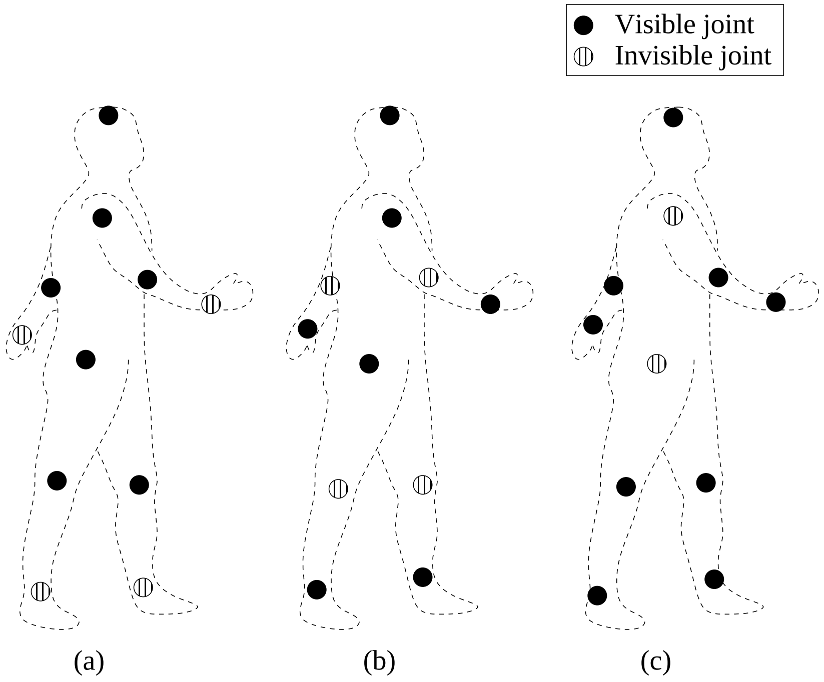


FIGURE 2.5: Minimal sub-configurations for biological motion perception. In (a) extremities (wrists and ankles), (b) mid-limb elements (elbows and knees) and (c) central elements (shoulders and hips) are missing.

more frequently as an upright image of a person moving forward in a very strange manner rather than as an inverted image of a person moving backward. The moving-dot pattern of the Johansson's film was hardly ever recognised as being upside down or as an inverted image of a moving person. Instead, it invoked perception of very unfamiliar motions of a biological type. Very few participants perceived the moving spots as a person walking or running in an inverted position. A static illustration of an inverted display is presented in Figure 2.4.

Pavlova and Sokolov (2000) reported that the spontaneous recognition of an MLD walker improves abruptly when the image-plane is rotated from the inverted to upright orientation. Despite prior familiarisation with an MLD figure at all orientations, its detectability within a mask decreased with a change in orientation from the upright to a range of $90° - 180°$.

The information inherent in Johansson patterns is retained even in the upside-down temporally reversed condition. In the light of these findings, any theory of perception based on the structural information as suggested by Johansson's visual vector analysis and Cutting's centre of moment oriented perception needs to be modified, if not totally abandoned as invalid.

2.5 Processing Constraints in Perception

From a mathematical point of view, MLDs are ambiguous and do not represent any unique three-dimensional object. Bertenthal and Pinto (1993) maintain that some processing constraints are essential to the perception of visual information, because the structure of the projected image, being a 2D projection of the 3D world on the retina, is undetermined. From a mathematical perspective, this image is consistent with an in-determinant number of different interpretations. Yet, the human perceptual system perceives only one stable interpretation. This suggests that certain constraints, which reduce infinite number of possible interpretations to only one, are part of the perceptual process. According to Bertenthal and Pinto (1993), these constraints are dynamic symmetry, frequency entrainment and a periodic attractor. Dynamic symmetry means that opposite limbs connected at shoulder and hip joints move in alternation. Frequency entrainment implies that all limbs of human body undergo pendular motion, having the same frequency but differing in the phase. Periodic attractor means that each joint angle displays its own characteristic pattern during the gait cycle. Significant findings of their work are:

- Relative inter-limb phase relationships were perceptually more influential than constancy of length of limbs. Violation of local rigidity constraints by varying limb lengths, as suggested in computational models (Webb and Aggarwal 1982; Hoffman and Flinchbaugh 1982), did not degrade the perception but alteration of these phase relations did.
- Single point masks were less effective than the triad masks consisting of 3 elements undergoing motion similar to that of a limb. This suggested that the participants were sensitive to the relative motion, and masks having the same relative motion were more effective.

2.6 Minimal Sub-configurations for Visual Perception

To determine the minimal conditions for perceiving a unique impression of biological motion in MLD stimuli, researchers have used masked, scrambled or partial MLDs. These studies have failed to identify a set of factors sufficient for biological motion perception. But many interesting features concerning the capabilities and flexibility of the human perceptual system are revealed. Some of the interesting findings are reported here. In their studies, Ahlström, Blake, and Ahloström (1997) reported the following:

- Perception of MLDs is immune to variations in dot contrast polarity.
- Regular biological motion sequences can be easily discriminated from phase scrambled sequences. Like regular sequences, phase scrambled sequences contain the

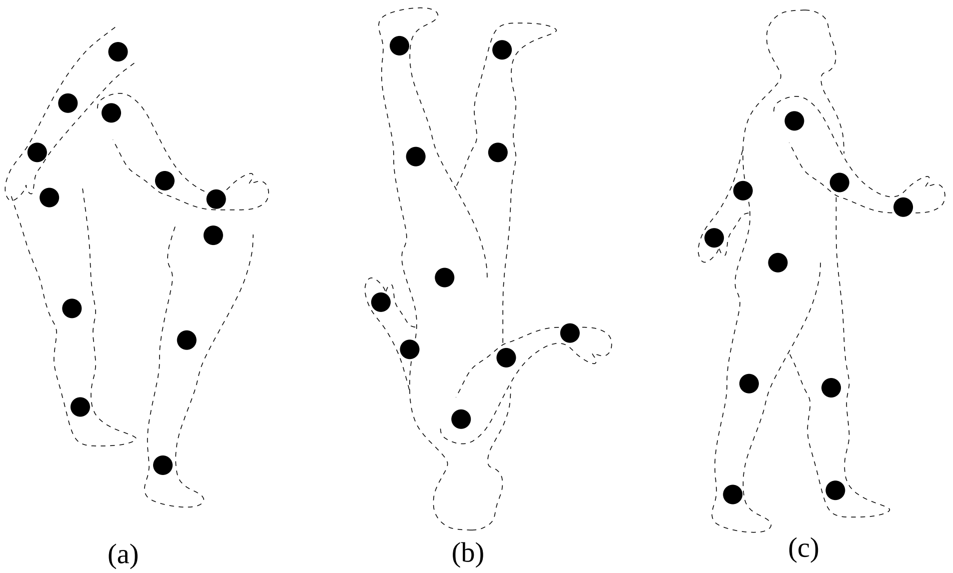


FIGURE 2.6: Minimal sub-configurations for (a) randomly-located limbs, (b) inverted figure and (c) upright figure.

same local dot motions but the starting point for the motion cycle of each dot is chosen randomly e.g. a dot associated with the wrist may start in frame 6 whereas the dot associated with the elbow may start with frame 14.

- Dots specifying only ankles were perceived as non-biological motion but addition of dots representing knees resulted in an impression of biological motion. This impression grew stronger as more dots were added.
- Superimposition of an inverted figure on otherwise normal biological-motion sequence resulted in a multi-stable perceptual grouping of dots. However, this perceptual multi-stability was absent when the dots describing an inverted figure were of different colours.

According to Pinto and Shiffrar (1999), the perception of human movement appears to be an integration of form and motion. The two principal components – dynamic symmetry among the limbs (equal and opposite motions of adjacent limbs) and the principal axis of organisation (primary structure about which the limbs are organised, i.e. torso) – play a fundamental role in production of human walking. Pinto and Shiffrar (1999) reported the following observations.

1. Detection of the figure missing elements on the extremities, i.e. ankles and wrists, did not differ significantly from the detection of the whole figure (see Figure 2.6(c)). Omission of the central elements – hips and shoulders – did significantly diminish performance. Omission of the mid-limb joints – knees and elbows – also impaired the performance. Figure 2.5 illustrates these sub-configurations with missing elements.

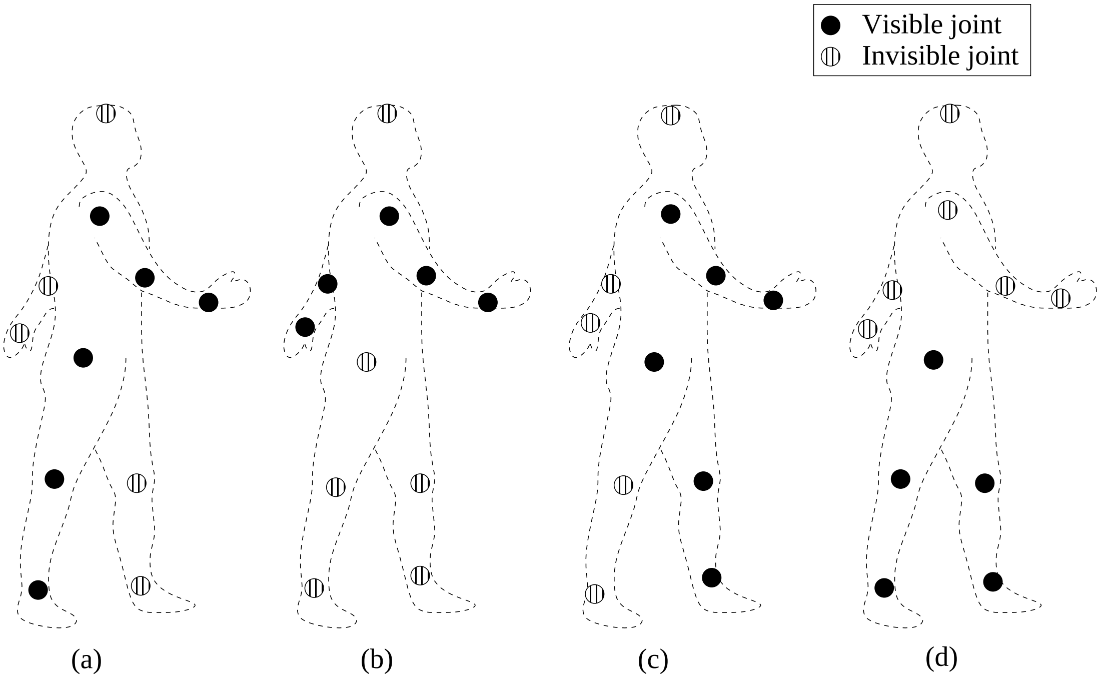


FIGURE 2.7: Minimal sub-configurations for biological motion perception. Of the major joints of four limbs, two arms and two legs, joints of any two limbs are shown. In (a) ipsilateral limbs, (b) contralateral limbs (arms), (c) diagonal limbs and (d) contralateral limbs (legs) sub-configurations are presented.

2. Detection of the upright figure even with missing elements was significantly better than that of the inverted figure (see Figure 2.6(b)).

3. The whole figure was detected with greater frequency and accuracy than the set of randomly-organised limbs (see Figure 2.6(a)).

4. Detection of ipsilateral (arm and leg on same side), contralateral limbs(both arms or both legs) did not significantly differ from that of diagonal limbs (arm and leg on opposite sides). All these sub-configurations are depicted in Figure 2.7.

5. Figures in which the light elements were positionally scrambled were not identified as human forms. It suggested that the spatial structure of the figure is essential for detection.

On the basis of observation 1, Pinto and Shiffrar (1999) concluded that a hierarchical vector analysis (Johansson 1973; Cutting 1981) by itself does not give a complete description of the visual perception. The three body configurations missing extremities, mid-elements and central elements maintained a hierarchical structure amenable to these analyses. However, the differences in human performance to these configurations cannot be explained by a simple hierarchical model. Rigid relations (Webb and Aggarwal 1982; Hoffman and Flinchbaugh 1982) alone cannot account for different performances as all these configurations have same number of rigid relations.

On the basis of observation 4, Pinto and Shiffrar (1999) concluded that while neither dynamic symmetry nor principal axis appeared necessary for human detection, the absence of both, as in the case of randomly organised limbs, reduced the detection to chance. If dynamic symmetry were necessary to human form detection, diagonal limbs would not be detected as these move in synchrony and no anti-phase information to indicate dynamic symmetry is available. If the elongated structure of principal axis were necessary, contralateral limb conditions would not be detected, as only legs or arms were shown.

Pinto and Shiffrar (1999) argued that the visual system responds equivalently to figures exhibiting any organisation of limbs consistent with the human form. Not only this, the visual system is capable of exploiting the configural information specifically indicative of the human form in the perception of MLDs. However, the figural coherence is not sufficient to explain the detection of human movement. If it were so, the inverted figure detection would not significantly differ from the upright one. In fact, this difference cannot be explained by models based on hierarchical vector analysis and rigid relations.

2.7 Biological Motion and Person Identification

Cutting and Kozlowski (1977) conducted MLD based experiments to assess the capability of the perceptual system for human identification. They reported that viewers could recognise themselves and others from the MLD sequences. Seven participants were shown six walking sequences of six subjects. The interesting features revealed in these studies can be summarised as follows:

- Although none of the participants could achieve 100% performance, the performance was above the chance rate for all particpants.
- The performance improved with time. The performance for the last three viewings (of all subjects) was better than the first three.
- The average performance was 27% for the first three presentations, actual performance ranging from 17%–39%. For the last three presentations, these figures were 59% and 22%–95%. The chance rate recognition was 16.7%.
- The poorest performance was exhibited by the participant who used the height as a clue.
- Successful identification rate was high (75%) when the observer was maximally confident and marginally better than the chance rate when the observer was minimally confident.

- Successful identification rate was better for the participants who used dynamic clues, such as speed, bounciness, rhythm, amount of arm swing, length of steps, for identification.

According to Stevenage, Nixon, and Vince (1999), the human visual system can learn to identify individuals on the basis of their gait signature under varying lighting conditions. In their experiments, 30 participants (15 male, 15 female) were shown image sequences of 6 subjects (3 male, 3 female) walking in simulated daylight, simulated dusk, as well as MLDs. The major findings are as follows:

- Participants learnt to identify the subjects correctly in average 8.3 trials. As all subjects were of about the same build, the participants had to depend upon gait.
- Adverse lighting conditions did not affect the learning rate to achieve 100% accurate identification.
- Learning was faster when the participants used dynamic features for identification.
- Even under adverse viewing conditions involving a single brief exposure of 2 s, humans could identify a target from a 'walking identity parade' at greater than the chance levels.
- Participants' confidence in the identification were correlated with their classification accuracy.

2.8 Implications for the Thesis

The human perceptual system is capable of perceiving human figures and their actions from MLD stimuli even in the presence of masks and random perturbations of the dots. Although the exact mechanism of this perception is not known, the dots undergoing some common motion appear to be perceived as a single unit. Relative and common motions seem more dominant than the absolute motion. As no single dot seems to be crucial to the perception process, it is difficult to determine which point or set of points plays a major role. The minimum presentation time required for the motion to be perceived indicates a possibility of some form of neural integration. However, as this duration is very small, it can be presumed that the perceptual process abstracts a small set of invariants from the moving-dot displays.

For the complex patterns like biological motion, it is likely that the abstraction of information is done in hierarchical manner. However, the exact point about which this abstraction starts is not very clear. According to Johansson, the abstraction process begins from a point with minimum perceptual speed. However, the actual point may still be affected by the eye fixation. According to Cutting, the first point to be extracted is the centre of moment, which lies somewhere in the middle of the torso. Pinto and Shiffrar (1999), however, contradict the structure-based hierarchical theories. Hence, a model-free approach (i.e. no data pre-processing) to the problem may be the requisite solution. This approach is discussed in the next chapter.

Chapter 3

Experimental Data and Methodology

As discussed earlier, one of the objectives of the presented work is to determine if a machine can detect a biological motion sequence in an MLD display in a manner similar to humans. Another objective is to see if the machine can also classify people on the basis of MLD information only. Figure 3.1 summarises the objectives of the work presented in this thesis.

Our approach to biological motion detection is model-free in the sense it is driven entirely by the input data without any pre-processing (i.e. data are not modelled). It is inspired by the fact that humans can detect such motion even in MLDs. It differs from other approaches as we have not employed any image processing techniques on the input data. It is assumed that the data contains all information necessary for the detection. It is anticipated that any structure, if it exists in data, will be learnt by the machine. The purpose of this preliminary study is to verify our assumption.

Three different types of machine learning models, explained in Appendix C, were used. Multi-layer neural networks with back-propagation learning have been used to solve temporal problems (Sejnowski and Rosenberg 1987). The advantage with this learning method is that analysis is not as complicated. As the complexity of problem from the machine perspective is unknown and these networks are suitable as good exploratory tools, neural networks with back-propagation learning were used. One of the most common problems with neural architectures is it may get stuck in one of the many local minima and may not converge converge to a global optimum. Support vector machines, however, provide an elegant solution to this problem. The obtained solution is unique and stable. Another advantage of this method is that use of kernels mean that the data modelling can be part of the learning model; the modelling can be varied by choice of the kernel. k-NN model is used to evaluate the performance of other models.

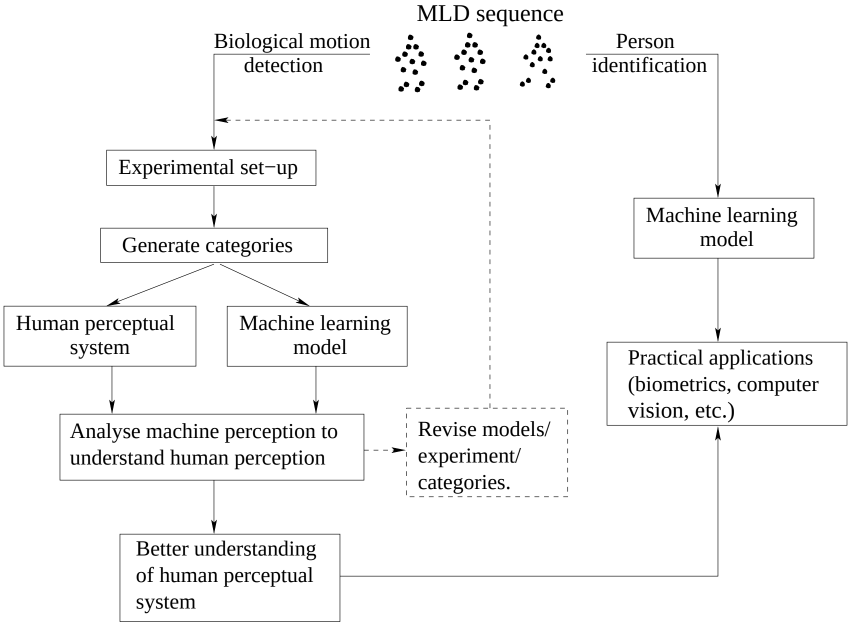

MLD sequence
Biological motion detection
Person identification
Experimental set−up
Generate categories
Human perceptual system
Machine learning model
Analyse machine perception to understand human perception
Revise models/ experiment/ categories.
Better understanding of human perceptual system
Machine learning model
Practical applications (biometrics, computer vision, etc.)

Figure 3.1: Flow-chart for illustration of the aims and objectives of our work.

Although the problem is temporal in nature, machine models like Hidden Markov Models or recurrent neural networks have not been used as these are difficult to train and analyse. In fact, for the detection problem, a recurrent neural network using Elman topology (MATLAB function *newelm*) was tried with no success. In view of slow convergence and large computation time, it is difficult to say if the network would converge or not.

Input to all the machine models is MLD sequences as both tasks (i.e. detection and classification) are inspired by the fact that humans can do these on this limited information. Section 3.1 discusses the various datasets used. For machine modelling of human motion perception, it is necessary to determine which of the MLD sequences invoke perception of biological motion and which do not. Although the real world can present a multitude of different types of MLD sequences, we have restricted these categories to a set of 12. This set is determined from the works reported in Chapter 2. These categories are discussed in Section 3.2. A pre-requisite for machine training is labels (positive or negative) of all categories in the context of biological motion. This labelling is also discussed here. These categories were presented in two different modes described in Section 3.3.

As the only available information from the MLDs is about the position of moving dots. The dot coordinates are used as input to the machine detectors. But humans can perceive a moving shape only when the display is dynamic, so this necessitates that the detectors be provided with temporally integrated MLD information. Hence a common framework which enables all the machine learning systems to deal with the temporal sequences needs to be established. This is explained in Section 3.4. Approaches to detection and classification problems are presented in Sections 3.5 and 3.6 respectively.

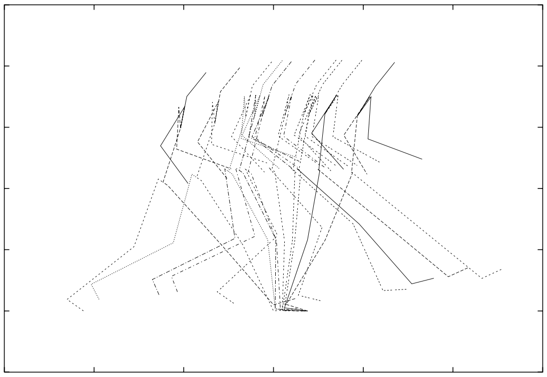


Figure 3.2: Walking sequence of a synthetic subject.

3.1 Experimental Datasets

Three different datasets were employed. Initial feasibility studies were done on a synthetic dataset. Later, manually labelled dataset was used. Both of these datasets were used only for the classification problem. For the biological motion detection problem, a precisely labelled dataset was used. The following sub-section gives a brief description of each of the datasets.

3.1.1 Synthetic Data

Ten walking sequences of four synthetic figures (refer to Appendix A) were generated. Figure 3.2 shows such a sequence. Each sequence consisted of one gait cycle and a gait cycle is made up of 20 frames. As all synthetic figures start their walking cycle with the same phase, no phase synchronisation was needed. However, data-points were normalised so that they lie inside a unit hypercube in an n-dimensional space, n being the dimensionality of the input. Of the ten sequences per synthetic subject, five were used for generation of training data-points and the rest were used for generation of testing data-points. Synthetic subjects were tested for leg joints only as the arm motion was not properly modelled. This dataset was employed only for classification (Laxmi, Carter, and Damper 2002b).

3.1.2 Manually Labelled Data (MaLD)

The initial studies were conducted with video recorded image sequences of four subjects. There are 4 image sequences for each subject, 2 walking left-to-right and 2 walking right-to-left. Only right-to-left walking sequences were used. Segmentation of the gait cycle was done manually. Each segmented sequence consists of a 3 step cycle or $1\frac{1}{2}$ gait cycles. These were manually labelled. Only seven joints – shoulder, elbow, wrist, hip, knee, ankle

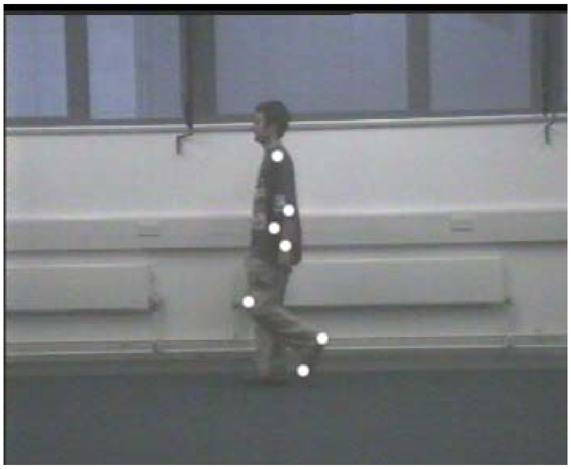


Figure 3.3: Manual labelling of an image sequence.

and toe – were labelled. As normal human gait is symmetrical, only one leg and one arm were labelled. Another reason for labelling limbs on only one side was the difficulty in labelling due to the self-occluding nature of walking in humans. This dataset was employed for classification problem only. Figure 3.3 shows the manual labels for a frame in the image sequence.

All the frames in all the sequences were labelled manually. Seven joints – shoulder, elbow, wrist, hip, knee, heel and foot – were labelled for each frame. The hip joint was more or less an approximate position as it was obscured through clothing. Poor quality of video, coupled with the self-occlusion characteristic of human walk, made the process prone to noise. To keep the process unbiased, no information of previously labelled frame was displayed. To maintain labelling consistency, all frames were labelled in one go.

Only two walking sequences, each consisting of 3 to 4 step cycles were available for labelling. Start and end frames for each subject were synchronised so that all subjects enter the start frame with same walking phase. To make the number of frames equal in each walking sequence, each sequence was re-sampled. Re-sampling was done by MATLAB function *resample*. A fully connected back propagation network was employed. Of the two available sequences, one was used for training and another for testing. Training and testing were done by considering: a) all the seven joints – shoulder, elbow, wrist, hip, knee, ankle, toe, and b) only leg joints (hip, knee, ankle, toe). The results of these experiments are presented in Laxmi, Carter, and Damper (2002a).

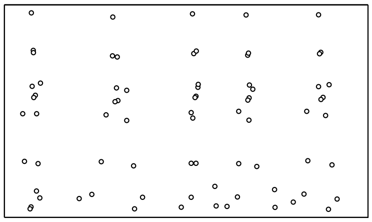
3.1.3 Infrared-marker Labelled Data (GTRI Data)

The GTRI dataset, courtesy Georgia Tech Research Institute, USA, contains accurate 3D labelling with the help of infra-red markers. The dataset consists of labelled sequences of 21 subjects. For each subject, there are four sequences a) with no shoes, b) with no shoes and a five pound shoulder backpack, c) with street shoes and d) with street shoes and five pound shoulder backpack. There is only one sequence for each walking mode, i.e. with or without shoes or shoulder backpack. In all there are 84 sequences. This dataset has been employed in both classification and detection studies.

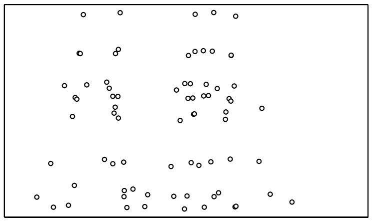
3.2 MLD Categories for Detection

For biological motion detection, 12 categories of MLD sequences were considered. From the perception studies as discussed in Chapter 2, MLDs of a walking or running human are perceived as 'biological motion', whereas a random configuration of dots is not. To determine if a machine detector is capable of learning the presence/absence of biological motion in an image sequence, the machine is presented with positive and negative examples of such motion. In all, 12 different categories of image sequences were considered. For every category, a snapshot of some frames from one of the corresponding image sequences is also shown. These categories are discussed in detail now:

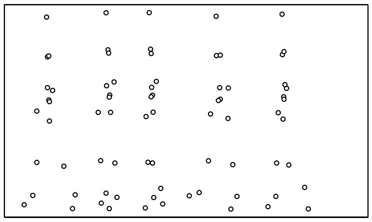
NOR This image sequence is the fronto-parallel view of a walking person. All the motion-perception studies discussed in Chapter 2 used sequences of this type. The image sequence was labelled positive as, in all the perception studies, discussed in Chapter 2, the participants recognised such displays as a human figure walking.



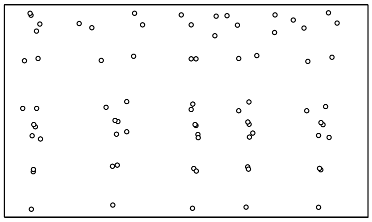
DIR In all image sequences, a person walks from right to left. This category was generated by a reflection in the vertical plane. As a result, the person walks from left to right. It was added to make biological motion detection independent of the direction of motion. The image sequence was labelled positive as biological motion detection is independent of direction of motion (Johansson 1976).



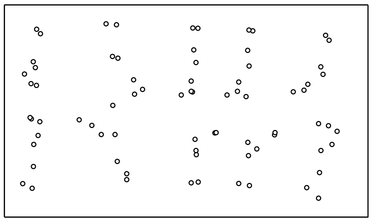
WBK This was obtained from fronto-parallel view by reversing the sequence. Hence, the last frame becomes the first frame and so on. This category preserves human shape, the human motion, although natural, is not commonly observed. The image sequence was labelled positive as it is reported to be identified as natural human motion (Johansson 1976).



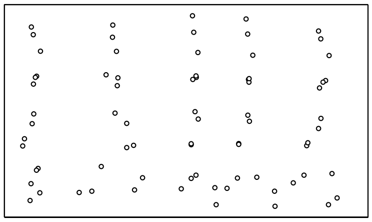
INV This is essentially a reflection of the fronto-parallel view with respect to the ground plane. Although inverted displays have relative spatial and motion relationships similar to normal ones, most of the structure based perception theories cannot explain the negative response of human perceptual system. Inverted display may, therefore, be a good negative example. The image sequence was labelled negative (see Section 2.4).



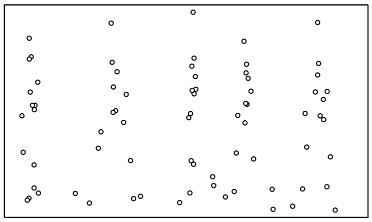
TOP This is the view obtained from the top. The image sequence is labelled negative. As there is no reported literature studies on this, this was assumed to be negative, as human shape and motion are different from the fronto-parallel view.



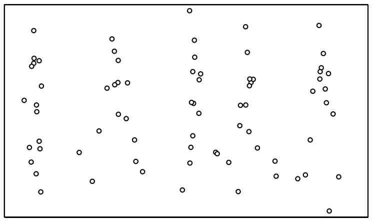
OBQ This is similar to the top view but viewing is done at an angle. Although not reported in the literature, both TOP and OBQ categories are included, as views from different angles are more likely to be available as input to a computer vision system. The image sequence was labelled negative as this is closely related to TOP category.



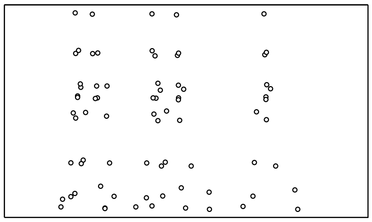
SPT A small perturbation was added to all the joint positions. The amount of perturbation retains the human figure, although inter-joint spatial relations get perturbed. Human perceptual system is capable of tolerating variations in spatial relations. A slight perturbation should have no adverse effect on biological motion perception. The image sequence was labelled positive as human shape is still visible and human motion is only slightly affected.



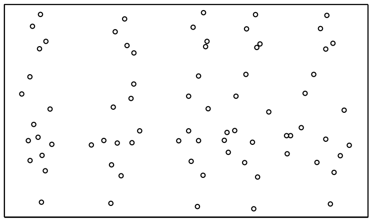
LPT A large perturbation was added to all the joint positions. No human shape is retained. There is only a certain range in which perceptual system can tolerate variations. A large perturbation is, therefore, likely to be a negative instance. SPT and LPT categories were included to determine the effect of degree of perturbation. The image sequence was labelled negative as both the human shape and motion are disturbed massively.



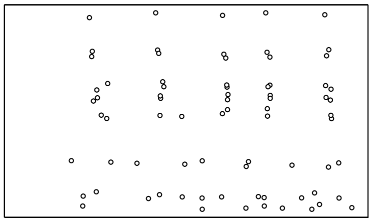
PER The frames in a sequence were selected in a random order. Thus the view retains the human shape but the natural order of walking is missing. It is expected that the machine learning models will be able to learn temporal associations better by the inclusion of this category. The image sequence was labelled negative as the sequence of human motion is permuted and not in a natural order.



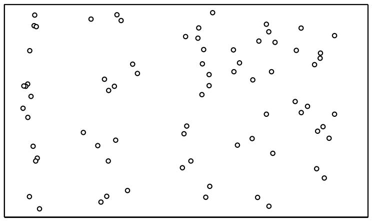
SSR This was derived by spatially scrambling the limbs – arms and legs. Only the positions of the limbs are randomised, relative motion is not disturbed at all. The image sequence was labelled negative (Pinto and Shiffrar 1999).



PSR This was also derived from the normal sequence. Relative phase relations of different joints are disturbed. Phase of each joint in the first frame corresponds to a randomly selected frame. For example, the knee joint may begin with a phase corresponding to frame 27, and elbow may start with frame 3 and so on. Consistency of human form and motion play a major role in motion perception. A random change in inter-limb spatial and/or phase relations is likely to affect the perceptual system adversely. Hence, SSR and PSR categories were included. The image sequence was labelled negative (Ahlström, Blake, and Ahloström 1997).



RAN This represents a random configuration of dots. It was obtained by putting random dots in the bounding rectangle of each frame of the fronto-parallel (NOR) sequence. As this configuration has neither shape nor motion consistent with humans, it is a good negative exemplar. The image sequence was labelled negative as it contains no aspect of human shape and human motion.



Category	Label
NOR	+
DIR	+
WBK	+
INV	−
TOP	−
OBQ	−
SPT	+
LPT	−
PER	−
SSR	−
PSR	−
RAN	−

TABLE 3.1: Labelling of various MLD categories.

Initially a label (positive or negative) for each category was determined from the literature or personal experience. Table 3.1 shows this labelling. A positive label means that the category is identified as natural human motion. A negative label is indicative of non-human motion.

3.3 MLD Presentation Modes

For both human and the machine, the input MLD data were presented in two different modes as discussed below.

Absolute: Translatory motion is retained. On a screen, the configuration of the dots moved from one end to another. The image sequence contained both absolute and relative motions of the dots.

Spot: No translatory motion is retained. The centroid of every frame was translated to the origin. On a screen, the configuration of dots appeared to be moving on a treadmill. The image sequence consisted of the relative motion only.

3.4 Spatio-Temporal Integration for Machine Input

For any frame in an MLD sequence, the joint positions describe the body configuration at time t, where t is the time index or the position of the frame, in the sequence. As for the detection problem, the joints are scanned in a top-down, left-right manner, the correspondence of joints from frame to frame is not preserved. In fact, the term joints does not make any sense in a random configuration of dots. Hence the term frame parameters will be more appropriate. Number of parameters per frame is twice the number of dots as each dot is represented by two coordinates – x and y. Thus for a given frame, its parameters will be represented by a $2M$ tuple $\langle x_1, y_1, \ldots, x_M, y_M \rangle$, where M is the number of dots per frame and (x_i, y_i) is the position of i-th dot. For this discussion, we will assume that the sequence consists of F frames, each frame containing M dots and hence $2M$ parameters.

The body configuration vector $\vec{c}(t)$ corresponding to a single frame with index t is given by

$$\vec{c}(t) \quad = \quad (x_1(t), y_1(t), \ldots, x_M(t), y_M(t))\,, \quad t = 1, \ldots, F$$

where $x_i(t),\ y_i(t)$ are the coordinates of the i-th dot in the frame.

The vector $\vec{c}(t)$ has temporal information of one frame only and dimensionality of $2M$, as each dot is represented by $\langle x, y \rangle$ coordinates. Integration of temporal information of N frames can be done by constructing a tuple $\langle \vec{c}(t),\ \vec{c}(t+1),\ \ldots,\ \vec{c}(t+N-1) \rangle$ by concatenating the configuration vectors of N consecutive frames. We will refer to this concatenated vector as a data-point in this work. For a given value of N, all possible data-points, each with temporal information spanning N frames, are generated. So, for $N = 5$, the first data-point was generated by concatenating configuration vectors for frames 1 to 5, the second data-point by concatenating these vectors for frames 2 to 6 and so on. In other words, each data-point is a snapshot of a fraction of a gait cycle. This provides temporal information to the classifier, and is a classical way to handle data sequences (Sejnowski and Rosenberg 1987). Thus each walking sequence is presented to a classifier as a collection of data-points or overlapping snap-shots of the sequences, where each snap-shot lasts for a given temporal span.

For a sequence consisting of F frames and M dots per frame, number of data-points with grouping of N frames is $F - N + 1$ and dimensionality of each data-point is $2MN$. With an increasing value of N, the number of data-points per sequence decreases but the temporal information available per data-point increases. Figure 3.4 illustrates the process of generating data-points from a given image sequence.

3.5 Generation of Data-points for Detection

For every sequence in the dataset, 12 sequences, one for each category, are generated. Data-points are generated for each sequence by combining frame parameters of N frames as discussed in previous sections. Frame parameters are dot positions scanned in a raster manner (top-down, left-right) as shown in Figure 3.5. The number against each dot specifies its position in the scan order.

For practical considerations, the data-points were normalised to unit hypercube and then were subjected to mean removal for the neural network detectors. The normalisation

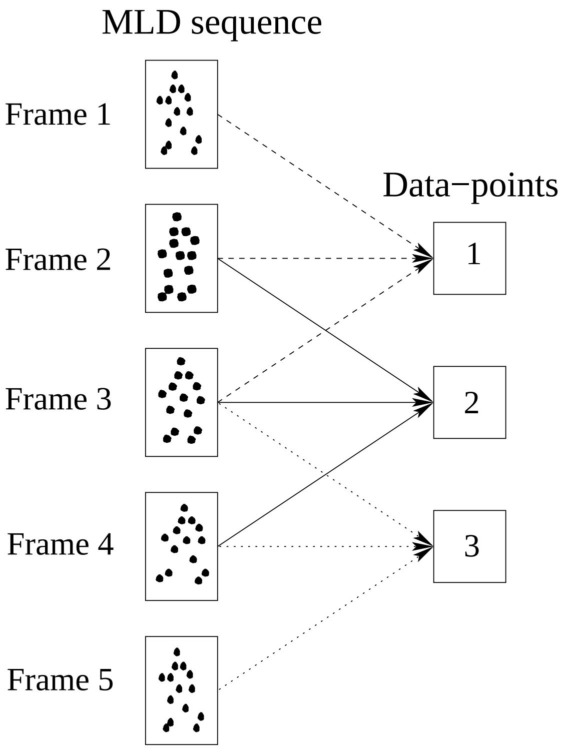


Figure 3.4: Temporal integration of frames across an MLD sequence to generate data-points.

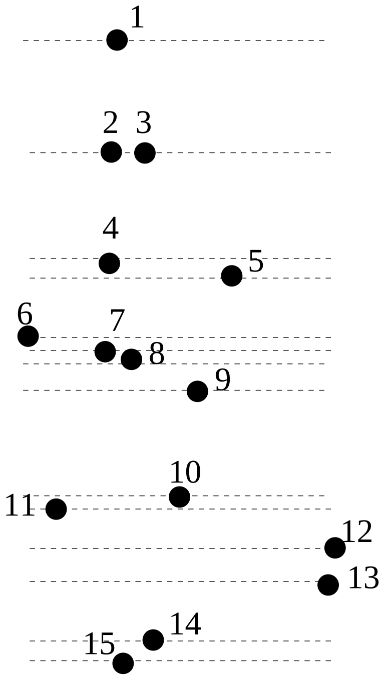


Figure 3.5: Dots in an MLD sequence are scanned in a top-down, left-right ('raster') manner.

was done to avoid network saturation and mean removal helped faster convergence.

3.6 Generation of Data-points for Classification

For the classification problem, data-points are generated in a similar way to that for the detection. The only difference is that only the NOR category is used and the dots are not scanned in raster-scan order. In fact, a frame-to-frame dot correspondence is maintained. As all MLD sequences depict a human figure in motion, the term joint is more appropriate and will be used instead of dot in the following discussion. In other words, joints are always considered in the same order. As no other variations should contribute to the classification, walking sequences of all subjects are manually synchronised. This implies that all the subjects enter the first frame of the sequence at the same phase. This process can be easily automated by computing angles of rotation for the hip and determining minima (maxima) to mark the beginning and end of a gait cycle. Once gait cycles for all the walking sequences were identified, they are translated such that the respective x coordinates of the hip joint are aligned and the ground specification is same for all the sequences. Then data-points for different values of N are generated. Training and test data-points are generated from different walking sequences. A time tag is added to the configuration vector of each frame. As a result, the dimensionality of each data-point becomes $(2M+1)N$. The addition of the time tag rendered the model dependent on the natural order of walking. Thus the modified configuration vector is given by

$$\vec{c}(t) \quad = \quad (t, x_1(t), y_1(t), x_2(t), y_2(t), \ldots, x_M(t), y_M(t))\,, \quad t = 1, \ldots, F$$

In the earlier experiments (Laxmi, Carter, and Damper 2002a), the walking sequences were re-sampled, using the MATLAB function *resample*, so as to keep the same number of frames per gait cycle for each subject in the dataset. However, as each walking sequence contains about 2–3 gait cycles, re-sampling introduces oscillations at the end positions. In the ensuing experiments, as discussed in Chapter 7, no re-sampling was performed and the number of data-points generated from different sequences were different. Although each data-point has the same temporal span in terms of the number of frames and hence time, the temporal information with respect to gait cycle is not identical. Therefore, the time tag indicates the fraction of gait cycle, wherein a complete gait cycle has a value of 1. The body configuration vector, in this case, can be defined as follows.

$$\vec{c}(t) \quad = \quad (t/F, x_1(t), y_1(t), x_2(t), y_2(t), \ldots, x_M(t), y_M(t))\,, \quad t = 1, \ldots, F$$

For both training and testing of the machine model, data-points were normalised to

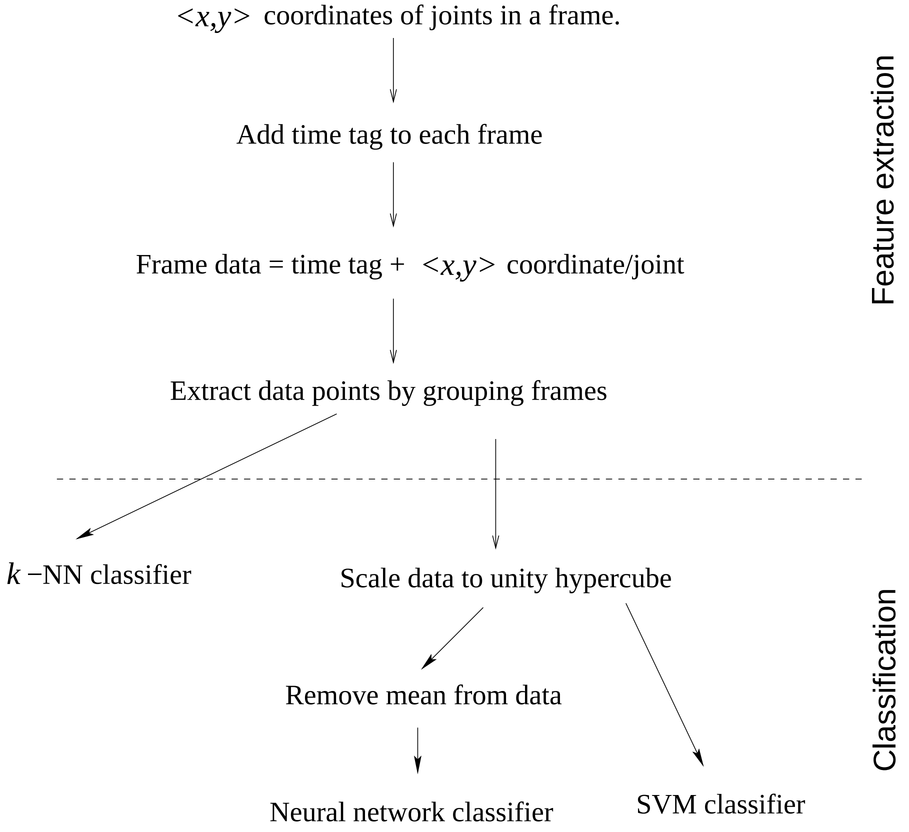

<x,y> coordinates of joints in a frame.
Add time tag to each frame
Frame data = time tag + <x,y> coordinate/joint
Extract data points by grouping frames
Feature extraction
k −NN classifier
Scale data to unity hypercube
Remove mean from data
Neural network classifier
SVM classifier
Classification

FIGURE 3.6: Steps involved in the classification process of a labelled image sequence.

the unit hypercube. In case of neural network classifiers, they were subjected to mean removal. Then, they were applied to the neural network classifier (see Figure 3.6). For the classification problem, it is assumed that the scene contains only one person walking fronto-parallel to the camera. Availability of joint coordinates was assured so as to avoid occlusion and missing data related situations.

In the next chapter, we test if the machine can learn to detect human motion or not. For the purpose of this initial investigative study, we assign a category for human or non-human motion as discussed earlier.

Chapter 4

Can A Machine Detect Human Motion?

One of the main objectives of our work is to ascertain if a machine is capable of discriminating a human motion sequence from a non-human one as humans do. Machine modelling of any aspect of human behaviour is a difficult and tricky problem. One of the major challenges is how to feed information about human experience and prior learning. Unlike humans, the machine is working in a closed world. As main aim of the machine learning is to identify aspects of motion responsible for human perception, isolation of a proper subset of prior learning is a difficult task in itself. In this work, we are trying to determine if machine learning can lead to identification of factor(s) triggering detection of biological motion by humans. At present, this detection will be restricted to the categories discussed earlier. Although human perceptual system needs no training for detection, the machine, nevertheless, is fed with raw data, i.e. data with no pre-processing, as this data is sufficient to invoke vivid perception in humans.

Chapter 2 summarises what constitutes a human/non-human motion from human perspective. Assignment of labels to the various categories has already been discussed in Section 3.2. A positive label means that the category is perceived as human motion, whereas a negative label implies a non-human motion. Data-points for both training and testing sets, discussed in Section 4.1, were generated as described in previous chapter. Section 4.2 presents results obtained with k-NN and ANN detectors followed by a discussion of these in Section 4.3. Different implementations of SVMs (Gunn 1997; Joachims 1999; Rifkin 2000) were not able to handle large amount of data. Therefore, only ANN and k-NN were used for the detection problem.

4.1 Training and Testing Sets

Only the GTRI dataset as discussed in Section 3.1.3 was used. Twelve categorised sequences (one sequence per category as discussed in Section 3.2) were generated for each sequence. For different values of N, data-points were generated from these $12 \times 84 = 1008$ sequences. Of 21 subjects, only categorised sequences of 9 subjects were used for generating data-points for training, the rest were used for generating testing data-points. Training and test datasets were mutually exclusive.

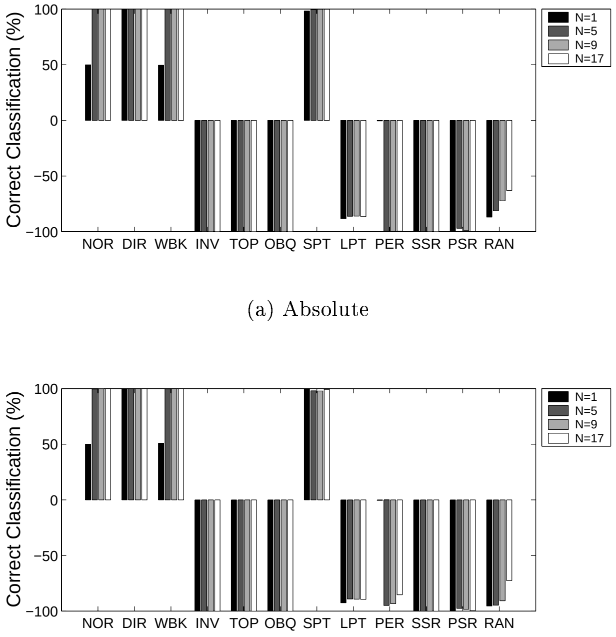
4.2 Preliminary Results

For a given detector, the capability of detecting biological motion was investigated in two different modes – (a) absolute and (b) on the spot. In the first case, the translatory motion was retained as such and no data modification was done. In the second case, the centroid of each frame was moved to origin. Thus frame-to-frame translatory motion was absent. The moving figure now appeared to be walking on a treadmill. This was added as humans can recognise the motion even in such circumstances. Translatory component can also bias the detection, especially in k-NN which uses a Euclidean distance based metric.

4.2.1 k-Nearest Neighbour Detector

In the k-NN detector, for each data-point, its k nearest data-points were determined. One of these data-points was randomly selected and designated as the nearest neighbour. If the labels of the data-point and its nearest neighbour match, it was assumed to be correctly classified, otherwise not. The process was repeated for different values of N, frames to be combined for temporal integration in data-point generation (as described in Section 3.4). Figure 4.1 displays the number of correctly classified data-points for each category having different values of N. For categories labelled positive, the classification accuracy is represented in the region 0 to 100; for others the -100 to 0 region is used. Any value near zero means that the machine model is not able to learn the respective category.

For lower values of N, especially $N = 1$, we observe that k-NN detector confuses the samples from NOR, WBK and PER categories, indicated by near 50% response for NOR and WBK and none for PER. This is an expected result, as when only one frame is considered, there is no difference between these categories. Another interesting feature of k-NN detector is that for higher values of N (15 in this case), RAN category is perceived as less negative. Figure 4.2 indicates how well the samples of a given category match with other categories in terms of nearest neighbour criterion. A bright square means



(a) Absolute

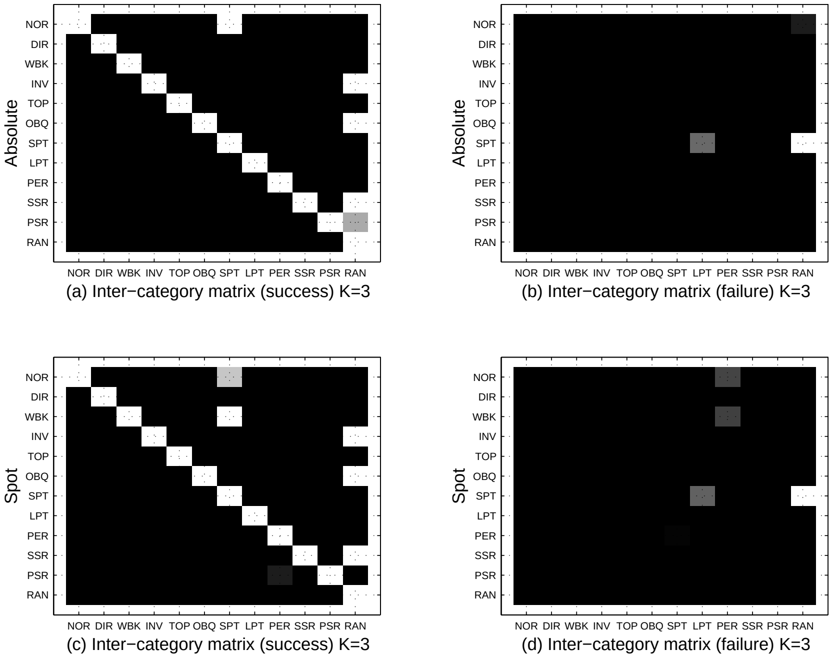
(b) Spot

FIGURE 4.1: Inter-category performance of k-NN as a biological motion detector for (a) absolute and (b) spot modes in the preliminary study.

near 100% match and a dark square means no match. The matrix reveals that, except for RAN category, every sample of a category is nearest to another sample of the same category. RAN category interacts with INV, TOP, SSR and PSR categories. Any assignment which labels any of these categories positive is likely to affect the classification accuracy adversely. The performance of k-NN detector in absolute and spot modes is the same except for PER category.

4.2.2 Artificial Neural Network Detector

A neural architecture with two hidden units and a single output unit was trained as per the labelling in Table 3.1. For each value of N, five randomly initialised, neural architectures were trained and tested. For each program run, the number of correctly classified data-points for each category and different values of N were computed. Figure 4.3 shows the average category-wise performance on these five runs. For the ANN detector, the average performance in the spot mode is worse than that in the absolute mode. This is because the network did not always converge, especially at the higher values of N. Whenever the convergence took place, each category was detected with $\geq 90\%$ accuracy. The network was assumed to be converged when the classification accuracy remained $\geq 95\%$ for 5 successive epochs.



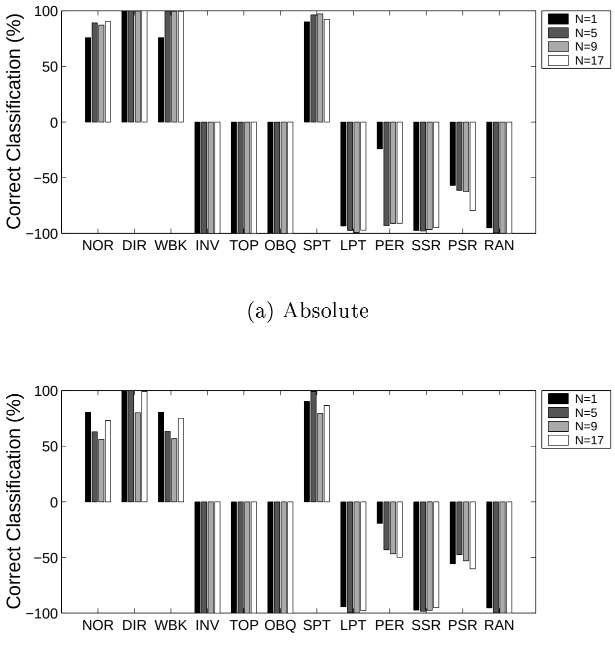
(a) Inter−category matrix (success) K=3

(b) Inter−category matrix (failure) K=3

(c) Inter−category matrix (success) K=3

(d) Inter−category matrix (failure) K=3

FIGURE 4.2: The figure illustrates inter-category confusion matrix for k-NN detector. The intensity of a square is an indication of the number of data-points of a category along the row classified as a category along the column. A bright square means near 100% match and a dark square means no match.



(a) Absolute

(b) Spot

FIGURE 4.3: Average category-wise performance of ANN, 2 hidden units, as a biological motion detector in (a) absolute and (b) spot modes in the preliminary study. Five, randomly initialised, neural architectures were used to determine the average.

4.3 Discussion

An overall high accuracy within each category for absolute mode suggests that ANN might be a suitable model for biological motion detection, at-least for the set of 12 categories considered in this work. A simple neural architecture as used in this study cannot, however, compare with the capability of a human brain. The aim of this study was to show that the data has sufficient information for even a simpler machine to learn the biological motion labelling of a limited set of categories. It needs to be seen if the features learnt by the machine lead to any understanding of human perception or not. To test if the machine performance compares with the human one or not, we need to ascertain human responses for the same MLD data. Reported literature studies cannot be used as a basis for determining human responses mainly because an MLD walker in our case has more joints. Also no account of TOP and OBQ categories are available. Hence an experiment to obtain the human responses on these MLD categories was conducted. The experimental set-up and its outcomes are discussed in the following chapter.

Chapter 5

Biological Motion Detection: Human Perception

In the earlier reported results with ANN detectors (Laxmi, Carter, and Damper 2002b), fronto-parallel view (Johansson 1973) and upside-down view (Sumi 1984; Pinto and Shiffrar 1999; Pavlova and Sokolov 2000) were labelled positive and negative respectively. In addition to upside-down, the top view was also labelled as a negative instance. Although the studies on biological motion perception give a fairly good account of capabilities of human perception, there is still an ambiguity regarding inverted displays. Sumi (1984) reported that most of the participants failed to report a backward upside-down biological motion when Johansson displays were played upside down. Instead an upright walker moving in a strange way was perceived. Pinto and Shiffrar (1999) has reported the results in the presence of a mask. According to Pavlova and Sokolov (2000), upside-down or inverted displays are multi-stable. Thus it is not clear if the humans consider inverted displays as natural human motion or not in the context of "biological motion". Also there is no account available on how human beings respond to views from the top. We conducted an experiment to determine what constitutes a "natural human motion" from the perspective of humans. The experiment also provides a fair comparison between human and machine performance as the identical dataset has been used for both. However, it needs to be said that any comparison between respective performance of machine and human can be considered fair only within a closed world. Even then, in a stricter sense, comparison is not exactly fair as humans do have prior experience.

This chapter is structured as follows. Section 5.1 describes the experimental set-up in detail. Responses of 93 participants and an inference based analysis of these results is presented in Section 5.2. In Section 5.3, labelling of various categories in the context of biological motion is reviewed. An overall summary of results is presented in Section 5.4.

5.1 Experimental Set-up

For each of four sequences of 21 subjects in the GTRI data (see Section 3.1.3), all 12 categories, as discussed in Section 3.2, were generated. As a result, the total number of sequences were 1008. Each sequence lasted 150 frames or 2.55 s. Each frame was displayed for 17 ms or inter-frame time as recorded in the GTRI data. In other words, video display was run at the same rate at which the data was filmed. The main aim was to determine what constituted human motion from a participant's point of view without revealing much about the experiment. Participants were shown a given sequence in one of the four different modes as explained below.

(a) Absolute: Image sequences were shown without any pre-processing. The translatory motion was retained. All 15 joints – head, shoulders, elbows, wrists, hips, knees, ankles and toes – were displayed.

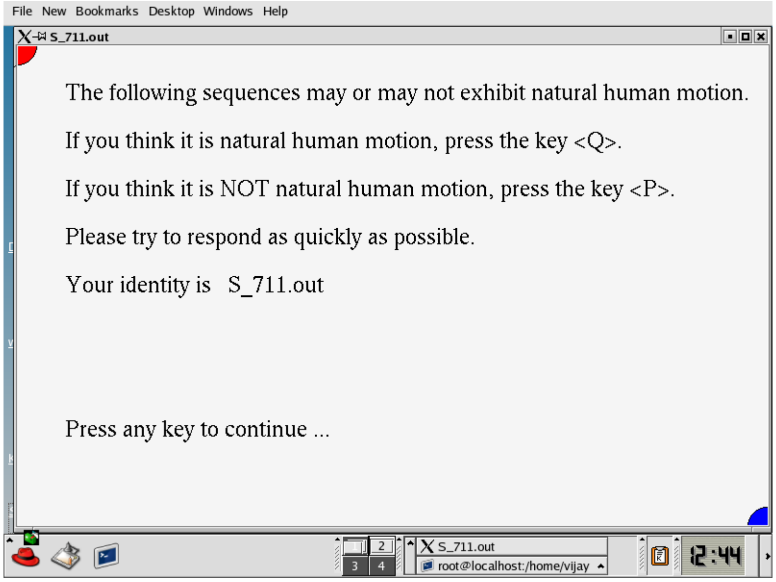
(b) Spot: Translatory motion was removed by aligning centroid of each frame to origin. This mode was used as k-NN classifier exhibited slightly better performance in this mode.

(c) Partial: Only half the joints (in fact the exact number is 7) were displayed. These joints were selected randomly.

(d) Centroid: Only one point representing the centroid of the body configuration was displayed.

As it was intended that the human responses should be obtained without revealing the actual categories being shown, it was necessary to introduce some sort of distraction. One possible solution is not to show the same number of dots each time. Hence only half the joints (randomly selected) were shown. As another extreme, only one dot per frame was shown. This dot corresponded to the centroid.

Within each mode, twelve instances of each category in a randomly selected order were presented. In fact, three instances of each kind of walk, i.e. with or without footwear, with or without shoulder backpack, were shown. This was to ensure that variations in walking due to these factors remained consistent for all the participants. In all, the participants viewed 576 sequences (4 modes, 12 categories/mode and 12 sequences/category). These were displayed in six sessions. Within each session, sequences were shown one after another with an inter-sequence blanking period of 300 ms. A participant could take a short break after each session. This was to prevent fatigue. For each sequence, the participants were required to press a key in response if the motion being displayed was natural human motion. If no key was pressed for the entire duration of a sequence, a time-out was recorded. For every participant, response time interval in terms of number of frames and the response were recorded for each displayed sequence. Figure 5.1 shows snapshots of an experimental run.


File New Bookmarks Desktop Windows Help
S_711.out
2
3
4
S_711.out
root@localhost:/home/vijay
12:44

The following sequences may or may not exhibit natural human motion.

If you think it is natural human motion, press the key <Q>.

If you think it is NOT natural human motion, press the key <P>.

Please try to respond as quickly as possible.

Your identity is S_711.out

Press any key to continue ...

FIGURE 5.1: A snapshot of the experiment to collect human responses.

Category	Number of participants identifying the category
NOR	86
DIR	9
WBK	51
INV	41
TOP	38
OBQ	1
SPT	9
PER	6
RAN	18
Spot	6
Partial	13
Other human motion	68
Non-human motion	41

TABLE 5.1: Number of participants identifying a given category are tabulated against the respective row. Categories LPT, SSR and PSR are not recognised as such.

To prevent any biasing, participants were not given any prior knowledge of the experimental set-up and were requested not to discuss it with any other participant. Of 93 participants, 37 were students or research staff from the University of Southampton, UK, and the rest were the undergraduate engineering students from Malaviya National Institute of Technology, India. After the experiment, each participant was asked to write down the perceived categories of the sequences on an information sheet (Appendix F). Table 5.1 summarises these responses. Against each category, the number of persons identifying the category is displayed. A detailed account of participant-wise responses is presented in Appendix G. Participants also filled in a questionnaire (Appendix E). The purpose of this was to assess the technical background of the participant.

A self-written code in Python (van Rossum 1999) was used to display the sequences and record the response with relevant information, such as actual sequences shown, the category and mode of each sequence, time to respond, set of joints displayed in partial mode.

5.2 Analysis of Human Responses

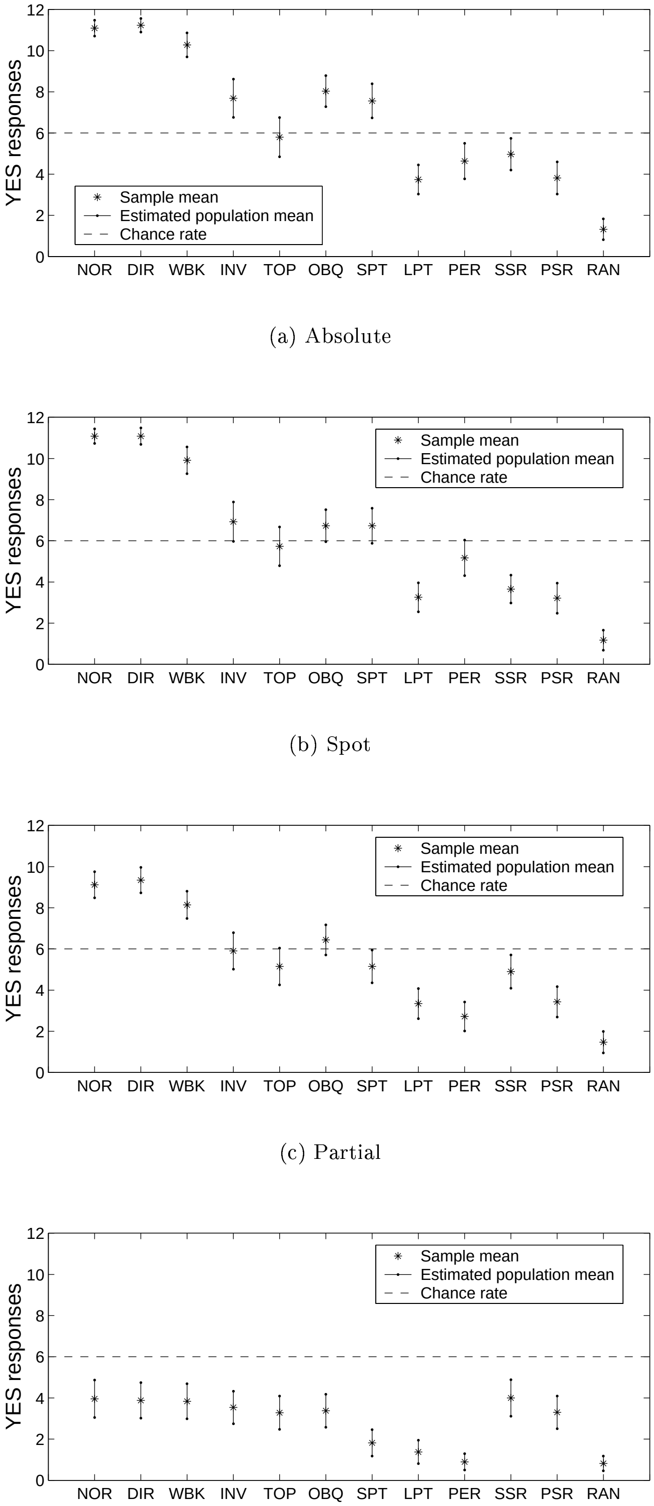
For the purpose of analysis, a time out, i.e. participant failing to respond, was considered as a negative response. As each category was shown 12 number of times, a value of 6 positive responses was equal to the chance rate as the only permissible responses were 'yes' or 'no'. Figures 5.2(a)–5.2(d) summarise the number of 'yes' keys pressed by all the participants for all four modes. Here the sample mean is shown as an asterisk and the vertical bar about this mean is the range of estimated population mean. Population mean estimates were calculated at a significance level of 0.05 (95% confidence interval) from sample means by a two-tailed t-test (refer Appendix D).

As seen from the plot, population mean estimates are always below the chance rate for the centroid mode irrespective of the category. This indicates that humans cannot make confident judgement about human motion on the basis of information of only one moving dot. Building a machine model for the partial mode has limitation of not enough information. For 15 dots, as in the present case, 7 dots can be chosen in $\binom{15}{7} = 6435$ ways. However, as 93 participants have been shown only $93 \times 12 = 1116$ cases, there is not enough information. Therefore, for the subsequent work and results only absolute and spot modes are relevant. Hence partial and centroid modes will not be discussed any more. These have been presented only for the sake of completeness in the context of the conducted experiment.

As some of the participants had sufficiently good knowledge about computer vision and image processing techniques and probably MLDs, these were classified as 'experts'. Figures 5.3(a)–5.3(d) summarise the mean responses of the experts and the rest of the participants for the absolute and spot modes. Mean response time for various categories is shown in Figures 5.4(a) and 5.4(b). The response time is in number of frames and actual time can be obtained by multiplying with inter-frame time or 17 ms. These response times, however, are not used for any subsequent analysis.

5.3 Relabelling of Biological Motion Categories

From the machine learning point of view, it needs to be inferred whether a given category is perceived as biological motion or not. A simple one-tailed t-test (Appendix D) is done



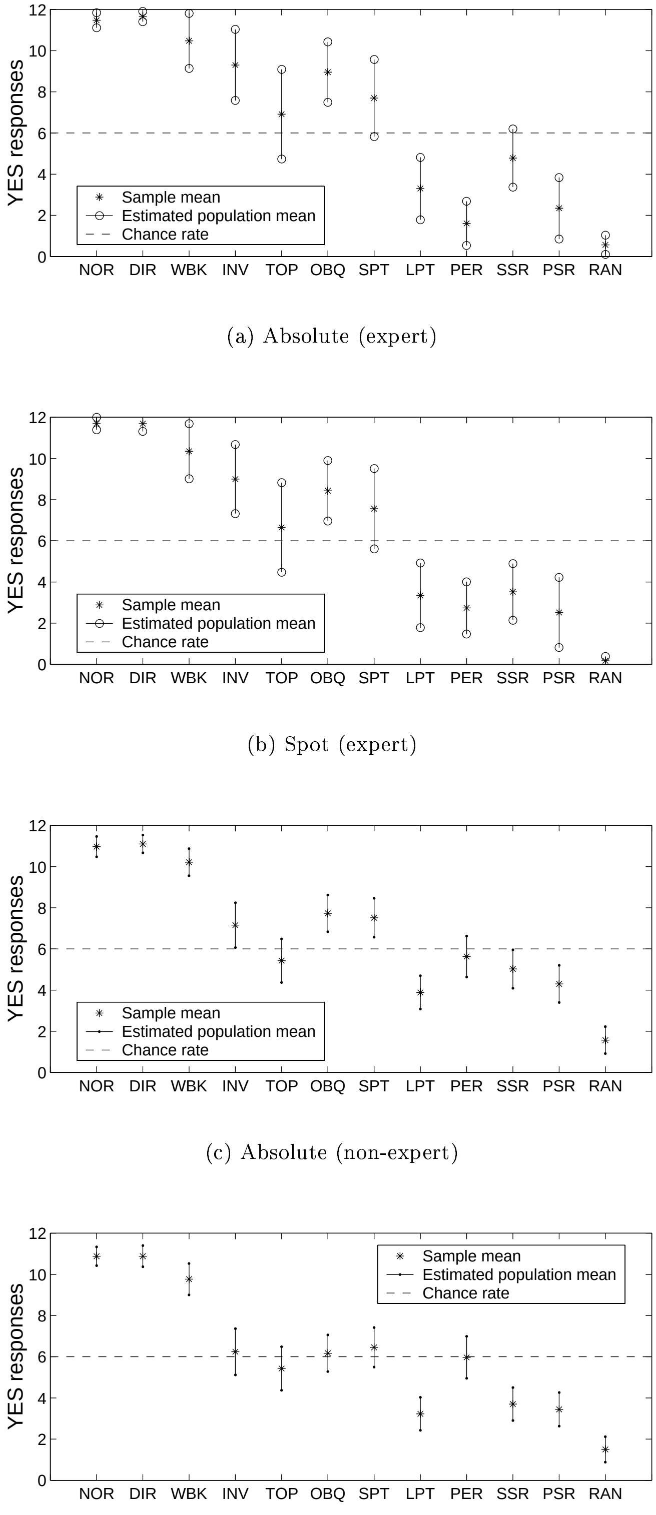
(a) Absolute

(b) Spot

(c) Partial

(d) Centroid

FIGURE 5.2: Human responses for (a) absolute, (b) spot, (c) partial and (d) centroid modes. Asterisk denotes sample mean; vertical bar about this mean is inferred population mean at a significance level of 0.05 (95% confidence interval).



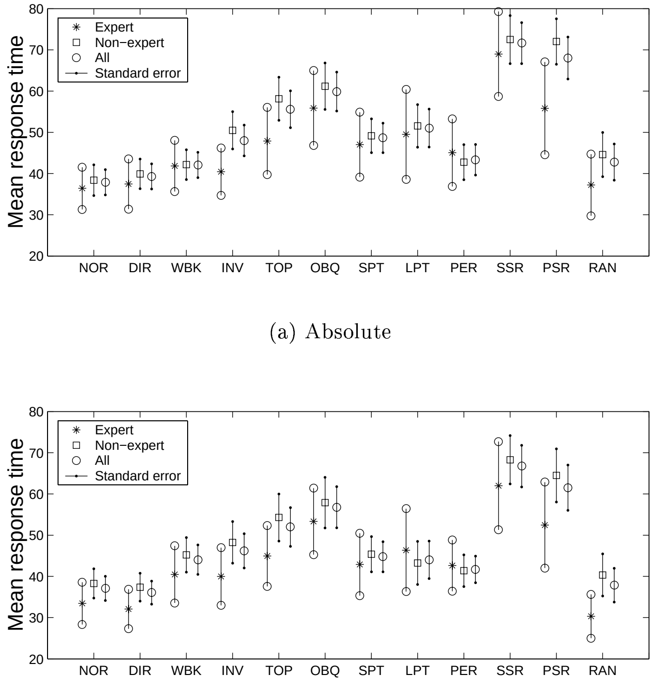
(a) Absolute (expert)

(b) Spot (expert)

(c) Absolute (non-expert)

(d) Spot (non-expert)

Figure 5.3: Responses of experts for (a) absolute and (b) spot modes. The asterisk is the sample mean and the vertical bar about this point is inferred population mean at 0.05 significance level. Responses of non-expert participants for absolute and spot modes are shown in (c) and (d) respectively.



(a) Absolute

(b) Spot

FIGURE 5.4: Mean response times of participants for (a) absolute and (b) spot modes. The asterisk is the sample mean and the vertical bar about this point is inferred population mean at 0.05 significance level.

Category	Null hypothesis: $\mu = 6$. Alternative hypothesis:					
	$\mu < 6$			$\mu > 6$		
	Expert	Non-Expert	Overall	Expert	Non-Expert	Overall
NOR	-	-	-	Reject	Reject	Reject
DIR	-	-	-	Reject	Reject	Reject
WBK	-	-	-	Reject	Reject	Reject
INV	-	-	-	Reject	Reject	Reject
TOP	-	-	-	-	-	-
OBQ	-	-	-	Reject	Reject	Reject
SPT	-	-	-	Reject	Reject	Reject
LPT	Reject	Reject	Reject	-	-	-
PER	Reject	-	Reject	-	-	-
SSR	Reject	Reject	Reject	-	-	-
PSR	Reject	Reject	Reject	-	-	-
RAN	Reject	Reject	Reject	-	-	-

TABLE 5.2: Summary of the results for testing if the population mean lies above or below the chance rate at 0.05% level of significance in absolute mode. 'Reject' means that the null hypothesis is rejected with 95% confidence; '-' means that the null hypothesis is not rejected.

Category	Null hypothesis: $\mu = 6$. Alternative hypothesis:					
	$\mu < 6$			$\mu > 6$		
	Expert	Non-Expert	Overall	Expert	Non-Expert	Overall
NOR	-	-	-	Reject	Reject	Reject
DIR	-	-	-	Reject	Reject	Reject
WBK	-	-	-	Reject	Reject	Reject
INV	-	-	-	Reject	-	Reject
TOP	-	-	-	-	-	-
OBQ	-	-	-	Reject	-	Reject
SPT	-	-	-	-	Reject	Reject
LPT	Reject	Reject	Reject	-	-	-
PER	Reject	-	Reject	-	-	-
SSR	Reject	Reject	Reject	-	-	-
PSR	Reject	Reject	Reject	-	-	-
RAN	Reject	Reject	Reject	-	-	-

TABLE 5.3: Summary of the results for testing if the population mean lies above or below the chance rate at 0.05% level of significance in spot mode. 'Reject' means that the null hypothesis is rejected with 95% confidence; '-' means that the null hypothesis is not rejected.

to determine if the population mean is above or below the chance rate (6 in the present case). If for a given category, the perceived population mean is above (below) the chance rate, most of the participants have detected (failed to detect) biological motion and the category represents a positive (negative) instance. Tables 5.2 and 5.3 represent the results of hypothesis testing for absolute and spot modes respectively. The results have been obtained for expert, non-expert and all participants. Any 'Reject' entry in these tables means that the population mean is not equal to the chance rate. In this case, the population mean is either above or below the chance rate. The former implies that the category was detected as biological motion and the latter is indicative of a non-human motion. Hence a directional alternative hypothesis, $\mu > 6$ ($\mu < 6$) was formulated to determine with 95% confidence if the category represents human(non-human) motion. Considering an entry of PER category in Table 5.2, we find that the null hypothesis is rejected by experts but not by non-experts. This implies that the experts consider this category as non-human motion. Although the fact that the null hypothesis is not rejected by non-experts indicates that the mean response is not above chance. This is indicative of the fact that the non-experts do not cateorise PER as human motion, but the level of rejection of PER as positive biological motion is not as high as in the case of experts. Results of these tables can be expressed in terms of positive or negative biological motion for absolute and spot modes as shown in Table 5.4. Once it has been ascertained (with 95% confidence) if the population mean lies above or below the chance rate, a two-tailed t-test (Appendix D) has been used to obtain this biological motion labelling. Any empty entry in this table is indicative of a population mean near the chance rate. Hence the labelling of the category remains inconclusive in such cases.

Category	Absolute			Spot		
	Expert	Non-Expert	Overall	Expert	Non-Expert	Overall
NOR	+	+	+	+	+	+
DIR	+	+	+	+	+	+
WBK	+	+	+	+	+	+
INV	+	+	+	+		+
TOP						
OBQ	+	+	+	+		+
SPT	+	+		+	+	+
LPT	−	−	−	−	−	−
PER	−		−	−		−
SSR	−	−	−	−	−	−
PSR	−	−	−	−	−	−
RAN	−	−	−	−	−	−

Table 5.4: Positive and negative instances of biological motion.

5.4 Discussion on Human Perception

The main findings of the human responses can be summarised as follows.

- As the distribution plots of absolute and spot modes are similar, it is suggestive of the fact that human perceptual system is sensitive to relative rather than absolute motions.

- Response for upside down display is not as negative as expected. In the feasibility studies, it was assumed to be a negative instance. However, the results tend to be indicative of a positive instance. This is significantly different from Sumi (1984), Pinto and Shiffrar (1999) and Pavlova and Sokolov (2000). This may be due to the fact that, unlike Johansson's displays, feet joints are displayed. Or the very fact that participants were asked whether the motion was human or not may have biased the results.

- Top view is not detected as human motion by majority of participants except for the experts. However, on the basis of the available results, it remains inconclusive if the category should be labelled positive or negative. Oblique view fares well as compared to the top one. As for different presentations, a random oblique angle is chosen. For values approaching 90°, oblique view approaches fronto-parallel view. Asymmetry in this view as compared to the top one may invoke a much better perception of human form.

- Human structure, as seen from the side, seems to play a major role in deciding if a given sequence displays biological motion or not. This appears to be true from the observation that oblique view is more detectable than the top one. Another point in favour of this observation is that non-expert participants do not greatly reject the permuted category either in absolute or in spot mode.

- Even for fronto-parallel view (NOR category), the response is not 100% positive as reported by Johansson (1973).
- A wide variation in response indicates that the non-expert participants are less assertive. In spot mode, their mean responses are more or less centred around the chance rate except for NOR, DIR, WBK and RAN categories.
- Responses of experts and non-experts differ. This is also substantiated by analysis of variance (ANOVA), which shows that the responses for experts are different from those of non-expert ones at 0.05 level of significance. A category-wise ANOVA indicates that responses for these two sets of participants vary in
  - INV, PER and PSR categories for absolute mode.
  - NOR, INV, OBQ, PER and RAN categories for spot mode.

The labelling of the different categories, as determined on the basis of the human responses, is significantly different from the one in Table 3.1. It is evident from these findings that the machine models, presented in Chapter 4 need to be re-checked against these results. Chapter 6 discusses these results on machine perception and contrasts it with human perception. As both human and machine responses are evaluated for the same dataset, this provides an unbiased basis for the comparison.

Chapter 6

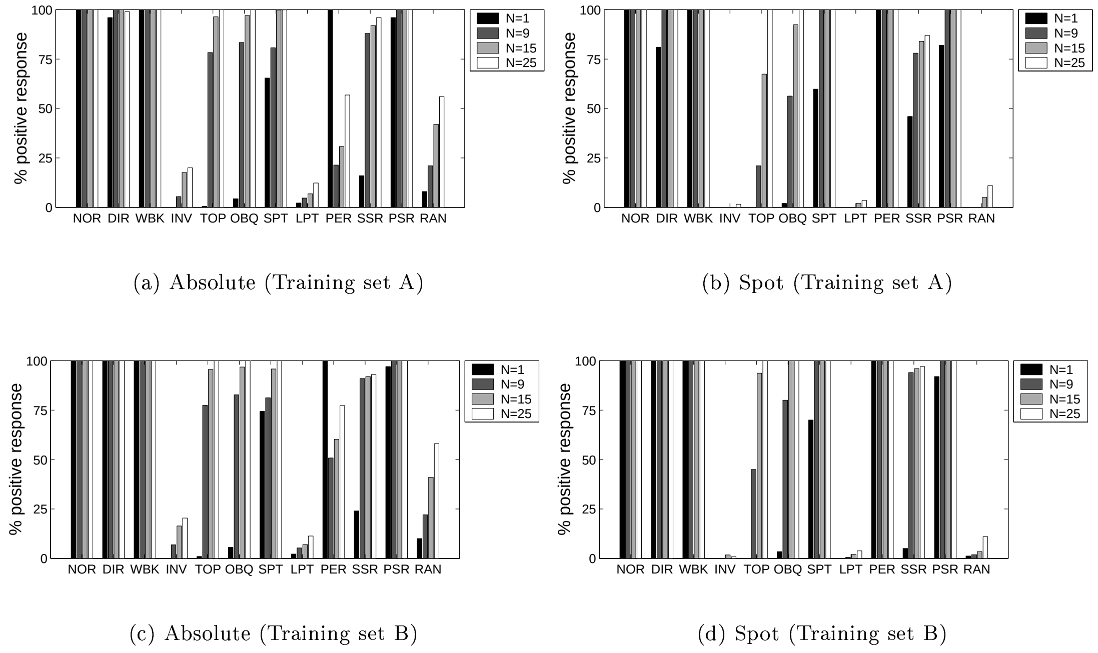
Biological Motion Detection: Machine Perception

As discussed in last chapter, the human responses for biological motion for the 12 categories considered in this work, differ substantially from those presented in Table 3.1. This necessitates verifying that the machine models can indeed learn to emulate the human behaviour on these new results presented in Table 5.4. This chapter presents the results of the machine perception in context of the human responses, as discussed in the last chapter. The main aim is to check if the machine perception is close to the human perception in the closed world of 12 categories. Although the humans can not only detect motion but categorise it, e.g. walking, running, hopping, jumping, etc. Johansson (1973). It is also reported that humans can also perceive the emotional state of the subject in an MLD sequence. The aim is not to model the human perceptual system in its entirety, but to determine what triggers the detection of human motion in an MLD sequence. To this end, the human responses, discussed in the previous chapter, form the basis of the machine learning. The aim is to obtain a machine performing equivalently to humans only from the motion identification point of view.

A brief structure of the chapter is as follows. Two different types of training sets used are described in Section 6.1. Sections 6.2 and 6.3 present machine perception of biological motion for 12 categories discussed earlier for k-NN and ANN detectors. A discussion on these results is presented in Section 6.4.

6.1 Training and Testing Sets

The 12 categories considered for this problem range from a well structured NOR category (well-structured in terms of shape and motion) to completely random RAN category. Human responses for these two categories are most positive and most negative



(a) Absolute (Training set A)

(b) Spot (Training set A)

(c) Absolute (Training set B)

(d) Spot (Training set B)

FIGURE 6.1: Motion detection: k-NN performance on training sets A and B. Performance for all 12 categories for both absolute and spot modes, described in Section 3.3, are shown by bars. For each category, the performance is displayed for different values of frames per data-point.

respectively. For all other categories the responses vary. If the data has all the information about the motion itself, the machine should be able to partition data at these two extremes. An equivalent machine should also be able to generalise to human performance (i.e. perform as humans even on limited training sets). If the response of the machine matches with the human performance, its decision criterion may be the one used by humans. This is the approach adopted for training machine. Along with NOR, DIR and WBK categories are also voted positive by humans. Here two different types of training sets have been considered.

- Training set A: Of 12 categories discussed in Section 3.2, only 2 categories for all the sequences of 9 subjects were used to generate training set A. The chosen categories were – NOR as positive instance and RAN as negative instance.
- Training set B: Of 12 categories discussed earlier, only 4 categories for all the sequences of 9 subjects were used to generate training set B. The chosen categories were – NOR, DIR, WBK as positive instances and RAN as negative instance.
- Test set consisted of rest of the data-points. In either case, A or B, the training and test sets were mutually exclusive.

6.2 k-Nearest Neighbour Detector

For k-NN detector, each data-point is assigned the label of the nearest category in the respective training set. Results of k-NN detector are presented in Figures 6.1(a)–6.1(d) for training sets A and B respectively. Each bar represents the fraction of data-points of the respective category voted as positive biological motion. Responses for INV, SSR and PSR categories are in gross error when compared with the human response (see Figure 5.2). As human response for TOP category remains indeterminant in the context of biological motion, the machine response to this category is of no practical significance. For a fixed dataset (as in present case), the performance of k-NN is invariant for a given metric (Euclidean in this case). Hence the responses are more categorical than the human responses. Interestingly, the machine was not able to discriminate between NOR and PER categories in the spot mode.

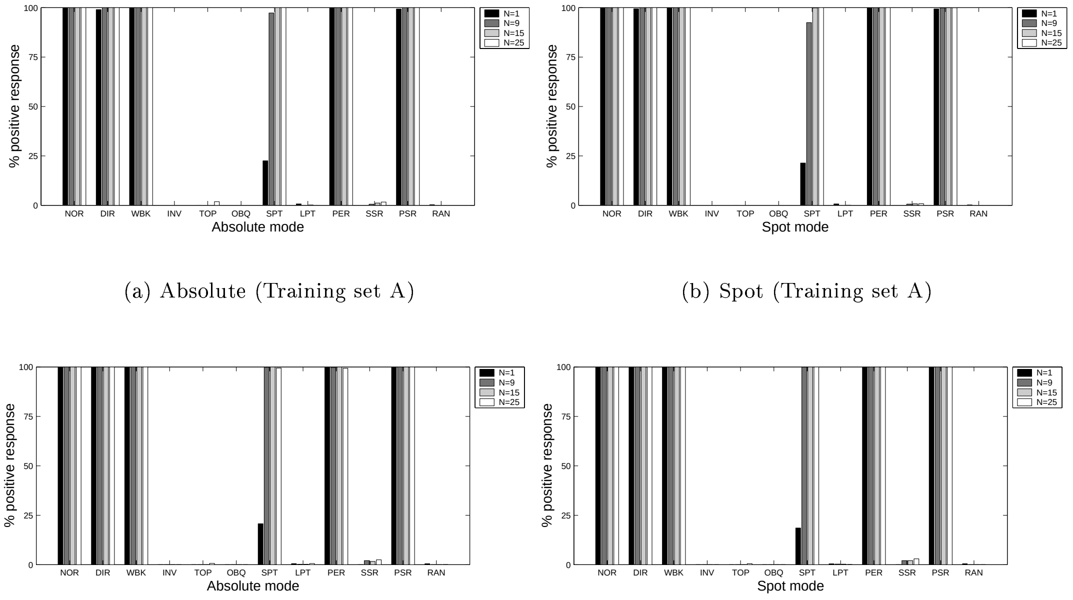
6.3 Artificial Neural Network Detector

A non-recurrent feed-forward network with back-propagation learning and two hidden units is used. The network is trained with one of the training sets and then its response is obtained for the entire dataset. As an MLD sequence consists of a moving shape, the shape and/or its motion pattern are probably most influential on the perception. Initially ANN detector is tested with the absolute coordinates of dots. No pre-processing of data is done, as data have enough information for the perceived human responses. However, if machine performance departs from the human performance, humans may not be using the same decision criterion. In the following discussion, the performance of the detector is averaged over five program runs; the neural architecture is randomly initialised for each run. The neural architecture is assumed to have converged if the classification accuracy remains high (i.e. $\geq 95\%$) for 5 successive epochs.

6.3.1 Absolute Coordinates

First of all tests are done on raw data with no preprocessing. Results for different values of N, frames per data-point and two training sets as discussed above are shown in Figure 6.2.

Even if we consider that such a simple machine may not have the capacity to rotate and may yield incorrect results for INV category, the machine performance still differs from the human one mainly in OBQ, PER and PSR categories. Machine learning seems to be insensitive to phase or temporal ordering. When trained only on the extreme categories, PER and PSR are perceived closer to NOR than to RAN. A possible explanation could be simplicity of the machine. However, an increase in hidden units does



(a) Absolute (Training set A)

(b) Spot (Training set A)

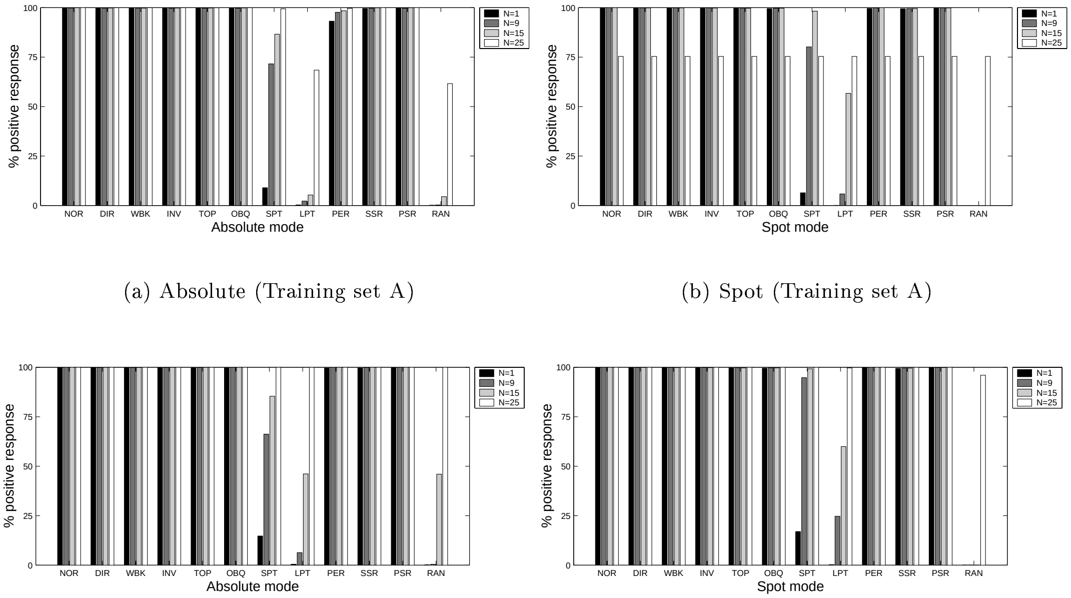
(c) Absolute (Training set B)

(d) Spot (Training set B)

FIGURE 6.2: Motion detection: ANN performance, using absolute coordinates, on training sets A and B.

not improve the performance. Rapid convergence does indicate that the machine can find discriminating features within the training set quite quickly. A possible hypothesis, which explains the machine behaviour, is that it is learning the range of y-positions occupied by the dots and, is insensitive to the change in x-positions. This may be because the machine has to learn almost the same x-range for all the categories.

If the hypothesis were true, INV, TOP and OBQ merit no positive responses as these are structurally different from NOR category and the scanned dots occupy altogether different positions and hence different ranges. As PER category has same structure as NOR but different temporal ordering, the hypothesis holds good. Also for PSR category, although different dots have different initial phase, the range of positions over the entire gait cycle remains same. Another point favouring the hypothesis is that for SPT category, the response builds up as temporal information per data-point increases. Almost 100% response of the trained machine, with the data in which the dots lie within the range, also confirms the hypothesis. However, expected near zero response is not obtained when the machine is tested on the data in which dots never occupy the same range as the positive data.



(a) Absolute (Training set A)

(b) Spot (Training set A)

(c) Absolute (Training set B)

(d) Spot (Training set B)

FIGURE 6.3: Motion detection: ANN performance, using relative coordinates, on training sets A and B.

6.3.2 Relative Coordinates

To test if the smoothness of the motion was the criterion used by the humans, absolute differences between the coordinates of the two nearest dots in successive frames were used instead of the absolute value. Here jth frame parameter is given by $\left|Z_{(t+1,j)} - Z_{(t,j)}\right|$ where $j = 1,\ \ldots,\ 2M$ and $t = 1, \ldots,\ F$. Here Z is x or y coordinate of the respective dot and t is the frame index. The results are shown in Figure 6.3.

If the machine is learning the ranges only, any smoothly varying shape, irrespective of the pattern, should get a high response. This does happen for all categories undergoing smooth transitions such as NOR, DIR, WBK, INV, TOP, OBQ and PSR. As PER category also receives a high response, this gives credence to the hypothesis that the machine is learning only the ranges of Y coordinates. A high response for the SPT category indicates that the machine can tolerate small violations in the smoothness of the motion. As the response for this category increases with the increase in temporal information per data-point (N), it indicates that the range of variations for the tolerance increases with N. This may explain the high response for the LPT category.

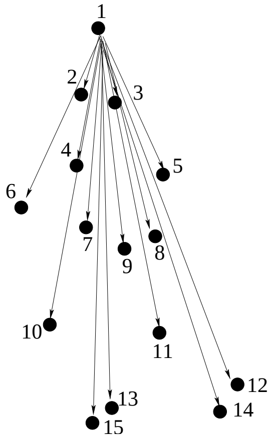


FIGURE 6.4: Shape vector: Concatenation of positional vectors of dots in each frame in the data-point. Positional vector is $\langle r, \theta \rangle$ relative to the top-most point of the same frame. The angle θ is computed relative to the vertical.

6.3.3 Shape Vector

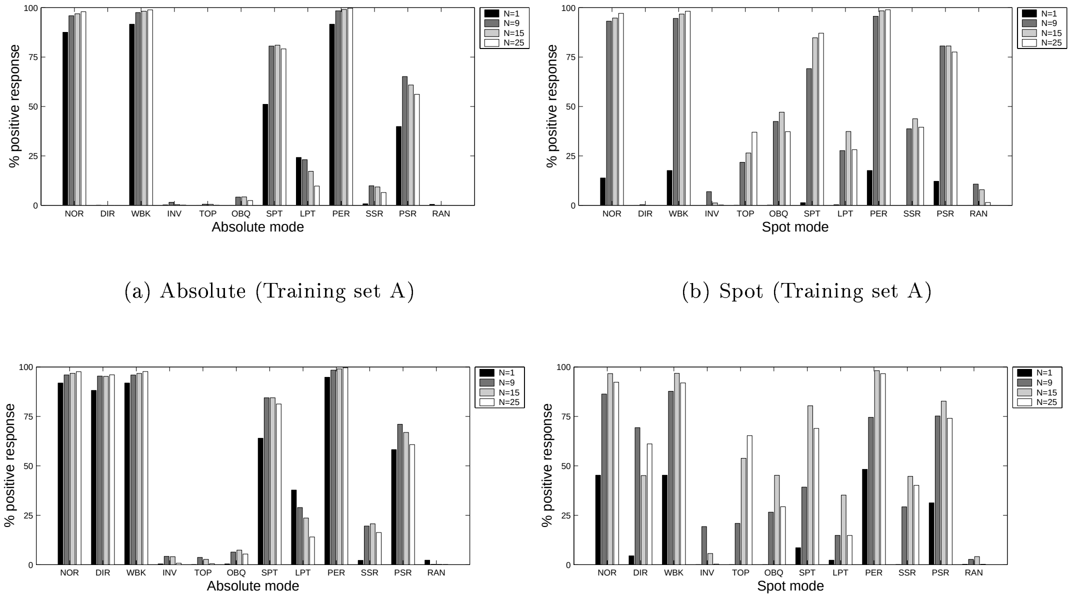
Figure 6.5 represents the results when the machine is presented with the shape information. Instead of using Cartesian $\langle x,\ y \rangle$ coordinates, polar $\langle r,\ \theta \rangle$ coordinates, with respect to the top-most dot of the same frame, were obtained for each point as illustrated in Figure 6.4. Top-most dot of every frame in the data-point is still retained in its rectangular coordinates. Interestingly when trained with set A, the responses become directional. As angles are computed as $\tan^{-1}\left(\frac{dy}{dx}\right)$, changing direction also means reversing the signs of the angles. As PSR and PER still retain the human contours, the responses are predictably high.

6.3.4 Relative Shape Vector

This is similar to the shape vector. The only difference is that, for each dot, the positional vector is computed relative to the top-most point of the first frame in the sub-sequence constituting the data-point. It is illustrated in Figure 6.6. The top-most point of the first frame is still retained in Cartesian coordinates. Machine response is similar to that obtained for the shape vector in absolute mode. The only difference is a slight fall in response for PER category at higher values of N. In spot mode, the performance is worse.

6.3.5 Randomly Scaled Data

Another way to test if the hypothesis is true is to train a machine on data in which the height of the MLD sequences vary randomly. This will prevent the machine from latching onto any quick shortcuts such as ranges. The response, as illustrated in Figure 6.8



(a) Absolute (Training set A)

(b) Spot (Training set A)

(c) Absolute (Training set B)

(d) Spot (Training set B)

Figure 6.5: Motion detection: ANN performance, using shape vector, on training sets A and B.

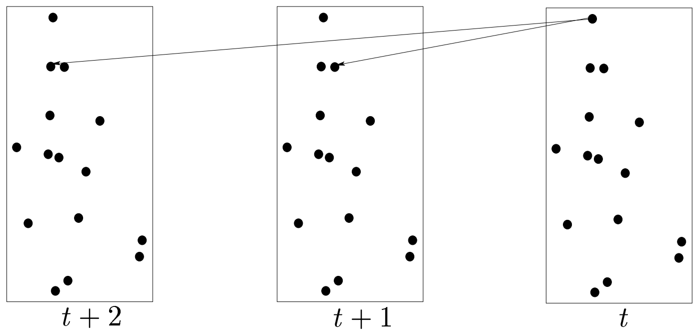
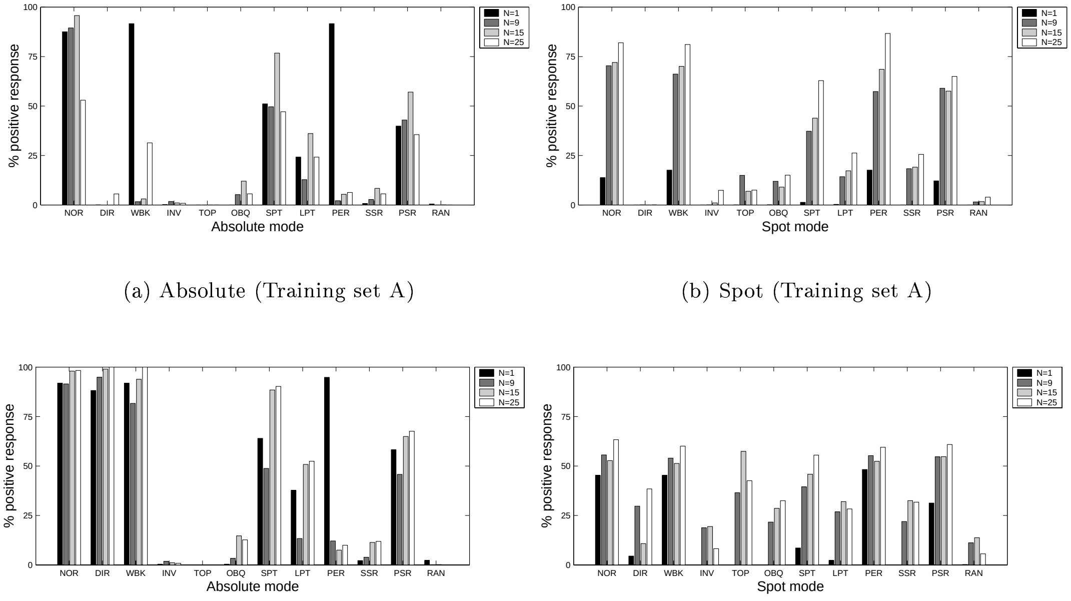


Figure 6.6: Relative shape vector: Concatenation of positional vectors of dots in each frame in the data-point. Positional vector of each dot is relative to the top-most point of the first frame in the data-point.

is similar to that shown in Figure 6.2 except for larger response for LPT category. It still supports the hypothesis, as randomly scaling the data has a the effect on widening the ranges. Another possibility is that this machine is learning something afresh. However, as both machines rapidly converged, it suggests that the data may have some inherent partition which is exploited by the machines. The insensitivity of the machine to temporal ordering may be due to the fact that all the training examples have sequenced temporal ordering, i.e. the pattern of dots move from one side to another without any spatial discontinuation in X axis. There are no jumps as would be seen in a permuted sequence. So unless explicitly trained on a sequence with permuted temporal ordering, the machine is unable to use temporal ordering as a decision criteria. In the



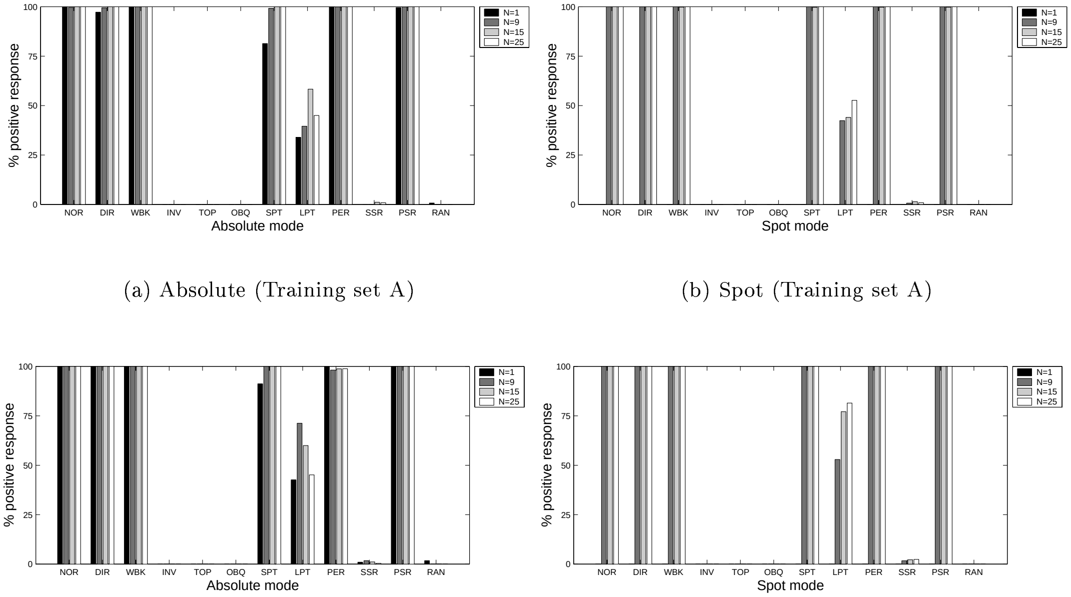
(a) Absolute (Training set A)

(b) Spot (Training set A)

(c) Absolute (Training set B)

(d) Spot (Training set B)

FIGURE 6.7: Motion detection: ANN performance, using relative shape vector, on training sets A and B.



(a) Absolute (Training set A)

(b) Spot (Training set A)

(c) Absolute (Training set B)

(d) Spot (Training set B)

FIGURE 6.8: Motion detection: ANN performance, using randomly scaled data, on training sets A and B. Each MLD sequence is scaled by a random factor to prevent machine from learning only y-ranges.

spot mode, all the frames of the sequence occupy almost the same position and hence temporal ordering has no spatial discontinuity.

6.4 Discussion on Machine Perception

Machine performance differs from human perception. Hence at this stage, it is difficult to say if an equivalent machine model will yield useful information about human perception of biological motion. The main findings of the machine perception can be summarised as follow.

- For any category, a high response indicates that the machine detects this as human motion. A low response, on the other hand, indicates that the category is nearer to negative instance (i.e. RAN category) in the machine space.
- Machine perception deviates from human perception in INV (upside down), TOP, OBQ, PER (permuted) and PSR (phase scrambled) categories. As machines may not have the capacity of rotation as humans do, a low response for INV category was understandable. A near-zero response for TOP category may be due to an altogether different structure.
- A high response for the PSR category is indicative that the machine is insensitive to the phase. As the k-NN detector also considers this category closer to NOR (i.e. positive category) than RAN (i.e. negative category), it is suggestive that the decision boundary only on the basis of training sets A or B does not have enough information for the machine to respond in a manner equivalent to the human response. Alternatively an equivalent machine would have classified these categories as negative.
- A high response for the PER category in the spot mode even by the k-NN detector also suggests a strong interaction of the various categories in this mode. This may explain the worse performance of the machine in the spot mode. Interestingly this trend is also observed in the human response.
- For the ANN detector, a high response for the PER category even in the absolute mode indicates that the machine is not able to learn the temporal associations and instead is making decisions probably on y-ranges of various dots. This argument is further strengthened by the high response in the PSR category. Most probably, the machine is learning the range of heights occupied by various dots. Presentation of the data in different ways seems to support this argument. In short, the machine seems to pick short-cuts.
- When the data is randomly scaled so as to alter the y positions of the dots and, hence, prevent the machine from learning 'short-cuts', the responses for PER and

PSR categories are slightly decreased but are still sufficiently high. This indicates that the machine does not learn the temporal/phase associations as the requisite information may be missing from the training set. The fact that there is no temporally permuted negative category and that all the categories in the training set have the same temporal ordering of frames suggests that there is not enough discriminatory information available in the data.

- Although the inclusion of the temporally permuted RAN category may help machine learn the temporal association, the exemplar category (which can let the machine discriminate between the NOR and the PSR categories) for the phase learning still needs to be determined. In short, presently the training set is too restrictive.
- Machine perception is more categorical and less variable. A wide variation in human responses may be because of different participants using different criteria, which cannot be modelled by a single neural architecture.

Chapter 7

Motion Classification: A Good Biometric?

A secondary aim of the work presented is to ascertain if MLD data can be used to establish gait as a biometric or not. As discussed in Section 2.7, humans do have ability to learn to recognise people, albeit in a small set of subjects. Human gait involves translatory and/or rotational movements of the various parts of the body. This has motivated the researchers to explore the potential of human gait as a biometric. This chapter highlights human gait as a biometric from the computational and bio-mechanical perspective.

This chapter is structured as follows. First a brief discussion on biometrics in general is presented in Section 7.1. Gait as a biometric is discussed next in Section 7.2. This is followed by a brief review of various gait-based person identification techniques presented in Section 7.5. A brief account of gait cycle and various components of gait is presented in Section 7.3. Similarities and dissimilarities in gait patterns of different individuals are highlighted in Section 7.4. Section 7.6.2 discusses the training and test sets used. Sections 7.6.3, 7.6.4 and 7.6.5 present classification results using k-NN, ANN and SVM classifiers. A discussion on these results is presented in Section 7.7.

7.1 Biometrics

Determining the correct identity of an individual is important, especially in security applications. Conventional approaches to personal identification are provided by means of passwords, keys, identity cards or similar possessions. The problem with these approaches is that the possessions can be lost, stolen, copied, forgotten or misplaced. Except for passwords, other possessions mentioned here are physical objects and can be faked. As these possessions do not incorporate any personal attributes of an individual, anyone, after acquiring the control of these, can abuse the privileges of the authorised user. Another approach, which is gaining popularity, is to use biometrics or the physical characteristics of an individual for identification. The main advantage of this approach is that a biometric cannot be stolen, misplaced or forgotten. Also it is more difficult to copy or fake a biometric. In fact, copying is nearly impossible in the case of some biometrics like DNA, fingerprint and iris patterns. The main limitation of this approach is public acceptance. The limitation arises due to the fear from lack of knowledge of how much personal information is revealed by a given biometric and consequent infringement of privacy. Any human physiological or behavioural characteristic can be a biometric (Jain, Bolle, and Pankanti 1999) provided it has the following desirable properties:

- **Universality:** Every person has this characteristic.
- **Uniqueness:** No two persons are same in terms of the characteristic.
- **Permanence:** The characteristic remains invariant with time.
- **Collectibility:** It is easy to measure the characteristic quantitatively.
- **Performance:** The characteristic can be employed to achieve a satisfactory identification accuracy level.
- **Acceptability:** This refers to the extent people are willing to accept the biometric.

From implementation point of view, the list of desirable characteristics can be extended to

- **Availability:** How high is the possibility that the requisite characteristic is available?
- **Computational cost:** The computational cost associated with the identification system will determine the applications where the underlying biometric can be used. For implementing security at ports, an on-line system which can return a decision in real-time is essential. For a crime investigation, however, an off-line system would suffice and does not impose too much restriction on computation time.
- **Circumvention:** How easy it is to fool the system by fraudulent techniques?

Many biometrics are in use. Most notable among these are high performance biometrics like DNA, fingerprints and iris patterns. These biometrics satisfy the criteria of universality, uniqueness, permanence and performance. DNA and iris patterns have low public acceptability because of the fear that these may reveal unintended personal information. DNA also suffers from invasive collection methods. Although not invasive, collectibility of iris patterns is, nevertheless, constrained. Automated systems employing iris patterns are easier to implement and are in use with military applications. Fingerprints have long association with forensic applications, which has resulted in reduced social acceptance.

In recent years many biometrics have been suggested and investigated. These include hand geometry, retina, signature, voice, infrared print of body/face, ear, etc. Successful automated identification systems based on voice have been reported and are being used. The major limitation of these systems is high computational cost.

In security and surveillance applications, the only information available may be a small amount of footage of video obtained from CCTV cameras. The only biometrics available may be face, if not masked, and gait. Strictly speaking, face and gait are not invariant with time as these undergo slight changes with ageing. High public acceptability, high universality, easy collectibility and often being the only means of availability has directed many research efforts to explore means for establishing these methods as reliable biometric parameters for personal identification systems.

7.2 Gait as a Biometric

Many biometrics (Jain, Bolle, and Pankanti 1999) have been investigated and used in different applications. Most notable of these are DNA, fingerprints and iris. Although any of these identifies an individual with a very low probability of error, their use in general is restricted due to lack of public acceptability. Using human gait as a biometric is appealing because of its universality and easy collectibility. After all we all need to walk. In surveillance applications, images from CCTV camera(s) may be too blurred to make a correct identification. Under such circumstances the motion is the only clue to the identity of a person. Using gait as biometric (Nixon, Carter, Cunado, Huang, and Stevenage 1999) offers the following advantages.

- Human gait data can be collected by non-invasive means. Equipment for acquiring these characteristics is not too expensive anymore. Video cameras, installed at locations under surveillance, will suffice. As most of the gait studies are based on the assumption of availability of fronto-parallel image sequences, the positioning of camera may need careful adjustment.

- Human gait can be acquired by non-contact means. No touching of equipment is required. This may be a deciding factor in public acceptability as personal hygiene issues may override a system involving contact with the equipment.

- Acquiring gait data is not restrictive as some of other biometrics like retina, face require the person to look into the camera.

- Gait data can be acquired from a distance. This factor may play a crucial role in surveillance applications.
- Cooperation of an individual is not a pre-requisite for acquiring gait data. This may be an important factor in crime prevention, as it will reduce the probability of making conscious efforts to alter one's gait to fool the system.
- Public acceptability may be higher. Although DNA, fingerprint, iris and retina patterns are well established biometrics, they may not be good for implementing identification systems acceptable to a large population.
- It may be difficult to copy or mimic the gait of other people. Also an individual may not always remember to change his gait especially when his efforts are more toward quickly getting away from the scene of a crime or untoward incident.
- This may be the only information available to identify a person.

Gait biometric also has some limitations. A few are listed below.

- An image sequence of a few gait cycles may need to be processed to determine the gait signature of an individual. Presently it is computationally expensive.
- Gait is affected by footwear.
- Gait can be obscured by clothing.
- Affliction of feet or leg, pregnancy and drunkenness affect gait.
- Any deviation from normal gait, e.g. running, hopping, etc. can affect the gait signature.
- Gait is also affected by the emotional state of a subject.

7.3 Components of Human Gait

Human gait is a complex periodic phenomenon. It involves synchronous and coordinated movements of almost all parts of the body. This entails the ability to support the upright body, maintain balance in the upright position and execute the stepping movement, which results in a forward motion in the plane of progression. The fundamental unit of human gait is the basic walking cycle, also referred to as the gait cycle. This cycle is defined as the walking sequence between two consecutive occurrences of the same body configuration. Conventionally, a walking cycle is the time interval between the successive instances of initial floor-to-floor contact of the same foot. As in normal walking, the heel of the foot/shoe is the first to contact the floor, heel-strike marks the beginning and end of a gait cycle.

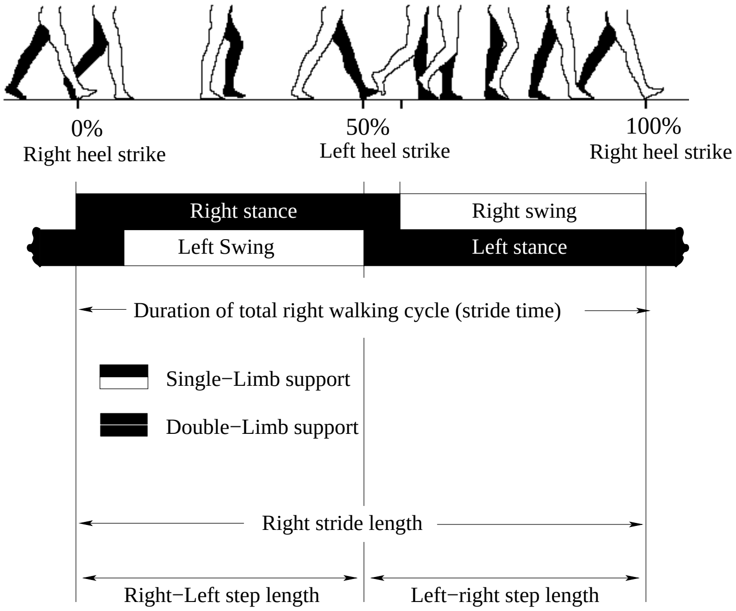


FIGURE 7.1: A complete gait cycle highlighting step cycle and various phases of the gait cycle. Redrawn from Murray (1967) with permission from The Lippincott, Williams & Wilkins Co.

In the entire process of walking, the moving body is supported by first one leg and then the other. To provide this support and maintain the body in the upright position, one of the feet is always on the ground or the walking surface, and the other leg swings forward to create the next step. When the swinging leg touches the ground, the body weight is shifted to this leg. For a brief amount of time, both feet are on the ground and the body weight is transferred from one foot to another. Two legs alternately switch these roles of weight support and forward movement. The result of these coordinated but anti-phase movements of the two legs is that the trunk is continuously translated forward over two alternating bases of support at a remarkably constant linear horizontal velocity.

A complete gait cycle can be divided into three phases distinctly (see Figure 7.1). The period of contact with the ground in which the leg acts as a supporting base to the rest of the body is called ‘stance phase’. This phase is followed by a brief period of ‘double-limb support’ when both feet touch the ground and the weight transfer takes place. In this phase, one leg is beginning the stance phase and the other is ending a stance phase. The next phase is the ‘swing phase’ when the forward movement takes place. When one leg is in stance phase, another is in swing phase except for the double-limb support phase. Within each walking cycle there are two periods of single-limb support and two periods of double-limb support. As the heel and toe are the first and the last points of contact respectively, events of heel-strike and toe-off define the beginning and end of the stance phase respectively. If the gait is symmetrical, the two double support periods will be equal in duration.

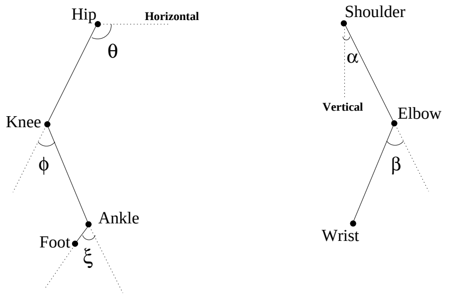


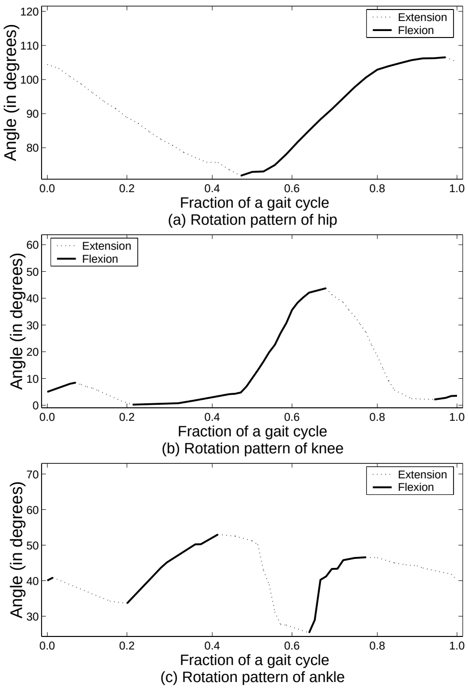
FIGURE 7.2: Angles to describe angular displacements of arms and legs.

In the course of walking, the entire body, including legs, undergoes various three dimensional movements. The body moves forward in the plane of progression, moves slightly from side to side in the lateral plane as weight is shifted from one leg to another, rises and falls in the vertical plane. Lateral plane is the horizontal plane and vertical plane is mutually perpendicular to both plane of progression and lateral plane. The translatory displacement of the entire body through the space is achieved by the angular displacements of various segments of the body about the axes that lie in the proximity of joints. Murray (1967) identified about 20 simultaneous movements in a normal human walk. These can be broadly classified into angular displacements of legs, movements of leg extremities namely heel and toe, angular displacements of arms, pelvic and thoracic movements, movements of neck and head. All accompanying figures were drawn using information from a given sequence of the MLD data.

7.3.1 Angular Displacement of Legs

Angular displacements of legs (Figure 7.3) bring about forward movement of the body and, hence, constitute the most significant and essential part of human gait. Both legs move in anti-phase, alternately supporting the body and propelling it forward in succession. The pendulum-like angular displacement of legs take place about joints – thigh rotating about hip, shank rotating about knee and foot rotating about ankle or toe. The following paragraphs describe these displacements for the leg providing support at the beginning of the gait cycle. The angles shown are calculated as shown in Figure 7.2(a). Although the directions of rotations are constantly shifting, three major leg joints rarely rotate simultaneously in the same direction.

1. **Rotation about hip (θ):** The first half of the cycle is characterised by continuous hip extension during the stance phase as the trunk moves smoothly over the supporting leg. During the swing phase, after the other leg has provided a supporting base, the hip begins to flex preparatory to the swing phase. This hip flexion directs the swinging leg forward for the next step.



(a) Rotation pattern of hip

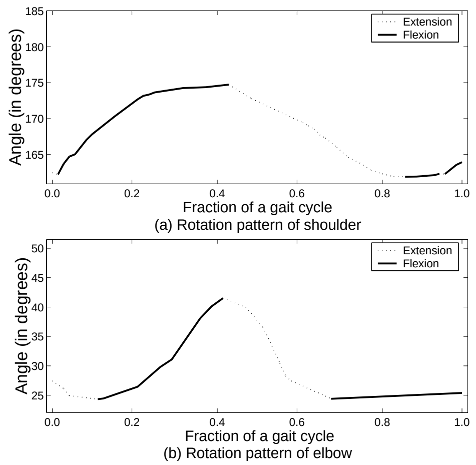
(b) Rotation pattern of knee

(c) Rotation pattern of ankle

FIGURE 7.3: Angular displacements of the leg beginning the stance phase. Rotation of (a) hip with respect to horizontal plane, (b) knee with respect to hip and (c) foot with respect to knee are shown.

2. **Rotation about knee (ϕ):** The pattern of knee rotation is more complex and shows two periods of flexion alternated with two of extension within each walking cycle. The supporting leg enters stance phase at heel strike with the knee joint in nearly full extension. The knee, then, begins to flex and continues to do so until the foot is flat on the ground. This flexion decreases the amplitude of the vertical trajectory of the trunk as it moves forward over the supporting leg. In the later half, the knee rotation pattern exhibits large rapid excursions into flexion and into extension. The flexion excursion provides foot-floor clearance early in the swing phase, and the extension excursion projects the extremity forward for the next step. Effects of this complex rotation pattern is to flatten the arc through which the centre of the mass of the body is translated. This results in a smooth motion. Otherwise, this would result in a jarring effect on the body.

3. **Rotation about ankle (ξ):** Like the knee rotation, the ankle rotation exhibits two waves of flexion and two of extension within each walking cycle. Near the beginning of the cycle, the ankle of the weight-bearing leg is relatively flexed and the heel is projected forward preparatory for floor-contact. Ankle extension permits the forefoot a controlled but rapid descent to the floor. Once the entire foot has made contact with the floor, the normal ankle abruptly reverses from extension to flexion and continues to do so, reaching its greatest amplitude of flexion as the body moves over this supporting extremity. After the body has passed over the



(a) Rotation pattern of shoulder

(b) Rotation pattern of elbow

FIGURE 7.4: Rotation patterns of the arm ipsilateral to striking leg. (a) Rotation pattern of shoulder with respect to vertical plane, (b) elbow with respect to shoulder.

supporting base, the ankle extends gradually shifting the contact area from the entire foot to the forefoot. The ankle, then, abruptly reverses into flexion after the toe leaves the floor and remains relatively flexed to provide foot-floor clearance in the swing phase.

7.3.2 Angular Displacements of Arms

Although not necessary for normal walking, the arms show definite participation in the total pattern of the gait (Figure 7.4). Arms swing forward and backward in phase with the contralateral leg and in opposition to the ipsilateral lower limb.

1. **Shoulder (α):** The rotation pattern of the shoulder shows one excursion of extension, when the arm is directed backward from the vertical, and one excursion of flexion, when the arm is directed forward. At the time of the heel-strike, the ipsilateral shoulder is near maximal extension. During the first half of the cycle, the shoulder flexes forward. Maximum flexion is reached at the midpoint in the cycle, when the contralateral heel strikes. Then the shoulder extends backward as the ipsilateral leg swings forward.

2. **Elbow (β):** Elbow pattern is similar to the shoulder in that the first half of the cycle is characterised by flexion and the next half by extension.

7.3.3 Vertical Pathways of Heel and Toe

These pathways (Figure 7.5) are essentially the movements of the heel and toe in the plane of progression.

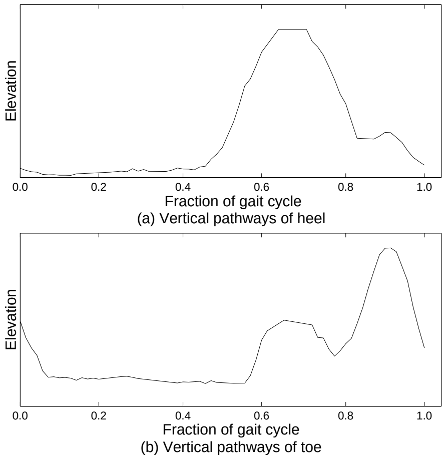


FIGURE 7.5: Vertical pathways of (a) heel and (b) toe of the leg touching the ground.

1. **Heel:** At the beginning of the cycle, the heel of the supporting leg makes the initial floor-contact and remains there for the first half of the stance phase. In the later half, the contact-area is transferred to the toe and the heel is lifted off the floor. The ascent of the heel becomes more rapid after the contralateral supporting base is provided. After this, the heel begins a steady descent as the foot swings forward.

2. **Toe:** Unlike the heel, the vertical pathway of the toe shows two peaks within each cycle. At the time of heel-strike, the toe is still off the ground and is in a phase of controlled descent. Once toe-floor contact is established, the toe remains there throughout the entire stance phase. Early in the swing phase before the leg assumes a forward direction, the toe reaches a minor wave of elevation and then descends a critical low point. The major peak of elevation is attained late in the swing.

7.3.4 Spatial Displacements of the Body

The angular movements of the leg joints result in the translation of the trunk. This translation consists of a series of left-right and up-down sinusoid movements. These movements (Figure 7.6) can be distinctly divided into

1. **Vertical oscillations:** The entire body gently oscillates through two vertical peaks and two valleys within each walking cycle. The valleys occur during double-limb support. The peaks occur during the single-limb support when the trunk is directly over its single support base.

2. **Lateral oscillations:** The lateral pathways of the trunk, measured at the head, also show two peak lateral deflections – one to the right and another to

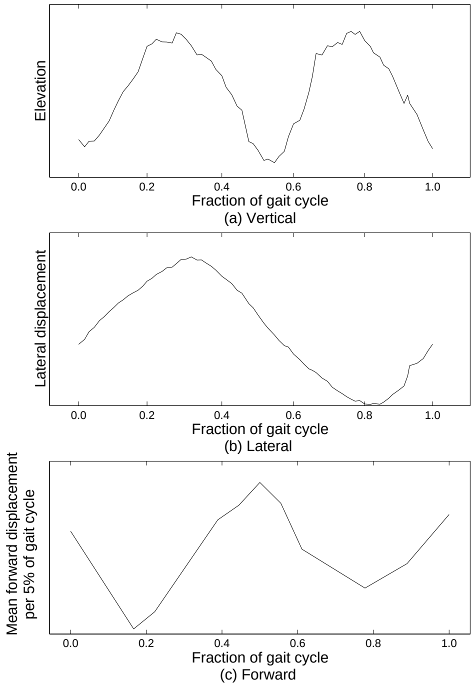


FIGURE 7.6: (a) vertical (b) lateral and (c) forward pathways of the body as viewed at light attached to head.

the left. During the double-limb support, the head is in a more central position.

3. **Forward pathways:** The pathway in the forward direction, as measured at the neck, is remarkably smooth. Forward movement is not constant but proceeds in two gentle waves of increased and decreased velocity. The forward speed decreases slightly as the trunk climbs to its highest and most lateral peaks and increases slightly as the trunk descends to the lower and more central positions.

7.3.5 Transverse Rotations

Transverse rotations are constituted by rotations of pelvis and thorax (see Figure 7.7). While thorax rotates clockwise, the pelvis rotates counter-clockwise and vice versa.

1. **Pelvic rotation:** The pelvis rotates about a vertical axis alternately to its right and left relative to the plane of progression. These rotations occur alternately at each hip joint. As a leg swings forward, the pelvis on that side pivots forward. The pelvis on the other side, however, assumes an increasingly backward direction.

2. **Thoracic rotations:** The thorax rotates in clockwise and anti-clockwise directions opposite to the pelvis. Thoracic rotation occurs alternately at each shoulder joint and the total amplitude is less than that of the pelvic rotation. The simultaneous but opposing thoracic and pelvic rotation in the lateral plane probably

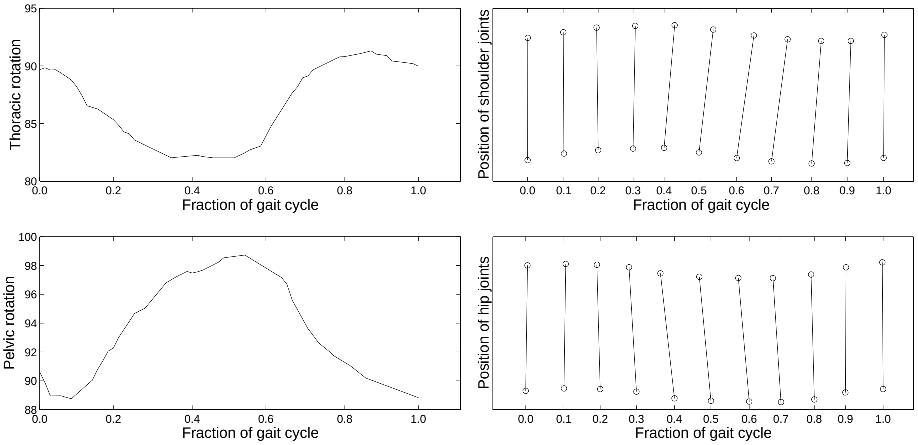


FIGURE 7.7: Thorax and pelvis rotate in opposite directions. The angles of rotations are relative to the plane of progression. Thoracic rotation is obtained by line joining shoulders joints and pelvic rotation by joining hip joints as shown in (b).

contributes to the smoothness of forward progression by providing counterbalancing restraints against excessive motion of the entire torso.

3. **Rotations of thigh and shank (leg):** These rotations are not quite obvious. In contrast to the thoracic rotations, thigh and shank rotate in phase with the pelvis. Rotary displacements progressively increase from pelvis to thigh, and thigh to shank. At the beginning of the swing phase, pelvis, thigh and shank rotate internally toward the supporting leg. This is continued during the double-limb support phase and into mid-stance. At mid-stance there is an abrupt reversal in the direction of rotation and this reversal continues until the beginning of the next swing phase.

4. **Rotations in the ankle and foot:** During the swing phase of walking, the segments of the lower limb including foot are free in space and can rotate internally without restriction. During the stance phase, the foot is on the floor and external rotation of the leg occurs because mechanisms exist in the ankle and foot that permit the leg to rotate externally while the foot remains stationary.

7.4 Characteristics of Walking Pattern

In the study of walking patterns, Murray, Drought, Kory, and Wisconsin (1964) found that for a given subject, gait components such as phase durations (stance, swing and double-support), vertical trajectories of heel and toe, etc., did not undergo any significant variation in repeated trials. Although patterns of rotation were strikingly similar for normal and fast-speed walking of the same subjects, the reversals in the direction of rotation (flexion to extension and extension to flexion) occurred earlier, durations of supportive (stance) phases decreased, out-toeing angle (angle the foot makes with the line of progression) decreased and the stride width (distance of the foot from the line of progression) increased at faster speed walking. From one subject to another, pelvic and thoracic rotations and the amplitude of the lateral pathway of the head, unlike that of the vertical pathway, were found to be more variable. Upper limb rotation patterns were the most variable gait components. According to Inman, Ralston, and Todd (1981), rotations of thigh and shank show marked differences and constitute one of the factors that provide distinctive characteristics to each individual's appearance when walking.

The above observations indicate that, for a given individual, the gait pattern does not vary within the same trial and repeated trials. The speed does affect the gait pattern. They also suggest that individuals may have different gait patterns. According to Inman, Ralston, and Todd (1981), bipedal walking seems to be a learnt activity and it is not surprising that each of us displays certain personal peculiarities superimposed on the basic pattern of bipedal locomotion. If all gait movements are considered, gait may be unique (Nixon, Carter, Cunado, Huang, and Stevenage 1999). However, from a computational perspective, measuring the gait components, which are highly variable from person to person, e.g. pelvic and thorax rotations, is difficult even from an overhead view of the subject. This is because of the self-occluding nature of human walking. Extracting these components from real images poses even more challenge.

7.5 Human Gait and Person Identification

As discussed earlier, human gait exhibits inter-subject variability and can be used as a biometric. This has focused much research in this direction. Recently a number of successful techniques to identify people from the way they walk have been reported. Niyogi and Adelson (1994a) used the braided pattern in XT-slice of an image sequence to signal the presence of human motion; characteristics of this pattern could also be used as a gait signature. Niyogi and Adelson (1994b) also described another method based on spatiotemporal surfaces, the surface being a combination of standard parametrised surface – the canonical walk – and a deviation surface specific to the individual walk. Meyer (1997a, 1997b, 1998a, 1998b) discussed the applicability of optical flow and hidden Markov models for gait-based identification. Huang (1999) discussed how principal component analysis can be used to extract a gait signature. On his pendulum based model, Cunado (1999) discussed extraction of gait parameters using a velocity Hough transform, an evidence gathering technique suited for temporal sequences. Little and Boyd (1998a) reported a moment based approach for gait-based person identification. Shuttler, Nixon, and Harris (2000) extended the concept to velocity moments which can also handle temporal information. Hayfron-Acquah, Nixon, and Carter (2001) discussed the applicability of symmetry operators. Foster, Nixon, and Prugel-Bennett (2001) described how statistical measures, after application of various area masks, could be employed as a gait signature.

Most of the studies extract motion features and use these to identify a moving subject. Motion features generally used are optical flow vectors, moments, eigenvectors, etc.. The methods employed may be model free (Little and Boyd 1995) or model based (Cunado 1999; Yam, Nixon, and Carter 2001). Little and Boyd (1997, 1998b) show that model free method can also be applied to moving dot displays.

7.6 Results for Motion Classification

Many image processing techniques, or their adaptations, have been applied to gait-based person identification with mixed results. However, our aim is to test if the machines can use only MLD information to achieve this. All the datasets described in Section 3.1 were used for the classification problem. Data-points for both training and testing were generated as explained in Section 3.6. Three different types of classifier – k-NN, artificial neural network and support vector machines – were used. These classifiers are described in detail in Appendix C. Results with ANN classifier on synthetic data were presented in Laxmi, Carter, and Damper (2002b). The results of ANN and SVM classifiers on the manually labelled dataset were discussed in Laxmi, Carter, and Damper (2002a). In earlier works, the input data was re-sampled such that all the sequences had same number of frames. The re-sampled sequences were, then, used for generating data-points for the training and the testing. The sequences for the GTRI data were, however, not re-sampled as re-sampling results in oscillatory effects. In this section, only results on the GTRI data are presented.

7.6.1 Assumptions

The current classification method is based on the following assumptions, some of which may not hold true in a real-world scenario.

- The scene contains only one person walking fronto-parallel to camera.
- Data are not missing. Information about the joint coordinates is available.
- The walking sequence is in the correct order.
- Correspondence information about any joint in all frames is available. In other words, for each frame, information about ‘which’ joint is ‘which’ is available.

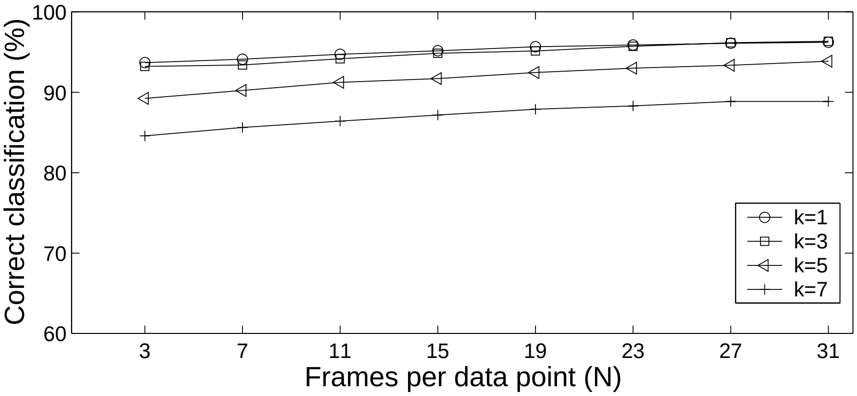


FIGURE 7.8: Motion classification: Performance of k-NN classifier.

- No normalisation of data except scaling to unit hypercube is required. Thus classification is based on both static, e.g. height, and dynamic features. However, this also makes classification sensitive to distance from camera.

- Data are free from noise, i.e. no missing or extrapolated data-points.

7.6.2 Training and Testing Sets

For the classification problem, individual gait cycles were manually extracted from the sequences in the GTRI data. The extraction was done in such a manner that all these cycles have the same phase. As each sequence consisted of about 2/3 complete gait cycles, the number of gait cycles per subject ranged from 5–8. For each frame, x, y and z coordinates of 15 joints – head, shoulders, elbows, wrists, knees, ankles and toes – were available. However, as an image is a projection on a 2D plane, only x and y coordinates were used. Only one gait cycle was used for the training. Similarly testing was done on another gait cycle. Training and testing data-points were obtained from two different gait cycles.

7.6.3 k-Nearest Neighbour Classifier

k-NN classifier was applied only to the GTRI data. Figure 7.8 displays the performance of a k-NN classifier. Number of frames grouped for temporal integration, N, is along the x-axis; the y-axis represents the number of test points correctly classified. Any given point on the plot represents the classification accuracy for all test data-points, each point spanning only N frames.

7.6.4 Artificial Neural Network Classifier

A fully-connected, two-layer, non-recurrent, feed-forward neural network with back propagation learning was used. In the following discussion, a given neural network architecture is represented by $m{:}n$ where m refers to number of hidden units and n refers to output units. Number of input units equals the dimensionality of a data-point (the configuration vector integrated over N frames as discussed in Section 3.4 and is $(2M + 1)N$ where M is the number of joints. As each subject was given a binary code, the number of output units (n) equals number of bits in the binary code. The neural network was assumed to be converged if its classification accuracy remains $\geq$ 95% for 5 successive epochs during the training.

Figure 7.9 displays the performance of the neural network classifier on this dataset of 21 subjects. Each subject was assigned a 5-bit binary code. The architecture of neural network was 16:5, i.e. 16 hidden units and 5 output units. Figures 7.9(a), 7.9(b) and 7.9(c) display the percentage of correctly classified data-points when (a) all joints, (b) joints on one side (head joint plus joints of ipsilateral arm and leg), and (c) only leg joints are considered. Each frame has 15 joints – head, shoulders, elbows, wrists, hips, knees, ankles and toes. The results are from five program runs with random initial weights of the neural network. The standard error is mean ± standard deviation. Dimensionality of the input for (a), (b) and (c) is respectively $31N$, $17N$ and $17N$, N being the frames being combined. For all 21 subjects, number of frames in the gait cycle vary from 58–71, the average gait cycle is about 68. The best performance remains in the range of 85%–90%. The best performance was obtained when only ipsilateral joints, one arm and one leg on the same side, were considered.

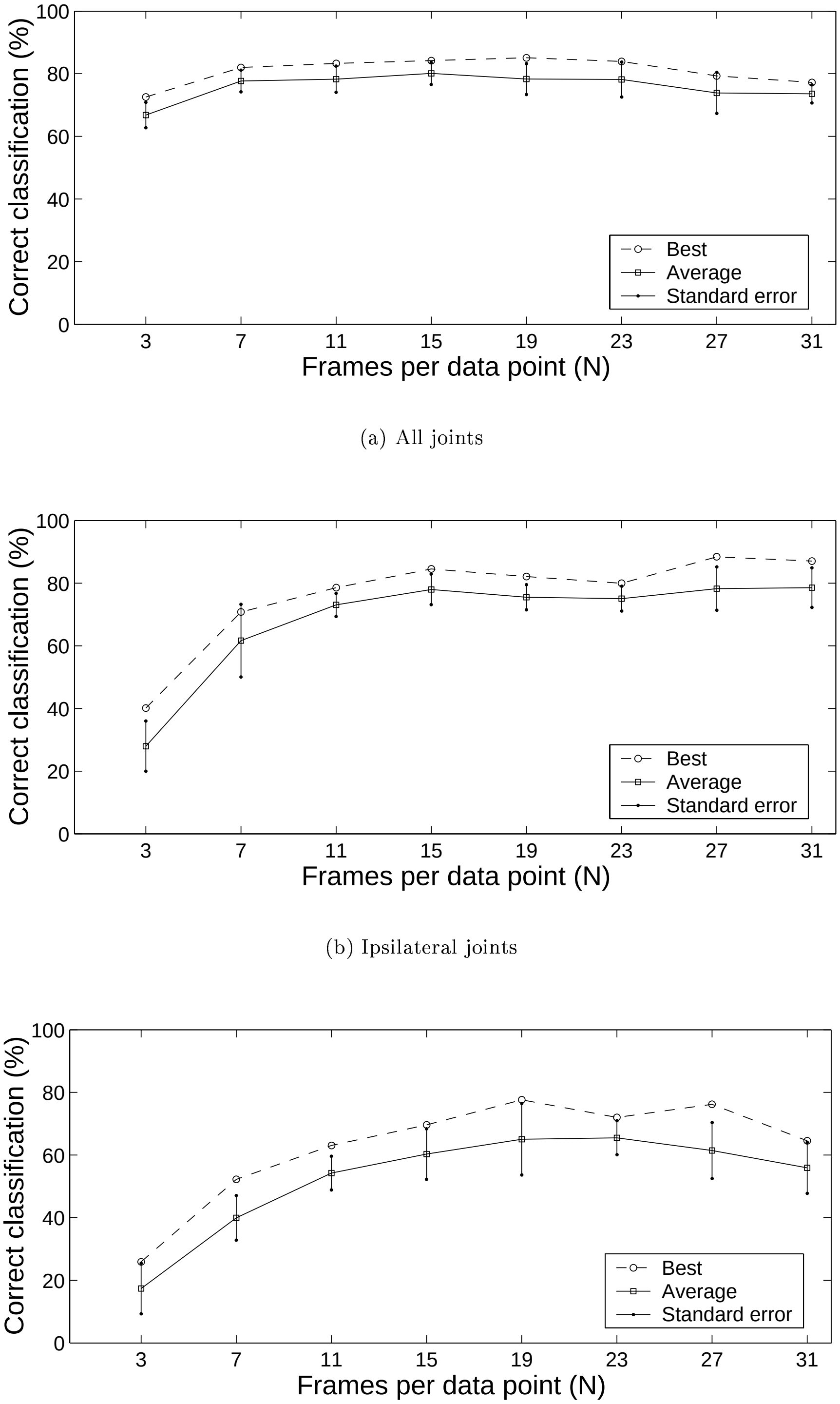
7.6.5 Support Vector Machine Classifier

One-against-one and one-against-rest (Section C.3.4 of Appendix C) multi-class support vector machine classifiers were used. Two kernels – linear and polynomial (degree 2) – were used.

Figures 7.10 and 7.11 display the performance of one-against-one and one-against-rest SVM classifiers. Only 8 joints, head joint plus ipsilateral joints (i.e. joints of one arm and leg on the same side) were considered. In either case, a linear kernel and polynomial kernels were used.

7.7 Discussion

Performance of k-NN classifier is good even when a data-point represents only one-frame. This may be due to the fact that different people walk with different speeds and hence

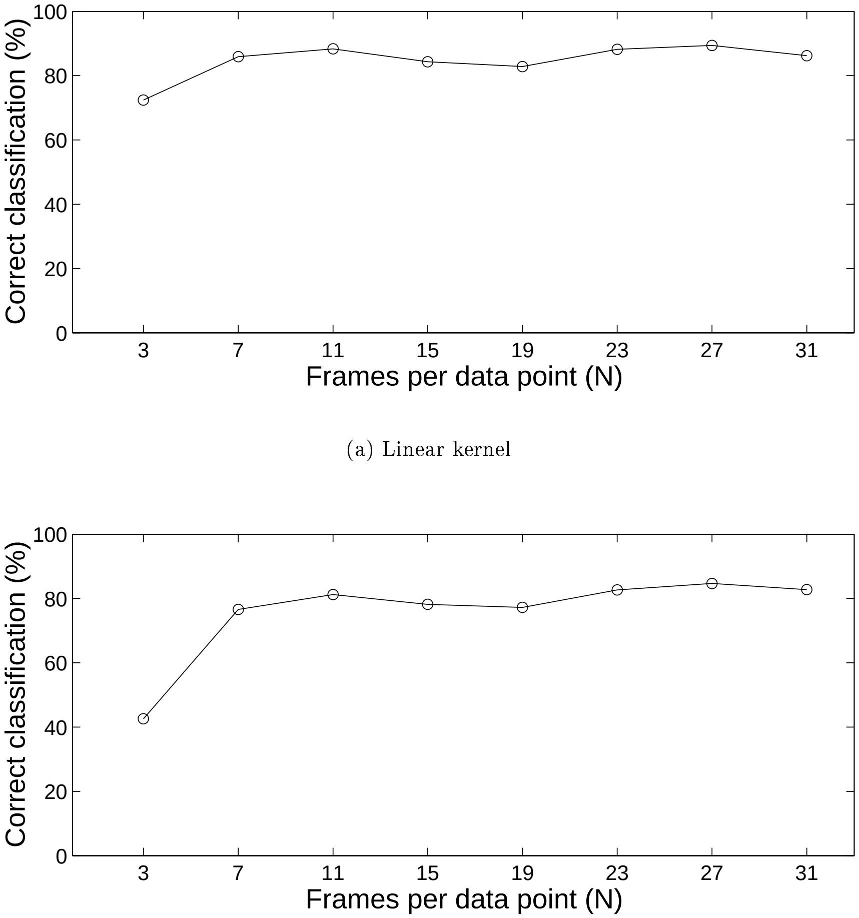


(a) All joints

(b) Ipsilateral joints

(c) Leg joints

Figure 7.9: Motion classification: Performance of the neural network with 16 hidden units and 5 output units. In (a), all the joints, in (b) only 7 joints for one leg and one arm on same side, and in (c) only leg joints are used.



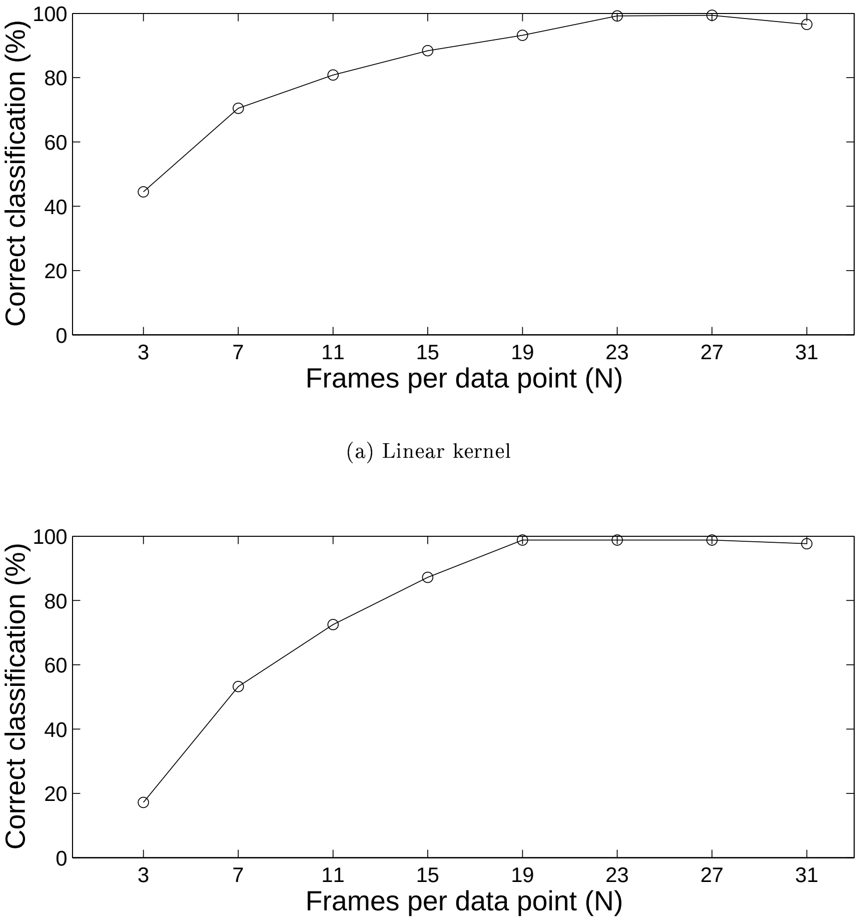
(a) Linear kernel

(b) Polynomial kernel

Figure 7.10: Motion classification: Performance of one-against-one multi-class SVM classifier. In (a) a linear kernel and in (b) a polynomial kernel with degree 2 were used.

cover different distances. The performance of the classifier does improve with increased value of N, number of frames grouped per data-point. Performance of ANN classifier is best when only ipsilateral limbs (one arm and one leg on the same side) are considered. Unlike k-NN classifier, performance is bad at smaller values of N. Performance of SVM classifier improves as N increases. The best performance is achieved using polynomial kernel with one-against-rest. Approximately 100% classification accuracy is achieved at about half the gait cycle. A smaller dip at the higher values of N may be due to insufficiency of the data-points. These results support the idea that the gait can indeed be used as a biometric.

In practical scenario, however, the joint information is not available. Although the data may be manually labelled, it is subject to errors. Another possibility could be to apply a corner detection algorithm to the sequences and obtain the information about various



(a) Linear kernel

(b) Polynomial kernel

FIGURE 7.11: Motion classification: Performance of one-against-rest multi-class SVM classifier. In (a) a linear kernel and in (b) a polynomial kernel with degree 2 were used.

joints. This can work only in an image with good contrast between the moving subject and the background. Even in a well segmented image, the joint information can still be affected by the footwear and the clothes. Once the joint information is extracted, the correspondence of the joints from one frame to another can be either manually supplied or automated using a suitable adaptation of the technique devised by Song, Goncalves, Bernardo, and Perona (2001).

Chapter 8

Conclusions and Future Work

This chapter concludes the work presented in this thesis with a framework for future work.

8.1 Conclusions

Although a machine can learn to detect biological motion, the machine perception differs from the human perception. Even though we were able to develop an insight into the machine models, attempts to directly relate and elicit an explanation of human perception on the basis of machine behaviour remains inconclusive. One of the major challenges to human perceptual understanding is the presence of a wide variation in the human responses regarding INV, TOP, PER categories. As most of the participants do recognise instances of INV category as upside down human figures, structure-based hierarchical perception theories put forward by Johansson (1973) and Cutting, Proffitt, and Kozlowski (1978) seem to find some merit in the explanation of this phenomenon.

In the present case, the machine does not have enough negative categories to learn about what precisely constitutes a human-like structure and/or motion. It appears that the set of 12 categories taken into consideration does not constitute a mathematically closed set from the point of view of machine learning. The training set does not have enough information for the machine to generalise to the level of human performance. As a result, machine performance is more categorical and less variable. In our experiments, human form seems to play an important role, as even the permuted sequence does not get as many negative votes as expected and SPT category gets less than expected positive response. This raises the question about the validity of asking each participant to respond to the same category as many as 12 times. Probably, a large number of responses about different categories, when participants were kept in the dark about the categories being shown, made it difficult for the participant to be too assertive about the decisions. Also, when asked to judge human or non-human motion, participants assumed that some of the sequences shown were of animals. This explains why the participants in our experiment did not respond 100% positive even for NOR category. The factor by which this assumption influenced the participants in making decisions about other categories remains elusive. Another possible reason for machine perception not matching human performance may be due to the fact that, while former works in a Euclidean space, humans might be using some other coordinate framework, possibly the central perspective framework suggested by Johansson (1973).

The applicability of k-NN, ANN and SVM as motion classifiers and person identification is supported by the experiments conducted. In earlier works, the input data was re-sampled such that all the sequences have same number of frames. The re-sampled sequences were, then, used for generating data-points for the training and the testing. The sequences for the GTRI data were, however, not re-sampled as re-sampling results in oscillatory effects.

As the number of frames per data-point (N) increases, classifier performance improves. This is possibly due to the increased availability of better temporal information and utilisation of dynamic features by the classifiers. A slight dip at the higher values of N may be due to an insufficient number of data-points. As datasets were not normalised, except for scaling to unit hypercube, static features might have also contributed to the classification. However, a noticeable difference in performance at low and high N indicates that the static features do not contribute exclusively. Performance of machine classifiers is better than that of the untrained humans. But this is when machines have information about dot-to-dot correspondence across all the frames in the sequence.

In the case of support vector machines, scaling of data affects the computation of Lagrangian multipliers. In the absence of scaling, even the largest Lagrangian multiplier for the linear kernel case is too small to be affected by any change in trade-off factor C. This affects the computation of the support vectors and the bias resulting in poor performance. However, it is observed that scaling of the data alleviates these problems. The scaling is done so as to fit the data into a unit hypercube.

Support vector machines, especially one-against-rest with polynomial kernel, perform better than the neural classifiers. A significant finding is that the classification is possible even with a fraction of the cycle provided the training is done on the entire gait cycle. Generally this fraction is approximately $\leq 50\%$ of the gait cycle. Our technique demonstrates that the gait data has sufficient variability for the person identification systems. However, the effect of variations in day-to-day walking, even under similar circumstances such as the same terrain, footwear, etc., need to be investigated.

8.2 Future Work

Although the results presented in this work remain largely inconclusive in the context of biological motion detection, the approach is definitely not a complete failure. The experimental set-up needs to be revised the approach should bring some fruitful results. A possible extension may be to include non-biological motion sequences, which are not directly derived from human motion such as the animal motion, motion of inanimate objects, etc. Permuted sequences of all the categories should be added and, within each category, there should be sequences of walks from both directions. This may help the machine learning system to be both time and direction sensitive. The main aim should be to look for a negative category close enough to the NOR category in euclidean space. This will help the machine learning system to achieve a better decision boundary. The present work suggest that PSR may be one of the candidates. To ascertain this, categories intermediate to NOR and PSR may be generated by varying the degree of randomness, and human responses for these may be obtained to determine if this is the case.

In the present work, all the joints were considered. It may be worthwhile to get the human response only for reduced number of joints. The aim would be to look for the minimal sub-configurations which do not evoke the perception of biological motion. However, if the human responses were taken for different number of joints, we need a machine learning system which can deal with data of varying dimensions. A possibility would be to apply other machine learning models such as hidden Markov models and recurrent neural networks.

Exposure time can also be reduced to force participants to make quick decisions. A category should be presented for not more than 4 instances to preserve the decision independence. In the current work, no masks have been employed for the sake of simplicity in machine performance analysis. However, masked data may force the machine to seek constructs responsible purely for the motion perception. A random mask applied to an organised structure (e.g. fronto-parallel view, upside-down view, etc.), may force the machine to learn discrimination between dots undergoing smooth and random motions. This may force the machine model to learn the motion trajectories and/or temporal inter-dot relationships rather than just the positional ranges of the dot.

In the classification task, the robustness of machine classifiers needs to be established. Also, some mechanism to solve correspondence problem need to be developed and incorporated. A possible solution to this can be found in Song, Goncalves, Bernardo, and Perona (2001). View-invariance learning capability (the performance when the image sequence is not fronto-parallel) of the machine is yet to be verified.

This work has been done on MLDs where location of each dot is determined by precise labelling using infra-red markers. But, in real life applications, this labelling needs to be done by application of corner detection algorithms. The current approaches can be extended for experimentation on the data obtained in such a manner.

Appendix A

Generation of Synthetic Data

A stick model (see Figure 1.1) is used to generate walking sequences. Synthetic data generations requires:

- Body dimensions, i.e. length of various limbs (Table A.1).
- Angles of rotation of various joints (Figures 7.3).
- Maximum deviation permissible in angles for same and different persons (Table A.2).
- A method to simulate walking.

The angles of rotation were taken from Murray (1967). Varying these angles by a large amount corresponds to generating data for different people. Once the angles for a given person were obtained, changing these randomly by a small amount amounts to generating walking sequences of the same person at different time instances. In the data

	Measurement minimum (cm)	Measurement maximum (cm)
Head	15.875	19.05
Neck	5.08	6.985
Shoulder	30.48	35.56
Arms	48.26	60.96
Palm	12.70	15.24
Torso	43.18	53.34
Pelvis	35.56	44.45
Legs	81.28	101.6
Foot	16.51	22.86

TABLE A.1: Stick model: Dimensions of various limbs.

	Deviation parameter different persons (in degrees)	Deviation parameter same person (in degrees)
Hip	5	1
Knee	3.5	0.7
Ankle	2.5	0.5
Elbow	4	0.8
Wrist	7	1.2
Palm	3	0.6

TABLE A.2: Stick model: Angle deviation.

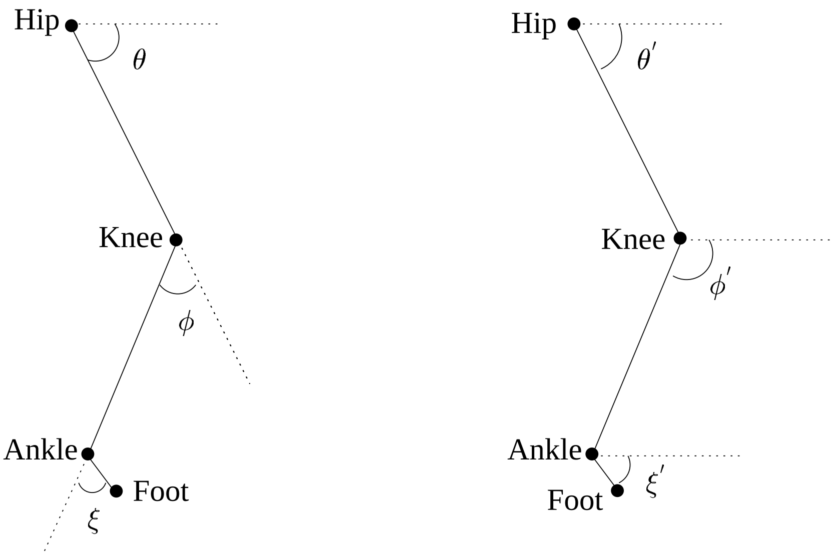


FIGURE A.1: Angles of rotations of leg joints are redefined relative to horizontal plane.

generation, a synthetic person walks along the x-axis from left to right. The lateral and vertical movements take place in xz and xy planes. At rest, the pelvis lies in the plane $z = 0$. To simplify computations, of joint coordinates, all angles were redefined relative to the horizontal plane (see Figure A.1). All angles are negative as these lie in the fourth quadrant. The following equations recalculate angles with respect to horizontal plane.

$$\begin{aligned} \theta' &= -\theta \\ \phi' &= -(\pi - \theta) \\ \xi' &= -(\pi - (\theta + \phi + \xi)) \end{aligned}$$

The method is based on the observation that at any time one of the legs is in supporting phase while the other is swinging. For the supporting leg, either heel or foot is always on the ground. The sequence is :

- Determine which of the legs is in support phase. The following relationship between previous, current and next instances of hip rotation angle at any given time determines the phase.

$$\begin{array}{ll} \theta'_{t+1} > \theta'_t > \theta'_{t-1} & \text{Swinging phase} \\ \theta'_{t+1} < \theta'_t < \theta'_{t-1} & \text{Supporting phase} \\ \theta'_{t+1} < \theta'_t > \theta'_{t-1} & \text{Swing to support changeover} \\ \theta'_{t+1} > \theta'_t < \theta'_{t-1} & \text{Support to swing changeover} \end{array}$$

- For the supporting leg, determine if heel or foot is on ground. This is done by determining which of the two is closest to the ground.
  - If heel is on ground, generate coordinates of foot.

$$\begin{aligned} x^{\text{support}}_{\text{heel}}(t) &= x^{\text{support}}_{\text{heel}}(t-1) \\ y^{\text{support}}_{\text{heel}}(t) &= y_{\text{ground}} \\ z^{\text{support}}_{\text{heel}}(t) &= z^{\text{support}}_{\text{heel}}(t-1) \\ x^{\text{support}}_{\text{foot}}(t) &= x^{\text{support}}_{\text{heel}}(t) + l_{\text{foot}} \cos(\xi'(t)) \\ y^{\text{support}}_{\text{foot}}(t) &= y^{\text{support}}_{\text{heel}}(t) - l_{\text{foot}} \sin(\xi'(t)) \\ z^{\text{support}}_{\text{foot}}(t) &= z^{\text{support}}_{\text{heel}}(t) \end{aligned}$$

  - If foot is on ground, generate coordinates of heel.

$$\begin{aligned} x^{\text{support}}_{\text{foot}}(t) &= x^{\text{support}}_{\text{foot}}(t-1) \\ y^{\text{support}}_{\text{foot}}(t) &= y_{\text{ground}} \\ z^{\text{support}}_{\text{foot}}(t) &= z^{\text{support}}_{\text{foot}}(t-1) \\ x^{\text{support}}_{\text{heel}}(t) &= x^{\text{support}}_{\text{foot}}(t) - l_{\text{foot}} \cos(\xi'(t)) \\ y^{\text{support}}_{\text{heel}}(t) &= y^{\text{support}}_{\text{foot}}(t) + l_{\text{foot}} \sin(\xi'(t)) \\ z^{\text{support}}_{\text{heel}}(t) &= z^{\text{support}}_{\text{foot}}(t) \end{aligned}$$

- Generate coordinates of knee and hip.

$$\begin{aligned} x^{\text{support}}_{\text{knee}}(t) &= x^{\text{support}}_{\text{heel}}(t) - \frac{l_{\text{leg}}}{2} \cos(\phi'(t)) \\ y^{\text{support}}_{\text{knee}}(t) &= y^{\text{support}}_{\text{heel}}(t) + \frac{l_{\text{leg}}}{2} \sin(\phi'(t)) \\ z^{\text{support}}_{\text{knee}}(t) &= z^{\text{support}}_{\text{heel}}(t) \\ x^{\text{support}}_{\text{hip}}(t) &= x^{\text{support}}_{\text{knee}}(t) - \frac{l_{\text{leg}}}{2} \cos(\theta'(t)) \\ y^{\text{support}}_{\text{hip}}(t) &= y^{\text{support}}_{\text{knee}}(t) + \frac{l_{\text{leg}}}{2} \sin(\theta'(t)) \\ z^{\text{support}}_{\text{hip}}(t) &= z^{\text{support}}_{\text{knee}}(t) \end{aligned}$$

- Add lateral swing and vertical swing to determine the hip position of the swinging leg.

$$\begin{aligned}
x_{\text{hip}}^{\text{swing}}(t) &= x_{\text{hip}}^{\text{support}}(t) + l_{\text{pelvis}} \sin(\omega_{\text{lateral}} t) \\
z_{\text{hip}}^{\text{swing}}(t) &= z_{\text{hip}}^{\text{support}}(t) - l_{\text{pelvis}} \cos(\omega_{\text{lateral}} t) \\
y_{\text{hip}}^{\text{swing}}(t) &= y_{\text{hip}}^{\text{support}}(t) + l_{\text{pelvis}} \sin(\omega_{\text{vertical}} t)
\end{aligned}$$

- Generate pelvis coordinates

$$\begin{aligned}
x_{\text{pelvis}}(t) &= \frac{x_{\text{hip}}^{\text{support}}(t) + x_{\text{hip}}^{\text{swing}}(t)}{2} \\
y_{\text{pelvis}}(t) &= \frac{y_{\text{hip}}^{\text{support}}(t) + y_{\text{hip}}^{\text{swing}}(t)}{2} \\
z_{\text{pelvis}}(t) &= \frac{z_{\text{hip}}^{\text{support}}(t) + z_{\text{hip}}^{\text{swing}}(t)}{2}
\end{aligned}$$

- Generate positions of knee, ankle and foot joints of the swinging leg.

$$\begin{aligned}
x_{\text{knee}}^{\text{swing}}(t) &= x_{\text{hip}}^{\text{swing}}(t) + \frac{l_{\text{leg}}}{2} \cos(\theta'(t)) \\
y_{\text{knee}}^{\text{swing}}(t) &= y_{\text{hip}}^{\text{swing}}(t) - \frac{l_{\text{leg}}}{2} \sin(\theta'(t)) \\
z_{\text{knee}}^{\text{swing}}(t) &= z_{\text{hip}}^{\text{swing}}(t)
\end{aligned}$$

$$\begin{aligned}
x_{\text{heel}}^{\text{swing}}(t) &= x_{\text{knee}}^{\text{swing}}(t) + \frac{l_{\text{leg}}}{2} \cos(\phi'(t)) \\
y_{\text{heel}}^{\text{swing}}(t) &= y_{\text{knee}}^{\text{swing}}(t) - \frac{l_{\text{leg}}}{2} \sin(\phi'(t)) \\
z_{\text{heel}}^{\text{swing}}(t) &= z_{\text{knee}}^{\text{swing}}(t)
\end{aligned}$$

$$\begin{aligned}
x_{\text{foot}}^{\text{swing}}(t) &= x_{\text{heel}}^{\text{swing}}(t) + l_{\text{foot}} \cos(\xi'(t)) \\
y_{\text{foot}}^{\text{swing}}(t) &= y_{\text{heel}}^{\text{swing}}(t) - l_{\text{foot}} \sin(\xi'(t)) \\
z_{\text{foot}}^{\text{swing}}(t) &= z_{\text{heel}}^{\text{swing}}(t)
\end{aligned}$$

The advantage of this method is that it adds a vertical swing to the pelvis without any need to compute it. Also the pelvis exhibits acceleration and retardation. This makes this model more realistic. The limitation is that the double support phase is not modelled properly. As a result, some of the sequences result in both feet off the ground or one of the feet beneath the ground. Such sequences were not used in the subsequent results. An example of synthetic sequence generated by the above method is shown in Figure 3.2.

The arm motion is modelled the same way. But the angles of rotations are best guesses because of non-availability of real data. However, arm motion is not an essential part of human motion as people may let both arms swing or swing only one arm, or fold them. Hence only leg data is used in the studies done.

Appendix B

Optimisation

A generic optimisation problem is minimisation or maximisation of a function with or without constraints. As solution space of maximisation of a function $f(\mathbf{x})$ is equivalent to negative of that of minimisation of $-f(\mathbf{x})$, we will discuss minimisation only.

B.1 Unconstrained Optimisation

Let us consider determination of minima of a function $f(x)$ without any constraints. Suitable conditions can be obtained by considering Taylor series expansion of the function.

For a one variable function $f(x)$, Taylor series expansion about the point x^* is given by

$$f(x^* + \Delta x) = f(x^*) + \left.\frac{df}{dx}\right|_{x=x^*} \Delta x + \frac{1}{2}\left.\frac{d^2 f}{dx^2}\right|_{x=x^*} (\Delta x)^2 + \ldots \tag{B.1}$$

As Δx is small, higher order terms can be ignored. Generally first order and second order terms are used to ascertain if the point x^* is a minimising point, i.e. $f(x^* + \Delta x) > f(x)$ and $f(x^* - \Delta x) > f(x)$. This can be true only if first order derivative $\frac{df}{dx}$ is zero, its sign being dependent on sign of Δx and second order derivative $\frac{d^2 f}{dx^2}$ is negative. However, if the second order derivative is also zero, higher order terms are used to determine the minimising point. Thus the necessary and sufficient conditions for minimising point are

- The very first non-zero term should be an even order derivative, i.e. $\frac{d^2 f}{dx^2}, \frac{d^4 f}{dx^4}, \frac{d^6 f}{dx^6}$ and it should be negative.

For a multi-variable function $f(\mathbf{x})$, $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, Taylor series expansion is

$$\begin{aligned}
f(\mathbf{x}^* + \Delta\mathbf{x}) &= f(\mathbf{x}^*) + g^*\Delta\mathbf{x} + \Delta\mathbf{x}^T H^* \Delta\mathbf{x} + \ldots \\
g^* &= \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*} \\
&= \left[\left.\frac{\partial f}{\partial x_1}\right|_{x_1=x_1^*}, \left.\frac{\partial f}{\partial x_2}\right|_{x_2=x_2^*}, \ldots, \left.\frac{\partial f}{\partial x_n}\right|_{x_n=x_n^*} \right]^T \\
H^* &= \left(\begin{array}{cccc} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \ldots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \ldots & \frac{\partial^2 f}{\partial x_n^2} \end{array} \right)_{\mathbf{x}=\mathbf{x}^*}
\end{aligned}$$

Sufficient conditions for local minima at $\mathbf{x}^*$ is

- $\nabla f|_{\mathbf{x}=\mathbf{x}^*} = 0$
- $H|_{\mathbf{x}=\mathbf{x}^*}$ is positive definite, i.e. $\Delta\mathbf{z}^T \cdot H \cdot \Delta\mathbf{z} > 0$ for all $\Delta\mathbf{z}$

B.2 Constrained Optimisation

Let us consider an optimisation problem with a single constraint.

$$\begin{aligned}
\text{minimise} \quad & y_o = f(x_1, x_2, \ldots, x_n) \\
\text{subject to} \quad & g(x_1, x_2, \ldots, x_n) = \gamma
\end{aligned}$$

As the constraint shows dependency among the input variables, this dependency can be removed by expressing x_n in terms of other variables and substituting in the objective function. This gives

$$\begin{aligned}
x_n &= H(x_1, x_2, \ldots, x_{n-1}) \\
y_0 &= f(x_1, x_2, \ldots, x_{n-1}, H(x_1, x_2, \ldots, x_{n-1}))
\end{aligned}$$

y_0 is now unconstrained and the minima can be obtained by setting

$$\begin{aligned}
\partial y_0/\partial x_j &= 0 \;\; j = 1, 2, \ldots, n-1 \\
\Rightarrow \frac{\partial f}{\partial x_j} + \frac{\partial f}{\partial x_n}\frac{\partial H}{\partial x_j} &= 0 \;\; j = 1, 2, \ldots, n-1 \qquad \text{(B.2)}
\end{aligned}$$

However, from $g(x_1, x_2, \ldots, x_n) = \gamma$, we get

$$\begin{aligned}
\frac{\partial g}{\partial x_j} + \frac{\partial g}{\partial x_n}\frac{\partial H}{\partial x_j} &= 0 \;\; j = 1, 2, \ldots, n-1 \\
\Rightarrow \frac{\partial H}{\partial x_j} &= -\left. \frac{\partial g}{\partial x_j} \middle/ \frac{\partial g}{\partial x_n} \right. \;\text{ if } \frac{\partial g}{\partial x_n} \neq 0 \;\; j = 1, 2, \ldots, n-1 \qquad \text{(B.3)}
\end{aligned}$$

Substituting the results of equation B.3 in equation B.2, we get

$$\begin{aligned}\frac{\partial y_0}{\partial x_j} = \frac{\partial f}{\partial x_j} - \left[\frac{\partial f}{\partial x_n} \cdot \frac{\partial g}{\partial x_j}\Big/ \frac{\partial g}{\partial x_n}\right] &= 0 \;\; j = 1, 2, \ldots, n-1 \\ \Rightarrow \frac{\partial f}{\partial x_j} - \alpha \frac{\partial g}{\partial x_j} &= 0 \quad j = 1, 2, \ldots, n-1 \\ \text{where} \quad \alpha = \frac{\partial f}{\partial x_n}\Big/ \frac{\partial g}{\partial x_n} \quad &\text{is Lagrange multiplier.}\end{aligned} \tag{B.4}$$

Now consider a new unconstrained problem derived from the constrained one by including Lagrange multipliers as follows

$$y_0 = f(x_1, x2, \ldots, x_n) - \alpha\left[g(x_1, x_2, \ldots, x_n) - \gamma\right]$$

Setting the derivative of the new objective function with respect to x_j equal to zero (a necessary condition to derive minima)

$$\frac{\partial y_0}{\partial x_j} = \frac{\partial f}{\partial x_j} - \alpha \frac{\partial g}{\partial x_j} = 0$$

gives the necessary condition for optimisation of the constrained problem. Also setting the derivative with respect to α to zero

$$\frac{\partial y_0}{\partial \alpha} = g(x_1, x_2, \ldots, x_n) - \gamma = 0$$

results in the original constraint. Thus Lagrangian multipliers give us a way to treat constrained problems in the same vein as unconstrained ones. Number of Lagrangian multipliers are same as number of constraints.

B.3 General Constrained Optimisation

In the previous discussion, the constraints were equality constraints. For a general nonlinear optimisation problem

$$\begin{aligned} \text{Minimise} \quad & f(\mathbf{x}) & \\ \text{subject to} \quad & h_j(\mathbf{x}) = 0 & j = 1, 2, \ldots, m \\ & g_j(\mathbf{x}) \geq 0 & j = m+1, \ldots, p \end{aligned}$$

Karush-Kuhn-Tucker(KKT) conditions

$$
\begin{aligned}
h_j(\mathbf{x}) = 0 \;\; j = 1, 2, \ldots, m \\
g_j(\mathbf{x}) \geq 0 \;\; j = m+1, \ldots, p \\
\beta_j \cdot g_j(\mathbf{x}) = 0 \\
\beta_j \geq 0 \\
\frac{\partial f(\mathbf{x})}{\partial x_k} + \sum_{j=1}^{m} \alpha_i \left[\frac{\partial h_j(\mathbf{x})}{\partial x_k} \right] - \sum_{j=m+1}^{p} \beta_j \left[\frac{\partial g_j(\mathbf{x})}{\partial x_k} \right] = 0 \quad k = 1, 2, \ldots, N
\end{aligned}
$$

are sufficient to determine local minima.

Appendix C

Machine Classifiers

Three different types of machine classifiers were used. These are

1. k nearest neighbour (k-NN)
2. Artificial neural network (ANN)
3. Support vector machine (SVM)

Initial classification studies to assess the feasibility of the problem were done using non-recurrent feed-forward back propagation neural network. These results were, then, checked against SVM formulation. The following paragraphs describe the workings of these classifiers in detail. k-NN classifiers were used to determine the relationship of the data-points in the feature space.

C.1 k-Nearest Neighbour Classifier

A k-NN classifier is the simplest to implement and most widely used for pattern classification. To ascertain the class of a given point x_t, its distances from all the points constituting input feature space, $\psi = \{x_1, x_2, ..., x_S\}$ are calculated. Of the first k points in ψ, closest to x_t in terms of distance parameter, one is randomly chosen. The class assigned to the point x_t is same as the chosen point. This classifier works well in the circumstances where the input space consists of clusters, a different cluster for a different class. Therefore, a given point is closer to the one contained in the cluster of its own class. The distance parameter to determine the closeness between two points is generally Euclidean distance in a multi-dimensional space.

The complexity of the above Algorithm is $S^2 \times D^2$, where S is number of points in input space and D is dimensionality of each point. The first term (S^2) indicates number of

Algorithm 1 k-NN Classifier.

```
Require: Input data-points, class of each data-point, value for $k$.
  Number of input data-points: $S$. Dimensionality of each data-point: $D$.

  for each data-point $p$ do
    Determine its distances from all other points.
    Sort the distances in ascending order.
    Of first $k$ distances, select one randomly.
    Assign the class of the chosen data-point.
  end for
```

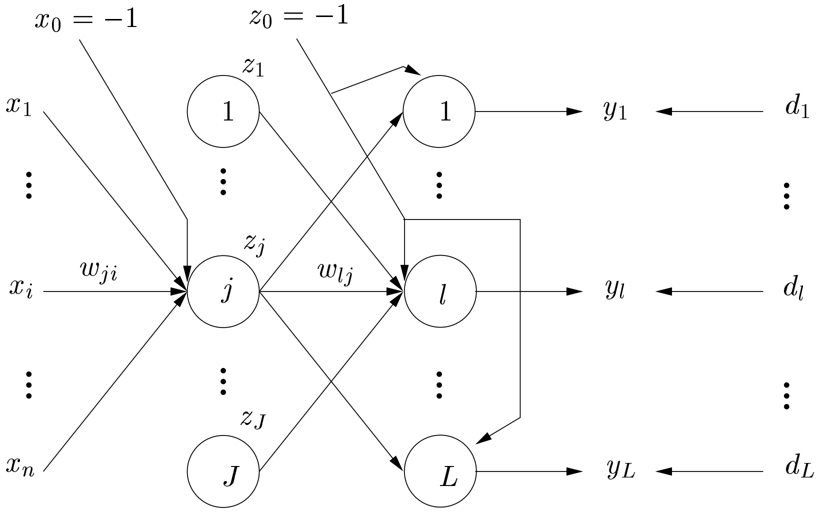


FIGURE C.1: Architecture of a multilayer, fully-connected, feed-forward neural network employing back propagation learning. $\mathbf{x}, \mathbf{w}, \mathbf{y}$ and $\mathbf{d}$ represent input, weight, output and target values respectively. Each node also has a negative bias at its input.

distances to be computed and the second term (D^2) is the cost of each distance computation for Euclidean metric, which is the most common distance metric. However, the actual computation time (not the computation complexity) can be reduced by maintaining a list of k shortest distances for (each data-point) and updating this list only when a better metric value is obtained. Any value less than the largest value in the list will replace the latter. Also a distance computation is discarded in middle if the partial metric value exceeds the largest of values in the list. Algorithm 1 describes the basic k-NN algorithm.

C.2 Neural Network Classifier

The neural network classifier used in this thesis is a non-recurrent, multilayer, feed-forward network with back-propagation learning. Such a network is generally fully connected and has two/three layers. The architecture of such a network is shown in

Figure C.1. The network consists of many nodes. Each node has many incoming and outgoing connections, each connection having some weight. These nodes can be divided into three categories – input, hidden and output node. Each input node has a single incoming connection and is connected to one parameter of the input data. Output node has only one outgoing connection. Except for input node, the inputs to a node are summed together and squashed by application of a sigmoid or similar function.

Algorithm 2 Neural network with back-propagation learning.

```
Require: Dimensionality of each input ($N$), preset threshold.
  Set $\lambda = \frac{1}{\sqrt{N}}$
  Initialise all weights $\mathbf{W}$ to small random values within the range $[-\lambda, \lambda]$.
  Randomly select a training pattern $(\mathbf{x^p}, \mathbf{t^p})$, where $\mathbf{x^p}$ and $\mathbf{t^p}$ denote $p$th training
  input and target vectors.
  repeat
    $O_j^0 = x_j$ /* Assign data to input layer */
    for each hidden layer $q = 1, \ldots, Q-1$ do
      for each unit $j = 1, \ldots, J$ do
        $H_j^q = \sum_i O_i^{q-1} w_{ji}^q$ /* Compute net input */
        $O_j^q = f\left(\sum_i O_i^{q-1} w_{ji}^q\right)$ /* Compute output */
      end for
    end for
    for each unit $j = 1, \ldots, L$ in the output layer $q = Q$ do
      $H_j^q = \sum_i O_i^{q-1} w_{ji}^q$ /* Compute net input */
      $O_j^q = f\left(\sum_i O_i^{q-1} w_{ji}^q\right)$ /* Compute output */
      $\delta_j^q = (O_j^q - t_j^p) f'(H_j^q)$ /* Compute deltas */
    end for
    for all hidden layers $q = Q-1, \cdots, 1$ do
      $\delta_j^{q-1} = f'(H_j^{q-1}) \sum_i \delta_i^q w_{ji}^q$ /* back-propagate error */
    end for
    for all weights do
      $\Delta w_{ji}^q = \eta \delta_i^q O_j^{q-1}$
      $w_{ji}^q = w_{ji}^q + \Delta w_{ji}^q$ /* Adjust weights */
    end for
  until Classification accuracy for successive five epochs remain above the preset threshold.
```

In this discussion, we will refer to the connections as a layer. The first layer of weighted connections connects input nodes with hidden nodes and the last layer connects hidden nodes to output nodes. The number of hidden node layers is $Q-1$, where Q is number of layers. There are some connections from output nodes to hidden nodes. These connections are used for back propagation of error.

The network is first trained to learn the classification patterns. Once trained, the network stores the information about these patterns in weights of the connections. The entire input data is presented to the network in a random order. This constitutes a epoch. For a given input, output is computed and compared with the target value associated with this input. This target value is encoding of the class to which this input belongs. If the output value does not match the target, the weights of the output nodes are adjusted, through application of some gradient-based error criterion function, to bring these closer. However, no target values are available for the hidden nodes. The error is, then, back-propagated and weights of successive layers are updated. The training is continued until the classification accuracy remains higher than a certain threshold (95%) for five successive epochs. The actual algorithm used is as discussed in Patterson (1996) and is detailed in Algorithm 2.

C.3 Support Vector Machine Classifier

Support vector machine is a method of separating two classes. This method determines an optimal separating plane in feature space. Unlike conventional approaches, this method aims at minimising test error rather than training error. A brief introduction to structural risk minimisation (Burges 1998; Schölkopf, Burges, and Simola 1999) has been presented to contrast this methodology with conventional approaches to machine learning.

C.3.1 Machine Learning

Learning is a process of building a model from incomplete information so that it can be used to predict the outcome of some unknown input as accurately as possible. In all learning methods, the underlying assumption is that observations used for training and the inputs to be expected in future along with associated outputs are generated from same probability distribution $P(\mathbf{x}, y)$. If $\mathbf{x} \in \mathbb{R}^N$ is the input, $y \in \mathbb{R}$ is the output and α is a set of parameters of the function f to be estimated, learning is determining solution to the equation

$$f\left(\mathbf{x}, \alpha\right) = y \tag{C.1}$$

from a finite number of observations or training data $(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \mathbb{R}$ such that f will correctly classify unseen examples or test data. A learning machine must choose from a given set of solution functions $f_1\left(\mathbf{x}, \alpha\right), \ldots, f_2\left(\mathbf{x}, \alpha\right)$ the one which is the best approximate.

C.3.1.1 Empirical Risk Minimisation

In conventional approaches as in neural network, learning is achieved through minimisation of empirical risk or training error. Equation (C.1) has two unknowns, f and α, the approach used is to fix f and then adjust α in such a manner that some objective function of training error (mean squared error, sum of squared errors) is minimised. However, as the error criterion is an n-dimensional surface, the search for a global minimum may get stuck in some local minima or flat region. Also for any finite training data, it is possible to find two different functions f_1 and f_2 satisfying $f_1(\mathbf{x}_i, \alpha_1) = f_2(\mathbf{x}_i, \alpha_2)\,, \quad i = 1, \ldots, l$ and $f_1(\mathbf{x}_{l+1}, \alpha_1) \neq f_2(\mathbf{x}_{l+1}, \alpha_2)$. So these two functions agree on the entire training set but may not agree on an unknown input. In other words, even when two functions have same training error, they may not exhibit same test error. Hence only the minimising training error or the empirical risk

$$R_{emp}(\alpha) \quad = \quad \frac{1}{2l}\sum_{i=1}^{l}|f(\mathbf{x_i}) - y_i|$$

does not imply a small test error or risk.

$$R(\alpha) \quad = \quad \int \frac{1}{2}|f(\mathbf{x}) - y|\,dP(\mathbf{x}, y)$$

$P(\mathbf{x}, y)$ is the probability distribution from which $(\mathbf{x}_i, y_i)$ are drawn.

C.3.1.2 Structural Risk Minimisation (SRM)

Statistical theory or VC theory put forward by Vapnik (1999) defined the following bound on the empirical risk with a probability of $(1 - \eta)$

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h\left(\log\frac{2l}{h} + 1\right) - \log\frac{\eta}{4}}{l}} \tag{C.2}$$

where h is the VC dimension of a family of functions $\{f(\mathbf{x}, \alpha)\}$. The VC dimension (Section C.3.1.3) of a family of functions is the maximum number of points that can be separated in all possible ways by one or the other member function. If h is known, the best choice of α can be calculated. VC dimension is indicative of learning capacity of a family of functions.

The second term in equation (C.2) is known as the confidence interval. For a given training set (l fixed) and given η, this term monotonically increases as h increases. However, large h means large learning capacity and hence low training error. Structural risk minimisation (SRM) is then a trade-off between accuracy (first term) and complexity of the approximation (second term) (see Figure C.2). As the confidence interval depends on the chosen family of functions and the empirical risk depends on a particular member function of the class, SRM aims to find the subset of the functions that minimises actual bound. VC dimension h being an integer imposes a structure on the class of functions by partitioning the class into nested subsets of the functions (Figure C.3). So if S_1, S_2,

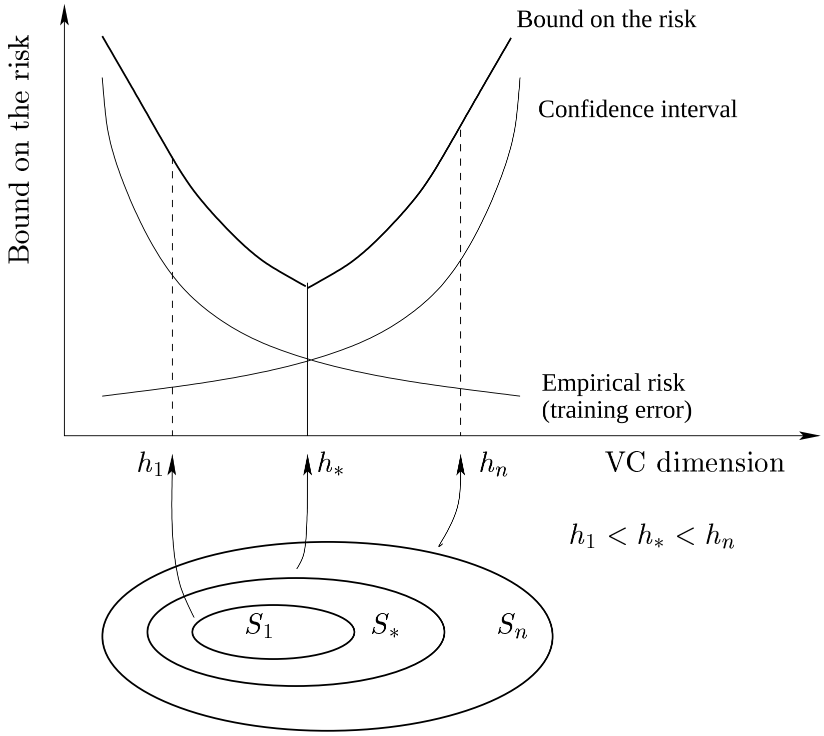


FIGURE C.2: In structural risk minimisation, learning capacity of a machine can be traded for bound on its test error. As the learning capacity increases, so does the bound and hence generalising capability decreases.

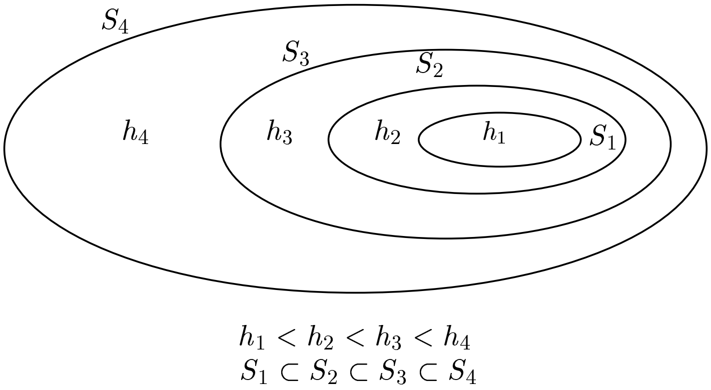


FIGURE C.3: Learning capacity partitioned by VC dimension.

...define these subsets on a class of functions $\{f(\mathbf{x}, \alpha)\}$ such that

$$S_1 \subset S_2 \subset \dots S_n \subset \dots$$
$$h_1 \leq h_2 \dots \leq h_n \leq \dots$$

where h_1, h_2 are VC dimensions of structures S_1 and S_2 respectively, SRM chooses an appropriate element S_k that minimises the bound. The Figure C.2 illustrates the principle.

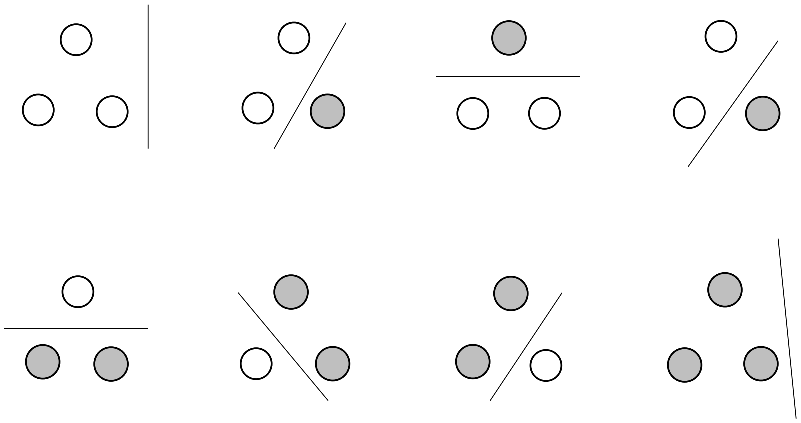


FIGURE C.4: VC dimension of a family of lines.

C.3.1.3 VC Dimension

The VC dimension of a family or class of functions is defined as the maximum number of points that can be separated in all possible ways by one member function or another. The VC dimension of a line is three. As can be seen from Figure C.4, different member lines from a set of lines can separate 3 points in all possible ways. However, four points cannot be separated using lines alone. This is because these points may not be coplanar. Hence a family of 3D planes is needed for separating 4 points. Similarly for a n-dimensional input, as in SVM, an n-dimensional hyperplane is required.

C.3.2 Linear Support Vector Machines – Separable Case

As stated earlier, SVM essentially solves a bi-class problem. These two classes are given labels $+1$ and -1 respectively. We will refer instances of the first class as positive and those of the other class as negative in the following discussion. The class of functions considered is a family of hyperplanes described by the following equation, $\mathbf{w}$ and b being unknowns.

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \quad \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}$$

As many planes can classify the data correctly (Figure C.5), the problem is recast to determining the plane with largest orthogonal distance from nearest positive and negative instances. Such a plane is referred to as *the optimal hyperplane*. If d_+ and d_- denote these distances, all positive and negative instances will satisfy the following equations.

$$\begin{aligned}
\mathbf{w} \cdot \mathbf{x}_i + b &\geq d_+ \text{ for } y_i = +1 \\
\mathbf{w} \cdot \mathbf{x}_i + b &\leq d_- \text{ for } y_i = -1
\end{aligned}$$

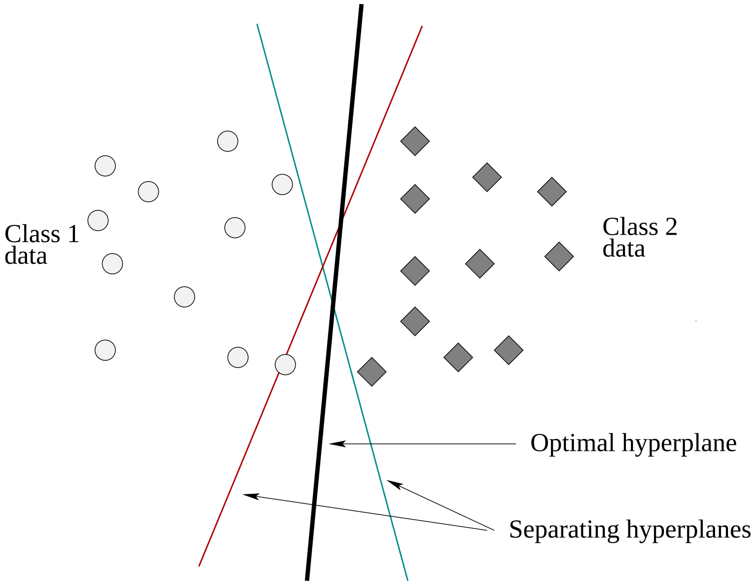


FIGURE C.5: Optimal separating hyperplane.

Although many planes can partition the data correctly, only one plane is optimal in terms of distance between nearest instances of two classes.

The point nearest to the plane will satisfy the following equations.

$$\mathbf{w} \cdot \mathbf{x}_1 + b = d_+ \tag{C.3}$$

$$\mathbf{w} \cdot \mathbf{x}_2 + b = d_- \tag{C.4}$$

Subtracting equation (C.4) from equation (C.3) gives the margin

$$\begin{aligned} \mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2) &= (d_+ - d_-) \\ \Rightarrow \text{margin} &= \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_1 - \mathbf{x}_2) = \frac{(d_+ - d_-)}{\|\mathbf{w}\|} \end{aligned}$$

The optimal plane should not be biased toward any class and should be symmetrical to both the classes. This can be true if $d_+ = -d_- = d$.

$$\Rightarrow \text{margin} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_1 - \mathbf{x}_2) = \frac{2d}{\|\mathbf{w}\|}$$

However, as d is merely a scaling factor, it can be replaced by unity and this gives rise to the following canonical form (Figure C.6).

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b &\geq +1 \text{ for } y_i = +1 \\ \mathbf{w} \cdot \mathbf{x}_i + b &\leq -1 \text{ for } y_i = -1 \end{aligned}$$

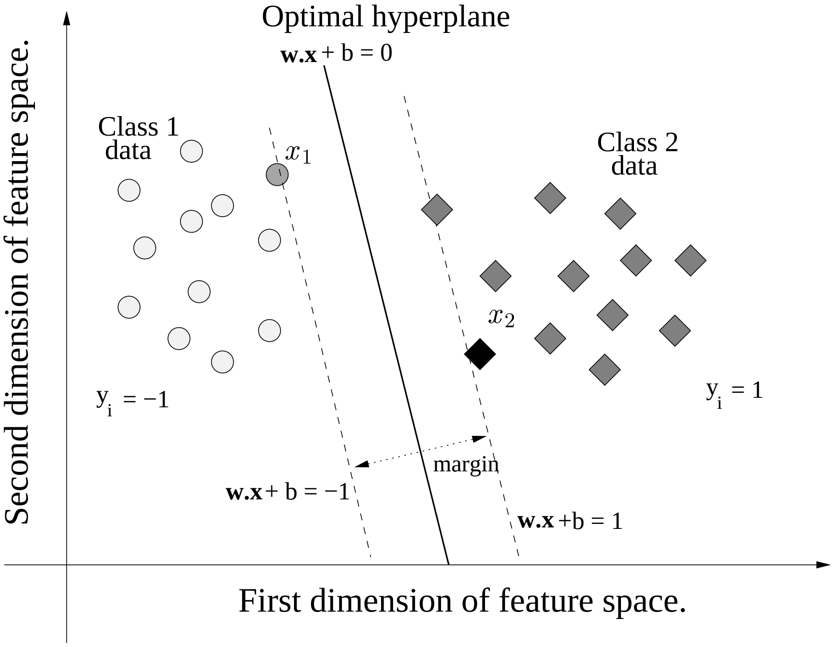


FIGURE C.6: Support vectors.

The above two equations can be combined into one as

$$y_i\left(\mathbf{w}\cdot\mathbf{x}_i+b\right)-1\geq 0 \quad \forall i$$

Now determination of the optimal hyperplane is simply an optimisation problem and can be stated as

$$\begin{aligned}\text{minimise}\quad \tau\left(\mathbf{w}\right) &= \frac{1}{2}\|\mathbf{w}\|^2\\ \text{subject to}\quad y_i\left(\mathbf{w}\cdot\mathbf{x}_i+b\right) &\geq 1, \quad \mathrm{i}= \ 1,\ldots,l\end{aligned}$$

As the above is a constrained problem, it can be transformed into an unconstrained optimisation problem using Lagrangian multipliers, α in the following equations (refer Appendix B.2).

$$L\left(\mathbf{w},b,\alpha\right)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{l}\alpha_i\left(y_i\cdot\left(\mathbf{x}_i\cdot\mathbf{w}+b\right)-1\right) \tag{C.5}$$

Now the problem can be solved by equating derivatives with respect to primal variables $\mathbf{w}$ and b respectively

$$\begin{aligned}&\frac{\partial}{\partial\mathbf{w}}L\left(\mathbf{w},b,\alpha\right)=0\\ &\Rightarrow \mathbf{w}=\sum_{i=1}^{l}\alpha_i y_i\mathbf{x}_i\end{aligned} \tag{C.6}$$

$$\begin{aligned}\frac{\partial}{\partial b} L(\mathbf{w}, b, \alpha) &= 0 \\ \Rightarrow \sum_{i=1}^{l} \alpha_i y_i &= 0\end{aligned} \tag{C.7}$$

Substituting the results of equation (C.6) and equation (C.7) in equation (C.5), we arrive at the Wolfe dual of the original problem, which can be solved by any quadratic programming method.

$$\begin{aligned}\text{maximise} \quad W(\alpha) &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ \text{subject to} \quad \alpha_i &\geq 0, i = 1, \ldots, l \\ \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i &= 0\end{aligned} \tag{C.8}$$

Karush-Kuhn-Tucker (KKT) complementary conditions (Appendix B.5), expressed as the following equation, are sufficient and necessary for the solution to be optimal.

$$\alpha_i [y_i (\mathbf{w} \cdot x_i + b) - 1] = 0, \quad i = 1, \ldots, l \tag{C.9}$$

Referring to equation (C.7), we find that the optimal solution **w** consists of input vectors for which the corresponding Lagrangian multiplier is non-zero. These input vectors are also called support vectors. As these satisfy KKT conditions, the following should be true for each of them.

$$y_i (\mathbf{w} \cdot x_i + b) - 1 = 0, \quad i = 1, \ldots, l$$

In short, support vectors lie on the margin.

Training phase: Training a support vector machine is solving equation (C.8) and determining α. Then the bias b can be computed from equation (C.9) for any i for which $\alpha_i \neq 0$. But it is numerically safer to take an average.

Testing phase: Testing is done by determining the class of each test vector by using the following decision function.

$$f(\mathbf{x}) = sgn \left(\sum_{i=1}^{l} y_i \alpha_i \cdot (\mathbf{x} \cdot \mathbf{x}_i) + b \right)$$

C.3.3 Support Vector Machines – Non-separable case

If the input data are not linearly separable, as most practical problems exhibit, the above method will fail. In this case, the problem can be solved either by incorporating some penalty function or by mapping the input data to a higher dimensional space.

C.3.3.1 Penalty Function

If the input data are not linearly separable, any hyperplane is bound to result in some misclassification error. The optimal plane, in this case, is one which tends to maximise the margin and minimise misclassification error at the same time. As an increase (decrease) in margin means an increase (decrease) in the misclassification rate, the best an optimal hyperplane can do is to find the best trade-off between the two. A commonly used penalty function is $f_\sigma(\xi) = \sum_i^l {\xi_i}^\sigma, \ \sigma > 0$ where slack variables $\xi_i \geq 0, i = 1, \ldots, l$ control the misclassification rate. Generally $\sigma = 1$ is set to simplify the numerical calculations. The equations that each input data should satisfy now become

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b &\geq +1 - \xi_i \text{ for } y_i = +1 \\ \mathbf{w} \cdot \mathbf{x}_i + b &\leq -1 + \xi_i \text{ for } y_i = -1 \\ \xi_i &\geq 0 \quad \forall i \end{aligned}$$

The objective function is modified to

$$\begin{aligned} \text{minimise } \tau(\mathbf{w}, \xi) &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \xi_i \\ \text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) &\geq 1, \quad \text{i} = \ 1, \ldots, l \end{aligned}$$

The corresponding Lagrangian is given by

$$L(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i - \sum_{i=1}^{l} \alpha_i\{y_i \cdot (\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i\} - \sum_i \beta_i\xi_i$$

where α and β are Lagrangian multipliers. Setting the derivative of the above Lagrangian with respect to $\mathbf{w}$, b, ξ equal to zero

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \alpha, \beta) = 0 \\ &\Rightarrow \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \\ &\frac{\partial}{\partial b} L(\mathbf{w}, b, \alpha, \beta) = 0 \\ &\Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0 \\ &\frac{\partial}{\partial \xi_i} L(\mathbf{w}, b, \alpha, \beta) = 0 \\ &\Rightarrow \alpha_i + \beta_i = C \end{aligned}$$

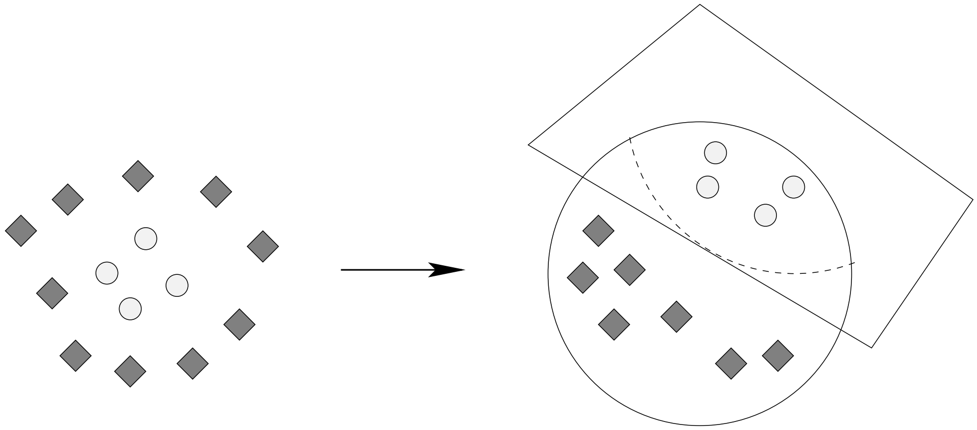


FIGURE C.7: Kernels may transform linearly non-separable input data into linearly separable data in a higher dimensional feature space.

The corresponding Wolfe-dual equation is given by

$$\begin{aligned}
\text{maximise } W(\alpha)\, i &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \left(\mathbf{x}_i \cdot \mathbf{x}_j\right) \\
\text{subject to } 0 &\leq \alpha_i \\
\alpha_i &\leq C \\
\sum_{i=1}^{l} \alpha_i y_i &= 0, \ \text{i} = 1 \ , \ldots, l
\end{aligned}$$

C.3.3.2 Kernels

If the input space is mapped to a higher dimensional space, also called feature space, through a nonlinear mapping ϕ, the mapped data may be linearly separable (Figure C.7). However, as the input data appear as dot products in the equations that need to be solved, one may use kernel methods without any explicit mapping. A kernel function should satisfy the following

$$k(\mathbf{x}, \mathbf{x}_i) = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i)$$

Some of the kernels generally used are

- Polynomials of degree d

  $k(\mathbf{x}, \mathbf{x}') = \{(\mathbf{x} \cdot \mathbf{x}') + 1\}^d$

- Two layer neural networks

  $k(\mathbf{x}, \mathbf{x}') = \tanh\left(\kappa\left(\mathbf{x}, \mathbf{x}'\right) + \Theta\right)$, where κ and Θ are gain and threshold respectively.

- Radial basis functions
  $k(\mathbf{x}, \mathbf{x}') = e^{\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}}$

In the work presented in this thesis, only linear and polynomial kernels have been used.

C.3.4 Multi-class Support Vector Machines

Multi-class support vector machines Weston and Watkins(1998 and Weston (1999) can be implemented using binary SVM classifiers. The most common techniques to implement multi-class classification using support vector method are

1. **Error correcting code classifier:** Each class is assigned an n-bit code as per some error correction coding scheme. A classifier is constructed for each bit. Testing is done by determining the code constructed from the output of each classifier. The class whose code is nearest to the computed code is assigned to the test input.

2. **One-against-rest:** If k is the number of classes, k binary class SVM classifiers are constructed, one classifier for each class versus the rest. While training a classifier, the training data-points for the given class are considered as positive instances and the rest as negative instances. During testing, the output of each classifier is computed as per decision function $f_i(\mathbf{x}) = \mathbf{w}_i \cdot \mathbf{x} + b_i, \quad i \in \{1, \ldots, k\}$, where k is the number of classes. The class assigned is the one corresponding to the classifier with maximum output.

$$\begin{aligned} f_i(\mathbf{x}) &= \operatorname{sgn}\left(\mathbf{w}_i \cdot \mathbf{x} + b_i\right), \quad i \in \{1, \ldots, n\} \\ V(i) &= \sum_{i=1}^{n} \delta_{f_i(\mathbf{x}),+1} \\ c &= \arg\max_k [V(k)], \quad k \in \{1, \ldots, n\} \end{aligned}$$

   where $\mathbf{w}$ and b are the weight and bias parameters respectively of the optimal separating plane, $\delta_{m,n}$ is the Kronecker delta, equal to 1 when $m = n$ and equal to 0 otherwise, and $V(i)$ represents the number of votes assigned to class i. In case of tie, the class having the greater value of $\mathbf{w}_i \cdot \mathbf{x} + b_i$ is assigned to the test input.

3. **One-against-one:** In this method only two classes are considered at a time resulting in $k \cdot (k-1)/2$ classifiers. To determine the class of an unknown input, the output of each classifier is computed. If the output of i-versus-j is 1, the counter corresponding to i is incremented. The class with maximum count is assigned to

the test input.

$$
\begin{aligned}
f_{i,j} &= \operatorname{sgn}\left(\mathbf{w}_{i,j} \cdot \mathbf{x} + b_{i,j}\right) \quad \begin{cases} i, j \in \{1, \ldots, n\} \\ j \neq i \end{cases} \\
V(i) &= \sum_{j=1}^{n} \delta_{f_{i,j}(\mathbf{x}),+1} \\
V(j) &= \sum_{i=1}^{n} \delta_{f_{i,j}(\mathbf{x}),-1} \\
c &= \arg\max_{k} [V(k)], \quad k \in \{1, \ldots, n\}
\end{aligned}
$$

In case of a tie, one of the classes is randomly chosen.

Appendix D

Statistical Inference

Statistical inference is the technique of obtaining knowledge about a large population from a small sample. A population is any exhaustive finite or infinite set of units about which inferences are required. For all practical purposes, the population is assumed to be infinite. A sample is a subset of population and should be representative of the population. A small representative sample will yield better estimates with less margin of error than a large non-representative sample. Random sampling is the most widely accepted method of selecting a representative sample. Inferential statistics is based on the assumption that the sample was obtained randomly from the populations. Most of the discussion in following paragraphs is referenced from (Glass and Hopkins 1996).

For the following discussion, a sample will be represented by $X = X_1, X_2, \ldots, X_n$ where n is the size of sample. Size of population will be represented by N. Population mean and variance will be represented by μ and σ^2 respectively. Sample mean and variance will be denoted by $\bar{X}$ and s_X^2 respectively. As variance is square of standard deviation, latter for a sample and a population will be described by s_X and σ. For a sample of size n, the degree of freedom or ν equals $n - 1$. The notation x_i will be used to describe how a given value X_i differs from the population mean μ. Therefore, $x_i = X_i - \mu$.

D.1 Statistics and Parameters

For a given population, any desired characteristic can be described in terms of the descriptive parameters like mean, percentiles, median, variance and standard deviation. For a given sample, these parameters are referred to as statistics. A statistic is an estimate of the respective parameter in the population. If the statistic is unbiased, its expected value is equal to the parameter it estimates. Mean represents the average value. The statistics are representative of how the values are distributed in a given sample. Mean, median are also called as measures of central tendency as these describe average or representative values. Variance is a measure of variability as it is indicative of how the values in a sample depart from the mean.

Mean of a sample and population are given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
$$\mu = \frac{\sum_{i=1}^{N} X_i}{n}$$

A p-th percentile is the value below which $p\%$ of the sample values lie. Thus a p-th perccentile will partition the sample into two sets – the first having values less than the statistic and another with values larger than the statistics. The respective size of the two sets are roughly $\frac{p \cdot N}{100}$ and $\frac{(100-p)\cdot N}{100}$.

Median is the value that divides the respective sample into two partitions of same size, all the members of the first partition having value less than the median and all the members of the second partition having value larger than the median. A median is also equal to 50% percentile and also denoted as P_{50}.

A sampling error is the difference between a statistic of the sample and respective parameter of the population. A small sampling error means that sample statistic is a good estimator of the population.

$$\text{Sampling error} = \text{Statistic - Parameter}$$

For a skewed data distribution, sample median is a better descriptive measure than the sample mean. Mean is preferred as an inferential statistic as the sampling error tends to be smaller for the sample mean.

A variance is a measure of to what extent the values in a sample differ from the mean. A small variance means that most of the values lie close to the mean value. Variance for a sample and a population are computed as

$$\begin{aligned} s_X^2 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\nu^2} \\ &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)^2} \\ \sigma^2 &= \frac{\sum_{i=1}^{n} (X_i - \mu)}{N^2} \end{aligned}$$

For a sample variance, degree of freedom instead of the sample size is used as in this form, sample variance is an unbiased estimator. Although s_X is a biased estimator and underestimates σ, the bias is about 1% for $n = 20$ and hence negligible for most practical purposes.

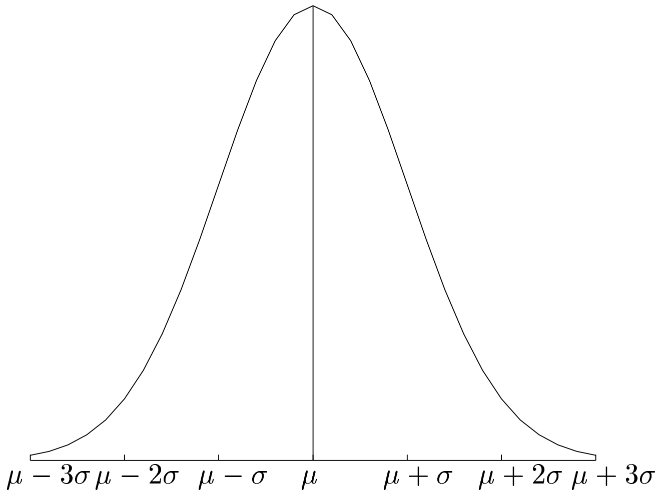


FIGURE D.1: A normal distribution curve.

D.2 Standard Error of the Mean

The standard deviation of the distribution of the sample means is called the standard error of the mean. This will be represented by $\sigma_{\bar{X}}$. For a given population, the standard deviation of its sample means is the deviation from the population mean. Regardless of sample size, the sample means are normally distributed. This is from the central limit theorem which states that the sampling distribution of means from random samples, of size n each, approaches a normal distribution regardless of the shape of the population distribution. This normal distribution has mean μ and variance σ^2/n. The standard error of the mean is related to the population variance as

$$\sigma_{\bar{X}}^2 = \frac{(1 - n/N)\sigma^2}{n}$$

As sample size n is much smaller than population size N, the above equation simplifies to the following one.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

D.3 Population Mean from Sample Mean

As stated earlier, the distribution of sample means is a normal distribution with mean μ and variance $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. If population mean and variance are known, this distribution curve can be obtained for a given value of n. However, in this case, one does not need the population mean estimate. In practise, only one sample mean is available and population mean needs to be estimated from this. As population mean is only an estimate, an associated parameter is the probability that this estimate is in an error.

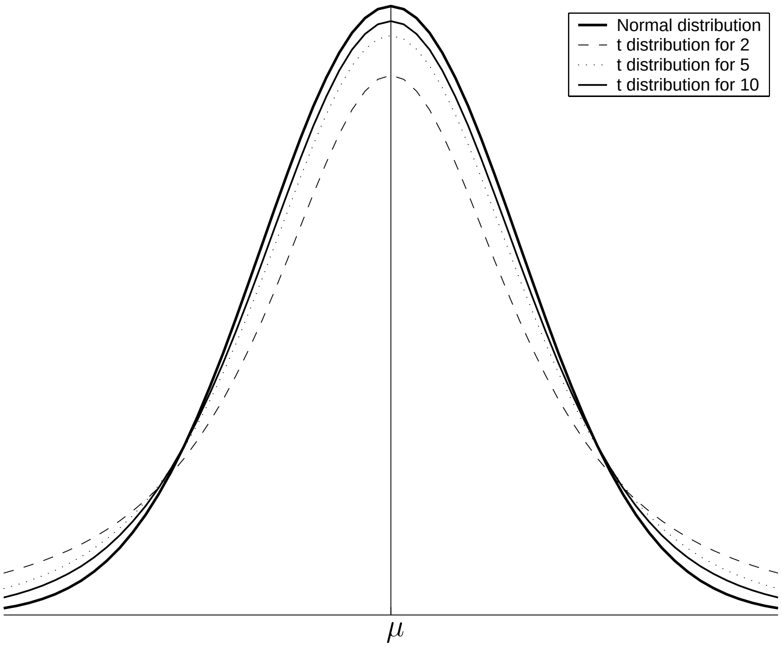


FIGURE D.2: A t-distribution curve.

The latter is called significance level and represents the probability that the population mean estimate is incorrect.

As the distribution of sample means is normal or Gaussian, sample means are distributed around the population mean as shown in Figure D.1. For any normal curve with mean μ and variance σ^2, 68% of the values lie in the interval $(\mu - \sigma, \mu + \sigma)$. This region is also called 68% confidence interval. So for any sample mean, say M_1, in this range, the population mean is within interval $(M_1 - \sigma, M_1 + \sigma)$. As the probability that a given sample mean lies in 68% confidence interval is 68%, for this sample value the estimated population mean range of $M_1 \pm \sigma$ has 32% error probability. In other words a 68% confidence interval corresponds to 0.32 level of significance. To reduce the error, confidence interval needs to be broadened. Generally a 95% or 99% confidence interval, corresponding to $\mu \pm 2\sigma$ or $\mu \pm 2.5\sigma$, is used.

Population Variance Unknown: The above discussion is valid only when the population variance or σ^2 is known. In practise, this is an unknown and the standard error of mean is approximated by $s_{\bar{X}} = s_X/\sqrt{n}$. In this scenario, t-distribution (Figure D.2) instead of normal distribution is used to estimate the population mean. Unlike normal distribution, the shape of t-distribution is also affected by degree of freedom, ν. As $\nu \to \infty$, t-distribution approaches the normal one.

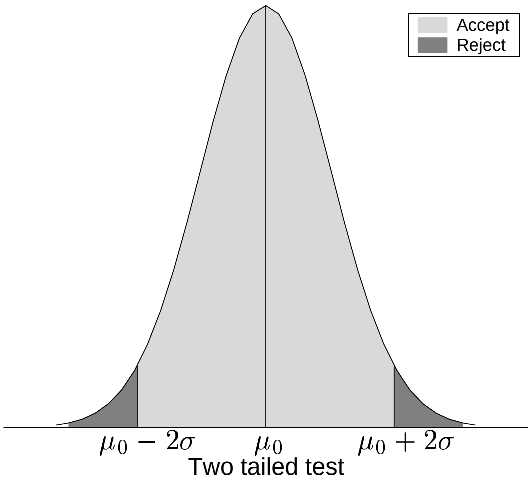


FIGURE D.3: Acceptance and rejection regions for two-tailed test. Null hypothesis is $\mu = \mu_0$. Alternative hypothesis is $\mu \neq \mu_0$.

D.4 Hypothesis Testing

The purpose of the hypothesis testing is to decide if a statement about a statistic is acceptable at a given level of significance. The statistical hypothesis to be tested is also called null hypothesis. As only inference used in this work is about population mean, the null hypothesis for this section is $\mu = k$, where k is some specified value. There are four steps involved in the hypothesis testing.

1. State the hypothesis to be tested. $\mu = k$.
2. Specify the probability of error or significance level, denoted by α. Generally $\alpha = 0.05$ implying that there is 5% probability of incorrectly concluding that the hypothesis is false when it is true.
3. Find the probability, p of obtaining a sample mean $(\bar{X})$ that differs from hypothesised population mean μ by an amount at lease as large as observed sample mean. For a normal (σ known) or t-distribution (σ unknown), this can be obtained by determining the area of the sample mean distribution curve satisfying the above condition.
4. If $p < \alpha$, the hypothesis is rejected, otherwise it is accepted.

Alternative hypothesis: An alternative hypothesis is associated with a null hypothesis. In case a null hypothesis is rejected, an alternative hypothesis can be accepted with the same error. For a null hypothesis $\mu = k$, there are three possible alternative hypotheses. These are

1. $\mu \neq k$.

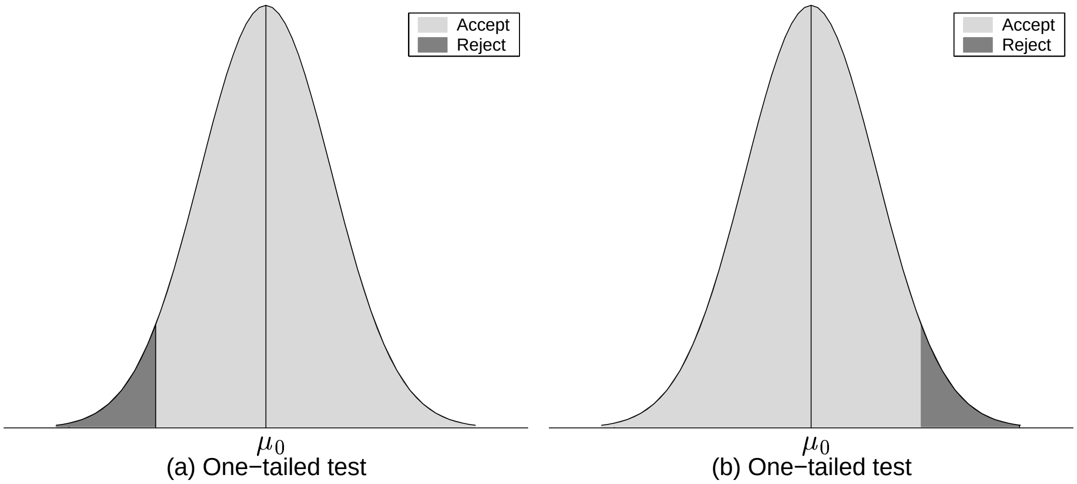


Figure D.4: Acceptance and rejection regions for one-tailed test. Null hypothesis is $\mu = \mu_0$. Alternative hypothesis is (a) $\mu > \mu_0$ and (b) $\mu < \mu_0$.

2. $\mu < k$.

3. $\mu > k$.

In the first case, sample values lying toward either tail of the distribution curve will contribute toward rejection of null hypothesis. Hypothesis testing with formulation of this alternative hypothesis, therefore, requires two-tailed test. For any sample values lying within the darker region, null hypothesis will be rejected (see Figure D.3). However, for the other alternative hypothesis formulations, one-tailed test would suffice. However, the tail of the distribution curve resulting in hypothesis rejection will depend upon if μ is less than or greater than k as illustrated in Figure D.4 (a) and (b).

Appendix E

Questionnaire

1. Identity number: ....

2. Age: ....

3. Gender: MALE / FEMALE

4. Profession: ....

5. Any background in (please tick all applicable)

   (a) Computers

   (b) Computer Vision/Image Processing

   (c) Cognitive Sciences

   (d) Psychology

   (e) Physio-therapy

   (f) Bio-mechanics

   (g) Relevant (please give details)

   ....

6. Participated in similar experiments YES / NO
   If yes, please give details

   ....

   ....

   ....

7. Decision based on

   (a) Periodic motion

   (b) Configuration of moving dots (shape)

   (c) Others

   ....

   ....

   ....

   ....

Appendix F

Information Sheet

Please categorise the image sequences on the basis of what you perceive these to be. You can add more.

1. ....

2. ....

3. ....

4. ....

5. ....

6. ....

7. ....

8. ....

9. ....

10. ....

Appendix G

Human Responses

The following table describes the categories identified by each participant in detail. For each entry a √ indicates that the respective participant identified the category. Some of the entries may be in slight error as it was difficult to classify some of the responses (e.g. does "walking on roof" signify that the participant had identified an upside-down MLD walker or not?).

In this table, each row of the table corresponds to a given participant. The column (1) indicates if the respective participant is an expert or not. A participant with a good computer vision/psychology background was considered to be an expert. Column (2) indicates familiarity with MLDs. The entry can range from 1–5, 5 being good familiarity and 1 means almost none. Columns (3) and (4) indicate if the participant has categorised spot and partial modes respectively. Columns (5) and (6) indicate if the participant has registered the presence of other human or non-human motion. Other human included "running", "dancing", "skating", "people walking in groups", etc. Non-human motion included "animal motion", "swarm of insects", "robotic motion", "some machine", etc.

Table G.1: Detailed summary of human responses.

(1)	(2)	NOR	DIR	WBK	INV	TOP	OBQ	SPT	PER	RAN	(3)	(4)	(5)	(6)
	2	√			√	√				√			√	
	1	√		√	√								√	√
	1													
	1	√											√	
	1											√	√	
	2	√												
	3	√	√		√	√							√	√
	1	√		√	√	√							√	√
	2	√			√	√						√	√	
	3	√		√	√					√				√
	1	√				√							√	

Detailed summary of human responses (continued).

(1)	(2)	NOR	DIR	WBK	INV	TOP	OBQ	SPT	PER	RAN	(3)	(4)	(5)	(6)
	1	√	√	√		√						√	√	√
	1													
	1	√												
	1	√		√		√							√	√
	1	√		√									√	√
	1	√		√	√	√					√		√	
	1	√		√		√							√	
	3	√			√	√								
	1	√												
	1	√		√	√								√	
	1	√	√	√						√		√	√	
	1	√		√	√	√		√					√	
	1	√		√	√								√	
	1	√		√	√	√								
	1	√							√					
	2	√		√	√	√							√	
	5													
	1	√			√									√
	1	√											√	
	1	√		√									√	
	1	√		√									√	
	1	√												
√	1	√											√	√
	1	√		√		√							√	√
	1	√											√	
	2	√			√	√						√	√	√
	1	√	√		√								√	
	1	√		√	√	√							√	√
	2	√	√			√							√	
	3	√												√
	1	√		√	√						√		√	
√	1	√	√			√							√	
	1	√		√									√	√
	1	√	√		√	√		√					√	√
	1	√			√								√	
	1	√		√		√					√		√	√
	3	√		√				√				√	√	√
	1	√		√							√		√	√
	3	√		√	√							√	√	√
√	1	√				√				√			√	
	2												√	√
	1	√		√	√	√								√

Detailed summary of human responses (continued).

(1)	(2)	NOR	DIR	WBK	INV	TOP	OBQ	SPT	PER	RAN	(3)	(4)	(5)	(6)
	1	√	√							√				√
	1	√												
	1	√		√	√					√			√	√
	1	√				√							√	
√	3	√		√	√			√				√		√
	3	√		√		√						√	√	
√	3	√		√	√	√			√		√			
√	4	√	√	√	√	√				√				
√	3	√				√							√	√
√	2	√		√	√								√	
	1	√			√								√	
	1	√		√	√								√	
	1	√		√						√			√	√
√	1	√		√				√		√			√	√
√	2	√		√	√			√	√	√				√
	1	√				√							√	
	1	√		√	√	√						√	√	√
	1	√											√	
√	4	√		√	√					√			√	
√	1	√		√	√								√	
√	2	√		√									√	√
	1	√		√		√							√	
√	1	√		√	√								√	√
	3	√		√									√	
	1	√		√	√	√							√	
√	1	√		√	√	√				√			√	√
√	1	√		√	√	√		√	√			√	√	
√	1	√								√				√
√	4													√
	3	√											√	
	1	√		√						√			√	√
√	4	√		√		√				√				
√	3	√			√					√			√	√
	1	√											√	
	1	√												√
	3	√		√	√	√				√		√	√	√
√	1													
√	3	√		√	√	√	√	√	√	√		√	√	√
	1	√		√	√	√					√		√	√
√	1	√		√		√		√	√				√	√

Bibliography

Ahlström, V., R. Blake, and U. Ahloström (1997). Perception of biological motion. *Perception 26*, 1539–1548.

Barclay, C. D., J. E. Cutting, and L. T. Kozlowski (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics 23*(2), 145–152.

Baumberg, A. M. and D. C. Hogg (1993). Learning flexible models from image sequences. Technical Report 93:36, School of Computer Studies, University of Leeds.

Bertenthal, B. I. and J. Pinto (1993). Complementary processes in the perception and production of human movements. In L. B. Smith and E. Thelen (Eds.), *A Dynamic Systems Approach to development: Applications*, Chapter 8, pp. 209–239. Cambridge: MIT Press.

Bobick, A. F. and J. W. Davis (1996a). An appearance-based representation of action. In *IEEE International Conference on Pattern Recognition*, Vienna, Austria, pp. 307–312.

Bobick, A. F. and J. W. Davis (1996b). Real-time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision*, Sarasoto, Florida, pp. 39–42.

Bregler, C. (1997). Learning and recognising human dynamics in video sequence. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 568–574.

Bregler, C. and J. Malik (1998). Tracking people with twists and exponential maps. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 8–15.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*(2), 121–167.

Cédras, C. and M. Shah (1995). Motion-based recognition: A survey. *Image and Vision Computing 13*(2), 129–155.

Chen, Y. Q., R. I. Damper, and M. S. Nixon (1997). On neural-Network implementations of k-nearest neighbor pattern classifiers. *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications 44*(7), 622–629.

Cunado, D. (1999). *Automatic Gait Recognition via Model-based Moving Feature Analysis.* PhD thesis, University of Southampton.

Cutting, J. E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance 7*(1), 71–87.

Cutting, J. E. and L. T. Kozlowski (1977). Recognizing friends by their walk : Gait perception without familiarity cues. *Bulletin of Psychonomic Society 9*(5), 353–356.

Cutting, J. E. and D. R. Proffitt (1981). Gait perception as an example of how we may perceive events. In R. Walk and H. L. Pick (Eds.), *Intersensory Perception and Sensory Integration*, Chapter 8, pp. 249–273. New York: Plenum.

Cutting, J. E. and D. R. Proffitt (1982). The minimum principle and the perception of absolute, common and relative motions. *Cognitive Psychology 14*, 211–246.

Cutting, J. E., D. R. Proffitt, and L. T. Kozlowski (1978). A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance 4*, 357–372.

Davis, W. and A. F. Bobick (1997). The representation and recognition of human movement using temporal templates. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 928–934.

Devijver, P. A. and J. Kittler (1982). *Pattern Recognition: A Statistical Approach.* Englewood Cliffs, NJ: Prentice Hall.

Fisher, R. A. (1936). UCI repository of machine learning databases. `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/`.

Foster, J. P., M. S. Nixon, and A. Prugel-Bennett (2001). New area based metrics for gait recognition. In *Proceedings of Third International Conference of Audio- and Video-Based Biometric Person Authentication ABVPA*, Halmstad, Sweden, pp. 312–317.

Glass, G. V. and K. D. Hopkins (1996). *Statistical Methods in Education and Psychology.* London: Allyn and Bacon.

Gunn, S. R. (1997). Support vector machines for classification and regression. Technical Report, Image, Speech and Intelligent Systems (ISIS) Research Group, Department of Electronics and Computer Science, University of Southampton, UK. MATLAB toolbox available for download from `http://www.ecs.soton.ac.uk/~srg/`.

Hayfron-Acquah, J. B., M. S. Nixon, and J. N. Carter (2001). Automatic gait recognition by symmetry analysis. In *Proceedings of Third International Conference of Audio- and Video-Based Biometric Person Authentication ABVPA*, Halmstad, Sweden, pp. 272–277.

Hoffman, D. D. and B. E. Flinchbaugh (1982). The interpretation of biological motion. *Biological Cybernetics 42*, 195–204.

Huang, P. S. (1999). *Automatic gait Recognition via Statistical Approaches.* PhD thesis, University of Southampton.

Inman, V. T., H. J. Ralston, and F. Todd (1981). *Human Walking.* Baltimore: Williams and Wilkins.

Jain, A. K., R. Bolle, and S. Pankanti (Eds.) (1999). *Biometrics – Personal Identification in Networked Society.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Joachims, T. (1999). Making large-scale SVM learning practical. See Schölkopf, Burges, and Simola (1999), Chapter 11, pp. 169–184.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics 14*(2), 201–211.

Johansson, G. (1975). Visual motion perception. *Scientific American 232*(6), 76–89.

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research 38*, 379–393.

Kozlowski, L. T. and J. E. Cutting (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics 21*(6), 575–580.

Laxmi, V., J. N. Carter, and R. I. Damper (2002a). Biologically-inspired human gait classifiers. In *IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'02)*, Tarrytown, New York, USA, pp. 17–22.

Laxmi, V., J. N. Carter, and R. I. Damper (2002b). Biologically-inspired human motion detection. In *10th European Symposium on Artificial Neural Network (ESANN'2002)*, Bruges, Belgium, pp. 95–100.

Laxmi, V., J. N. Carter, and R. I. Damper (2002c). Support vector machines and human gait classification. Presented in BMVA Meeting, London.

Laxmi, V., J. N. Carter, and R. I. Damper (2003). Biologically-inspired motion detection and classification: Human and machine perception. Presented in BMVA Meeting on Biologically Inspired Computer Vision Approaches, London.

Little, J. J. and J. E. Boyd (1995). Describing motion for recognition. In *IEEE International Symposium on Computer Vision*, Coral Gables, Florida, pp. 235–240.

Little, J. J. and J. E. Boyd (1997). Global v/s structured interpretation of motion: Moving light displays. In *IEEE Non-rigid and Articulated Motion Workshop, CVPR'97*, Puerto Rico, pp. 18–25.

Little, J. J. and J. E. Boyd (1998a). Recognizing people by their gait - the shape of motion. *Videre - The Electronic Journal of Computer Vision 1*(2), 1–32.

Little, J. J. and J. E. Boyd (1998b). Shape of motion and the perception of human gaits. In *IEEE Workshop on Empirical Evaluation Methods in Computer Vision, CVPR'98*, Santa Barbara, California.

Meyer, D. (1997a). Human gait classification based on hidden Markov models. In *3D Image Analysis and Synthesis '97*, Erlangen, pp. 139–146.

Meyer, D. (1997b). Model based extraction of articulated objects in image sequences for gait analysis. In *Fourth International Conference on Image Processing*, Volume 3, Santa Barbara, California, USA, pp. 78–81.

Meyer, D. (1998a). Features for optical flow based gait classification using HMMs. In *Image and Multidimensional Digital Signal Processing '98*, Alpbach, Austria, pp. 75–79.

Meyer, D. (1998b). Gait classification with HMMs for trajectories of body parts extracted by mixture densities. In *British Machine Vision Conference*, Southampton, UK, pp. 459–468.

Murray, M. P. (1967). Gait as a total pattern of movement. *American Journal of Physical Medicine 46*(1), 290–329.

Murray, M. P., A. B. Drought, R. C. Kory, and M. Wisconsin (1964). Walking patterns of normal men. *Journal of Bone and Joint Surgery 46-A*(2), 335–359.

Nixon, M. S., J. N. Carter, D. Cunado, P. S. Huang, and S. V. Stevenage (1999). Automatic gait recognition. See Jain, Bolle, and Pankanti (1999), pp. 231–249.

Niyogi, S. A. and E. H. Adelson (1994a). Analyzing and recognizing walking figures in XYT. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, Washington, pp. 469–474.

Niyogi, S. A. and E. H. Adelson (1994b). Analyzing gait with spatiotemporal surfaces. In *IEEE Workshop on Non-Rigid Motion and Articulated Objects*, Austin, Texas, pp. 64–69.

Patterson, D. W. (1996). *Artificial Neural Networks – Theory and Applications.* Singapore: Prentice Hall.

Pavlova, M. and A. Sokolov (2000). Orientation specificity in biological motion perception. *Perception and Psychophysics 62*(5), 889–899.

Pinto, J. and M. Shiffrar (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica 102*, 293–318.

Restle, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review 86*(1), 1–24.

Rifkin, R. (2000). SvmFu 3: Software for support vector method in C++. `http://five-percent-nation.mit.edu/SvmFu/index.html`.

Schölkopf, B., C. J. Burges, and A. J. Simola (Eds.) (1999). *Advanced in Kernel Methods Support Vector Learning.* MIT Press.

Sejnowski, T. J. and C. R. Rosenberg (1987). Parallel networks that learn to pronounce English text. *Complex Systems 1*(1), 145–168.

Shuttler, J. D., M. S. Nixon, and C. J. Harris (2000). Statistical gait recognition via velocity moments. *IEE Colloquium - Visual Biometrics*, 11/1–11/5.

Song, Y., L. Goncalves, E. Bernardo, and P. Perona (2001). Monocular perception of biological motion in Johansson displays. *Computer Vision and Image Understanding 81*(3), 303–327.

Stevenage, S. V., M. S. Nixon, and K. Vince (1999). Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology 13*, 513–526.

Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception 13*, 283–286.

Takahashi, K. and J. Ohya (1999). Comparison of neural network based pattern classification methods with application to human motion recognition. In *Fifth International Conference on Engineering Applications of Neural Networks*, Warsaw, Poland.

van Rossum, G. (1999). Python 1.5.2 – a programming language. Public domain software available for download from `http://www.python.org`.

Vapnik (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks 10*, 988–999.

Webb, J. A. and J. K. Aggarwal (1982). Structure from motion of rigid and jointed objects. *Artificial Intelligence 19*, 107–130.

Weston, J. (1999). *Extensions to the Support Vector Method.* PhD thesis, Royal Holloway, University of London.

Weston, J. and C. Watkins (1998). Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX, UK, 1998.

Yacoob, Y. and M. J. Black (1998). Parameterized modeling and recognition of activities. In *Sixth IEEE International Conference on Computer Vision*, Mumbai, India, pp. 120–127.

Yacoob, Y. and L. S. Davis (1998). Learned temporal models of image motion. In *Sixth IEEE International Conference on Computer Vision*, Mumbai, India, pp. 446–453.

Yam, C.-Y., M. S. Nixon, and J. N. Carter (2001). Extended model-based automatic gait recognition of walking and running. In *Proceedings of Third International Conference of Audio- and Video-Based Biometric Person Authentication ABVPA*, Halmstad, Sweden, pp. 278–283.