2016

Combining Local and Global Features in Automatic Affect Recognition from Body Posture and Gait

Benjamin Stephens-Fripp
University of Wollongong

Combining Local and Global Features in Automatic Affect Recognition from Body Posture and Gait

A thesis submitted in fullment of the

requirements for the award of the degree

Master of Philosophy

from

University of Wollongong

by

Benjamin Stephens-Fripp, B.E.(Mechatronics)(Hons1),
Grad Dip Ed.

School of Electrical, Computer and Telecommunications Engineering

December 2016

ACKNOWLEDGEMENTS

I would like to first and foremost thank my friends and family for their support throughout my studies, particularly my wife for her continual encouragement.

I would like to thank my supervisor Prof. Fazel Naghdy for his help and guidance throughout this degree. I am grateful for the many hours spent reviewing my work, having discussions with me and offering advice. I also thank my co-supervisors Dr David Stirling and Assoc. Prof. Golshah Naghdy for their extra support, insight and advice provided to me during this time.

I would like to express my gratitude to Amir Hesami for help recording the motion capture data and guidance in extracting the initial data.

CERTIFICATION

I, Benjamin Stephens-Fripp, declare that this thesis, submitted in fulfillment of the requirements for the award of Master of Philosophy, in the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, is entirely my own work unless otherwise referenced or acknowledged. This manuscript has not been submitted for qualifications at any other academic institute.

Benjamin Stephens-Fripp

Date: 21 December 2016

ABSTRACT

There has been a growing interest in machine-based recognition of emotions from body gait and posture, and its combination with other modalities. Applications such as human computer interaction, social robotics, and security have been the driving force behind such trend. The majority of the previous work in automatic affect perception deploys only either local features or global features. Whilst a combination of both types of features are deployed in applications such as object recognition and facial recognition, the literature does not reveal any study in affect recognition from body language using combined global and local features. In this thesis, such gap is addressed by examining how deploying a combination of local and global features can improve the recognition rate in automatic classification of emotions using gait and posture.

The motion data used in the study comprising kinematic parameters associated with the gait and posture of a number of actors expressing a set of emotions, were recorded electronically using an inertia motion capture system. A combination of local and global features proposed by Kapur et al. and Zacharatos et al., respectively, were used in the classification process using WEKA classification system. Additional global features of shape flow and shaping, horizontal and vertical symmetry were added to the combination feature set to increase the performance of the classifier.

The results obtained in the analysis demonstrate that deploying a combination of local and global features leads to a more robust and reliable method for automatic affect recognition from body language as it improves accuracy across a range of classifiers. This research also demonstrates that the inclusion of the additional features, which represent additional Laban Movement Analysis components, increases the maximum classification accuracy from 88.5% to 92.3%.

Achieving better automatic affect recognition rates can lead to increased application of the approach, improved usefulness and reliability of such systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Problem Statement and Rational

I. Affect Recognition

We constantly rely on our ability to recognise emotions in body language, facial expressions, and vocal sounds. The perceived emotions form our judgement of how people communicate with us and how we respond to them. Over the last few decades, many researchers have been exploring how such capability can be built into machine. Machine based affect recognition has the potential to enhance both human-robot and human-computer interactions. It would allow machines to be more effective in the area of social robotics and create more entertaining interactions in the computer gaming industry. The ability of the intelligent machines to recognise and respond to the user's behaviour allows for more acceptance of a computer or robot, and engagement with them. It can also enable security cameras to anticipate potential threats and inappropriate behaviours.

II. Body Movements and Emotion

A. Human Recognition

There is a growing amount of literature that indicates our body movements can communicate a large amount of information. Kozlowski et al. [4] showed that viewers can determine the sex of an individual by viewing only Point Light (PL)

displays on major joints of the body. Cutting and Kozlowski [5] demonstrated that subjects can identify themselves and their friends by observing PL displays as shown in Figure 1. Body posture has also been demonstrated to be a more reliable method of decoding affect at a distance than facial expressions [6].

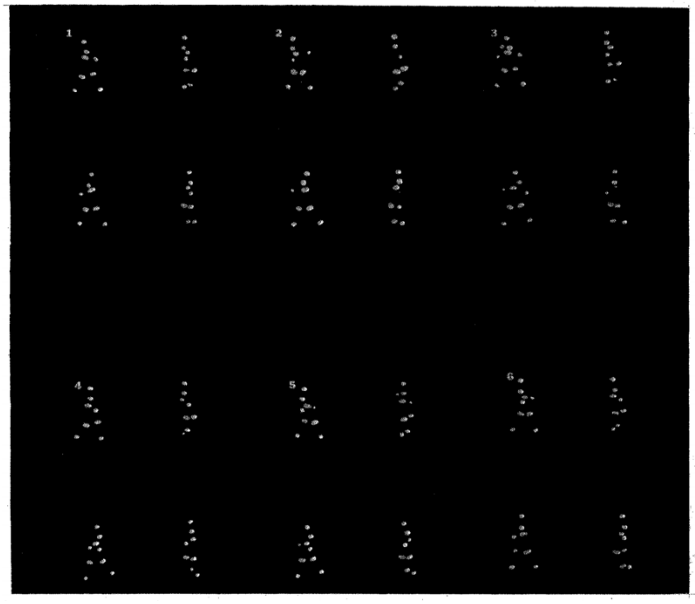


Figure 1 - Point Light Display of Posture [5]

It has also been shown that our body language can reveal our emotions. Brownlow et al. [7] found that observers were able to distinguish between happy and sad dance movements by observing only PL displays.

De Meijer [8] showed 85 adult subjects 96 recordings of body movements performed by three actors. The subjects rated each recording based on the emotional categories it conveyed. There were 12 emotional categories comprising of joy, grief, anger, fear, surprise, disgust, interest, shame, contempt,

sympathy, antipathy and admiration. De Meijer concluded that body movements revealed specific emotional states. This was not just based on one specific movement such as raising a fist, but a combination of movements performed by various body segments.

Walbot [9] also examined the connection between patterns of the body movements and postures, and the emotions portrayed. The study deployed a coding schema to analyse 224 video recordings of six actors and concluded that in fact there were body movement and posture characteristics specific to certain emotions.

The work conducted by Johansson [10] represents an early study of gait analysis for body motion identification. Light markers were mounted on joints, as shown in Figure 2, to recognise the movement of different body parts and to determine the number of sensors needed to recognise the body movement. The subject was correctly identified by using reflective dot markers on ten different points of the body and tracking the markers using a TV camera. Only five reflective points were utilised to identify leg motion. This study demonstrated that patterns of different joints provided all the essential information for immediate identification of human motion.

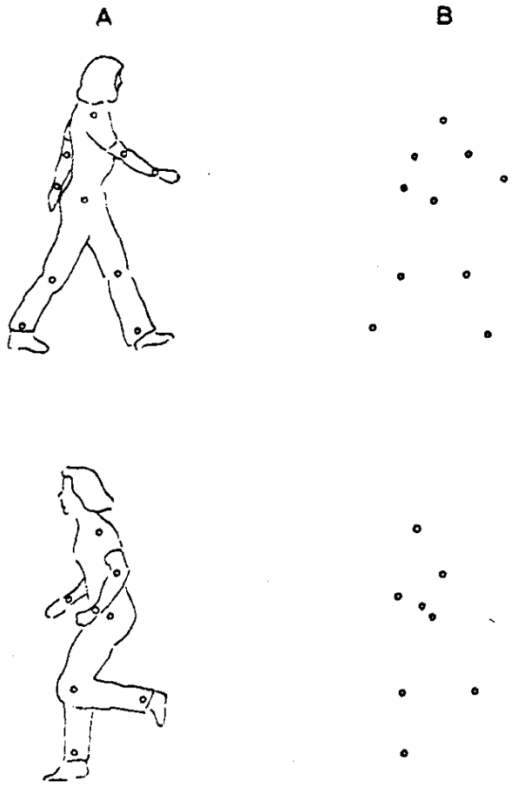


Figure 2 - Light Point Displays used to analyse walking and running [10]

The work conducted by Atkinson et al. [11] is another early study of humans' ability to recognise emotion through gait. Ten trained but unrehearsed actors expressed the emotions of happiness, sadness, fear, anger and disgust. The actors were covered in black with 13 two-centimetre-wide strips of white reflective tape placed on their bodies. They were given the workspace of two large paces around them and were given freedom to walk in any direction whilst being filmed. Two versions were created: a full video or Full Light (FL), and a white strip information video or PL. Emotions were identified from PL information, but the FL video had a higher recognition accuracy than PL

observations. The authors then compared the effectiveness of moderate intensity of emotions against exaggerated and much exaggerated emotions in affect recognition. They concluded that the more exaggerated the emotion, the more easily it could be identified.

Gross et al. [12] also studied human ability to recognise emotions, investigating two factors that could be used to qualitatively detect emotions: effort-shape and body-limb movements. A motion capture system was deployed utilising 31 lightweight spherical markers taped over anatomical landmark points recorded by a high-speed camera. Sixteen actors' front and side views were recorded, as they displayed sad, angry, joyous, content and neutral emotions while they walked. A series of emotion memories were utilised to induce the emotional response in the actors prior to walking. In stage one, untrained observers were able to identify the same emotional memories through gait observations with an accuracy of 76%. Stage two demonstrated that each emotion communicated a unique combination of the effort shape analysis features.

Other psychological studies [13], [14], [15] also confirm the human ability to recognise affective states by observing the body movements.

B. Machine Recognition

The literature reveals interest in the study of automatic affect perception in applications such as human computer interaction, social robotics, and security.

There has been a large amount of research conducted on recognition of emotions through facial expressions. According to de Gelder [16], in 2009 95% of the literature on emotion in humans had been focused on facial expressions. As discussed in section 1.IIA, emotions are not only conveyed through facial expressions, but also through body expression. There is now a growing interest in automatic gait analysis due to its wide range of potential applications in areas such as personal identification [17], deception recognition [18], and detection of illnesses such as multiple sclerosis [19]. Niewiadomski et al. [20] also demonstrated that laughter can be identified by analysing the full body movement.

In their survey paper, Kleinsmith and Bianchi-Berthouze [21] discussed the conflicting views that were reported in the literature on the importance of facial expressions versus body expressions in communicating emotions. They cited a study by Ekman and Friesen [22] which studied emotional deception in facial expressions and body movements. Ekman and Friesen used the term “non-verbal leakage” to describe clues towards deception that was unintentionally conveyed. Ekman and Friesen concluded that facial expressions were easier to

conceal this leakage and therefore people could lie about their emotions. Since it is easier to hide deception in facial expressions, body expressions are identified as potentially a better media for emotion recognition. Kleinsmith and Bianchi-Berthouze's paper also suggested that analysing body expressions could provide clues into understanding facial expressions, leading to higher recognition accuracy.

In Gross et al.'s study [12] on affect recognition from gait, kinematic analysis of the data obtained from the motion capture systems was deployed to quantify both body and limb motions during walking. Differences were shown in the gait measurements and joint movement between different emotions. For example, sad emotions typically contained slower movement and less movement of arms and elbow joints, and less trunk rotation. Angry walkers also had a more flexed trunk and elevated shoulders than joyful or content walkers, even when they had a similar walking speed. Nevertheless, there were many movements that were common to different emotions that could lead to difficulties in discriminating between them. Careful selection of features could help reduce the number of false positives on selected emotions. Gross et al. also suggested that the effort-shape might provide more information when combined with the kinematics data.

III. Cross Cultural Similarities and Differences in Emotion

Ekman and Friesen studied whether emotions conveyed by facial expressions were culture specific [23]. The subjects selected had limited contact with western culture, hence, they were not influenced by media and did not know the meaning of various gestures in western culture. Happiness, sadness, anger, surprise, disgust and fear were explored. In order to overcome the language barrier and equivalent words for emotions not existing in the subject's culture, a story expressing an emotion was read to the subjects and they were asked to point to one of the three face pictures that best represented the emotions portrayed in the story. The results for adults and children, males and females, showed support for the hypothesis that particular facial behaviours were universally associated with particular emotions irrespective of culture.

Kleinsmith et al. [24] tested the cross-cultural similarities and differences of emotion perception through body postures of people from Japan, Sri Lanka and the United States of America. They deployed 13 actors (11 Japanese, one Sri Lankan and one American) who adopted a posture to represent anger, fear, happiness and sadness. These postures were recorded using a motion capture system with 32 markers on the actor's body utilising eight cameras. Non-gender, non-culture specific computer avatars without facial expressions were then created from the captured motions. The 108 affective postures were presented to observers in a different randomised order for each participant. The observers

(25 Japanese, 25 Sri Lankan and 20 Caucasian Americans) were asked to rate the intensity of the emotions they perceived and to identify which emotion label best represented the posture. For each emotion they had two nuances of the same emotion, i.e. anger (angry, upset), fear (fearful and surprise), happiness (happy, joy) and sadness (sad, depressed). When postures from all three cultures were combined, the observers were able to recognise the emotions with accuracy between 54% and 56% for each of the three different groups of observers. When they only observed members of their own culture, the Japanese had a success rate of 90%, the Sri Lankans 88% and the Americans 78%. Therefore, although there were differences in the way cultures expressed emotions in their body movement, there was still a moderate level of agreement between them.

1.2 Aim, Objectives and Hypothesis

The primary aim of this thesis was to develop a more effective approach to machine based affect recognition using human gait. According to the literature, gait analysis was applied in affect recognition by a number of research groups, but the majority of the methods proposed deployed only either local features or global features. Local features are the characteristics associated with specific locations in a pattern or an image. Global features, on the other hand, represent the characteristics associated with all the points in a pattern or an image. In this

research, we examined how using a combination of local and global features could improve the recognition rate.

The motion data used in the study was obtained by an inertia motion capture system. The raw joint data was imported into Matlab and processed to derive the required features. A combination of local and global features were deployed to recognise emotions expressed by actors in a series of experiments. The global features were suggested by Zacharatos et al. [25], and the local features were the same as the featured used by Kapur et al [26]. Additional global features of Shape flow and shaping [8], horizontal and vertical symmetry [10] were added to the combination feature set to increase the performance of the classifier. Once the features were generated four different feature data sets were developed for the following scenarios:

1. Local Features only
2. Global Features only
3. Local and Global Features
4. Local, Global and Additional features.

These four feature data sets were imported into the WEKA classification system [27] and run using a variety of algorithms proposed in previous literature. The results were critically analysed and compared.

Overall, the results showed that a combination of local and global features outperformed the affect recognition rate against scenarios in which either local features or global features were deployed. The addition of additional global

features further increased the accuracy of the classifier.

1.3 Contributions of the Work and Research Outcomes

This thesis represents a systematic and thorough study of machine-based affect recognition using gait analysis within the scope of a Master of Philosophy degree. While the degree offered the candidate numerous opportunities to acquire generic skills on how to systematically manage and conduct a research and thoroughly apply scientific method to infer facts from observations and experimental work, it has resulted in a number of tangible outcomes that can be listed as the contribution of this work:

a) A rigorous and systematic literature review on machine-based affect recognition using gait analysis and posture was conducted. The literature review, broad in its scope and rich in its depth, represents a unique collection of the previous work in this area and proved to be a major source of learning and training for the candidate.

b) A critical comparison of various methods of machine-based affect recognition using local features obtained in gait analysis and posture recognition was conducted. The data used in the study was obtained using an inertial motion capture suit, described later in the thesis.

c) A similar study was conducted on global features using our database.

d) As the major contribution of this work, a combination of global and local features was applied to the gait data and showed that the affect

recognition rate can be improved using this novel approach. The global features were suggested by Zacharatos et al. [25], and the local features were the same as the featured used by Kapur et al [26]. Additional global features of Shape flow and shaping [8], horizontal and vertical symmetry [10] were added to the combination feature set to increase the performance of the classifier.

The thesis has produced the following publications that are currently under review:

The results of the literature review were compiled as a survey paper and submitted to Journal of Social Robotics and is currently under review.

- Automatic Affect Perception Based on Body Gait and Posture: A Survey *(Benjamin Stephens-Fripp, Fazel Naghdy, David Stirling, Golshah Naghdy)*

The methodology, results and analysis of combining local and global features into a single classifier was compiled as an article and submitted to to IEEE Transactions of Affective Computing and is currently under review.

- Combining Local and Global Features in Automatic Affect Recognition from Gait *(Benjamin Stephens-Fripp, Fazel Naghdy, David Stirling, Golshah Naghdy)*

1.4 Outline and Structure of Thesis

This thesis is structured in six chapters

Chapter 1 provides a background and rational on the study. It spells out the primary aim of the project and provides an overview of the approach deployed. The contributions of the work are highlighted in this chapter and the structure of the thesis is described.

Chapter 2 presents a literature review of the previous work that employs body language in automatic affect recognition. The characteristics of each study comprising data collection method, features and classifiers used, testing method and accuracy are provided.

Chapter 3 provides the theoretical framework behind our research. A background on the Laban Movement Analysis model deployed in this work is provided, alongside a justification of combining local features into a single classifier. We present the methods that our approach is based upon, and describe the classification algorithms and performance validation techniques deployed in this project.

Chapter 4 describes the experimental set up deployed in this study, outlining the data collection process, including the software and hardware utilised. The

software packages and methods deployed to export the motion data, extract features and run classification algorithms are also presented.

Chapter 5 outlines validation process and the results of classification using each of the different feature sets with a variety of algorithms and tenfold cross validation. A comparison of our results is then made against similar results reported in the literature.

Chapter 6 discusses the impact of the outcomes produced in this work and their significance. The limitations of our research is then presented and the potential future work is discussed.

2 LITERATURE REVIEW

2.1 Introduction

In this chapter the results of the literature review conducted to identify major previous work in affect recognition using body language are reported. The database and keywords used and the constraints imposed on our literature search are outlined in section 2.2. In section 2.3 different methods of data collection for machine based affect recognition are outlined. 2.4 to 2.7 are mainly focussed on utilising machine based affect recognition from gait. For each method, the source of the data used, data and features extraction methods, the processing techniques and classifiers deployed are studied and the results produced are compared against other methods. Common trends between the studies and gaps in literature are finally identified in Section 2.8.

2.2 Scope of Literature Review

In order to identify the relevant literature, a search was performed on three databases: IEEE Xplore, Scopus and Web of Science. Different combinations of combinations of words (emotion/affect, recognition/detect, gait/posture/body/gestures) were used as keywords.

Results were refined to only include studies concerned with machine-based affective recognition from body language in human beings; as opposed to

recognising emotions in robots. For the purpose of this research, body language was defined as visual cues other than facial recognition alone. Gestures from hands and arms were therefore included within the category of body language. For projects using a multimodal approach, the review was limited to studies that deployed body gait, posture or gestures as one of the information sources. For each relevant publication, the references cited within them were also examined in addition to the other publications by the authors of those papers.

Nayak et al. [28] classify a simple activity in recognition as one which involves a single person with minimal background noise. Currently, emotion detection studies are limited to recognising emotions as simple activities. That is, they are restricted to viewing one person, generally within a controlled environment/background. A long-term goal, however, is to recognise emotions within any environment with the ability to take into account interaction with other people.

2.3 Data Collection Methods

In the studies cited in this chapter, various instruments are deployed to capture the data used in the analysis such as video cameras, optical and inertia motion captures, Kinect sensors and walking pressure sensors. Collected motion data is then fed into a classifier as either raw data or pre-processed data. Pre-processed data takes the raw data and either performs an algorithm to identify key data

points, combine data points and/or to give extra weighting to key data. The classifier is designed and trained to identify the key features representing a particular emotion. The classifiers used are further explained in Section 2.4.

In recent years, there has been a growing number of studies exploring the effectiveness of local features (raw data points) in automatic recognition of human emotions manifested in gait and body movement. The data collection methods used can be broadly categorised into two groups: perceptive and responsive systems. The responsive systems use sensors such as motion capture suits to capture joint movements, whereas the perceptive systems do not require the subject to wear any specialised equipment. Examples of perceptive systems include image processing from video cameras, gait force measurement from pressure plates, and the Kinect depth camera systems. Responsive systems capture as much data as possible, but since they require the subject to wear multiple sensors they are impractical in natural real world environments, such as security camera analysis and HRI situations. Perceptive systems are more suitable for these real world scenarios, but they generate less data than active systems.

The detection methods used within the literature are video detection, Microsoft Kinect system, optical motion capture system and force platform. These techniques are explained below.

I. Video Detection

In the FABO database created by Gunes and Picardi [29], body postures were obtained from video recordings. They utilise a series of single frames rather than treating it as a time series of multiple frames. Only the major points in the body, such as head, shoulder and hands are tracked. The CamShift technique [30], shown in Figure 3, is one way of tracking these points.

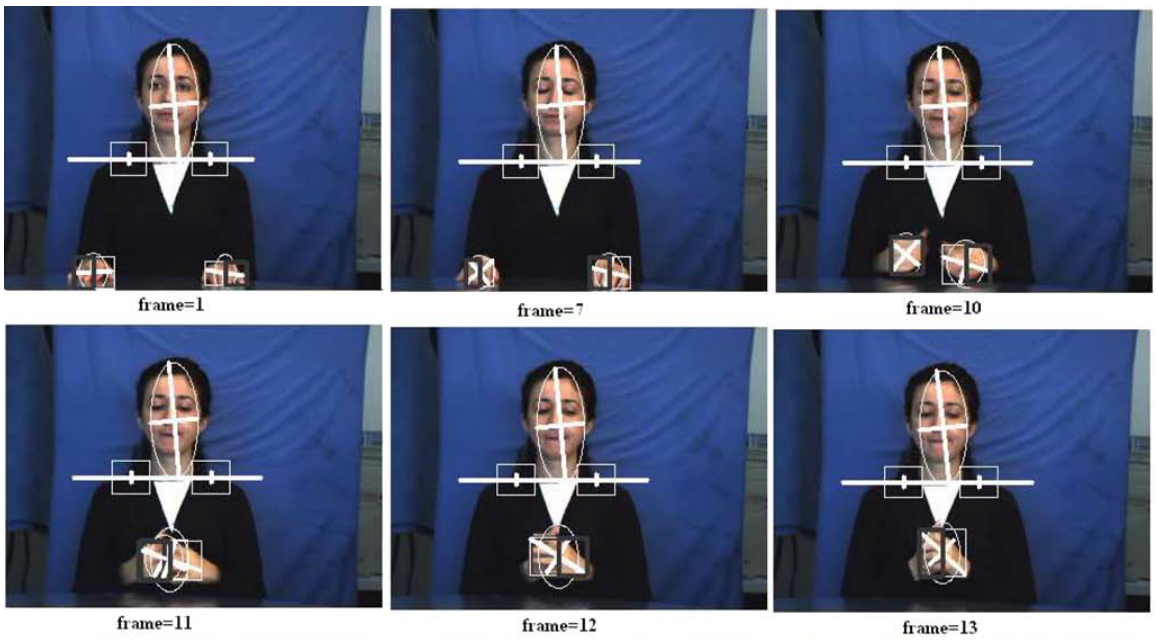


Figure 3 - Body Posture from Video using Camshift [29]

An overview of the studies reviewed that utilise video recognition are outlined in Table1

Authors	Emotions Studied	Dataset	Classifier	Truth Comparison	Success Rate	Sensors
Gunes & Picardi [29]	anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness	10 Actors	Feature level fusion: Adaboost with Random forest of ten trees Decision Level Fusion: Face - Adaboost with C4.5 Body – Random forest of ten trees	Intended Emotion	Feature level fusion – 82.65% Decision level fusion – 78%	Two video cameras, Face & Body
Shan et al. [31]	anger, anxiety, boredom, disgust, joy, puzzle and surprise	23 Actors	SVM Combined with CCA	Intended Emotion	Face -79.2% Body -72.6% Combined – 88.5%	Two video cameras, Face & Body
Chen et al. [32]	anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness	FABO database using 284 videos	SVM with RBF kernel	Intended Emotion	Combined – 73%	1 Camera on face & body
Chen & Tian [32]	anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness	FABO database using 255 videos	One vs one	Intended Emotion	Combined - 77.3%	1 Camera on face & body
Kessous et al. [33]	anger, despair, interest, pleasure, sadness, irritation, joy and pride	Ten non-actor subjects	Bayes Net (WEKA)	Intended Emotion	Facial – 48.3% Body – 67.1% Voice – 57.1% Combined – 74.6%	Two video cameras, Face & Body, microphone on shirt
Park et al. [34]	Happy, Angry, Surprised, Sad	4 Professional Dancers	Time Delayed MLP	Intended Emotion	73%	Video Camera
Sanghvi et al. [35]	Engaged & Not Engaged	5 eight year olds	Various tested but best results from ADTree and OneR classifiers	3 Trained Human Coders	82%	2 Video Cameras (Frontal & Lateral)

Table 1 - Studies using Video Analysis

II. Kinect System

Microsoft Kinect utilises a video camera and a depth sensor. This provides greater accuracy and ability to track joints compared to using a video signal alone. 3D location of joints can then be extracted by the Kinect system, shown in Figure 4 [36]. A summary of the studies utilising Microsoft Kinect and reviewed is this chapter are outlined in Table 2.

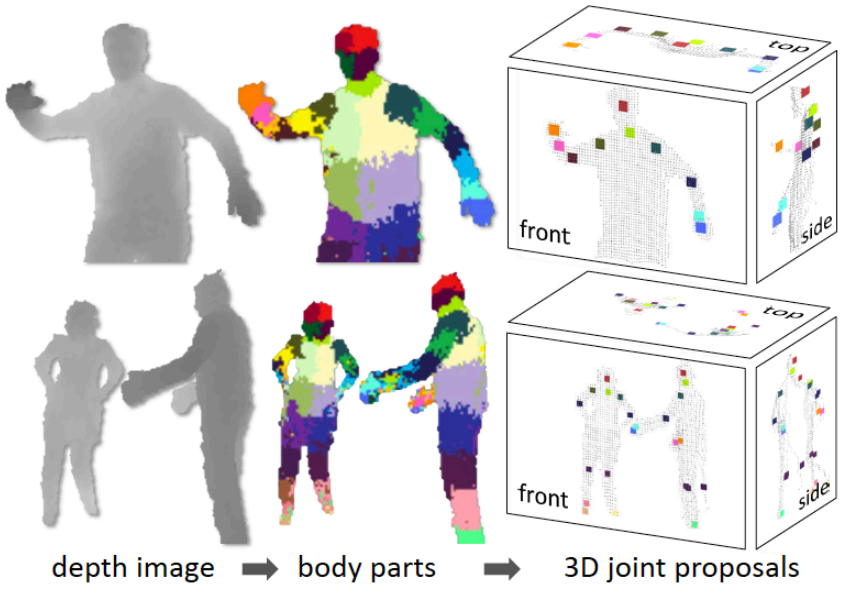


Figure 4 - Feature Extraction from Kinect system [36]

Authors	Emotions Studied	Dataset	Classifier	Truth Comparison	Success Rate	Sensors
Woo Hyan et al. [37]	Rejoicing & Lamenting	1 Participant	N/A	N/A	Two graphs of Space, Weight and Time were easily distinguishable for entire frames	Kinect
McColl et al. [38]	Valence & Arousal	8 elderly individuals	WEKA toolbox using various classifiers, best individual performances were: RBFN, Adaptive Boosting with Naïve Bayes	Human Observer	V - 77.9%, A – 91.4% V – 70.0% A 93%	Kinect
Garber-Barron and Si [39]	Triumph, Concentration, Defeat and Frustration	Eleven participants playing Wii	Bagging Predictor	Human Observers	66.5% joint & limb rotation, & body posture 55% joint rotation 61% limb rotation 62% body posture	Kinect System
Xiao et al. [40]	have question, object, praise, stop, succeed, weakly agree, call, drink, read and write	Twenty three subjects	kNN	Intended Emotion	97%	Kinect System with cyberglove II

Table 2 - Studies using Kinect

IV. Optical Motion Capture System

Optical Motion Capture Systems utilise multiple light markers attached to the body, as shown in Figure 5, which are tracked via infrared cameras. These additional required sensors make these setups impractical in natural real world scenarios such as security camera systems and HRI, but they provide accurate data points for testing and comparing feature extraction and classification methods. Video Cameras alone often rely on crude methods of tracking, such as silhouette extraction, that don't provide data on individual joints. Using infrared cameras in Optical Motion Capture system, however, make it possible to track individual joints in the x, y and z axes and therefore obtain more detailed data on body motion.

Figure 5 - Optical Motion Capture System [41]

A summary of the reviewed works using optical motion capture systems is provided in Table 3.

Authors	Emotions Studied	Dataset	Classifier	Truth Comparison	Success Rate	Sensors
Venture et al. [42]	Neutral, Joy, Anger, Sadness	4 Professional Actors	Similarity index	20 Human Observers 90% Agreement except Joy	78% for an individual, 69% for the group	Motion Capture
Kapur et al. [26]	Sadness, Joy, Anger, Fear	5 Participants (2 Professional Dancers)	Logistic regression, naïve bayes, decision tree, multilayer neural network, SVM	10 Human Observers 93% Agreement	85.6%-91.8% depending on classifier used	Motion Capture
Samadani et al. [43]	Sadness, Happiness, Fear and Anger	13 Demonstrators	HMM to calculate FS representations, which are used in k-NN	Actor's Intended Emotion	77%	Motion Capture
Samadani et al. [44]	Sadness, Happiness, Fear and Anger	13 Demonstrators	FSCPA-GRBF	Actor's Intended Emotion	53.6%	Motion Capture
Xu and Sakazawa [45]	Neutral, Happy, Angry, Sad	30 Demonstrators	SVM with weighted segments	Actor's Intended Emotion	77%	Motion Capture
Bernhardt and Robinson [46]	Neutral, Happy, Angry, Sad	30 Demonstrators hand knocking	SVM with polynomial kernels with weighting of limb speeds	Actor's Intended Emotion	50% without weighting 81% with weighting	Motion Capture
Kleinsmith [47]	Concentration, Defeat, frustration (removed from results), triumph	Eleven participants playing Wii	MLP	8 Human Observers on an avatar replication	66.7%	Inertial Motion Capture
Karg et al. [48]	Neutral, Happy, Angry, Sad Displeased, Content, Bored, Excited and Obedient	13 Actors	SVM	Intended emotion	69% (compared to human success of 63%) 95% if individual person is taken into account Pleasure – 88% Arousal – 97% Dominance – 96%	Optical Tracking
Zacharatos et al. [25]	Concentration, Meditation, Excitement & Frustration	13 Actors	WEKA – MLP	4 Human Observers	85.27%	Motion Capture
Lim and Okuno [49]	Happiness, Sadness, Fear, Anger	10 speech participants & 28 ankle participants	SciKit Learn Toolkit	Human Observers	63% - trained on voice in SIRE 72% - trained on gait in SIRE	Voice & motion capture data on ankle

Table 3 - Studies utilising Motion Capture

V. Force Platform Sensor

A force platform can be used to measure the ground reaction forces from gait along a designated path. The force platform setup used by Janssen et al. [50] is shown in Figure 6. Data obtained from the force platform can be analysed both independently and in conjunction with kinematic analysis of the markers mounted on the body of the subject.

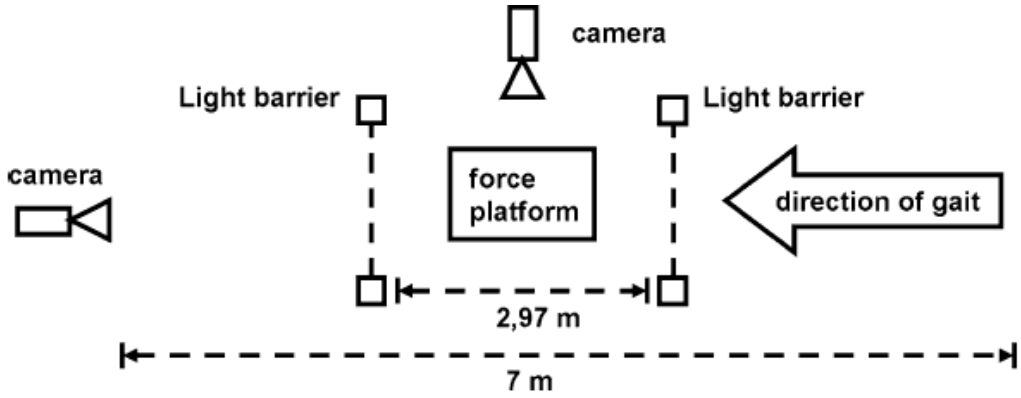


Figure 6 - Force Platform Setup [50]

2.4 Affect Recognition Based on Raw Data

I. Using Motion Capture

Bianchi-Berthouze and Kleinsmith [51] explored the use of an associative neural network called Categorisation and Learning Model (CALM) to learn over time. The VICON motion capture system captured data on twelve subjects performing angry, happy and sad emotions freely. A total of 138 gestures were collected. Emotional category labels were

collected showing an avatar based on motion capture data to 114 Japanese observers. The most frequently used emotion label was chosen for each of the gestures. Eighteen features were used, focusing on upper body gestures based upon the sphere of movement used in dance, consisting of normalized displacement of entire arm, normalized displacement of forearm, normalized extension of body and face orientation. The order of presentation was changed ten times, with each configuration repeated with five different sets of initial conditions. The average error was 0.043% with a standard deviation of 0.002.

Kapur et al. [26] demonstrated the high potential of automatically detecting emotions through the use of body movements. A VICON Motion System captured 14 reference point markers placed on five different subjects. The participants acted out four basic emotions (sadness, joy, anger and fear). To serve as a comparison against cognitive recognition, point light display on fourteen reference points were recorded and shown to ten subjects. The subjects identified emotions from the markers with an accuracy of 93%. Five different classifiers: logic regression, naïve bayes, decision tree, artificial neural network, and a support vector machine, were applied to the data. The classifiers identified the emotions with success rates between 85.6% and 91.8%. Artificial neural network and the support vector machine both produced the most accurate recognition rate. These rates were comparable to that of a human observer judging emotion based off point light displays. However, the study was limited to four acted emotions, and the deployment of a motion system that utilises six cameras; this not practical in real life scenarios.

Venture et al. [42] proposed the use of vector analysis and Principal Component Analysis (PCA) decomposition to detect emotions from gait. Four professional actors were recorded

displaying four basic emotions whilst walking in a straight line through a motion capture system's space. The affective states of neutral, joy, anger and sadness were repeated five times by each actor and were recorded via a motion capture system and video recording. A comparison was made between the detected emotion and one identified by 20 human observers viewing animations to determine the accuracy. Vector analysis, as well the animations produced from the performed emotions, indicate that the lower torso, waist rotations and head movements are the most important features in affect perception as leg and arm data can bias the recognition process. Hence, in the emotions recognition process Venture et al. only deployed six Degrees of Freedom (DOF) to describe the lower torso, three DOF to describe waist and three DOF to describe the head movements. Venture et al. then used a similarity index computation to test similarity between test data and the training data. Through the animation study they concluded that some movements better conveyed emotion than others. For this reason, they applied a weighting to joints that had more impact in conveying emotions, resulting in overall improvement in their results. Weighting resulted in an improved detection rate for all emotions except for sadness, which had the lowest accuracy. For a given subject, Venture et al. detected emotions with an average success of 78%. A global database was developed from a combination of data from all participants and fed into their classifier. As a result, joy and anger had a decrease in performance, there was no effect on the neutral emotion and improvement was observed in the recognition rate of sadness. The global database, however, had an overall negative effect on inter-subject recognition of emotions with an average total recognition of 69%. In that study only a relatively small number of subjects were used and there is a possibility that deployment of more subjects might produce a different result. Both male and female actors

were used in the study with no difference in recognition rates. The false negative classification seems to be for neutral states rather than the other emotions.

Lim and Okuno [49] developed a robot to study multimodal emotional intelligence (MEI) and trained it to recognise emotions in voice, gesture and gait from voice training alone. A unified model for all three modalities was deployed by considering the four properties of speed, intensity, irregularity and extent (SIRE) so that the emotional recognition was no longer context specific. Lim and Okuno assumed that human beings developed their recognition of affect displayed in body language by matching it to the corresponding emotion conveyed in the subject's voice. This principle was applied in training their MEI robot. Lim and Okuno suggested that SIRE systems could be trained to recognise gait using the voice alone. Recognition of happiness, sadness, fear and anger was performed using the Sci Learn Toolkit. Only ankle joint data was used for the gait modality. A recognition rate of 63% was achieved using voice only training, compared to 72% when gait data was used in both the training and testing process. Potential errors were identified when actors whispered whilst expressing fear. This study showed potential for use of high-level feature analysis instead of low level feature analysis to detect emotions, particularly when high success rate body language data is used.

A number of studies on affective recognition from body posture and movement rely on acted emotions. In contrast, Garber-Barron and Si [39] attempted to classify emotions in non-acted scenarios. They used the UCLIC Affective Body Posture and Motion database that contained information from eleven participants playing the Nintendo Wii sports game for a minimum of 30 minutes. This database contained the rotational angles of the joints along

the x, y and z axes. Triumph, concentration, defeat and frustration were recognised with an accuracy of 66.5%, using a combination of joint rotation data, limb rotation data and body posture cues. The success rate decreased by 7% and 4% when using only joint rotation data and only limb rotation data, respectively.

Kleinsmith et al. [47] also explored the feasibility of recognising affective states of players from non-acted scenarios while playing a video game. Participants played Nintendo Wii Tennis for 30 minutes while their body movements were recorded using a motion capture system. Three university students selected 103 usable affective body movements from the recording by viewing the movements as a simplistic avatar. Triumph, defeat, frustration and concentration examined utilising a Multilayer Perceptron (MLP) for automatic classification. Recorded movements were converted into a faceless, non-gender specific computer avatar to remove any bias when evaluated by eight human observers. Each observer evaluated all of the postures five times. The observer's views were divided into three subsets to compare human recognition of emotions against machine recognition. Subsets one and two were used to compare the agreement between the human observations, and subset three was used as the training data and subsequently tested against subset one. An agreement rate of 66.7% was found between the two views of human observation and machine recognition. There was difficulty with the recognition of frustration in the automatic classification; perhaps because of the small amount of training data available. With the frustration label removed, the method achieves a recognition rate of 66.3%. It is noted, however, that since there was no neutral category, concentration was often used as a fall back emotion when the observers feel that there was no other appropriate category.

II. Application of Video Detection

Barakova and Lourens [52] detected movements that express emotions. They examined the Laban sections of weight, time and flow; then translated combinations of these into sadness, joy, fear and anger. Fifteen 20 second recordings of waving patterns that demonstrated happiness, anger, sadness and nervousness were captured via two cameras. A Neural gas algorithm was deployed and 42 children were used to determine the ground truth. Barakova and Lourens achieved an overall accuracy of 63.8% in their machine recognition.

III. Application of Kinect

Xiao et al. [40] studied the use of upper body gestures in the context of virtual reality. A wearable immersion cyberglove II captured hand gesture data and a Microsoft Kinect captured data on the arm and head posture. The action and gestures of confident, have question, object, praise, stop, succeed, weakly agree, call, drink, read and write were studied. Twenty-three subjects were used, each expressing the eleven gestures. The data was split into training and testing set randomly, repeated 5 times and the average result and was compared to the intended emotion. An accuracy of 97% was achieved by using a kNN classifier.

McColl et al. [53] studied the context of social robotics to determine the level of accessibility based on the non-verbal interaction and states analysis (NISA) scale. One expert in the scale was used to code a comparison truth. Kinect system generated a 3D ellipsoid model of a

person's static pose using the trunk and arm orientation towards the robot. Weka data mining software was utilised with ten-fold cross validation. Naïve Bayes, logistic regression, random forest, k-nearest neighbour, Adaboost with Naïve Bayes, multilayer perceptron, support vector machine classifiers were used on 300 static poses from 11 different individuals. Adaboost technique with Naïve Bayes base classifier performed best with an accuracy of 99.3%.

IV. Application of Force Platform

Janssen et al. deployed neural networks to recognise emotions using gait data [50]. In the first experiment, the emotions of sadness, anger or happiness were prompted in their subjects by asking them to remember a time when they felt the emotion. The ground reaction force in x, y and z dimensions was recorded whilst the Subjects walked through the test zone. This data was then fed into a three-layer neural network. The system was trained on two thirds of the data and tested on the remaining third. For each individual they identified the emotion felt with an accuracy of 80%. In the second experiment, subjects listened to either calming or exciting music, or no music, and then walked through the test zone. The aim was to identify the emotion triggered by music. In this experiment, the same kinetic data was utilised as their first experiment, with the addition of kinematic data obtained from a vision system measuring the angles and angular velocities of the arm, hip, knee and ankle. Both the kinetic and kinematic data were fed into the same neural network. For a given individual, the proposed algorithm could recognise emotions at a rate of 77.8% for kinetic data and 73% for kinematic data, which they proposed were not statistically significantly different.

2.5 Affect Recognition Based on Processed Data

I. Dimensional Reduction

Data obtained from motion capture technology can be particularly large. This is computationally difficult and may contain data that is irrelevant and potentially misleading the classifier. Dimensional reduction techniques are usually applied to this type of data to simplify its structure. As stated by Samadani et al., “Statistical dimensionality reduction (DR) techniques transform high-dimensional data to a lower-dimensional subspace” [44] .

Samadani et al. [43] proposed a method of identifying emotions through low-level features. Data recorded by a motion capture system was used to both train and test the system. A Fisher Score (FS) Representation of each of the moments was calculated after the training was performed through Hidden Markov Models. The FSs were then transformed to find a lower dimensional subspace by using Supervised Principle Component Analysis (SPCA). Affective states were then detected using the k-Nearest Neighbour algorithm. The algorithm was trained and tested on the emotional states of sadness, happiness, fear and anger and was applied to both the full body set, and a hand and arm model. The full body dataset was collected from a motion capture system and thirteen demonstrators. The hand and arm dataset was collected independently from the full body data to prevent any confusion between them. University of Waterloo collected this data using a motion capture system with eight cameras. When the subject was part of the training data, the system achieved a success rate of 77% for the full body set and 79% for the hand and arm model. In the leave-

one-subject-out cross validation procedure, the result was dropped slightly to a success rate of 72% which was a high success rate with unseen candidates. In their studies, the authors did not combine the hand/arm model with the full body data set, and they did not incorporate any high-level motion analysis.

Samadani et al. [44] investigated the use of statistical dimensionality reduction techniques in emotion recognition from body movement. A fixed length representation of the features was obtained from sequential observations using the Basis Function Expansion method. A variety of dimensionality reduction techniques such as PCA, Fischer Discriminate Analysis (FDA), Functional supervised PCA (FSPCA) (with both a linear kernel and Gaussian radial basis function (GRBF)), and Functional Isomap was then applied. Samadani et al. tested their algorithm against a hand movement dataset and full body movement dataset. The hand movement was a small dataset consisting of opening and closing hand movements displaying sad, happy and angry emotions with five trials on the left hand and five on the right hand. The full body motion data contained 183 movements from 13 actors conveying sadness, happiness, fear and anger. Different techniques produced a large range of results with the Linear FSPCA producing the highest recognition rate of 96.7% on the hand movement. The algorithm did not perform as well on full body motion data with the highest recognition accuracy, produced by FSPCA-GRBF, being only 53.6% when using the leave-one-out cross validation method.

II. Using Temporal Segmentation

Motion time series can be broken down into stages, such as different stages of a knocking or walking. When analysing motion, not all stages of the motion will equally contribute to the classification process. Both Xu and Sakazawa [45] and Bernhardt and Robinson [46] explored the idea of segmenting motion. In their studies, knocking motion data was segmented into different stages, but then recombined with a different weighting given to the data associated with each segment. Their work, however, could also be applied to segmentation of walking. In both studies, Leave One Subject Out Cross Validation (LOSO–CV) tests were conducted and the weighted segment approach achieved a higher accuracy classifying the motion as a whole action. Hence, some aspects of motion influenced the affect identification more strongly. In analysing time series such as gait data, the general assumption was that not all stages of walking equally contributed to the classification process. For example, raising the leg could provide more clues about the mood than lowering the leg. Accordingly, the data could be segmented into components representing different stages of gait and only the raising segment of the leg could be weighted more heavily in the classification.

Xu and Sakazawa [45] assumed that body movements such as gestures have multiple phases and that none of these segments expressed an affective state equally. This meant that each segment must have its own weight. The method was developed and validated based on a University of Glasgow database. In this Temporal Lobe approach, the emotions associated with each segment were identified and were recombined together with a weighting given to each segment. Xu and Sakazawa achieved a 2.5% to 3.4% higher detection rate of gestures

by deploying the temporal lobe approach compared to traditional deployment of motion data.

Bernhardt and Robinson [46] also showed the benefit of giving weightings to different segments of motion data in emotion recognition. They utilised a collection of knocking performed by 30 individuals in neutral, happy, angry and sad affective styles contained within the University of Glasgow motion capture database. The motion energy was calculated by a weighted sum of the rotational limb speeds to detect the emotion of the individual. A set of accuracies ranging from 50% to 81% was achieved. This method, however, relied heavily on normalising the joint position data based on body size and known properties for that specific subject. For an unknown candidate, however, an estimation of body size for normalisation was made, which potentially decreased the accuracy. Only right handed knocking was utilised but this method could be applied to body language to identify emotions from walking styles.

2.6 Using Global Features

I. Types of Global features

Global features represent the overall characteristics of the posture rather than the properties of certain key points on the posture. Sanghvi et al. [35] utilised the quantity of motion and contraction index as global features. Quantity of motion was obtained by

subtracting the silhouette of the subject in the current frame from the previous frame. The difference in images represented how much movement had occurred. Contraction index was a measure of the expansiveness of the body and was determined by the area of a rectangular bounding box that surrounded the silhouette.

Another example is Laban Movement Analysis (LMA) [54], which has been extensively used in activity recognition systems but has potential for more use in affect recognition.

Hachimura et al. [3] deployed the LMA method (which has four major components: body, effort, shape and space), but with focus on the effort and space components only. Effort was broken down even further into weight, time, space and flow factors and shape was broken down into shaping and shape flow. They used machined based recognition to extract the LMA components of ballet motion and made a comparison against the analysis conducted by an LMA specialist. The study demonstrated that LMA components can be computed numerically, but they did not go on to use these values within Classification.

II. Global features using Motion Capture

Karg et al. detected emotions using human gait and compared different component analysis techniques and classifiers [48]. The Technische Universität München (TU München) gait database was utilised, which contained motion capture recordings of 13 male non-professional actors demonstrating neutral, happy, sad and angry emotions. Initially, the motion capture data was applied to an animated puppet in order to determine the accuracy

in determining human emotions purely from the gait, without any influence of facial expressions or physique. Human observers were able to identify emotions portrayed by the puppet gait with an average accuracy of 63%. Karg et al. used velocity, stride length and cadence, as well as the minimum, maximum and mean joint angles as features. The feature space was transformed using three different methods: principal component analysis (PCA), kernel PCA (KPCA) and linear discriminant analysis (LDA). Three different classifiers were applied to each transformation, naïve bayes, nearest neighbour and a support vector machine, to categorise the emotion based on the data. PCA with a support vector machine classifier achieved the highest accuracy at 69%. This was comparable to the accuracy of human recognition of emotions in the animated puppet. Taking into consideration the characteristics of the individual being observed, the emotion recognition had an accuracy of 95%. Following the same approach, Karg et al. also studied the ability to recognise pleasure, arousal and dominance (PAD) in the subjects as they expressed the emotions of displeasure, contentment, boredom, excitement and obedience. These emotions were chosen as they lied at the extremes of the PAD model. Using the same SVM from data from all joint angles, the system produced an accuracy of 88% for pleasure, 97% for arousal and 96% for dominance. However, there is no reported attempt to use PAD recognition models for classifying data into different emotions.

III. Kinect

Zacharatos et al. [25] applied Laban movement analysis to classify the emotions of candidates playing exergames. Thirteen players played sport games for 30 minutes on an Xbox with Kinect whilst being recorded through an eight-camera motion tracking system

and a separate video camera. Ground-truth was determined by four observers labelling the video footage. Out of the 309 clips recorded, only 197 were in agreement with the observers and were hence utilised. For the analysis, Zacharatos et al. only considered the space and time motion factors of LMA. Concentration, meditation, excitement and frustration were recognised with an overall classification accuracy of 85.27%. Motion clips were only used if they felt the subjects exhibited one of the four emotions being classified and if the four observers agreed on the portrayed emotion. The study did not take into account a range of other emotions that may be misclassified by the system.

In their study, Woo Hyun et al. [37] proposed using an LMA to distinguish between emotions. Microsoft Kinect was deployed to study 20 points on the body considering space, weight and time. Flow always appears in a state of motion so it was not used. Rejoicing and lamenting were found to be easily distinguishable from each other in space, weight and time. These two emotions are largely different in their nature and more study is needed to see how this system works with less extreme emotions.

McColl et al. [38] set out to improve social robots for use at meal times in long term care facilities. They recognised the need for a caregiver to detect the emotions of their patient at meal times so that they can respond and interact appropriately. Body Posture and movements in a seated position was utilised to determine affect. 3D data from a Kinect system was deployed to detect different body language features (e.g. speed of the body, bowing/stretching of the trunk) to classify the valence and arousal values of the subjects. McColl et al. utilised nine different learning techniques to compare their effectiveness, benchmarking them against the median value of twenty one human observers. The highest

accuracy for valence recognition obtained was 77.9% using a radial basis function network (RBFN) and 93.6% for the arousal recognition rate using adaptive boosting with Naïve Bayes.

IV. Video Analysis

Lourens et al. [55] studied one subject performing waving in angry, happy, sad and polite emotions, discovering that they each produced distinct acceleration profiles. A combination of skin colour tracking and motion analysis was used to view the movement of hand arm and head. It was shown that these emotions occupy distinct regions of weight, time, flow and time areas of LMA.

The Geneva Multimodal Emotion Portrayal (GEMEP) corpus developed at the University of Geneva was utilised by Glowinski et al. [56], containing 12 emotions expressed by ten actors. Three subsequent layers of processing was deployed, ranging from low-level physical measures (e.g., position, speed, acceleration of body parts) to overall gesture features (e.g., motion fluency, impulsiveness) and high-level information describing semantic properties of gestures (affect, emotion, and attitudes). Extraction of expressive features from human movement was carried out using the EyesWeb XMI Expressive Gesture Processing Library, utilising head and hand movements. The features of energy, spatial extent, symmetry, and forward-backward learning of head were used with PCA. They were able to rate the emotions into four different clusters of combinations of low/high valence and low/high arousal.

Park et al. [34] explored the application of Laban Movement Analysis to recognise emotions from dance image sequences. A camera captured four professional dancers freely performing various movements of dance portraying happiness, surprise, anger and sadness. They eliminated the background and extracted the number of dominant points on the boundary, the coordinates of centroid, the aspect ratio and the coordinates of rectangle, the velocity and acceleration of each feature. Singular value decomposition was applied to the features to distinguish those that were reliable. These features were then classified into the emotion categories using a time delayed multi-layer perceptron. They were able to classify the emotions with an average accuracy of 73%.

Sanghvi et al. [35] also used global features in affective recognition by social robots. They analysed human postures and body motion to measure the level of engagement of children playing chess with their companion which was an icat robot using an electronic chessboard. The icat interacted with the child appropriately by making a sad facial expression when the child made a good move and a happy facial expression when the child made a bad move. Sanghvi et al. recorded the gameplay via two cameras; one looking at the child in a lateral view and one in a frontal view. Five eight-year-old subjects playing two chess exercises at different levels of difficulty was used. Because of their age, the participants they were unable to accurately identify their own levels of engagement. Instead Sanghvi et al. used three coders to manually label the different sections of video as either engaged, not engaged or unsure. The unsure segments were discarded in order to remove sections of the video that could easily confuse the machine. In order to measure the levels of engagement, Sanghvi et al. used features of body lean angle, a slouch factor, quantity of motion and a

contraction index. A variety of classifiers were tested with ADTree and OneR classifiers which achieving the highest accuracy of 82%.

2.7 Combining Global Features

A combination of local and global features was deployed in object recognition [57], action recognition [58] and, more recently, in facial expression recognition with encouraging results [59]. In spite of rigorous search, no report of the application of this approach to automatic affective recognition from gait and posture was found in the literature.

Siddiqi and Vincent [60] deployed a combination of global and local features in writer identification of handwriting. They recognised that previous techniques were classified into global and local approaches, which contained different information. Global methods examined the overall look and feel of writing, whereas local methods utilised localised features of writing which were different to each user. By combining these two groups of features they were able to effectively identify the writer.

He et al. [61] examined the detection of license plates from video. Previous detection systems classified the license plates based on local Haar-like features that could identify an object within a complex backgrounds invariant to colour, illumination, position or size of the object. [61]. Using these Haar-like features, however, resulted in the classifiers becoming very large, leading in turn to complexity and instability. In order to overcome this deficiency, the authors deployed the use of global statistical features in combination with the local Haar-like features. The combination of global and local features produced a simpler, more efficient and flexible detection system.

Lising et al. [57] observed previous work on object recognition utilised either local or global features in isolation from each other. They reduced the error rate of classification by 20%, by combining both local and global features in a single classifier

Wang et al. [58] deployed a combination of local and global features in action recognition. Spatio-temporal cuboids were deployed for their local features, and silhouette projection histograms for their global feature. They showed that using local and global features on their own often did not provide sufficient information to distinguish variation amongst different motions. A more stable action recognition system was produced by combining all the features in the classification process.

Bosch et al. [62] examined the automatic detection of food items. Local colour, local entropy colour, Tamura perceptual features, Gabor filters, SIFT descriptor, Haar wavelets, Steerable filters, and DAISY descriptor for a patch around the point of interest in the food item were deployed as local features. The average of colour statistics, entropy statistics, and predominant colour statistic across the whole image were utilised as global features. A combination of local and global features again resulted in a more accurate classification system.

2.8 Multiple Modality Fusion

D'mello and Kory [84] performed a meta-analysis of the studies undertaken between 2003 and 2013. Accuracy of 90 studies were reported on, including both uni-modal and multimodal approaches to affective recognition so that the two approaches could be compared without taking into account the individual aspects of the studies. Their study included different combinations of multimodal systems using information from the face, voice, text, physiology, and body. Most of these studies used a combination of two or more modalities, but sometimes they used three or more. D'mello and Kory found that a multimodal approach to affective recognition consistently performed better than a uni-modal system by an average of 9.8%. Underperformance of one modality, however, can decrease the overall accuracy of the system as shown by Gunes and Picardi [63]. Therefore, improving body modalities will in turn increase multiple modality results and overall affect recognition accuracy. In this section, the focus will be on the approach taken in utilising the body modality within multimodal systems.

In their work, Gunes and Picardi [63] utilised information from the upper body posture to improve the recognition rate of emotions from facial recognition alone. They assumed that the subject had a frontal view, with the upper body, face and two hands within full view and not obstructing each other. The emotions of disgust, happiness, surprise, anger, happy-surprise, fear, sadness and uncertainty were studied. For upper body information, body action units were utilised containing classes of emotions that a posture, or combination of postures, could correspond to, e.g. extended body and/or two hands up could represent either anger or happiness.

The system would therefore give extra weighting to the recognition of either of these emotions portrayed in facial expressions. Body modality was used as an auxiliary mode in their system to combine with facial recognition. Facial recognition and body posture recognition were first trained separately and then trained together. A variety of classifiers were tested with BayesNet providing the best results for the face and C4.5 providing the best results for body posture. Gunes and Picardi were able to increase the recognition rate using facial information from 72.83% to 89.8%. They repeated the results with Adaboost and were able to recognise emotions from the face alone with an 87.54% accuracy compared to 94.66% when using both face and body modalities. It is interesting to note that although Gunes and Picardi improved upon their accuracy for using facial expressions alone, the combined success rate was lower than that with the body cues alone. This could be due to the significantly lower recognition of affect from facial expressions alone as compared to recognition using body posture.

Body gesture analysis was performed by extracting spatial-temporal features and using an SVM classifier. Facial recognition and body gesture analysis were combined using canonical correction analysis (CCA). In a single modality alone, the system achieved 72.6% accuracy from body gestures and 79.2% accuracy from facial recognition. When the two modalities were combined using canonical correction analysis, the system reached an accuracy of 88.5%.

Gunes and Picardi [29] also investigated the difficulty of combining emotional information from face and body modality when they had a temporal relationship but were not

necessarily synchronous. Body modality was found to follow the facial modality in time, even though they appear to occur simultaneously. They proposed that since each of the feature vectors from the face and body had distinct set phases (neutral-onset-apex-offset-neutral) in a set order, then they could phase synchronise the apex from each modality together. The authors were not able to identify a suitable database at the time and they created their own database (FABO). Then different actors using a scenario approach where they provided the actors with a short scenario that outlined an emotion-eliciting situation and then asked them to act as if they were in this situation. The actors' responses were recorded by two cameras, one for the face and another for the body against a plain coloured background to help the detection. Anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness were examined. Frames from the face and body modalities were first classified into temporal segments, and the feature vectors from the apex frames were used in the classification. Gunes and Picardi classified these emotions using a variety of both frame and sequence-based classifiers. Individual frames were classified, then either feature level or decision level fusion was performed. In feature level fusion, the apex feature vectors from the face and body were paired together and fed into a classifier for bimodal affect recognition. In decision level fusion, the two modalities were classified separately, then decision-level fusion provided the eventual bimodal affective recognition. Although Gunes and Picardi expected the face to be the primary modality, experiments prove this assumption wrong and they achieved a confidence level of 0.3 for the face modality and 0.7 for the body modality. For the body modality, they focussed on emotions generated with one or two hands, head, shoulders or combinations of these. For uni-modal approaches, they only obtained a success rate of 35.22% for facial expressions and 76.87% for body gestures. With

combined modalities, they were able to achieve an accuracy of 82.65% for feature level fusion and 78% for decision level fusion.

Shan et al. [31] also used the FABO database to study at the fusion of the combined facial and body modalities. The categories of anger, anxiety, boredom, disgust, joy, puzzlement and surprise were detected from videos of twenty-three participants. When using the combination of data from facial expression and body posture, the result increased to 88.5% compared to 79.2% from facial recognition alone

Chen et al. [64] also considered fusing together information from both facial expressions and body cues with a temporal relationship. An alternative method was proposed to compensate for complicated real time processing. A motion history image (MHI), consistent of a histogram of oriented gradients (HOG) and an image-HOG was produced. Instead of using the apex frame, they utilised data from the onset through the apex to the offset. After extracting MHI-HOG and Image-HOG, PCA was performed to reduce the feature dimension in each frame. Each frame was assigned a neutral divergence (the difference between the frame image and the neutral frame) to break the data into temporal segments. Chen et al. also applied a temporal normalisation over the whole range (from onset, apex, to offset) to overcome the significant variation in time resolutions of expressions. Classification was performed by a SVM with an RBF kernel. They also used the FABO database [29]. Two thirds of the data was used as training and the other third for testing. Chen et al. were able to achieve an accuracy of 73% for combined facial expressions and body gestures. Although this was a lower accuracy than that recorded by Gunes and Picardi, Chen et al. believed that it was a more appropriate approach for real-time processing as it did not rely on facial

component tracking, hand tracking and shoulder tracking. Fusing the two modalities increased the accuracy by 7% to 9% compared to the use of face or body modalities alone.

Chen and Tian [32] then proposed an alternative method of fusing together facial and body gesture information. They proposed a method using a margin constrained multiple kernel learning (MCMKL) based fusion approach in order to avoid any contamination from less discriminating features, as the margin could measure the discriminating power of each feature. After determining the base features, one vs one classifier is trained using the optimally combined kernel and evaluated on the FABO database [29]. The facial features image-HOG and MHI-hog are extracted as well as body gesture features of location, motion area, image-HOG and MHI-HOG. As applied in [64], each expression is segmented into onset, apex, offset and neutral phases, and a temporal normalisation procedure is undertaken, following by the MCMKL. Chen et al. found that this approach outperformed the concatenation fusion with an average of 1.3%, achieving an accuracy of 77.3%.

Kessous et al. [33] combined multiple modalities into an emotion recognition system. Their own database of ten people (non-actors) pronouncing a sentence while making eight different emotional expressions (anger, despair, interest, pleasure, sadness, irritation, joy and pride) was utilised. These eight emotions were chosen as they were equally distributed within the valence and arousal space. The demonstrators belonged to five different nationalities: French, German, Greek, Italian and Israeli. Two cameras were used, one for facial recognition and the other for body gestures, and a microphone on the subject's shirt recorded the voice. The Viola Jones algorithm was deployed for face detection. Kessous et al.'s system measures facial animation parameters (FAPs) tracking points and compares the

deformation against a neutral frame. These FAPs, along with their calculated confidence levels were used to provide the facial expression estimation. For body gestures, Kessous et al. used the EyesWeb expressive gesture processing library to extract the quantity of motion, contraction index of the body, velocity, acceleration and fluidity of the hands barycentre. For speech features, a set of features based on intensity, pitch, Mel frequency cepstral coefficient, Bark spectral bands, voice segmented characteristics and pause length were utilised. The same classifier, BayesNet from the WEKA toolbox, was used on all classification in order to compare unimodal, bimodal and multimodal system performance. Kessous et al. explored both the use of feature level fusion and decision level fusion for the bimodal and multimodal classification. For decision level fusion, two alternative methods were studied; using the emotion that had the highest probability in the three modalities and initially determining whether there was an agreement in emotions between more than one of the modalities before reverting back to the highest probability. When operating as a uni-modal system, the accuracy was 48.3% for facial recognition, 67.1% for body gestures, and 57.1% for speech recognition. The best results were obtained from the system operating as a multimodal system looking at information from speech, facial and body gestures combined with a feature level future fusion method. This resulted in an overall accuracy of 78.3 %. It is worth noting that the poorest emotion recognition is for despair, with an accuracy of 53.33%, whereas the other emotions each had a recognition rate of more than 70%. The decision level approach for multimodal recognition produced an accuracy of 74.6%. Bimodal approaches also achieved more accurate results than a uni-modal approach with an accuracy of 62.5% for speech and face modalities and 75% for speech and gesture modalities. However, the accuracy of 65% for facial expression and gesture did not represent an improvement on the gesture recognition approach only. Although their

recognition rate from facial expression and bimodal approaches was less than that reported in other studies, Kessous et al. proposed that this was caused by their more natural set-up. Eight emotions based on non-trained actors from multiple nationalities were studied, with the actors given no instruction on the type of facial expression to use - only the overall emotion. The outcome was a larger variety of facial expressions for any one emotion.

2.8 Discussion

The methods reported in various studies are so diverse that the differences in their approach are barriers to their effective comparison. The major inconsistencies between the studies reviewed in this chapter can be categorised as:

- The number and type of actors and observers
- Consistency in dataset conditions
- The approach used to define the correct emotion
- The emphasis on real time computing and obtrusive sensors
- The type and number of emotions

A number of studies were used to identify the emotions of subjects acting out various situations, though professional actors were not consistently used. The majority of the methods relied on the subject's ability to correctly communicate the intended emotion. While some actors, even not professional, could act out the emotions correctly, many did not. Variation in the quality of portrayed emotions could have also caused by inconsistent methods of inducing the emotion. Some studies used a story to evoke an emotion; others

relied on the actor recalling their own memories, while others left it up to the actor. Acted emotions, however, can be exaggerated and less subtle and the inconsistent performance of emotions by non-actors may be more indicative of how emotions are naturally portrayed. For the purpose of this project, the data was collected from professional actors. Each actor was asked to perform each emotion three times to ensure that any variability in emotions displayed was recorded. Future research, however, could replicate this project with natural emotions, such as the studies ([47], [39]) performed on emotions displayed whilst playing video games.

Both the number of actors and observers used in the databases and datasets deployed within literature are quite small compared to what is currently used within facial expression databases. These low numbers decrease the reliability of any result obtained, as any outlier of performance or opinion of the actors and observers has a large effect on the overall results. For this research, 71 recordings were deployed, using ten-fold cross validation to minimise the effect of overfitting to the data.

Different studies use a variety of style of emotions. This is dependent upon the context of their study. For example Calvo and D'Mello [65] suggest that emotions such as confusion, frustration, boredom, flow, curiosity and anxiety are more suited to student engagement environments. However, these would not be appropriate in the context of security. There is debate [66] as to whether the labelled categories that we place on emotions exist or if they are in a spectrum. It is unknown at this stage whether categorical or dimensional categories are more suited for affect recognition.

There is emphasis on real time analysis of acquired data without requiring the user to wear any special equipment in some studies, whereas other studies use wearable sensors or motion suits, and multiple cameras. Hence, intensive computation of data in real time is not possible. The latter methods provide a better outcome but the ultimate goal is to apply affective perception in real time. The differences in approach are barriers to more effective comparison of methods. In this work, the emphasis was on providing proof of concept rather than real time deployment of the methodology. Therefore, a motion capture device was deployed that required subjects to wear a specially designed suit with inertial sensors.

The cited works used different databases and datasets that increased the complexity of a comprehensive comparison between them. Several studies deployed the FABO database but they only used a specific selection from the database rather than the whole set. Since each study used its own dataset and data detection method, comparing the analysis and classification methods can be difficult. Studies that use a common data set and detection method need to be undertaken to enable comparison of the various processing options (including raw data) to determine their comparative effectiveness. Comparing classifiers within the same dataset and processing options should be considered to determine the more effective classifiers. They also incorporate different types and numbers of emotions. This thesis research aims to overcome some of these issues.

It is evident from the literature that machine based affective recognition from gait and posture is at an early stage of its development. The evidence produced by the studies presently supports the conclusion that using gait alone may not produce results which are as accurate as those obtained when multimodal information is used. But, the studies on

multiple modalities indicate that using body cues still provide strong performance and can be added to further improve facial and voice recognition. Underperformance of one modality, however, can decrease the overall accuracy of the system as shown by Gunes and Picardi [63]. Therefore, improving body modalities will in turn increase multiple modality results and overall affect recognition accuracy.

As seen in the literature, moderately strong performance can be achieved with local (Section 2.4 & 2.5) or global features by themselves (Section 2.6). A combination of local and global features was utilised in object recognition [57], action recognition [58] and, more recently, in facial expression recognition with encouraging results [59]. To date, however, this approach is not applied to automatic affective recognition from gait and posture.

In this thesis, the performance the classifiers using of a combination of local and global features was studied. More specifically, the local features of Kapur [26] and the global features used by Zacharatos [25] were deployed in our analysis. The effect of additional global features that could further improve the performance of the classifier were also considered. Overall, four datasets were deployed, Kapur's local features, Zacharatos' global features, Kapur and Zacharatos features combined, and the combined feature set with additional global features. In order make an effective comparison, all four feature sets were examined on the same data set and emotion categories. Multiple classifiers were applied to the four datasets and their performance were compared.

3 MODELLING

3.1 Introduction

The previous work on automatic affect recognition from gait are based on only local or global features, in isolation from each other. Local features are the characteristics associated with specific locations in a pattern or an image. Global features, on the other hand, represent the characteristics associated with all the points in a pattern or an image. Although applied in other areas, there are no reports in the literature on the automatic affect recognition from gait utilising a combination of local and global features. In this thesis, the effectiveness of combining these two categories of features in increasing the accuracy of affect recognition is explored. The local feature approach developed by Kapur et al. [26] (referred to as Kapur local method) and the global feature method proposed by Zacharatos et al.[25] (referred to as Zacharatos global method) are combined in our approach.

The Kapur local method is outlined in Section 3.2. The global methods deployed in this study rely heavily on Laban Movement Analysis [54], which is described in Section 3.3. The Zacharatos Global method is outlined in Section 3.4. The theoretical framework behind the expected improvement of combining local and global features into a single classifier is described in Section 3.4. Section 3.5 outlines the additional features that are introduced, and the reasoning behind the resulting expected improvement. Finally, a detailed description of the classification algorithms deployed, and validation techniques utilised, is provided in Sections 3.6 and 3.7.

3.2 Local Features

As mentioned earlier, our approach is inspired by the Kapur local method. Kapur et al. [26] demonstrated the high potential of automatically detecting emotions through the use of body movements. A VICON Motion System was used to capture 14 reference-point markers placed on five different subjects. The participants acted out four basic emotions (sadness, joy, anger and fear). The VICON Motion Capture system recorded human movement through the use of six infrared cameras that tracked the markers. The markers positions were reconstructed by VICON to give them a vector (v = [x, y, z]) of Cartesian coordinates representing the markers' positions in 3D space.

For each marker, a vector containing the first and second derivative of the position marker was reconstructed to represent the velocity and acceleration. The mean and standard deviations of the velocity and acceleration were calculated to examine variation in these parameters over 10 seconds of recording.

3.3 Laban Movement Analysis

Laban Movement Analysis (LMA) components are employed in our research as global features. LMA was proposed by Rudolf Laban for analysis of dance movements and is used in the literature for observing, describing, notating and interpreting human movement.

Moore and Yamamoto listed five general principles that underlie Laban's conception of human movement [67]:

1. *Movement is a process of change.*

It is not only a change in body position that communicates movement, but how the change has occurred is important in conveying information about movement.

2. *The change is patterned and orderly.*

Although the process of change of the body movement may appear to be disordered and random, there are distinct sequencing and patterns.

3. *Human movement is intentional.*

Human movement is purposeful and intentional to satisfy a need. Thus, there are clear energies and dynamics associated with the “effort” of the movement that can show the intention behind the movement.

4. *The basic elements of human movement may be articulated and studied.*

The basic elements forming a motion like alphabets in a language are the same. These elements can be used to observe and analyse any human movement.

5. *Movement must be approached at multiple levels if it is to be properly understood*.

In Laban's approach to the analysis of movement, all aspects of motion are analysed at once. It incorporates the analysis of body part position, where and how these parts are moved the energy of the movement, and the use of space.

Laban originally outlined that all the body movements can be categorised by their use of the body, the use of space and the use of dynamic energy. Hence, LMA is generally divided into four categories, Body, Effort, Shape and Space.

I. Body

Body describes the physical characteristics of the body motion. It examines the components that initiate movement, the final position of the body movement and the sequence of the movement.

II. Effort

Effort identifies the intention behind movements. For example, a fist in the air can represent anger, or can represent happiness. For this action, the body component would identify the same components and structure of the body parts involved with the movement, but Effort can distinguish how the movement takes place. The movements associated with different emotion are different in their power, control and timing of movement, which can be represented within Effort. Effort describes whether the movement is smooth, sharp, slow, fast, flowing etc. and is broken down into the motion categories of space, weight, time and

flow. Each motion category exists as a continuum between two extremes, as displayed in Table 4.

Space – attention to the surroundings	
Indirect	Flexible, meandering, wandering, multi-focus
Examples	Waving away bugs, slashing through plant growth, surveying a crowd of people, scanning a room for misplaced keys
Direct	Singe focus, channelled, undeviating
Examples:	Pointing to a particular spot, threading a needle, describing the exact outline of an object
Weight – attitude towards the impact of one's movement	
Light	Buoyant, delicate, easily overcoming gravity
Examples	Dabbing paint on a canvas, pulling out a splinter, describing the movement of a feather
Strong	Powerful, having an impact, increasing pressure into the movement
Examples:	Punching, pushing a heavy object, wringing a towel, expressing a firmly held opinion
Time – lack or sense-of urgency	
Sustained	Lingering, leisurely, indulging in time
Examples	Stretching to yawn, stroking a pet
Sudden	Hurried, urgent
Examples:	Swatting a fly, lunging to catch a ball, grabbing a child from the path of danger, making a snap decision
Flow – amount of control and bodily tension	
Free	Uncontrolled, abounded, unable to stop in the course of the movement
Examples	Waving wildly, shaking off water, flinging a frock into a pound
Bound	Controlled, restrained
Examples:	Moving in slow motion, tai chi, fighting back tears, carefully carrying a cup of hot liquid

Table 4 - Motion Factors and Effort Elements [68]

III. Shape

Shape characterises how the body form changes in the space. Shape has three components: Shape Flow, Directional Movement and Shaping/Carving. Shape Flow is a measure of the size that the torso grows or shrinks, and the opening and closing of body limbs. Directional Movement describes the direction of movement towards an object. Shape measures the changes in movement in the three planes of horizontal, vertical and sagittal.

Horizontal	
Indirect	Affinity with Indirect (i.e., deviating, circling)
Examples	Opening arms to embrace, sprawling in a chair, smoothing the wrinkles of a table cloth, a fisherman throwing out a net
Enclosing	Affinity with Direct (i.e., undeviating, pointing)
Examples:	Clasping someone in a hug, crossing one's arms as when feeling cold
Vertical	
Rising	Affinity with Light (decreasing pressure)
Examples	Reaching for something in a high shelf, showing off with a pompous bearing, looking over the shoulder
Sinking	Affinity with Strong (increasing pressure)
Examples:	Stamping the floor with indignation, pulling down a shade, a boxer ducking to avoid a punch
Sagittal	
Advancing	Affinity with Sustained (i.e., decelerating)
Examples	Reaching forward to shake hands, reaching forward to listen more carefully
Retreating	Affinity with Sudden (i.e., accelerating)
Examples:	Darting back, avoiding a punch, pulling one's hand back from a hot stove, shocked by a sad or surprising news

Table 5 - Shaping Dimensions and Affinities [4]

Although Shape has three components, Chi et al. [68] suggested that they could be merged into one three dimensional Shape term, comprising of horizontal, vertical and sagittal planes. Each dimension exists on a continuum between two different extremes, as shown in Table 5.

IV. Space

Space is a measure of how the body moves through the space. There are certain combinations of movements that can be practised to be more harmonious and aesthetically pleasing. Space can describe the area the body is moving within, the space being used, and the direction of the movement and where the movement is occurring.

3.4 Global Features

In Zacharatos et al.'s study [25], 13 players played sport games for 30 minutes on an Xbox with Kinect whilst being recorded through an eight-camera motion tracking system and a separate video camera. After recording, small motion clips less than two seconds long were extracted that represented one of the four mind states being investigated; meditation, concentration, excitement and frustration. Ground-truth was determined by four observers labelling the video footage and the actors intended emotion. Out of the 309 clips recorded, only 197 were in agreement with the observers and were hence utilised.

These authors employed Laban Movement Analysis (LMA), which describes movement based on the components of Body, Shape, Effort and Space. However, they only deployed the features of space and time, which is a subset of the LMA component effort.

The space feature vector was a combination of the change in head height and the prospective focus. The change in head height was calculated as a percentage and was referred to as the percentage of narrowing down, as shown in equation 1.

$$P_{ND} = (Y_{InitialHead} - \bar{Y})/Y_{InitialHead} \quad (1)$$

The dot product of the face features with the four extremity vectors represent the prospective focus of the movement relative to a given point, as shown in equations 2 and 3.

$$S = \{\vec{L}_{hand}, \vec{R}_{hand}, \vec{L}_{foot}, \vec{R}_{foot}\} \quad (2)$$

$$\forall x \in S, F \cdot x \quad (3)$$

Together, these parameters form the components of the Space feature set, as shown in Table 6. This Space feature set contains the percentage of narrowing down, and the prospective focus relative to the left and right hand, and the left and right foot.

Feature	***Description***
P_{ND}	$(Y_{InitialHead} - \bar{Y})/Y_{InitialHead}$
$DotL_{hand}Direct$	$\vec{F} \cdot \vec{L}_{hand}$
$DotR_{hand}Direct$	$\vec{F} \cdot \vec{R}_{hand}$
$DotL_{foot}Direct$	$\vec{F} \cdot \vec{L}_{foot}$
$DotR_{foot}Direct$	$\vec{F} \cdot \vec{R}_{foot}$

Table 6 - The Space Feature Vector

The velocity, acceleration, and jerk of the four extremities were utilised by Zacharatos et al. to represent the time that corresponds to the speed of the movement. Velocity, acceleration and jerk formed the feature set of the time as shown in Table 7.

These combined feature sets were fed into WEKA. Ten-fold cross validation was used with a Multilayer Perceptron classifier.

Feature	*Description*
$L_{hand}V$	Velocity (v) for Left Hand
$R_{hand}V$	Velocity (v) for Right Hand
$R_{foot}V$	Velocity (v) for Left Foot
$L_{foot}V$	Velocity (v) for Right Foot
$L_{hand}A$	Acceleration (a) for Left Hand
$R_{hand}A$	Acceleration (a) for Right Hand
$R_{foot}A$	Acceleration (a) for Left Foot
$L_{foot}A$	Acceleration (a) for Right Foot
$L_{hand}J$	Jerk (j) for Left Hand
$R_{hand}J$	Jerk (j) for Right Hand
$R_{foot}J$	Jerk (j) for Left Foot
$L_{foot}J$	Jerk (j) for Right Foot

Table 7 - The Time Feature Vector

3.5 Combining Global and Local Features

As discussed in Chapter 2, combining information from a variety of different modalities (body, face, voice etc.) can improve the accuracy of affect recognition. These different data sources bring together different types of information that, when combined, are greater than any of them individually. Machine based affect recognition from gait and body posture was

achieved in the previous work only through the use of either local features or global features. Local features examine movements within individual joints, whereas global features examine the movement of the whole body as a single entity.

As shown in a variety of contexts in Section 2.7, combining the local and global features results in improved classification models. The model of global features used within this study relies on LMA. The fifth principal stated by Moore and Yamamoto [67] underlying LMA indicates that movement can be understood better if it is approached at multiple levels. That is, we need to analyse motion based on all components of LMA. As discussed in Section 3.2, LMA is made up of four components: Body, Effort, Shape and Space; and they are best used in combination with each other.

In Zacharatos global method, only two subsections of the Effort LMA component were utilised. In Kapur local method, the features describing the actual positions of the body parts corresponding to the Body component of LMA were used. By deploying a combination of the Kapur local method and the Zacharatos local method, the Effort and Body LMA components are effectively combined resulting in a more complete LMA description.

3.6 Extra features

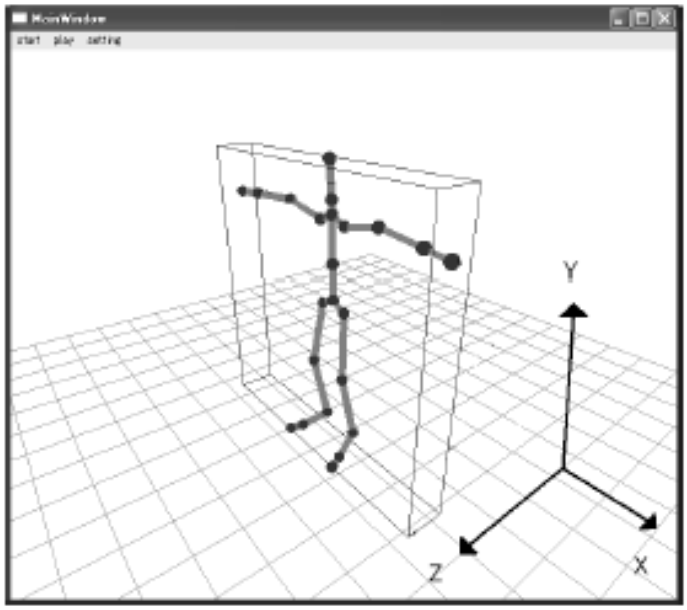
As discussed in Section 3.2, LMA contains four components and is best represented when all four components are utilised together. There are other components of LMA that were not used in Zacharatos global and Kapur local methods. In addition, each component of LMA

contains multiple subsections, which are not all utilised in the features deployed thus far. Following the extraction of the combination feature set, we employed additional features to further improve classification accuracy. Features utilised by Hachimura et al. [3], and Garber-Barron and Si [39] were deployed to articulate a more complete LMA description.

The LMA component of the Body, the space and time subset of Effort were already included within the existing features. The features used by Hachimura et al were deployed to represent the weight component of Effort, and the shape-flow component of Shape. The Horizontal and Vertical Symmetry utilised by Garber-Barron and Si were deployed to contribute towards the Space component of LMA. Hachimura et al. also utilised space and time subset of Effort, but this was not deployed to avoid duplication in the feature set used in this study.

Hachimura et al. [3] also utilised LMA and incorporated weight, space, time section of Effort, and the shape flow and shaping features of Shape. Only their approach to shape flow and shaping were incorporated in our approach as space and time were already included.
Weight was also ignored because of its similarity to the velocity local feature. This study proposes that the least computationally intensive method for calculating the shape flow is using a rectangular parallelepiped enclosing the whole body as shown in Figure 7. The size of each side of this rectangular prism is used as a measure of the shape flow. The shape-flow is calculated by (4).

Figure 7 - Rectangular Parallelepiped [3]



$$\text{shapeflow} = (\text{xpos}_{\max} - \text{xpos}_{\min}) \times (\text{ypos}_{\max} - \text{ypos}_{\min}) \times (\text{zpos}_{\max} - \text{zpos}_{\min}) \quad (4)$$

The shaping feature, deployed by Hachimura et al., is the variance of the x-y frame along the z-axis. It is calculated by determining the change in the torso's z coordinate.

Horizontal and Vertical Symmetry, as explored by Garber-Barron and Si [39] and estimated by (5) and (6), were deployed in the model.

$$\text{Horizontal Symmetry} = \left| \frac{\lfloor \text{centre}_x - \text{left}_x \rfloor - \lfloor \text{centre}_x - \text{right}_x \rfloor}{\lfloor \text{right}_x - \text{left}_x \rfloor} \right| \quad (5)$$

$$\text{Vertical Symmetry} = \left| \frac{\lfloor \text{centre}_y - \text{left}_y \rfloor - \lfloor \text{centre}_y - \text{right}_y \rfloor}{\lfloor \text{right}_y - \text{left}_y \rfloor} \right| \quad (6)$$

3.7 Classifiers

The classifiers are algorithms applied to the motion data for analysis and classification. More specifically, in affect recognition, these algorithms determine what features, or combinations of features, can be used to most accurately predict the associated emotion category. In this section, a brief outline of the most commonly used classifiers in machine based affect recognition is provided; alongside their advantages, disadvantages and an example of their application. The following classifiers were chosen since they have been used in previous studies, as shown in Table 8: BayesNet [69], Naïve Bayes [70], Multi-Layer Perceptron [71], RBF Network [72], SMO [73] , IBk [74], J48 [75] and Random Forest [76]. WEKA was used to apply the different classifiers and to compare their accuracy.

Classification Algorithm	***Deployed by***
BayesNet	Kessous et al. [32]
Naïve Bayes	Kapur et al. [26] and McColl et al. [38].
Multilayer Perceptron (MLP)	Kapur et al. [26], Kleinsmith et al. [47], Zacharatos et al. and [25] Park et al. [33]
Radial Basis Function (RBF) Network	McColl et al. [38].
Sequential Minimal Optimisation (SMO)	Kapur et al. [26], McColl et al. [38], Karg et al. [48], Shan et al. [30], Xu and Sakazawa [44]
IBk	McColl et al. [38] and Xiao et al. [39]
J48	Kapur et al. [26]
Random Forest	McColl et al. [38], Gunes and Picardi [28]

Table 8 - Summary of classifiers used

I. BayesNet

BayesNet classifier is an implementation of a Bayesian Network algorithm. A Bayesian Network is a graphical model that represents a set of variables that have dependence upon each other. A simple Bayesian network is shown in Figure 8, with the dependence relationships shown by the arrows.

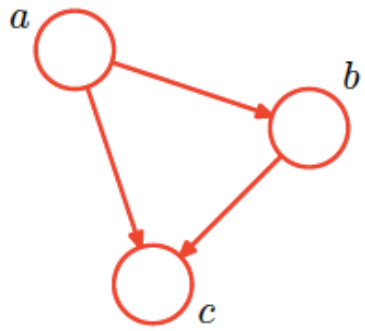


Figure 8 - Bayesian Network [77]

In this example, node b is dependent upon its parent node a, and node c is dependent upon its parent nodes a and b.

The joint probability distribution of this network in Figure 8 can be determined using the product rule of probability based upon the product of conditional probabilities, as demonstrated by (7) [77].

$$\Pr(a, b, c) = \Pr(c|a, b)\Pr(b|a)\Pr(a) \quad (7)$$

This can be generalized for a larger Bayesian network with any number of nodes. The joint distribution of a Bayesian network with K nodes is shown in Equation 8 [77], where pa_k denotes the set of parent notes of x_k, and the input data $x = \{x_1, \ldots, x_k\}$.

$$\Pr(x) = \prod_{k=1}^{K} Pr(x_k|pa_k) \tag{8}$$

BayesNet classifier was implemented in WEKA with one parent and a simple estimator alpha value of 0.5.

II. Naïve Bayes

Naïve Bayes classifier utilises Bayes Theorem under the assumption that features are independent, as shown in Figure 9. It requires only a small amount of training data, is fast and easy, and performs reasonably well. However, since it assumes that features are independent, no learning takes place as a result of interaction between features. For example, the approach can identify emotions when the fist is closed, or the arm is raised, but cannot identify emotions when both the fist is closed and the arm is raised.

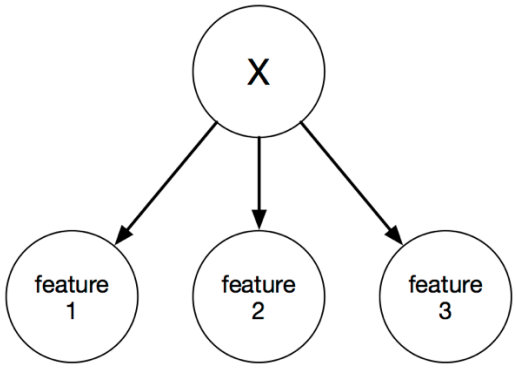


Figure 9 - Naïve Bayes Classifier [78]

This classifier is based on the Bayes rule of conditional probability. As the name assumes, it naively assumes that there is independence between the different events. Because of this, the probabilities are able to be multiplied together as shown in (9) [79]. Although this is a simplistic assumption in real life, it is still effective [79].

$$Pr[H|E] = \frac{Pr[E|H]\ Pr\ [H]}{Pr\ [E]} \tag{9}$$

Naïve Bayes was deployed in emotion recognition by Kapur et al. [26] and McColl et al. [38].

III. Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP) is an artificial network which employs a hidden layer to connect the input features to the output layer, as shown in Figure 10. The MLP hidden layer contains neurons in the shape of a sigmoid function.

MLP can utilise weighted components and biases, and is able to learn non-linear models. This classifier, with the appropriate training data, can become a good generalised classifier with a high fault tolerance. However, when training an MLP classifier, it sometimes settles into a local minimum of the error instead of finding the global minimum error.

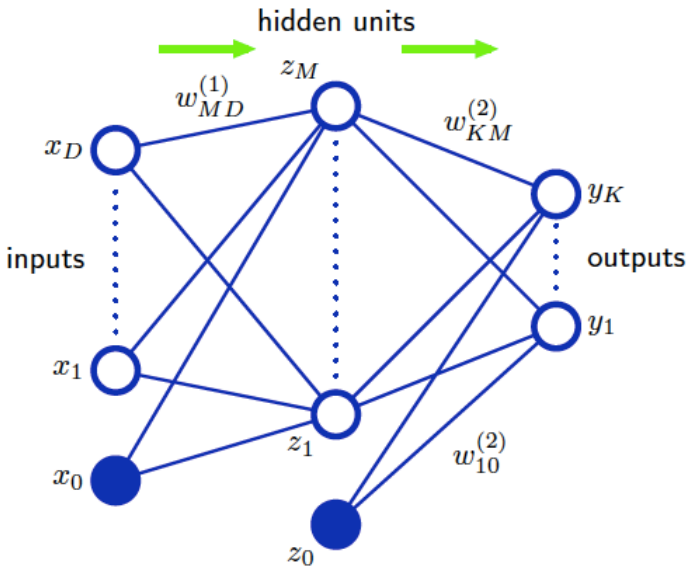


Figure 10 - Multilayer Perceptron [77]

In MLP, initially, the input variables are linearly combined with input biases and weights. The result is then transformed using hidden units comprising differentiable non-linear activation functions, such as a logistic sigmoid or the 'tanh' functions. These hidden units are again linearly combined together and transformed through an appropriate activation function, resulting in a set of outputs. The overall equation for a multiclass problem is demonstrated by Bishop [77] and shown here by (10).

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{jk}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \quad (10)$$

Where

- $w_{jk}^{(2)}$ and $w_{ji}^{(1)}$ are weights
- $w_{j0}^{(1)}$ and $w_{k0}^{(2)}$ are biases
- $\{x_i\}$ is a set of input variables
- [80] is a set of output variables
- σ is a logistic sigmoid function

MLP was deployed in emotion recognition by Kapur et al. [26], Kleinsmith et al. [47] and Zacharatos et al. [25]

The MLP algorithm was implemented in WEKA with a learning weight of 0.3, momentum of 0.2, and a training time of 500.

IV. Radial Basis Function (RBF) Network

RBF Network is also an artificial network using a hidden layer. However, it employs spherical Gaussian functions as the boundaries in the hidden layer. When these hidden layers are recombined, adjustable weights can also be applied. The RBF Network is illustrated in Figure 11. Similar to MLP, it is effective at generalising trends and performs well on unseen data. It can, however, be quicker than MLP in training, but slower in execution when the network is trained.

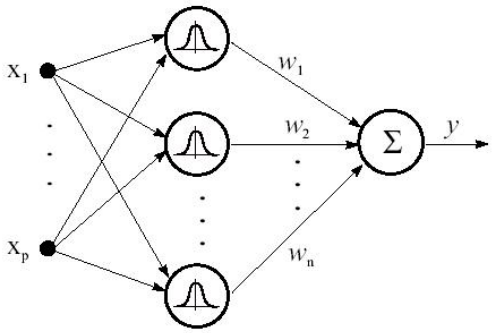


Figure 11 - RBF Network [81]

The RBF Network was deployed in gait recognition by McColl et al. [38].

Each basis function depends on the radial distance from the centre u_j so that

$$\emptyset_j(x) = h(\| x - u_j \|) \tag{11}$$

The goal is to find a smooth function f(x) for a set of input vectors (x_1, ….,x_n) and corresponding target values {t_1, … , t_n) so that f(x_n) = t_n for n = 1,…,N. This is achieved through radial basis functions centred on every data point, as shown in equation 12 [77].

$$f(x) = \sum_{n=1}^{N} \omega_n h(\| x - x_n \|) \tag{12}$$

where the value of the coefficients ω_n are found by the least square method.

RBF Network was implemented in WEKA with a minimum standard deviation of clusters of 0.2, and the number of clusters of 2.

V. Sequential Minimal Optimisation (SMO)

SMO is an algorithm based on the Support Vector Machine (SVM) classifier that is speeded up by breaking SVM down into a series of smaller optimization problems. SVM Classifiers work by maximising the distance between the decision boundary and the data points, as shown in Figure 12.

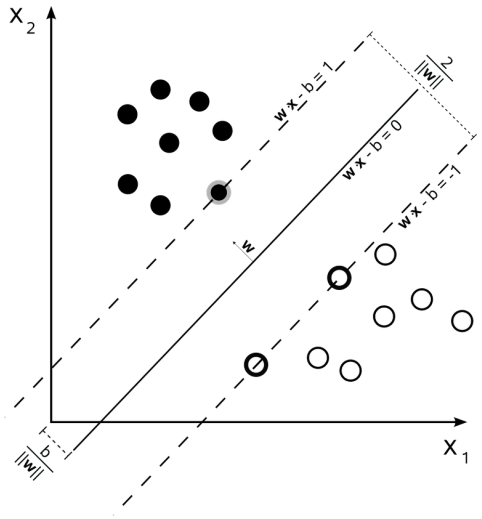


Figure 12 - SVM Boundary Maximisation [82]

Maximising the margin is achieved by solving equation 13 [77].

$$\underset{w,b}{arg\,max}\left\{\frac{1}{\|w\|}\underset{n}{min}[t_n(w^T\emptyset(x_n)+b)]\right\} \tag{13}$$

SMO has good generalisation, but is sensitive to the initial constraint parameters. Although SMO is faster than SVM, it can be slower than other classifiers in both training and running.

SVM classifiers were deployed in gait recognition by Kapur et al. [26], McColl et al. [38] and Karg et al. [48].

VI. IBk

IBk is an implementation of the k-nearest neighbour algorithm, whereby the class assigned is based on the most common class amongst the k closest neighbours of the training data. Figure 13 illustrates implementation of a one nearest neighbour algorithm (left), and a four nearest neighbour algorithm (right). It is a very simple classifier that can work well with basic classification. This, however, means that since IBk doesn't learn anything from the training data, it is not good at generalisation and doesn't perform well on unseen data.

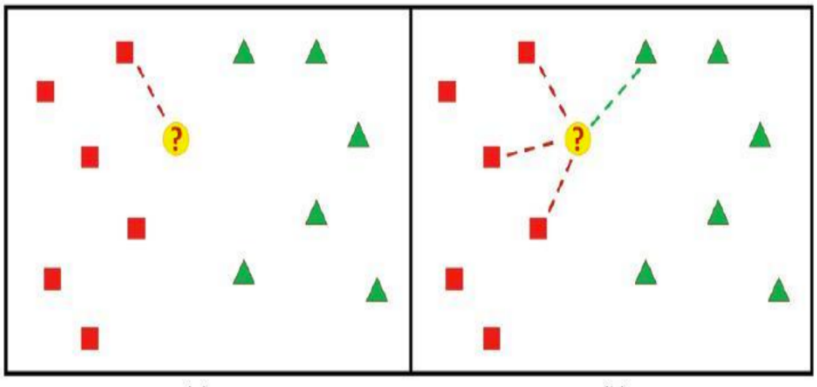


Figure 13 - KNN Classifier [83]

KNN classifier was deployed in gait recognition by McColl et al. [38].

Let *k* be the number of nearest neighbours and D be the set of training examples
for each test example $z = (x', y')$ **do**
Compute $(d(x', x)$, the distance between z and every example, $(x,y) \in D$.
Select $D_z \subseteq D$, the set of k closest training examples to
$y' = \arg\max_v \sum_{(x_i,y_{i)\in D_z}} I(v = y_i)$
end for

Figure 14 - K-nearest neighbour classification algorithm [1]

The k-nearest neighbour classification algorithm as explained by Tan et al. [1] is shown in Figure 14. The k number of nearest neighbours is determined by calculating the distance between each test point [z=(x,y)] and every training data point [$(x,y) \in D$]. Majority class voting is performed on these k nearest neighbours to determine the class of the test point, as shown in equation 14 [1].

$$Majority\ Voting = y' = \underset{v}{argmax} \sum_{(x_i,y_{i)\in D_z}} I(v = y_i) \quad (14)$$

IBK was implemented in WEKA with KNN value of 1 and no distance weighting.

VII. J48

J48 is developed based on the C4.5 Decision tree algorithm, whereby tree-like graph decisions are used to group the data into different classes. The depth of the tree is limited by the number of attributes. Figure 15 illustrates a sample decision tree used to classify a plant based on its petal width and petal length.

J48 classification is easy to visualize and explain, but should be rebuilt when more training data is incorporated. The method is also prone to over-fitting.

The J48 classifier was utilised in gait recognition by Kapur et al. [26].

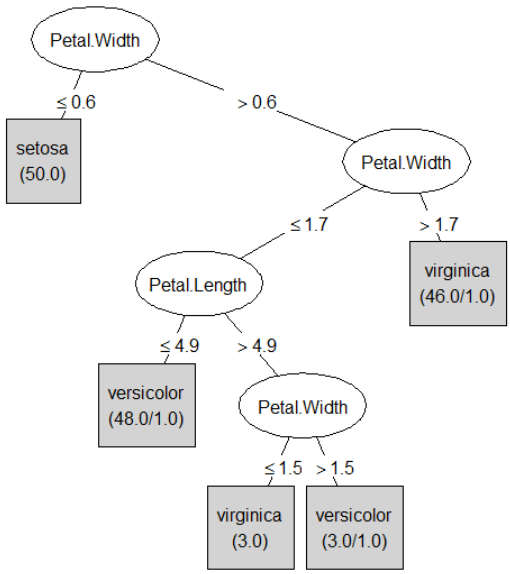


Figure 15 - Decision tree visualisation [84]

A decision tree is built recursively by splitting up the training data into subsets. If all records in the subset belong to the same class, they are grouped together in a leaf node. If, however, all records in the subset belong to more than one class, they are split up based on a test condition, creating multiple child nodes. This process is then repeated for each child node. An algorithm for building a decision tree is shown in Figure 16.

```
TreeGrowth (E,F)
   if stopping_cond(E,F) = true then
         leaf = createNode().
         leaf.label = Classify(E).
         return leaf.
   else
         root = createNode()
         root.test_cond = find_best_split(E,)
         let V = {v|v is a possible outcome of root.test_cond }
         for each v ∈ V do
            Ev = {e | root.test_cond() = v and e ∈ E}.
            child = TreeGrowth(Ev, F).
            add child as descendent of root and label the edge ( root→child) as v
         end for
   end if
   return root
```

Figure 16 - Skeleton decision tree induction algorithm [1]

To determine the best split, the J48 algorithm utilises a criterion known as gain ratio, as shown in (15) [1].

$$Gain\ ratio = \frac{\Delta_{info}}{Split\ info} \tag{15}$$

Where

$Split\ info = -\sum_{i=1}^{k} P(v_i) \log_2 P(v_i)$, and

k = total number of splits

$P(v_i)$ = Probability of each node value

J48 was implemented in WEKA with two minimum instances per leaf, and three folds for pruning.

VIII. Random Forest

Random Forest is a classifier whereby a decision is made through a collection of decision trees from a random sub selection of data.

Since Random Forest employs multiple trees from a random selection of training data, it is resistant to over fitting. The large collection of trees, however, can make it slow for real time processing. The upper generalisation error bound of Random Forest converges towards being dependent upon the strength of the tree classifiers and the average correlation of trees, as shown in equation 16.

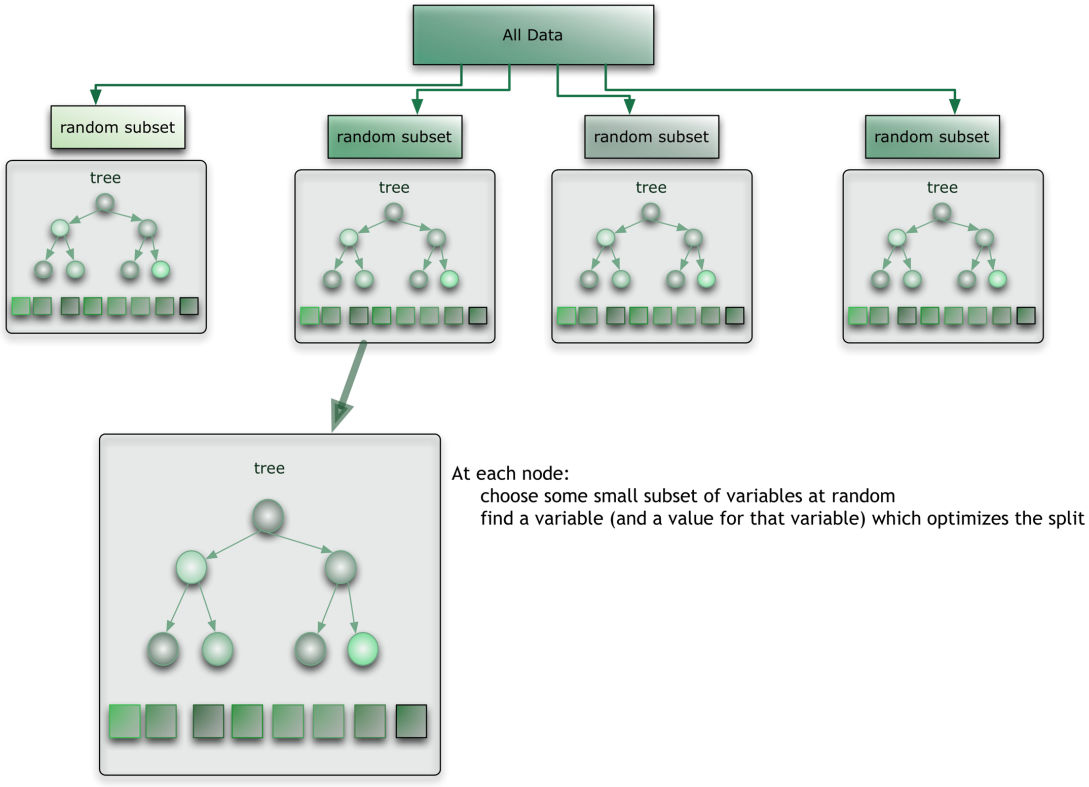


Figure 17 - Random Forest Visualisation [85]

$$Generalisation\ error\ \leq \frac{\bar{\rho}(1-s^2)}{s^2} \tag{16}$$

where $\bar{\rho}$ is the average correlation of the trees and s corresponds to the "strength" of the tree classifiers.

According to equation 16 [1], an increase in correlation between trees produces an increased generalisation error bound. By undertaking randomisation, the correlation reduces resulting in a lower generalisation error [1].

The Random Forest Classifier was deployed in gait recognition by McColl et al. [38]. Random Forest was implemented in WEKA with 100 Trees.

IX. J48 Graft

Although J48 Graft [86] does not appear in any of the literature on gait recognition, it was also tested in this work as an alternative approach as it also demonstrates substantial accuracy in this research. J48 Graft is similar to the J48 algorithm but utilises grafts which adds nodes to reduce the prediction error. Grafting is a post processing technique that is applied to decision trees to remove branches that either occupies space not contained within the training data, or contains misclassified data. The grafting process examines alternative branching at ancestor nodes of branches in question. If the replacement branch increases classification strength, then it is grafted to the existing decision tree. Although this

potentially increases the tree performance, at the same time it adds to the complexity of the decision tree [86].

J48 Graft was implemented in WEKA with a minimum of two instances per leaf.

3.8 Validation

A problem often encountered in classification training is that the classification models become over-fitted to the training data and do not perform satisfactory on unseen data. For our research, the ten-fold cross validation was chosen as it is considered a "standard way of measuring performance" [79]. In N-Fold Cross validation, the data is randomly split up into N parts, trained on (N-1) parts, evaluating the performance against the remaining section. The process is repeated for all N possible options for the held out group. A sample process of four-fold cross validation is shown in Figure 18, where the red box indicates the group left out for testing and the other 3 remaining boxes are used for training. As shown, each run is repeated where the allocated group removed for testing is changed. The overall performance is based on the average of the performance across each run.

Figure 18 - Four-Fold Cross Validation

A downside to this process is that the time taken for process is increased N times, where N is the number of subsections.

3.9 Summary

The Kapur local method and Zacharatos global method underlying our approach was outlined. A background to the LMA model was provided and the theoretical reasoning was presented to justify the combination of local and global features into a single classifier. Introduction of additional features were justified as they provide a more complete LMA. Finally, the classification algorithms and performance validation technique deployed in our approach was described.

4 EXPERIMENTAL SETUP

4.1 Introduction

The experimental approach undertaken in our research is presented in this chapter. This includes the hardware, software and techniques used to record and extract the features, and to classify the motion data. The hardware deployed to capture the motion data is described in Section 4.2, and the data recording process is outlined in Section 4.3. The software utilised to record and store the data is described in Section 4.4 whilst Section 4.5 explains the methods used to extract and store the features from the data, ready for classification. The classification software toolbox is presented in Section 4.6; and the calculations employed to extract the various feature sets are finally presented in Sections 4.7-4.10.

4.2 Motion Capture Device

An X-sens Moven (MVN) [2] inertial movement suit was deployed to capture the position, velocity and acceleration of 22 joints and 23 segments of the body during the experimental work, as shown in Figure 19.

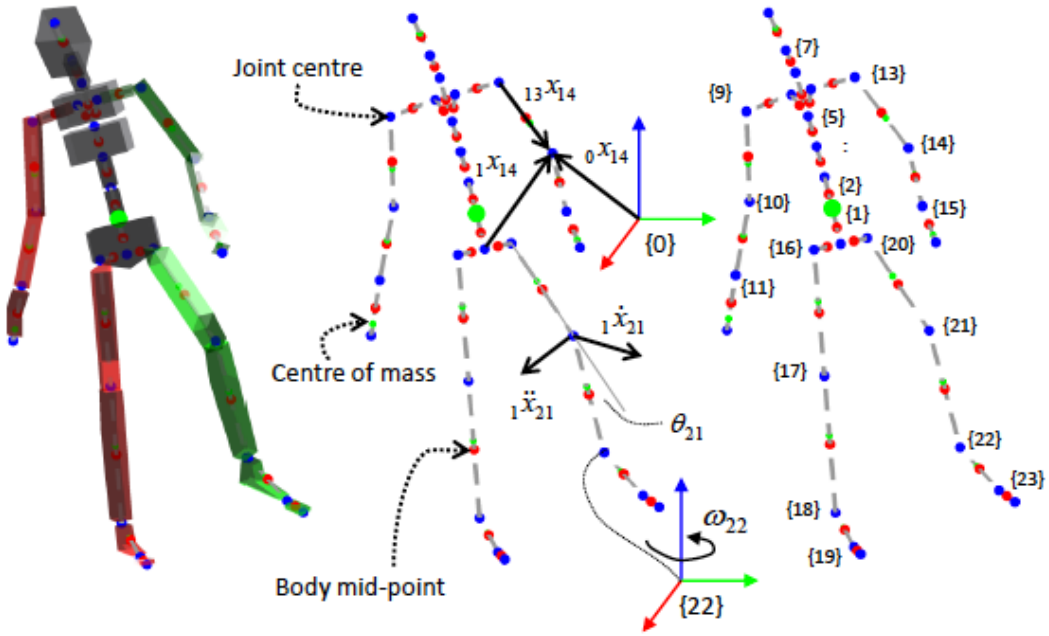


Figure 19 - Location of MVN Segments and Joints [87]. *Blue markers represent the joint's centre; green markers represent the segment's centre of mass; red markers represent the segment's mid-point.*

The MVN motion suit consists of a series of body mounted inertial sensors, (Figure 20), and does not require any emitters or external cameras [88]. The MVN system utilises 17 MTx sensors with two XBus Masters. Each MTx sensor is comprised of a 3D gyroscope, 3D accelerometer and a 3D magnetometer.

Figure 20 - Inertial Sensor [2]

Each sensor is connected to an XBus Master transmitter via daisy chains, with one cable going towards each limb. The sensors are placed onto segments that are surrounded by two joints. The locations of these sensors are outlined in Table 9 and are shown in Figure 21.

Figure 21 - MTx Sensor Placement [2]

The initial location and orientation of these sensors are determined by a combination of measured body dimensions and initial alignment calibrations. For each subject, measurements are taken of the body height, shoe size, arm span, hip height, knee height, ankle height, hip width, shoulder width and the shoe sole thickness. Based on these measurements, the location of the joints and sensors are estimated by the MVN system. The accuracy of the distances between the sensors and the joints can be increased by entering their measured values, rather than relying on estimations.

Location	Optimal position
Foot	Middle of Bridge of foot
Lower leg	Flat on the shin bone (medial surface of the tibia)
Upper leg	Lateral side above knee
Pelvis	Flat on sacrum
Sternum	Flat in the middle of the chest
Shoulder	Scapula (Shoulder Blades)
Upper Arm	Lateral Side above elbow
Fore arm	Lateral and flat side of the wrist
Hand	Backside of hand
Head	Any comfortable position

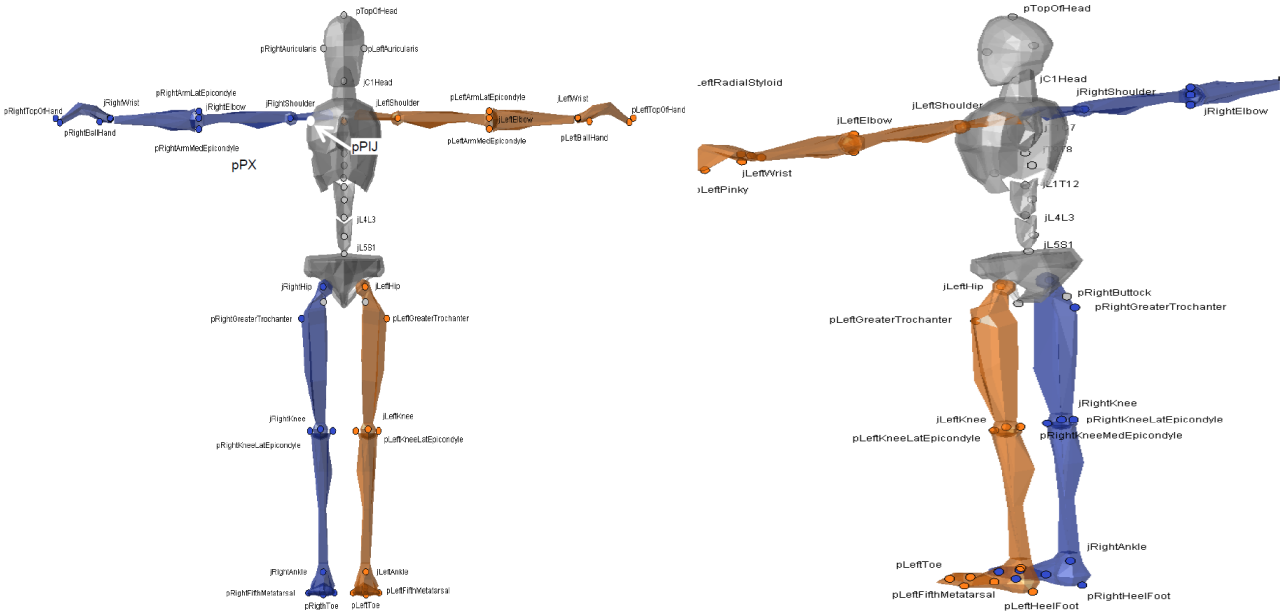
Table 9 - MTx Sensor Placement [2]

The orientation of the sensors is calibrated with either an n-pose or t-pose stance of the subject. In an n-pose, the subject stands up straight with their hands by their side; whereas in a t-pose position the subject stands up straight with their arms and hands out parallel to the ground, with their thumbs facing forwards. The calibration accuracy can be increased by incorporating a squat and/or hand touch movement during the calibration process.

Joint origins are determined based on the anatomical frame, and are described in the Cartesian coordinate frames. After the initial measurements and calibration, the system

translates data obtained from the sensors into a 23 segment biomechanical model. The segments are pelvis, L5, L3, T12, T8, neck, head, and right and left shoulder, upper arms, forearms, hands, upper legs, lower legs, feet and toes. Although there are only 17 sensors attached, the remaining segments are estimated based on a biomechanical model utilising constraints for connecting segments. The anatomical landmarks are presented in Figure 22.

Figure 22 - X-Sens MVN Anatomical Landmarks [2]



In addition to the recorded joint angles, the system estimates and provides the joint positions, linear and angular velocities, and accelerations. These are referenced to an initial global frame with a sampling frequency rate of 120Hz.

During the motion, integration drift can occur within the data due to sensor noise, sensor signal offset and sensor orientation errors. This is continuously corrected based on the biomechanical characteristics of the human body. Joint characteristics and contact points are used to constrain the position and the velocity. An additional magnetic field sensor is deployed to correct accumulation errors that may occur over time. The sensor fusion and correction process is illustrated in Figure 23.

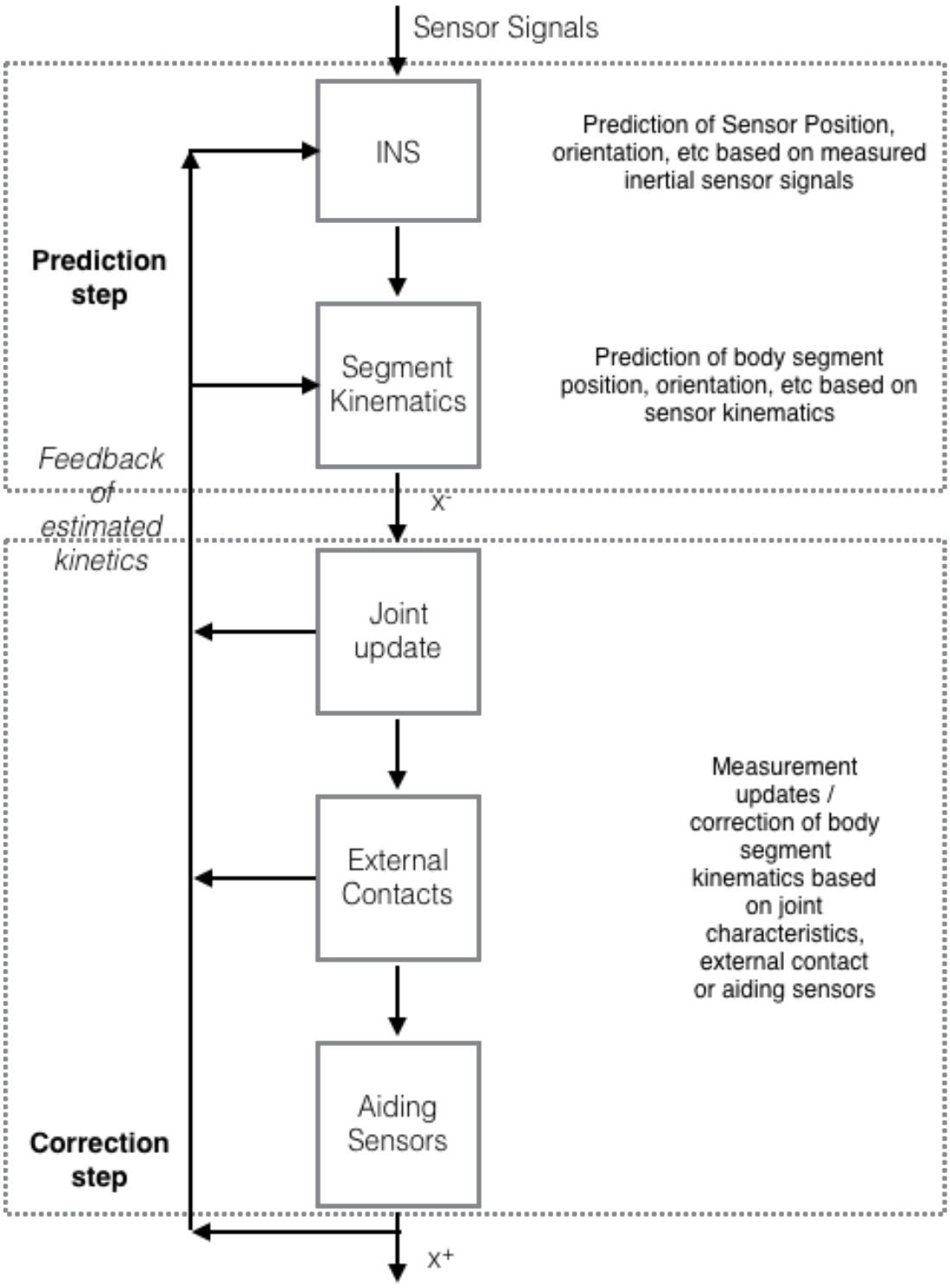


Figure 23 - Sensor Fusion and Correction [88]

The XSens MVN system deploys an Extended Kalman Filter (EKF) [89] to estimate and correct the drift in measurements. The EKF compares the estimated orientation based on accelerometer data, inbuilt magnetometers measuring the deviation from the earth's magnetic field, and a sample of the gravitational direction. This fusion and correction process results in an accuracy of five mm for position and three degrees for orientation [89].

4.3 Data Collection

The kinematics data utilised in this work was produced at the Centre for Intelligent Mechatronic Research (CIMR) at University of Wollongong, Australia [90]. Nine actors (five females and four males) were used to demonstrate grieving, neutral and happy affects whilst walking.

In the experimental work, the intended emotion was sent to the actor via an SMS message and recordings were done in three phases: the gait of the actor walking towards the table with the phone on it was the first phase, the actor reading the received SMS formed the second phase, and the actor walking whilst acting the designated emotion was the third phase, as shown in Figure 24. In each of these stages, gait data was recorded through the MVN system, as shown in Figure 25.

Figure 24 - A subject playing the emotions in the experimental work

Each emotion was performed three times, producing a total of 81 recordings. In this study only the motion data recorded in the third phase was used, which consisted of the acted emotion. There was, however, data corruption that occurred within three recordings (2 Joy and 1 Grief), resulting in 78 files available for classification training and testing. Although this thesis research only focuses on the third phase, the recordings were part of a broader database, with the other two phases being utilised in other works.

In training the classifiers, a ground truth must be established. Some studies [48], [43], [45] use the actor's intended emotion as the ground truth. Alternatively, the emotion observed by a human observer is considered as the true emotion when training the classifier [26], [25], [47]. Sometimes there is a disagreement between the emotion observed by the human

observer and the actual emotion portrayed. In this study the emotion sent to the actor via an SMS was employed as the ground truth.

4.4 MVN Studio

The motion data measured by the sensors is sent wirelessly to a computer and is recorded via the MVN Studio Program. This software package provides simulation of the motion capture files for verification, as shown in Figure 25. MVN Studio can trim recordings, view variance within the data attributes, provide a selection of different motion data characteristics to export, and offers various exported file formats.

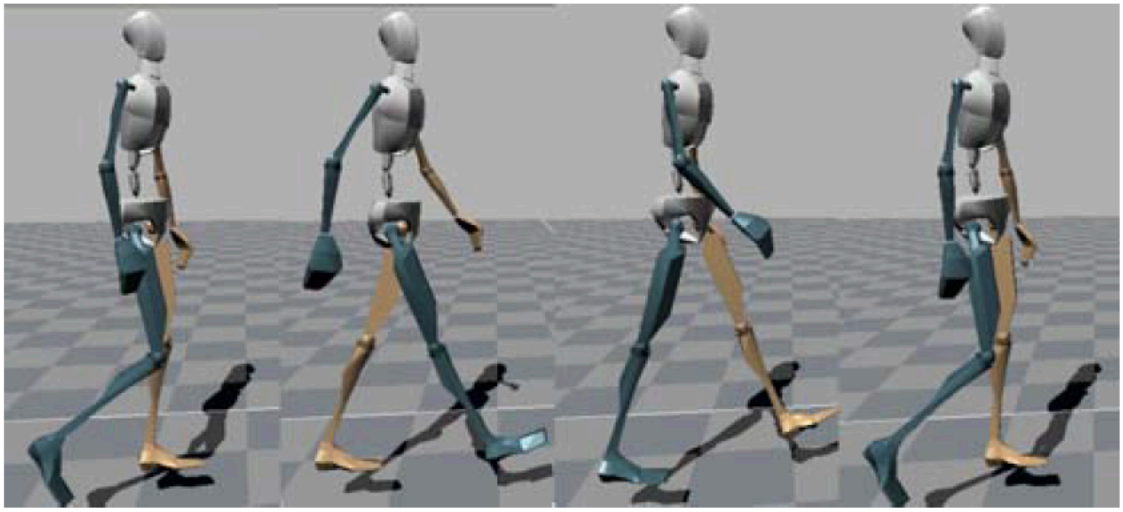


Figure 25 - Gait Cycle as reconstructed in MVN Studio [90]

MVN Studio exports all of the motion data into an MVNX file, which is based upon an XML format. Figure 26 shows the format of a section of an MVNX file. The sample data shown in Figure 25 represents the data associated with a four frames of motion from a gait cycle. The

data used in our experiments is sampled at every 10th frame of motion data produced by the motion capture device.

```
</joints>
<frames segmentCount= "..." sensorCount= "..." jointCount= "...">
  <frame time= "..." index= "..." type ="normal">
  <orientation>GBqseg1 GBqseg2 ...etc... GBqseg23</orientation>
  <position>Gposseg1 Gposseg1 ...etc... Gposseg23</position>
  <velocity>GVseg1 GVseg2 ...etc... GVseg23 </velocity>
  <acceleration>Gaseg1 Gaseg2 ...etc... Gaseg23 </acceleration>
  <angularVelocity>GWseg1 GWseg2 ...etc... GWseg23</ angularVelocity >
  <angularAcceleration>GAWseg1 GAWseg2 ...etc... GAWseg23</angularAcceleration>
  <sensorAcceleration>Sasen1 Sasen2 ...etc... Sasen17 </sensorAcceleration>
  <sensorAngularVelocity>SWsen1 SWsen2 ...etc... SWsen17</sensorAngularVelocity>
  <sensorMagneticField>Smsen1 Smsen2 ...etc... Smsen17</sensorMagneticField>
  <sensorOrientation>GSqsen1 GSqsen2 ...etc... GSqsen17</sensorOrientation>
  <jointAngle>jjnt1 jjnt2 ...etc... jjnt22</jointAngle>
   <jointAngleXZY>Jjnt1 Jjnt2 ...etc... Jjnt22</jointAngleXZY>
  <centerOfMass>GCseg1 GCseg2 ...etc... GCseg23 </centerOfMass>
  <marker> "name" </ marker >
  </frame>
```

Figure 26 - Sample MVNX format [2]

The MVN System records acceleration, angular velocity, magnetic field, and orientation of different body segments. MVN Studio processes the data captured by the sensors and can export the following parameters for each segment: orientation, position, velocity, acceleration, angular velocity, angular acceleration and centre of mass and joint angle. In this study, we chose to export only the position, velocity and acceleration of each segment into our MVNX file.

4.5 Motion Data Extraction

Matlab was utilised to import the MVNX files and extract the data on the position, velocity and acceleration of each segment, storing them in separate arrays. Initially, the MVNX file was loaded into Matlab. Each line of the file was read until the tags <position>, <velocity> or <acceleration> were found. The data contained within these tags were then assigned to the corresponding array, one frame at a time. For example, for the first frame of readings, the data was assigned to the first row of their corresponding arrays as shown below:

$$Pos\,[1,:]=\{p_{seg1x}, p_{seg1y}, p_{seg1z}, p_{seg2x}, p_{seg2y}, p_{seg2z}\ldots etc\ldots p_{seg23x}, p_{seg23y}, p_{seg23z}\} \quad (17)$$

$$Vel[1,:] = \{v_{seg1x}, v_{seg1y}, v_{seg1z}, v_{seg2x}, v_{seg2y}, v_{seg2z}\ldots etc\ldots v_{seg23x}, v_{seg23y}, v_{seg23z}\} \quad (18)$$

$$Accel[1,:] =\{a_{seg1x}, a_{seg1y}, a_{seg1z}, a_{seg2x}, a_{seg2y}, a_{seg2z}\ldots etc\ldots a_{seg23x}, a_{seg23y}, a_{seg23z}\} \quad (19)$$

Similarly, the second frame was assigned to the second row of the corresponding arrays:

$$Pos\,[2,:]=\{p_{seg1x}, p_{seg1y}, p_{seg1z}, p_{seg2x}, p_{seg2y}, p_{seg2z}\ldots etc\ldots p_{seg23x}, p_{seg23y}, p_{seg23z}\} \quad (20)$$

$$Vel[2,:] = \{v_{seg1x}, v_{seg1y}, v_{seg1z}, v_{seg2x}, v_{seg2y}, v_{seg2z}\ldots etc\ldots v_{seg23x}, v_{seg23y}, v_{seg23z}\} \quad (21)$$

$$Accel[2,:] =\{a_{seg1x}, a_{seg1y}, a_{seg1z}, a_{seg2x}, a_{seg2y}, a_{seg2z}\ldots etc\ldots a_{seg23x}, a_{seg23y}, a_{seg23z}\} \quad (22)$$

Since there were 23 segments, the position, velocity and acceleration arrays contained 69 columns. The number of rows varied depending on the number of frames within the recording. These arrays could then be used to calculate the appropriate features. The

calculated features were exported into a CSV file format were read by the classifier. In the CSV file, each row related to a different instance. In our case, each row represented a different recording. Each column corresponded to a different attribute used by the machine learning algorithm, with the last column designated to be the class or state of the attribute. In our case, the last column was used to store the “ground truth” emotion corresponding to that recording.

4.6 Classification

In our analysis, the Waikato Environment for Knowledge Analysis (WEKA) classifier toolbox [27] was used for classification (Figure 27). WEKA is a graphical user interface based software consisting of a collection of machine learning algorithms. It was developed by the University of Waikato on a Java Platform and is an open source software, freely available through general public license. WEKA is a workbench to data mine large amount of data and can load data files in the form of a CSV table file.

WEKA allows us to quickly experiment on a variety of datasets using a range of learning algorithms. The Graphical User Interface, WEKA Explorer shown in Figure 27, allows a classification model to be built, and its performance analysed, without the need for any code to be written.

The classification results; comprising of accuracy, error rates and a confusion matrix for each algorithm; were stored in a classification log, as shown in Figure 28. This process could be repeated for multiple classification algorithms.

In all of the classification models, ten-fold cross validation was deployed to obtain a more reliable accuracy.

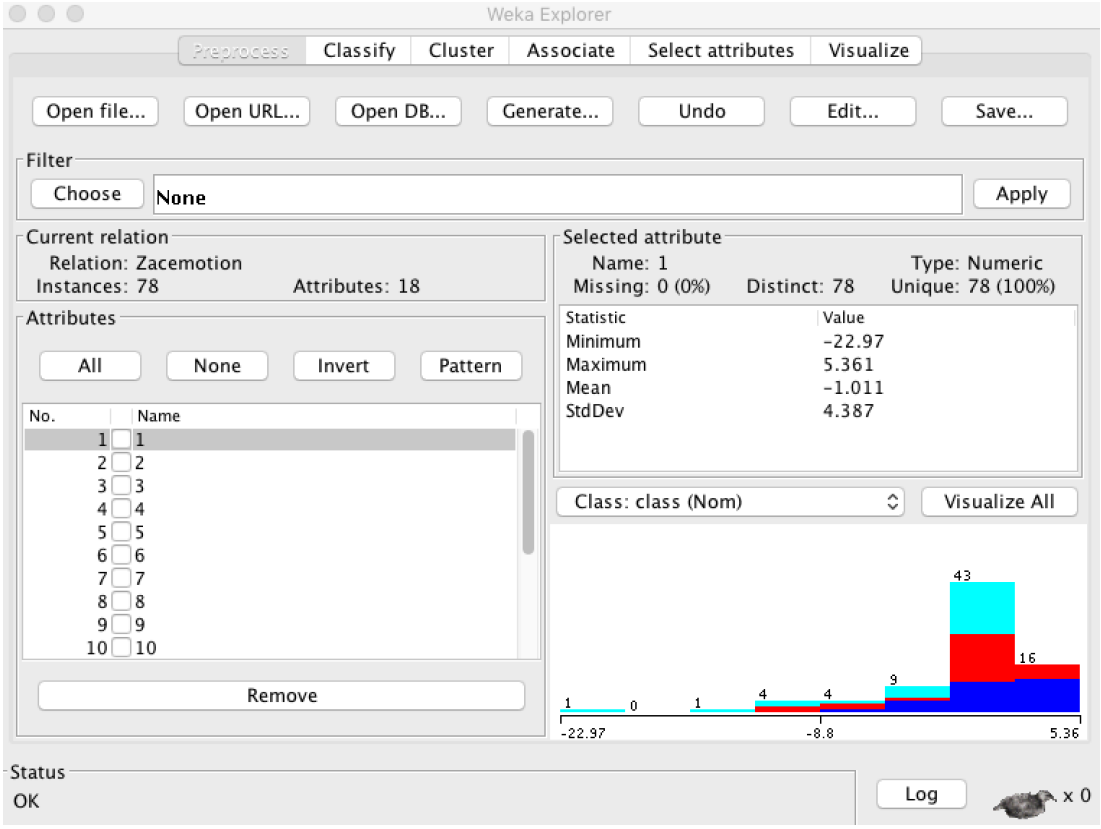


Figure 27 - WEKA Classification Toolbox

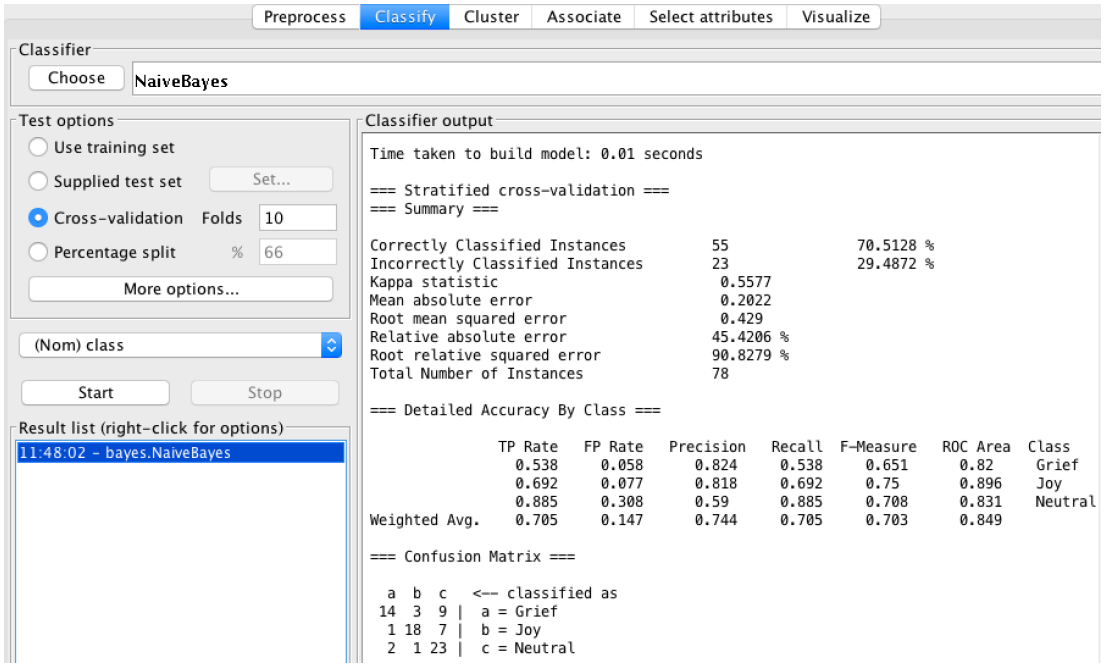


Figure 28 - WEKA classification Output

4.7 Extraction of Local Features

When applying Kapur local method, we deployed the mean of velocity and acceleration; and the standard deviation of position, velocity and acceleration of each of the markers. A sample matrix of three frames of velocity motion data is shown in equation 23. A similar structure was used for both the position and the acceleration.

$$v = \begin{matrix} \left[\mathrm{v}_{\mathrm{seg1x}}, \mathrm{v}_{\mathrm{seg1y}}, \mathrm{v}_{\mathrm{seg1z}}, \mathrm{v}_{\mathrm{seg2x}}, \mathrm{v}_{\mathrm{seg2y}}, \mathrm{v}_{\mathrm{seg2z}}, \ldots \mathrm{v}_{\mathrm{seg23x}}, \mathrm{v}_{\mathrm{seg23y}}, \mathrm{v}_{\mathrm{seg23z}},\right]_{time1} \\ \left[\mathrm{v}_{\mathrm{seg1x}}, \mathrm{v}_{\mathrm{seg1y}}, \mathrm{v}_{\mathrm{seg1z}}, \mathrm{v}_{\mathrm{seg2x}}, \mathrm{v}_{\mathrm{seg2y}}, \mathrm{v}_{\mathrm{seg2z}}, \ldots \mathrm{v}_{\mathrm{seg23x}}, \mathrm{v}_{\mathrm{seg23y}}, \mathrm{v}_{\mathrm{seg23z}},\right]_{time2} \\ \left[\mathrm{v}_{\mathrm{seg1x}}, \mathrm{v}_{\mathrm{seg1y}}, \mathrm{v}_{\mathrm{seg1z}}, \mathrm{v}_{\mathrm{seg2x}}, \mathrm{v}_{\mathrm{seg2y}}, \mathrm{v}_{\mathrm{seg2z}}, \ldots \mathrm{v}_{\mathrm{seg23x}}, \mathrm{v}_{\mathrm{seg23y}}, \mathrm{v}_{\mathrm{seg23z}},\right]_{time3} \end{matrix} \quad (23)$$

The mean and standard deviation of each column within the matrix was calculated, resulting in a 1 x 69 sized matrix for each of the five features stated above. These five matrices were exported into a .csv file side by side, resulting in 345 columns and one row used for the first

motion recording. This process was repeated for each recording, placing features from each successive file into a new row, resulting in a total of 345 columns and 68 rows from the 68 files. The ground state of the emotion of each recording was manually added into the 346th column, as the class for WEKA.

4.8 Extraction of Global Features

In applying the global Zacharatos method, the percentage of narrowing down, prospective focus of movement (as represented by the four extremity vectors) and the velocity, acceleration and jerk of both hands and feet were deployed. The first frame of the y-axis of the head was used as the initial head position in calculating P_{ND} according to (24).

$$P_{ND} = \frac{Pos_{Head-yaxis}(1) - Mean(Pos_{Head-yaxis})}{Pos_{Head-yaxis}(1)} \quad (24)$$

The eye direction was determined using the shoulder and head position, as shown in equations 25-34, for deployment in the face feature vectors.

$$V_x = pos_{headx} - pos_{Rshouldx} \quad (25)$$

$$V_y = pos_{heady} - pos_{Rshouldy} \quad (26)$$

$$V_z = pos_{headz} - pos_{Rshouldz} \quad (27)$$

$$W_x = pos_{headx} - pos_{Lshouldx} \quad (28)$$

$$W_y = pos_{heady} - pos_{Lshouldy} \quad (29)$$

$$W_z = pos_{headz} - pos_{Lshouldz} \quad (30)$$

$$Eye_x = V_y * W_z - V_z * W_y \quad (31)$$

$$Eye_y = V_x * W_x - V_x * W_z \quad (32)$$

$$Eye_z = V_x * W_y - V_y * W_x \quad (33)$$

$$F = [Eye_x, Eye_y, Eye_z] \quad (34)$$

The prospective focus of the movement was then calculated using equations 35-38 for each frame, with the results stored into an array.

$$DotLHandDirect = F \cdot Pos_{LHand} \quad (35)$$

$$DotRHandDirect = F \cdot Pos_{RHand} \quad (36)$$

$$DotLFootDirect = F \cdot Pos_{LFoot} \quad (37)$$

$$DotRFootDirect = F \cdot Pos_{RFoot} \quad (38)$$

The velocity, acceleration and jerk of the four extremities were calculated as a single vector, rather than one for each dimension. A sample calculation of velocity for the right hand is shown in Equation 39.

$$RHandVel = \sqrt{[mean(vel_{RHandx})]^2 + [mean(vel_{RHandy})]^2 + [mean(vel_{RHandz})]^2} \quad (39)$$

This velocity calculation was repeated for each of the other three extremities. Additionally, acceleration was calculated in a similar manner for all four extremities. There was no jerk data extracted from the motion capture data suit. Instead the first derivative was calculated for each of the extremities in each axis (e.g. Jerk in the x-axis of the Right hand is shown in Equation 40). This process of derivation was repeated in x, y and z directions for each of the four extremities.

$$Jerk_{RHandx} = \frac{\partial a}{\partial t} acel_{RHandx} \quad (40)$$

The $P_{nd,}$ mean values of each of the four prospective focus arrays, and the overall velocity, acceleration and jerk of the four extremities were combined into one matrix of size 1 x 17 for the global features. They were exported into a .csv file resulting in 17 columns and one row being utilised for the first emotion data recording. This process was repeated for each recording, placing features from each successive file into a new row. A total of 17 columns and 68 rows were filled from the 68 motion files. The ground state of the emotion of each recording was then manually added to the 18^{th} column to acted as class for the WEKA classification.

4.9 Extraction of Combined Features

The local features and global features were then combined into one matrix of size 1 x 362 for the combination feature set. They were exported into a .csv file resulting in 362 columns and one row being utilised for the first emotion data recording. This process was repeated for each recording, placing features from each successive file into a new row. A total of 362 columns and 68 rows were filled from the 68 motion files. The ground state of the emotion of each recording was then manually added to the 363^{rd} column to act as a class for the WEKA classification.

4.10 Extraction of Additional Features

Only the Weight and Space components were deployed as extra features. Weight was calculated as shown in equations 41-42.

$$V_i^{k^2} = (x_i^k - x_{i-1}^k)^2 + (y_i^k - y_{i-1}^k)^2 + (z_i^k - z_{i-1}^k)^2 \qquad (41)$$

The weight feature in the i-th frame was then obtained by adding the weight of each marker

$$\text{Weight}_i = \text{Weight}_i^{\text{Root}} + \text{Weight}_i^{\text{Toe}} + \text{Weight}_i^{\text{Finger}} \qquad (42)$$

where

$$\text{Weight}_i^{\text{Root}} = a_r V_i^{\text{Root}^2}$$

$$\text{Weight}_i^{\text{Toe}} = a_t(V_i^{\text{LToe}^2} + V_i^{\text{LToe}^2})$$

$$\text{Weight}_i^{\text{Finger}} = a_f(V_i^{\text{LFinger}^2} + V_i^{\text{RFinger}^2})$$

Root is defined as the centre of the body

This resulted in an array of the Weight for each frame of motion. The mean and standard deviation of this array was calculated and stored as a feature for each file.

Shape-flow was calculated based of the volume of space taken up by the actor in three dimensions, as shown in equation 43.

$$ShapeFlow = \left(Max_{xpos} - Min_{xpos}\right) * \left(Max_{ypos} - Min_{ypos}\right) * \left(Max_{zpos} - Min_{zpos}\right) \quad (43)$$

Shape-flow was created for each frame and stored in an array. The mean and standard deviation of this area was calculated and stored as a feature for each motion data file.

Shaping was determined by the change of the root marker in the y-axis and z-axis. The variance of the root marker's motion data was calculated, as shown in equations 44 and 45.

$$Shaping\_y = Variance(root_y) \quad (44)$$

$$Shaping\_z = Variance(root_z) \quad (45)$$

The Horizontal and Vertical Symmetry was calculated as shown in equations 46 and 47.

Horizontal Symmetry

$$= \left| \frac{\left\lfloor pos_{rootx} - pos_{\ LeftShoulderx} \right\rfloor - \left\lfloor pos_{rootx} - pos_{\ RightShoulderx} \right\rfloor}{\left\lfloor pos_{\ RightShoulderx} - pos_{\ LeftShoulderx} \right\rfloor} \right| \quad (46)$$

Vertical Symmetry

$$= \left| \frac{\left\lfloor pos_{rooty} - pos_{\ LeftShouldery} \right\rfloor - \left\lfloor pos_{rooty} - pos_{\ RightShouldery} \right\rfloor}{\left\lfloor pos_{\ RightShouldery} - pos_{\ LeftShouldery} \right\rfloor} \right| \quad (47)$$

The local features outlined in Section 4.7, global features outlined in Section 4.8, and the eight extra features outlined in Section 4.9 were combined into one matrix of size 1 x 370

for the extra feature set. The combined feature set was exported into a single .csv file, resulting in 370 columns and one row being utilised for the first emotion data recording. This process was repeated for each recording, placing features from each successive file into a new row. A total of 370 columns and 68 rows were filled from the 68 motion files. The ground state of the emotion of each recording was then manually added to the 371st column to act as a class for the WEKA classification.

4.11 Summary

In this chapter an overview of the experimental set up used in this thesis was provided. The main device was the XSens MVN motion capture system that deployed inertial sensors to measure and kinematics parameters association with motion. The nature of data produced by the motion capture device was discussed and the approach used to process it was described. We then presented the software packages utilised to export the data motion, extract features and run classification algorithms. Finally, the method of extracting the various feature sets was presented.

5 VALIDATION

5.1 Introduction

In this chapter the classification results from the various feature sets will be presented. The significance of these results, limitations of our work and potential directions for future research is also discussed. For each feature set, the accuracy from ten-fold cross validation is presented for all of the nine different classifiers deployed.

The results for local features, global features, combined feature set and the additional feature set are provided in Sections 5.2, 5.4, 5.6 and 5.8, respectively. Since the local and global feature results are produced based on the previous work, a comparison of our results are made in Section 5.3 and Section 5.5 against equivalent results reported in the literature. Section 5.7 discusses the results of combining local and global features in a single classifier and Section 5.9 examines the results of the additional features provided. The significance of our work is presented within Section 5.10, and the limitation of our work and ideas for future research is described in Section 5.11.

5.2 Local Feature Results

For the local feature set, we deployed the mean of velocity and acceleration; and the standard deviation of position, velocity and acceleration of each of the markers. A classification model was built in WEKA deploying the following classifiers: BayesNet, Naïve Bayes, MLP, RBF Network, SMO, IBk, J48, J48 Graft and Random Forest.

The accuracy of classification of these models using the local feature set with ten-fold cross validation is presented in Table 10, including the average and maximum performance.

The highest performance was produced by the IBk classifier with an accuracy of 87.2%. The lowest performance was from Bayes Net and Naïve Bayes, both with an accuracy of 71.8%. These local features resulted in an average accuracy of 87.2% across all of the nine classifiers.

Classifier	**Local Feature Accuracy (%)**
Bayes Net	71.8
Naïve Bayes	71.8
Multi-Layer Perceptron	84.6
RBF Network	73.1
SMO	82.1
IBk	87.2
J48	80.8
J48Graft	79.5
Random Forest	80.8
Average	79.1
Max	87.2

Table 10 - Local Feature Accuracy

5.3 Benchmarking against Kapur Local Method

The result from both Kapur et al.'s data and our data demonstrates that using raw joint data alone can lead to moderate performance across a variety of classifiers. Our average (79.1%) and highest accuracy (87.2%) was lower than their average (84.2%) and highest accuracy (91.8%). However, their lowest accuracy (66.2%) was lower than our lowest accuracy (71.8%). These results are difficult to compare due to the difference in data collection technique, number of markers and number of emotions. This highlights the need for comparison of different techniques on a dataset that contains the same data and the same emotions. Otherwise, as demonstrated in this research, the same techniques performed on two different data sets can result in two differing results and hence reduces the reliability of any comparisons made.

On both tests, however, Naïve Bayes was the lowest performing classifier, suggesting that it was not the most appropriate classifier when only using raw joint data. SMO and MLP performed strongly in both tests, suggesting that they well suited to these styles of features. IBk, however, outperformed both of SMO and MLP in our study, but this classifier was not tested by Kapur et al. This identifies the challenge when a small number of classifiers is deployed, as there may exist a higher performing classifier that is not applied.

5.4 Global Feature Results

For the global feature set, we deployed the percentage of narrowing down, prospective focus of movement (as represented by the four extremity vectors) and the velocity, acceleration and jerk of both hands and feet. A classification model was built in WEKA deploying the following classifiers: BayesNet, Naïve Bayes, MLP, RBF Network, SMO, IBk, J48, J48 Graft and Random Forest.

The accuracy of classification of these models using the global feature set with ten-fold cross validation is presented in Table 11, including the average and maximum performance.

Classifier	**Local Feature Accuracy (%)**
Bayes Net	71.5
Naïve Bayes	70.5
Multi-Layer Perceptron	73.1
RBF Network	73.1
SMO	68
IBk	82.1
J48	77
J48Graft	74.4
Random Forest	79.5
Average	74.4
Max	82.1

Table 11 - Global Feature Accuracy

The highest performance was produced by the IBk classifier with an accuracy of 82.1%. The lowest performance was from SMO, with an accuracy of 68%. These global features resulted in an average accuracy of 74.4% across all of the nine classifiers.

5.5 Benchmarking against Zacharatos Global method

The result from both Zacharatos' et al.'s and our data demonstrates that LMA alone can result in moderate performance across a variety of classifiers. Our average accuracy (74.4%) and highest accuracy (82.1%) are lower than Zacharatos' average accuracy (84.2%). However, they deployed a different category of emotions and utilised a different data collection technique. This again highlights the need for comparison of different techniques on a dataset that contains the same motion data and emotions.

In our data, SMO was the lowest performing classifier and IBK was again the most successful. Zacharatos et al. only tested the MLP classifier, but in our data there were four classifiers that produced a higher accuracy. This reinforces the need to consistently test a wide variety of classifiers.

5.6 Combined Feature Results

The accuracy of classification using a combination of local and global features with ten-fold cross validation is presented in Table 12. The classification model was built in WEKA

deploying the following classifiers: BayesNet, Naïve Bayes, MLP, RBF Network, SMO, IBk, J48, J48 Graft and Random Forest. The results for local and global features are also placed within the same table for comparison. The table contains the difference between the combined feature set and the best performer out of the local and global feature sets for each classifier. The average and maximum accuracy across all classifiers is presented for each feature set. The difference in average and maximum values across all classifiers, and the improvements upon the average is also displayed.

Classifier	**Feature Set 1 Local**	**Feature Set 2 Global**	**Feature Set 3 Combined**	**Difference**
Bayes Net	71.8	71.5	71.8	0
Naïve Bayes	71.8	70.5	71.8	0
Multi-Layer Perceptron	84.6	73.1	84.6	0
RBF Network	73.1	73.1	75.6	+2.5
SMO	82.1	68	85.9	+3.8
IBk	87.2	82.1	88.5	+1.3
J48	80.8	77	87.2	+6.4
J48Graft	79.5	74.4	85.9	+6.4
Random Forest	80.8	79.5	80.8	+0
Average	79.1	74.4	**81.3**	**+2.2**
	improvement on average			**+2.3**
Max	87.2	82.1	**88.5**	**+1.3**

Table 12 - Combined Feature Set Results

5.8 Discussion on Combining Features

There was no improvement achieved by combining features for Bayes Net and Naïve Byes. This might be due to the low performance of both classification algorithms using only Local features and Global Features independent from each other. Multi-Layer Perceptron Algorithm and Random Forest also showed no improvement by combining the features together. The largest improvement was produced by J48 and J48 Graft algorithms, resulting in an increase of 6.4%. By combining local and global features into a single classifier, the average accuracy increased by 2.3%. A new high performance was achieved at 88.5% through the deployment of IBk algorithm.

Combining the two types of features never resulted in a decrease of accuracy of the systems. On the contrary, for most of the classifiers, the combination resulted in an improved performance. Similar to what has been reported in the literature in other applications, combining local and global features results in a higher accuracy in automatic classification.

5.9 Additional Feature Results

For the additional features, as the combined feature set we deployed the Weight, shape flow, shaping, vertical and horizontal symmetry. A classification model was built in WEKA deploying the following classifiers: BayesNet, Naïve Bayes, MLP, RBF Network, SMO, IBk, J48, J48 Graft and Random Forest.

The accuracy of classification using a combination of the local and global feature sets with the additional features with ten-fold cross validation is presented in Table 13. The results for local features, global features, and the combined features are also placed within the same table for comparison. The table contains the difference between the additional feature set and the best performer in local and global features for each classifier, as well as the difference in accuracy of the additional feature set against the combined feature set. The accuracy of the average and maximum values across all classifiers is presented for each feature set. The difference in average and maximum values across all classifiers and the improvements on the average are also shown.

	Feature Set 1 Kapur	Feature Set 2 Zacharatos	Feature Set 3 Combined	Difference	Feature Set 4 My Combination	Improvement upon Set 1-2	Improvement upon Set 1- 3
Naive Bayes	71.8	70.5	71.8	0	71.8	0	0.0
Multi-Layer Perceptron	84.6	73.1	84.6	0	87.2	2.6	2.6
RBF Network	73.1	73.1	75.6	2.5	76.9	3.8	1.3
SMO	82.1	68	85.9	3.8	88.5	6.4	2.6
IBk	87.2	82.1	88.5	1.3	92.3	5.1	3.8
J48	80.8	77	87.2	6.4	87.2	6.4	0.0
J48Graft	79.5	74.4	85.9	6.4	84.6	5.1	-1.3
Random Forest	80.8	79.5	80.8	0	82.1	1.3	1.3
Average	79.1	74.4	**81.3**	2.2	83.8	3.8	1.3
		improvement on average		3.0		4.7	2.5
Max	87.2	82.1	**88.5**	1.3	**92.3**	**5.1**	**3.8**

Table 13 - Additional Feature Set Classification Results

5.10 Discussion on Additional Features

The new combination (feature set four) resulted in an increased performance from the majority of classifiers deployed. The IBk classifier, produced the highest performance when examining feature set with an accuracy of 92.3%.

With the inclusion of extra features, most of the classifiers showed an increase in their accuracy. Only J48Graft had a very small (1.3%) decrease in the classification rate. The new additional features increased the average performance by 2.5%, compared to the deployment of the combination feature set. This feature set also resulted in an increased average performance of 4.7%, compared to the best performance of deploying either local and global feature set by themselves.

These results support the statement by Moore and Yamamoto [67] that movement must be approached at multiple levels if it is to be properly understood. They also support the hypothesis that the classification performance increases from adding additional LMA components.

6 CONCLUSION

6.1 Overview of work

We presented a method of improving performance of affect recognition using body language. Within existing literature, the majority of systems deploy either local or global features independently. In this study, the impact of combining local and global features into a single classifier was explored in automatic affect recognition based on body language.

A motion capture suit was deployed to record 68 different walking movements from nine actors performing happy, neutral and grieving emotions. These data files were exported through MVN studio and imported into Matlab. Matlab extracted and calculated the required feature sets, then exported these into a CSV file ready for classification. Feature set one contained the Kapur local method features consisting of the mean of velocity and acceleration; and the standard deviation of position, velocity and acceleration of each of the markers. Feature set two contained the Zacharatos global method features, consisting of the percentage of narrowing down, prospective focus of movement (as represented by the four extremity vectors) and the velocity, acceleration and jerk of both hands and feet. Feature set three was an amalgamation of the Kapur local method features and the Zacharatos

global method features. Features set four, containing the expanded LMA set, produced an even higher accuracy across multiple classifiers.

A classification model was built in WEKA using ten-fold cross validation whilst deploying the following classifiers: BayesNet, Naïve Bayes, Multi-Layer perceptron, RBF Network, SMO, IBk, J48 and Random Forest.

The results of classification from the combination feature set three outperformed classification undertaken with only the local features or global features individually for a variety of classifiers. Several classification algorithms achieved an even higher accuracy when deploying feature set 4, which contained the additional LMA components.

6.2 Significance of Results

Combining the two types of features never resulted in a decrease in accuracy of the results. On the contrary, for most classifiers, the combination resulted in an improved performance. Hence, it can be confidently stated that using a combination of local and global features results in a more robust and reliable method for affect recognition using gait by improving accuracy across a range of classifiers.

Features set four, containing the expanded LMA set, produced an even higher accuracy across multiple classifiers. This supports the hypothesis that deploying LMA components together impacts positively upon the classification accuracies. Therefore, when deploying LMA as part of movement classification, it is beneficial to deploy as many LMA components as practical.

As discussed in Chapter 2, body language provides useful information in communicating affect. Although the use of multiple modalities in affect recognition tends to outperform a single modality being used by itself, improving any single one of them in isolation will result in improvement when combined with other information. This research demonstrates a different approach to choosing features deployed in classification than previously reported in the literature, producing more accurate results. Better automatic affect recognition rates can lead to increased application of the approach, as well as its usefulness and reliability.

Our results show that the same classifier can have a range of performance depending upon the style of feature set used. For example, Multi-layer Perceptron and SMO both had very poor accuracy with the Zacharatos global method, but performed well with the others. In addition, an approach can be better suited to different classifiers, as not all classifiers obtained the same performance increase,

which may relate to the assumptions made for each individual classifier. The previous affect recognition studies, however, appear to only report the performance of the classifier with the highest accuracy. The success of a classifier can depend on a number of factors including the size of the training data, number of emotion categories, method of data collection and the number of features used for classification. To determine the impact of these factors on various classifiers, the performance of a variety of classifiers should be reported, even when the accuracy of each classifier is poor.

6.3 Limitations and Future Work

There are a number of limitations associated with the research conducted in this thesis which can be overcome with further work. These limitations and possible methods to address them are described in this section.

Although it is assumed that the improvement from using multiple LMA features in combination with each other in a gait system would improve a multiple modality system, this has yet to be tested. Further work could compare a multiple modality system deploying a single body language feature set (feature set one or feature set two), and one deploying a combined local and global feature set (feature set three) or a multiple component LMA set (feature set four) for the body language modality.

Both feature set three and feature set four demonstrated a performance increase across multiple different classifiers when compared against the Kapur local method and the Zacharatos global method. This improvement, however, may be due to over fitting to our data. That is, even though ten-fold cross validation was used, since the data set was relatively small this feature combination might have been suited to this particular data set. Our dataset was small in comparison to those deployed in facial recognition, with only a total of 68 recordings. Further testing should, therefore, be conducted across different data sources, particularly ones with a larger collection of recordings. Feature sets and techniques explored in other studies may also face this problem and only be successful with their own data. A more reliable comparison would be achieved by testing different feature sets and techniques against multiple data sets.

As outlined in chapter 2, Moore and Yamamoto [67] state that the first principle underlying LMA is that movement is a process of change. The start and end positions are not only important but the pattern of change of body position helps to communicate emotion. We simplified this process by utilising mean values and variance of the features. This allowed us to feed in single values in each cell for classification. A more thorough approach would be to undertake discriminant analysis to determine the most distinguishing features, then feed these features into our classifier model as a time series, rather than utilising single values based on

the mean and variance. However, to isolate the improvement resulting from combining local and global features into a single classifier, discriminate analysis was not applied to the data. Instead only the mean and standard deviation of data was used rather than using it as a time series in building our classification model. Post processing techniques, such as segmentation and weighting [45], [46], or Dimensional Reduction [43], [44], have previously been deployed in affect recognition with success. These techniques, however, have yet to be applied to LMA which may improve its performance. These post-processing techniques may also benefit the combination feature sets three and four.

Our research only examined the three acted emotions of grieving, neutral and happy by nine professional actors. Acted emotions, however, can be exaggerated and less subtle and the inconsistent performance of emotions by non-actors may be more indicative of how emotions are naturally portrayed. Future research could replicate this project with natural emotions, such as the studies ([47], [39]) performed on emotions displayed whilst playing video games.

The three emotions studied are only a small portion of emotions that are displayed by humans and the results should be repeated with a set of recordings that contain a large number of emotions. Although it is unknown how scalable it is the number of emotions, based on the literature this approach would apply to a large number of emotions. In addition, all emotions that are contained within the database belong

to one of these three emotions. To further test the accuracy of the system, the classifier should be trained on these three emotions but tested on recordings that contain more than those three emotions. Rather than the system choosing the best fitting emotion, it would be required to place any emotion unsure of into a different category of unknowns.

Currently most of the studies use actors who can display extremes of emotions but the intensity of such extremes vary in different people. For example, sometimes we could feel a little bit angry and other times really angry. This could lead to differences in how much is communicated in our gait. In addition, the tendency to happiness and anger can occur at the same time. An alternative way of approaching affect recognition would be assigning a specific confidence rating to an emotion. For example, rather than determining an emotion as happy or sad, it might be better to identify it as 60% confidence of being happy, and 40% chance of being sad. Using a percentage confidence rating could allow recognition of mixed emotions rather than single extreme emotions. The focus of this study was on three distinct emotions of grieving, happy and neutral. Future work could investigate the deployment of confidence ratings, rather than distinct boundaries.

In our research, a motion capture device was deployed which required subjects to wear a specially designed suit with inertial sensors. Although a large amount of data is captured in this method to validate the concept, it is not a practical approach in

real time applications. Our research should be replicated with the deployment of perception systems, such as Kinect or video cameras, which do not require the subject to wear any specialised equipment.

The performance of combinations of other feature sets currently being deployed in affect recognition from gait should be analysed. The accuracy of these feature set combinations may change depending on the data source and the classifier deployed. Some classifiers appear to work better with different feature sets. Some classifiers, such as Naïve Bayes, appear to benefit insignificantly from the added extra features. There was only a difference of 1.5% between the Zacharatos global method and the Kapur local method features using these classifiers, and there was no improvement when deploying the combination feature set three, or with the extra information in feature set four. However, some studies ([38], [53]) recorded Naïve Bayes with Adaboost as their highest performing classifier. Their feature set may combine with a different complementary feature set to produce a higher performance with Naïve Bayes classification.

Since each study uses its own dataset and data detection method, it is difficult to compare the analysis and classification methods of our technique against other previous literature. Studies using a common data set and detection method need to be undertaken to enable comparison of various processing options (including raw data) to determine their comparative effectiveness. This thesis research

demonstrated its effectiveness by comparing the combined features and additional feature approach with the Kapur local method and the Zacharatos global method on the same data set. However, this work could be expanded by applying all of these methods on another publically available dataset for further comparison.

REFERENCES

[1] M. S. Pang-ning Tan, Vipin Kumar, *Introduction to Data Mining*. Pearson Education, 2006.

[2] XSens, "MVN User Manual," Document MV0319P, Revision H, ed. www.xsens.com, 2013.

[3] K. Hachimura, K. Takashina, and M. Yoshimura, "Analysis and evaluation of dancing movement based on LMA," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, 2005, pp. 294-299: IEEE.

[4] L. T. Kozlowski and J. E. Cutting, "Recognizing the sex of a walker from a dynamic point-light display," *Perception & Psychophysics,* vol. 21, no. 6, pp. 575-580, 1977.

[5] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the psychonomic society,* vol. 9, no. 5, pp. 353-356, 1977.

[6] R. D. Walk and K. L. Walters, "Perception of the smile and other emotions of the body and face at different distances," *Bulletin of the Psychonomic Society,* vol. 26, no. 6, pp. 510-510, 1988.

[7] S. Brownlow, A. R. Dixon, C. A. Egbert, and R. D. Radcliffe, "Perception of movement and dancer characteristics from point-light displays of dance," *The Psychological Record,* vol. 47, no. 3, p. 411, 1997.

[8] M. Demeijer, "THE CONTRIBUTION OF GENERAL FEATURES OF BODY MOVEMENT TO THE ATTRIBUTION OF EMOTIONS," (in English), *Journal of Nonverbal Behavior,* Article vol. 13, no. 4, pp. 247-268, Win 1989.

[9] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology,* vol. 28, no. 6, pp. 879-896, 1998.

[10] Johansson.G, "VISUAL-PERCEPTION OF BIOLOGICAL MOTION AND A MODEL FOR ITS ANALYSIS," (in English), *Perception & Psychophysics,* Article vol. 14, no. 2, pp. 201-211, 1973.

[11] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," (in English), *Perception,* Article vol. 33, no. 6, pp. 717-746, 2004.

[12] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Effort-Shape and kinematic assessment of bodily expression of emotion during gait," (in English), *Human Movement Science,* vol. 31, no. 1, pp. 202-221, 2012.

[13] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior,* vol. 28, no. 2, pp. 117-139, 2004.

[14] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition,* vol. 82, no. 2, pp. B51-B61, 2001.

[15] W. H. Dittrich, T. Troscianko, S. E. Lea, and D. Morgan, "Perception of emotion from dynamic point-light displays represented in dance," *Perception,* vol. 25, no. 6, pp. 727-738, 1996.
[16] B. de Gelder, "Why bodies? Twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 364, no. 1535, pp. 3475-3484, 2009.
[17] A. Kale *et al.*, "Identification of humans using gait," *Image Processing, IEEE Transactions on,* vol. 13, no. 9, pp. 1163-1173, 2004.
[18] S. Van Der Zee, R. Poppe, P. Taylor, and R. Anderson, "To freeze or not to freeze: A motion-capture approach to detecting deceit," in *Proceedings of the Hawaii International Conference on System Sciences, Kauai, HI*, 2015.
[19] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, and A. Abdelgawad, "Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 453-457.
[20] R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri, "Automated Laughter Detection From Full-Body Movements," *IEEE Transactions on Human-Machine Systems,* vol. 46, no. 1, pp. 113-123, 2016.
[21] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey," *Affective Computing, IEEE Transactions on,* vol. 4, no. 1, pp. 15-33, 2013.
[22] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry,* vol. 32, no. 1, pp. 88-106, 1969.
[23] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology,* vol. 17, no. 2, pp. 124-129, 1971.
[24] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers,* vol. 18, no. 6, pp. 1371-1389, 2006.
[25] H. Zacharatos, C. Gatzoulis, Y. Chrysanthou, and A. Aristidou, "Emotion recognition for exergames using Laban movement analysis," in *6th International Conference on Motion in Games, MIG 2013*, Dublin, 2013, pp. 39-43.
[26] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data," in *Affective Computing and Intelligent Interaction, Proceedings*, vol. 3784, J. Tao and R. W. Picard, Eds. (Lecture Notes in Computer Science, Berlin: Springer-Verlag Berlin, 2005, pp. 1-7.
[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, no. 1, pp. 10-18, 2009.
[28] N. Nayak, R. Sethi, B. Song, and A. Roy-Chowdhury, "Motion pattern analysis for modeling and recognition of complex human activities," *Guide to Video Analysis of Humans: Looking at People,* 2011.

[29] H. Gunes and M. Piccardi, "Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 39, no. 1, pp. 64-84, 2009.
[30] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal,* vol. 2nd Quarter, 1998.
[31] C. Shan, S. Gong, and P. W. McOwan, "Beyond Facial Expressions: Learning Human Emotion from Body Gestures," in *BMVC*, 2007, pp. 1-10: Citeseer.
[32] C. Shizhi and T. YingLi, "Margin-constrained multiple kernel learning based multi-modal fusion for affect recognition," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 2013, pp. 1-7.
[33] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," (in English), *Journal on Multimodal User Interfaces,* vol. 3, no. 1, pp. 33-48, 2010.
[34] H. Park, J. I. I. Park, U. M. Kim, and N. Woo, "Emotion Recognition from Dance Image Sequences Using Contour Approximation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 3138, ed, 2004, pp. 547-555.
[35] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, 2011, pp. 305-311: IEEE.
[36] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM,* vol. 56, no. 1, pp. 116-124, 2013.
[37] K. Woo Hyun, P. Jeong Woo, L. Won Hyong, C. Myung Jin, and L. Hui Sung, "LMA based emotional motion representation using RGB-D camera," in *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, 2013, pp. 163-164.
[38] D. McColl, G. Nejat, "Determining the Affective Body Language of Older Adults during Socially Assistive HRI," (in English), *2014 IEEE/Rsj International Conference on Intelligent Robots and Systems (Iros 2014),* Proceedings Paper pp. 2633-2638, 2014.
[39] M. Garber-Barron and S. Mei, "Using body movement and posture for emotion detection in non-acted scenarios," in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, 2012, pp. 1-8.
[40] Y. Xiao, J. Yuan, and D. Thalmann, "Human-virtual human interaction by upper body gesture understanding," in *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*, 2013, pp. 133-142: ACM.

[41] H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, "Automatic Emotion Recognition Based on Body Movement Analysis: A Survey," *Computer Graphics and Applications, IEEE,* vol. 34, no. 6, pp. 35-45, 2014.
[42] G. Venture, H. Kadone, T. X. Zhang, J. Grezes, A. Berthoz, and H. Hicheur, "Recognizing Emotions Conveyed by Human Gait," (in English), *International Journal of Social Robotics,* Article vol. 6, no. 4, pp. 621-632, Nov 2014.
[43] A. A. Samadani, R. Gorbet, and D. Kulic, "Affective Movement Recognition Based on Generative and Discriminative Stochastic Dynamic Models," *Human-Machine Systems, IEEE Transactions on,* vol. 44, no. 4, pp. 454-467, 2014.
[44] A. A. Samadani, A. Ghodsi, and D. Kulic, "Discriminative functional analysis of human movements," (in English), *Pattern Recognition Letters,* Article vol. 34, no. 15, pp. 1829-1839, Nov 2013.
[45] J. Xu and S. Sakazawa, "Temporal fusion approach using segment weight for affect recognition from body movements," in *2014 ACM Conference on Multimedia, MM 2014*, 2014, pp. 833-836: Association for Computing Machinery, Inc.
[46] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *2nd International Conference on Affective Computing and Intelligent Interaction, ACII 2007* vol. 4738 LNCS, ed. Lisbon, 2007, pp. 59-70.
[47] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic Recognition of Non-Acted Affective Postures," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 41, no. 4, pp. 1027-1038, 2011.
[48] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of Affect Based on Gait Patterns," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 40, no. 4, pp. 1050-1061, 2010.
[49] A. Lim and H. G. Okuno, "The MEI Robot: Towards Using Motherese to Develop Multimodal Emotional Intelligence," (in English), *IEEE Transactions on Autonomous Mental Development,* Article vol. 6, no. 2, pp. 126-138, Jun 2014.
[50] D. Janssen, W. I. Schollhorn, J. Lubienetzki, K. Folling, H. Kokenge, and K. Davids, "Recognition of emotions in gait patterns by means of artificial neural nets," (in English), *Journal of Nonverbal Behavior,* Article vol. 32, no. 2, pp. 79-92, Jun 2008.
[51] N. Bianchi-Berthouze and A. Kleinsmith, "A categorical approach to affective gesture recognition," *Connection science,* vol. 15, no. 4, pp. 259-269, 2003.
[52] E. I. Barakova and T. Lourens, "Expressing and interpreting emotional movements in social games with robots," *Personal and Ubiquitous Computing,* journal article vol. 14, no. 5, pp. 457-467, 2010.
[53] D. McColl, C. Jiang, and G. Nejat, "Classifying a Person's Degree of Accessibility from Natural Body Language During Social Human-Robot Interactions," *IEEE Transactions on Cybernetics,* vol. PP, no. 99, pp. 1-15, 2016.

[54] R. Laban, *Principles of Dance and Movement Notation*. New York: Macdonald & Evans, 1956.
[55] T. Lourens, R. van Berkel, and E. Barakova, "Communicating emotions and mental states to robots in a real time parallel framework using Laban movement analysis," *Robotics and Autonomous Systems,* vol. 58, no. 12, pp. 1256-1265, 12/31/ 2010.
[56] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a Minimal Representation of Affective Gestures," *Affective Computing, IEEE Transactions on,* vol. 2, no. 2, pp. 106-118, 2011.
[57] D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, and M. C. Benfield, "Combining Local and Global Image Features for Object Class Recognition," in *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, 2005, pp. 47-47.
[58] L. Wang, H. Zhou, S. C. Low, and C. Leckie, "Action recognition via multi-feature fusion and Gaussian process classification," in *2009 Workshop on Applications of Computer Vision, WACV 2009*, Snowbird, UT, 2009.
[59] H. Yu and H. Liu, "Combining appearance and geometric features for facial expression recognition," in *6th International Conference on Graphic and Image Processing, ICGIP 2014*, 2015, vol. 9443: SPIE.
[60] I. Siddiqi and N. Vincent, "Combining global and local features for writer identification," in *Proceedings of the 11. Int. Conference on Frontiers in Handwriting Recognition, Montreal*, 2008.
[61] X. He, H. Zhang, W. Jia, Q. Wu, and T. Hintz, "Combining global and local features for detection of license plates in a video," in *Proceedings of Image and Vision Computing New Zealand*, 2007, pp. 288-293.
[62] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining global and local features for food identification in dietary assessment," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 1789-1792.
[63] H. Gunes and M. Piccardi, "Fusing face and body gesture for machine recognition of emotions," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, 2005, pp. 306-311.
[64] C. Shizhi, T. YingLi, L. Qingshan, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 2011, pp. 7-12.
[65] R. A. Calvo and S. D. Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Transactions on Affective Computing,* vol. 1, no. 1, pp. 18-37, 2010.
[66] J. A. Russell, "Core affect and the psychological construction of emotion," (in eng), *Psychol Rev,* vol. 110, no. 1, pp. 145-72, Jan 2003.
[67] C. L. Moore and K. Yamamoto, *Beyond Words: Movement Observation and Analysis*. Taylor & Francis, 2012.

[68] D. Chi, M. Costa, L. Zhao, and N. Badler, "The EMOTE model for effort and shape," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 173-182: ACM Press/Addison-Wesley Publishing Co.
[69] R. R. Bouckaert, "Bayesian network classifiers in weka for version 3-5-7," *Artificial Intelligence Tools,* vol. 11, no. 3, pp. 369-387, 2008.
[70] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338-345: Morgan Kaufmann Publishers Inc.
[71] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Transactions on Neural Networks,* vol. 1, no. 4, pp. 296-298, 1990.
[72] G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation," *IEEE Transactions on Neural Networks,* vol. 16, no. 1, pp. 57-67, 2005.
[73] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods,* pp. 185-208, 1999.
[74] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning,* vol. 6, no. 1, pp. 37-66, 1991.
[75] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
[76] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.
[77] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media LLC, 2006.
[78] A. P. K. Weinberger. (2015, 14/08/2016). *CS4780/CS5780: Machine Learning*. Available: http://www.cs.cornell.edu/courses/cs4780/2015fa/iframe/iframe-5/index.html
[79] I. H. Witten, *Data Mining: Practical Machine learning tools and techniques*, 3rd ed. Elsevier Inc, 2011.
[80] E. L. Graczyk, M. A. Schiefer, H. P. Saal, B. P. Delhaye, S. J. Bensmaia, and D. J. Tyler, "The neural basis of perceived intensity in natural and artificial touch," *Science Translational Medicine,* vol. 8, no. 362, pp. 362ra142-362ra142, 2016.
[81] C.-T. Chen. (14/8/2016). *Now Comes The Time to Defuzzify Neuro-Fuzzy Models*. Available: http://neuron.csie.ntust.edu.tw/homework/93/Fuzzy/%E6%97%A5%E9%96%93%E9%83%A8/homework_1/D9202102/welcome.htm
[82] L. Garrett. (2016, 14/08/2016). *Extra Classifier Research*. Available: https://ds.lclark.edu/laurelgarrett/2016/02/03/extra-classifier-research/
[83] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *International Journal of Engineering Research and Applications,* vol. 3, no. 5, pp. 605-610, 2013.
[84] C. Danilo. (2010, 14/08/2016). *Decision tree algorithm*

Weka tutorial. Available: http://www.uniroma2.it/didattica/WmIR/deposito/dectree_weka_tutorial.pdf

[85] B. B. K. Lily Amadeo and C. O. Suman Kumar Lama. (2015, 14/08/2016). *Real-Time Human Pose Recognition in Parts from Single Depth Images*. Available: http://web.cs.wpi.edu/~cs548/s15/Showcase/CS548S15_Showcase_Decision_Trees.pdf

[86] G. I. Webb, "Decision tree grafting from the all-tests-but-one partition," in *IJCAI*, 1999, vol. 2, pp. 702-707.

[87] M. Field, "Acquisition and distribution of synergistic reactive control skills," Doctor of Philosphy, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2014.

[88] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens MVN: full 6DOF human motion tracking using miniature inertial sensors," *Xsens Motion Technologies BV, Tech. Rep,* 2009.

[89] D. Roetenberg, P. J. Slycke, and P. H. Veltink, "Ambulatory position and orientation tracking fusing magnetic and inertial sensing," *IEEE Transactions on Biomedical Engineering,* vol. 54, no. 5, pp. 883-890, 2007.

[90] A. Hesami, F. Naghdy, D. Stirling, and H. Hill, "Perception of human gestures through observing body movements," in *Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on*, 2008, pp. 97-102.