2017

Automatic Affect Perception Based on Body Gait and Posture: A Survey

Benjamin Stephens-Fripp
University of Wollongong, bsf147@uowmail.edu.au

Fazel Naghdy
University of Wollongong, fazel@uow.edu.au

David Stirling
University of Wollongong, stirling@uow.edu.au

Golshah Naghdy
University of Wollongong, golshah@uow.edu.au

Automatic Affect Perception Based on Body Gait and Posture: A Survey

Abstract

There has been a growing interest in machine-based recognition of emotions from body gait and its combination with other modalities. In order to highlight the major trends and state of the art in this area, the literature dealing with machine-based human emotion perception through gait and posture is explored. Initially the effectiveness of human intellect and intuition in perceiving emotions in a range of cultures is examined. Subsequently, major studies in machine-based affect recognition are reviewed and their performance is compared. The survey concludes by critically analysing some of the issues raised in affect recognition using gait and posture, and identifying gaps in the current understanding in this area.

Disciplines

Engineering | Science and Technology Studies

Publication Details

B. Stephens-Fripp, F. Naghdy, D. Stirling & G. Naghdy, "Automatic Affect Perception Based on Body Gait and Posture: A Survey," International Journal of Social Robotics, vol. 9, (5) pp. 617-641, 2017.

Automatic Affect Perception Based on Body Gait and Posture: A Survey

***Abstract*— There has been a growing interest in machine-based recognition of emotions from body gait and its combination with other modalities. In order to highlight the major trends and state of the art in this area, the literature dealing with machine-based human emotion perception through gait and posture is explored. Initially the effectiveness of human intellect and intuition in perceiving emotions in a range of cultures is examined. Subsequently, major studies in machine-based affect recognition are reviewed and their performance is compared. The survey concludes by critically analysing some of the issues raised in affect recognition using gait and posture, and identifying gaps in the current understanding in this area.**

I. Introduction

A significant amount of interest in the study of automatic affect perception in applications such as human computer interaction, social robotics, and security is now evident in the literature. In this paper, a range of approaches on automatic affect recognition reported in literature is critically reviewed. The emphasis is on machine-based human emotion perception through gait and posture.

In the past, a significant amount of research was conducted on the recognition of emotions through facial expressions. According to de Gelder [2] in 2009, 95% of the literature on emotion in humans were singularly focused on facial expressions. However, emotions are not only conveyed through facial expressions, but also through body expression. Currently, there is also a growing interest in gait analysis because of its potential wide range of applications, such as personal identification [3], deception recognition [4], and detection of illnesses such as multiple sclerosis [5]. Body posture can also be used to effectively decode emotions at a distance compared to facial expressions alone [6].

Kleinsmith and Bianchi-Berthouze's survey paper [7] discusses the range of conflicting views arisen in the literature on the importance of facial expressions versus body expressions in communicating emotions. This is primarily based on the study by Ekman and Friesen [8] that identifies emotional deception in facial expressions and body movements. Ekman and Friesen use the term "non-verbal leakage" to describe clues towards deception that is unintentionally expressed. Importantly, Ekman and Friesen conclude that facial expressions can easily conceal this leakage and therefore can be used to lie about emotions. Since facial expressions can easily hide real emotions, body expressions may potentially offer a better approach for emotion recognition. Kleinsmith and Bianchi-Berthouze also conclude that analysing body expressions can help in providing clues to understand facial expressions, leading to higher recognition accuracies.

The study reported in this paper presents an overview of the various approaches undertaken thus far in machine-based affect recognition from body language. Such focus allows us to deal with methods proposed in the reviewed studies in more depth than previous survey papers. Within each study, in addition to overall success, a variety of classification techniques utilised and the features deployed are examined. The paper concludes with a critical analysis of the reviewed papers and identifying the obstacles that prevent an effective comparison between the studies including variances in data collection methods, number and quality of subjects, number and type of emotions and the absence of common ground truth for comparison. The conclusion also highlights the high impact of the proposed methods and outlines several directions for future research.

It is generally challenging to conduct a comprehensive comparison between different methods deployed in the literature in this field due to wide variation in detection methods, data sets, emotion categories etc. Therefore, a more realistic approach is taken in this study to initially provide an introduction into the field of machine based affect perception based on body language and then review the approaches thus far developed and reported in the literature.

Although a number of articles [7], [9] and [40] suggest that affect recognition is best performed under a multimodal approach, the higher the accuracy achieved through body language, the better is the classification recognition. Results of the search were therefore 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 is 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 sources of information. As the intention of this paper was to provide an overview of different approaches deployed in connection with gait and posture, studies associated with multimodal methods were also included in the review but the focus was mainly kept on how the body language was deployed.

Five survey papers were included in the search results. In 2009, Zeng et al. [10] conducted a review of the literature on recognition of emotions based on visual and audio signals. The visual features were based predominately on facial expression, with a limited number of studies considering gait and posture.

Another review paper was published by Kleinsmith and Bianchi-Berthouze in 2013 [7]. The psychology of affective recognition and the importance of taking culture into account when expressing, labelling and detecting affect from body language was first examined. The last section of the paper cites 18 different studies on affective recognition including older less successful works, compared to 39 recent papers cited in this survey paper.

Another review paper was published by Karg et al. [11] in 2013, devoting only a small section of the paper to the review of the previous work. The major part of the paper explored affective notation systems, human recognition of emotion and systems that generate emotions.

Zacharatos et al. [12] also published a survey paper in 2014. This paper provides only a brief overview of the different methods used in affective recognition, multimodal systems, segmentation and different models of emotion and notation systems.

McColl et al. [13] developed a survey paper in 2016 on affect recognition in the context of Human Robot Interactions (HRI). Affect recognition using a variety of modalities including facial, voice, body and physiological was examined. Only a small section of the paper, however, was dedicated to studies on recognising affect from body language.

The remainder of the paper is structured as follows. In Section II, the cultural similarities in emotional recognition by people from different backgrounds, and the mechanism used by human beings to recognise emotions from gait, are explored. This is followed in Sections III and IV by a review of machine-based affective recognition methods using raw data and processed data, respectively. For each study, we investigate the classifier deployed and the outcomes produced. We also review different approaches that are currently used to improve classification rate. Section V provides a critical discussion on common themes and trends within the literature, including research gaps in automatic affective recognition from gait that should be addressed through further investigations.

II. Affective perception in human Beings

A. Body Movements and Emotion

There is a long established and increasing collection of the literature suggesting that a large amount of information can be derived from body motion and posture. Kozlowski et al. [14] demonstrated that viewers could determine the sex of a subject using point light displays on major joints, as shown in Figure 1. Cutting and Kozlowski [15] also confirmed that participants could identify themselves and their friends using body mounted point light displays.

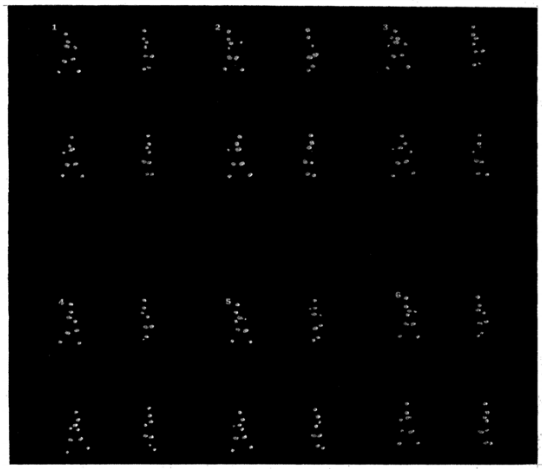

Figure 1 - Point Light Display of Posture [15]

The work reported in the literature indicates that body language can communicate emotions. Brownlow et al. [16] demonstrated that observers could distinguish between happy and sad dance movements using body mounted point light displays. De Meijer's [17] showed 96 recordings of body movements to 85 adult subjects and asked them to classsify those movements in twelve emotional categories. They concluded that certain body movements indicated specific emotional states. This was not just for one particular movement such as raising the fist, but the motion of a specific combination of the body parts conveying emotional states to human observers.

Walbot [18] also studied the connection between patterns of body movements and postures and the emotion displayed. A coding schema was deployed to analyse the 224 video recordings of six actors. Walbot concluded that specific body movements

and postural characteristics were framed to represent certain emotions.

Other psychological studies [19-21] also confirm that human perception can recognise affective states communicated only through the body movements. There has also been work examining the activity of the brain when gait patterns were observed [22, 23].

B. Cross Cultural Similarities and Differences in Emotions

Ekman and Friesen studied whether emotions conveyed by facial expressions were culture specific [24]. Although only facial expressions were observed, an insight was obtained into whether emotion expression crossed cultural boundaries. Subjects selected for the study 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. A number of emotions were explored including happiness, sadness, anger, surprise, disgust and fear. In order to overcome the language barrier and the fact that equivalent words for emotions might not exist 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 confirmed the hypothesis that a specific set of facial behaviours were universally associated with particular emotions irrespective of culture.

Recently, however, a pair of studies [25] challenge this hypothesis, as the data used has conflicted with Ekman and Friesen's results. Crivelli et al. tested 68 indigenous subjects from Papua New Guinea and Mozambique on their ability to recognise emotions through facial expressions. Although happiness achieved the highest result of 58% and 56% respectively, the other emotions only achieved a lower recognition rate, ranging from 7% to 53% for both studies.

Kleinsmith et al. [26] 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 used 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 using eight cameras. The captured motions were then used to create non-gender, non-culture specific computer avatars without facial expressions. The 108 affective postures were presented to observers in a different randomised order for each participant. The observers, consisting of 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 the postures from all three cultures were combined, the observers were able to recognise the emotions at a moderately successful level of 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%. This shows that there are differences in the way cultures express emotions in their body movement and there is only a moderate level of agreement between them. Other studies also confirm the thought that it is harder to identify emotions from facial and body expressions across cultures [27, 28]. One approach to addressing cross cultural issues is to apply different classification models to different cultures [29, 30].

C. Human Affective Recognition from Gait

The work conducted by Atkinson et al. [31] is an early study of humans' ability to recognise emotion from gait. Ten trained but unrehearsed actors were used to express certain emotions. The actors were covered in black with thirteen 2cm wide strips of white reflective tape placed on their bodies. They were given the workspace of two large paces around them and were free to walk in any direction, portraying each of the five emotions of happiness, sadness, fear, anger and disgust. Two filmed versions were created; one with the full video or full light (FL), and one with only the white strip information or Point Light (PL). It was easier to identify the emotions using only the PL information, but the observations made based on FL had higher recognition accuracy than PL observations. The authors also 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 easier it could be identified.

Gross et al. [32] also studied human ability to recognise emotions, and investigated 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 whilst walking. 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.

Kinematic analysis of the data obtained from the motion capture systems was deployed in the final section of the study to quantify both body and limb motions during the walk. Differences in the gait measurements and joint movement between different emotions were demonstrated in their results. For example, sad emotions typically resulted in slower movement and less movement of arms and elbow joints, and less trunk rotation. Angry walkers had an increased trunk flexion and shoulder elevation compared to joyful or content walkers even when they had a similar walking speed.

D. Context in Affective Recognition

Nayak et al. [33] defined 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. However, there is a growing body of literature suggesting that context provides important information in recognising human emotion.

Lankes et al. [34] examined the recognition of facial expressions from still, animated, and within game context. They found that within the game context provided the richest experience in perceiving emotions, followed by the animated. This demonstrated that context is helpful in understanding emotions as well as movement.

Body language is also highlighted as an important element in providing understanding for the context of facial expressions [35]. Similarly, the context of facial expressions helps to improve the recognition rate of emotions from body expressions [36].

By comparing emotion recognition with and without the scene context, Kret and deGelder [37] demonstrated that the surrounding social context aids in recognising emotion from body expressions. Similarly, Van den Stock et al. [38] also showed that background images can be helpful in recognising fear from body posture. Kret et al. also demonstrated that recognition of emotions of facial and body expressions were dependent upon their surrounding natural context [39]. The environment can even impact how humans walk with data suggesting that by changing the sound of footsteps subjects feel more positive, which impacts their gait [40].

Muller et al. examined [41] recognising emotions in a subject interacting with another person and the environment using body cues and audio recordings. However, they only achieved low accuracy recognition rates. Although the majority of current literature focuses on simple activities without taking into account context, a long-term goal is to recognise emotions within any environment and taking into account interactions with other people.

III. Affect Recognition Using Raw Data

In recent years, there have been a growing number of studies exploring the effectiveness of local features (raw data points) in automatically detecting human emotions manifested in gait and body movement. The methods used can be broadly categorised into two groups of 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 wearing of any specialised equipment. Examples of perceptive systems are image processing from video cameras, gait force measurement using plates and multiple sensor systems such as Kinect. Responsive systems capture as much data as possible, but since they require the subject to wear multiple sensors they are impractical in real world applications. Examples are security camera analysis and HRI situations. Perceptive systems are more suitable for real time applications but they generate less data than active systems.

A. Optical Motion Capture

In Optical Motion Capture Systems, a number of light markers are attached to the body and tracked through a set of infrared cameras, as shown in Figure 2. Such systems are not practical in real world scenarios such as security and HRI, but provide accurate data points that can be deployed in feature extraction and classification methods. Video Cameras alone often rely on crude methods of tracking, such as silhouette extraction, that do not provide data on individual joints. However, it is possible to track individual joints in the x, y and z axes using infrared cameras in Optical Motion Capture systems and obtain more detailed data on the body motion.

Figure 2 – Optical Motion Capture System [12]

Kapur et al. [42] demonstrated the high potential of automatically detecting emotions through the use of body movements. A VICON Motion Capture System captured fourteen reference point markers placed on five different subjects. The participants interpreted and subsequently acted or represented four basic emotions (sadness, joy, anger and fear). Each actor repeated the emotion 25 times, resulting in 500 recordings. To serve as a comparison against cognitive recognition, point light display mounted on 14 reference points were recorded and showed to ten subjects. The subjects identified emotions from the markers with an accuracy of 93%. An automatic classification model was deployed based on the motion data of the various body markers, incorporating the mean velocity and acceleration, and the standard deviation of the position, velocity and acceleration of each marker. Five different classifiers were applied to the data: logic regression, naïve bayes, decision tree, artificial neural network, and a support vector machine. The classifiers identified each actor's intended emotions with success rates between 85.6% and 91.8% using ten-fold cross validation. Artificial neural network and the support vector machine both produced the most accurate recognition rate. These rates were comparable to that of the human observer judging emotion based on point light displays. However, the study was limited to detecting four acted emotions and the deployment of a motion system that utilised six cameras, not practical in real life scenarios.

Lim and Okuno [43] 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. The authors 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. They 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 Scikit-Learn Toolkit [44] deploying a Gaussian Mixture Model. 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 demonstrates the potential of utilising high-level feature analysis instead of a low level feature analysis to detect emotions, particularly when body language data with high success rates are used.

Bianchi-Berthouze and Kleinsmith [45] explored the use of an associative neural network referred to as a Categorisation and Learning Model (CALM) to learn over time. Twelve subjects performed angry, happy and sad emotions freely whilst motion was recorded by a VICON motion capture system. A total of 138 gestures were collected. An avatar based on the motion capture data was shown to 114 Japanese observers to determine the emotion category labels, chosen as the most frequently used label between the observers for each gesture. Eighteen features were deployed, focusing on upper body gestures based upon the sphere of movement utilised in the dance. Normalized displacement of the entire arm, normalized displacement of the forearm, normalized extension of body and face orientation were examples of features deployed. The order of presentation was changed ten times, with each configuration repeated with 5 different sets of initial conditions. The average error was 0.043% with a standard deviation of 0.002.

B. *Inertial Motion Capture System*

Inertial Motion Capture Systems consist of a series of body mounted inertial sensors (Figure 3a) and do not require any emitters or external cameras. The sensors are placed onto the body segments surrounded by joints (Figure 3b). Sensors are able to record position, acceleration and velocity of the body parts, and via inference all the connecting joints. A summary of the studies deploying the Motion Capture is shown within Table 1 (raw data) and Table 3 (processed data).

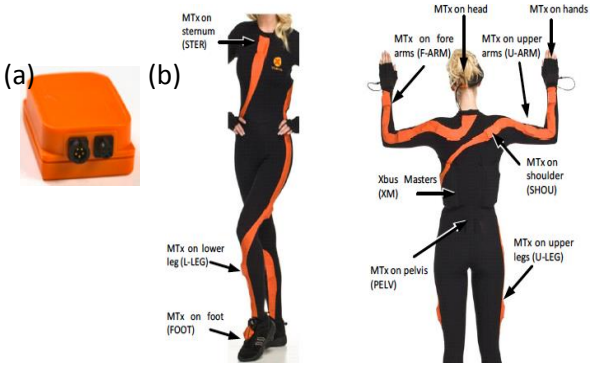

Figure 3 -[1] (a) Inertial Sensor (b) Inertial Sensor Placement

A number of studies on affective recognition from body posture and movement rely on acted emotions. In contrast, Garber-Barron and Si [46] attempted to classify emotions in non-acted scenarios. They used the UCLIC Affective Body Posture and Motion database which contained information from eleven participants playing the Nintendo Wii sports game for a minimum of thirty minutes, containing a total of 103 recordings. This database contained the rotational angles of the joints along the x, y and z axes. The Euler angles were used as features alongside their average rate of change, jerk and posture symmetry. Triumph, concentration, defeat and frustration were recognised with an accuracy of 66.5% when ten-fold cross validation was applied. 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 Gypsy 5 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 were examined utilising a Multilayer Perceptron (MLP) for automatic classification on the joint rotation features. 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. This process was repeated ten times. An agreement rate of 66.7% was found between the two subsets of views. 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 achieved a recognition rate of 66.3%. It was noted, however, that since there was no neutral category, concentration was often used as a fall back emotion when the observers felt that there was no other appropriate category.

A summary of these studies deploying motion capture data for raw data is shown in Table 1.

C. *Force Platform*

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. [48] is shown in Figure 4. Data obtained from the force platform can be analysed both independently and in conjunction with kinematic analysis of the markers mounted on the subject's body.

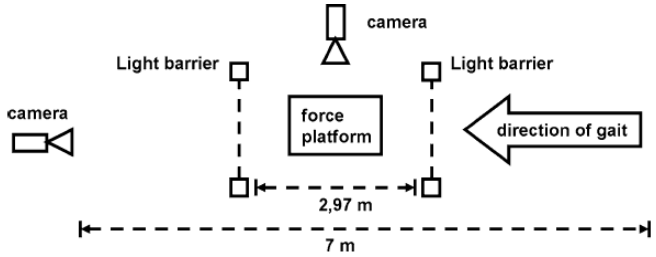

Figure 4 - Force Platform Setup [48]

Janssen et al. deployed neural networks to recognise emotions using gait data [48]. 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 that 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 significantly different. One of the recommendations made in this study was to combine the approach with features from either facial or vocal expressions to recognise emotions in an unknown subject.

Fawver et al. [49] concluded that there was a unique centre of pressure for different emotions in the preparation of walking phase, prior to forward movement. However, this has not been applied to automatic gait recognition. Initial work by Giraud et al. [50] showed the relationship between the changes in Centre of Gravity through video silhouettes, and the centre of gravity and centre of pressure on force plates to assess change in posture when reacting negatively and positively towards situations. Although this new method was not tested in automatic affect recognition, their data suggests it is a suitable alternative to required force pressure plates to analyse changes in pressure whilst walking.

D. Kinect

In addition to a video camera, the Kinect system utilises a depth sensor which provides greater accuracy and ability to track joints compared to the video signal alone. Kinect can convert a depth camera shot into 3D locations of joints as shown in Figure 5 [51]. A summary of the studies deploying the Kinect system with Raw data is shown within Table 2.

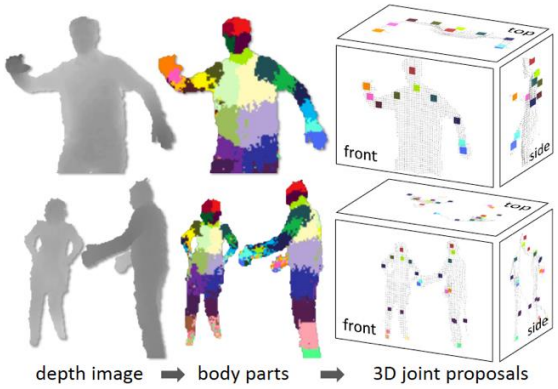

Figure 5 - Screenshot of Kinect system [51]

Table 1 - Studies deploying motion capture with raw data

Authors	Emotions Studied	Dataset	Classifier	Features Deployed	Truth Comparison	Success Rate	Sensors
Kapur et al. [42]	Sadness, Joy, Anger, Fear	5 Participants (2 Professional Dancers) Total of 500 recordings	Logistic regression, naïve bayes, decision tree, multilayer neural network, SVM	Mean velocity and acceleration; Standard Deviation of position, velocity and acceleration	10 Human Observers 93% Agreement	85.6%-91.8% depending on classifier used using 10 fold cross validation	VICON Motion Capture – 6 cameras tracking markers
Lim and Okuno [43]	Happiness, Sadness, Fear, Anger	10 speech participants 329 recordings & 28 ankle participants with 546 recordings	SciKit Learn Toolkit	Speed, Intensity, Irregularity and Extent of the Ankle Joint	Human Observers	63% - trained on voice in SIRE 72% - trained on gait in SIRE	Voice & motion capture data on ankle
Xiao et al. [52]	confident, have question, object, praise, stop, succeed, weakly agree, call, drink, read and write were studied	23 participants demonstrating each movement Total of 253 gestures	kNN	Hand Joints and Upper Body Joints	Intended Emotion	97%. Randomly split into training and testing. Repeated 5 times and averaged	Cyber Glove and Kinect System
Bianchi-Berthouze and Kleinsmith [45]	Angry, Happy and Sad	12 subjects, total of 138 gestures	CALM Network	Normalized displacement of each entire arm forearm and normalized extension of the body	Intended emotion	Repeated 10 times with 5 different initial conditions	VICON Kinect capture system
Li et al. [53]	Happy & Angry	59 Participants	Naïve bayes, Random forests and SMO classifiers	Joint Data	Intended Emotion	55% with ten fold cross validation	Two Kinnect Cameras
Garber-Barron and Si [46]	Triumph, Concentration, Defeat and Frustration	Eleven participants playing Wii (UCLIC Database) – total of 103 recordings	Bagging Predictor	Euler rotation angles of joints; their average rate of change and jerk; Posture symmetry	Human Observers	66.5% joint & limb rotation, & body posture 55% joint rotation 61% limb rotation 62% body posture using 10-fold cross validation	Inertial Motion Capture Data
Kleinsmith [47]	Concentration, Defeat, frustration (removed from results), triumph	Eleven participants playing Wii. Total of 103 postures	MLP	Joint Rotations Angles	Agreement from 8 Human Observers on an avatar replication	66.7%. Randomly split into 3 groups. 1/3 training, 1/3 testing. Repeated 10 times	Gypsy 5 Motion Capture

Xiao et al. [52] examined 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 upper body gestures of confidence, having a question, objection, praise, asking to stop, success, lightly agreeing, calling, drinking, reading and writing were studied. Twenty three subjects, each expressing the eleven gestures, were deployed resulting in a total of 253 recordings. The data was split into training and testing set randomly, and repeated 5 times. The results were averaged and compared against the intended emotion. A kNN classifier was applied to the data, achieving an accuracy of 97%.

Other studies have not produced as accurate results. Li et al. [53] deployed two Kinect cameras to automatically recognise either the happy emotion, or anger emotion of 59 participants. The authors applied the Naïve bayes, Random forests and SMO classifiers, with Random forest achieving the highest recognition rate of only of 55%. Since they only used raw joint data, they concluded that the angry and happy gait styles may have been too similar to effectively distinguish between the two emotions.

Table 2- Studies deploying Kinect with Raw data

IV. Recognition Using Processed Data

To improve upon the accuracy obtained by automatic affect recognition system, several studies have moved beyond using raw joint and segment data obtained from the data collection methods, with some utilising global features, temporal segmentation and/or Dimensional reduction to improve their accuracy. A summary of the studies deploying processed data are presented in Tables 3-5 for those using Motion Capture, Kinect and Video Analysis respectively.

A. 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 equally contribute to the classification process. Both Xu and Sakazawa [54] and Bernhardt and Robinson [55] explored segmenting motion. In their studies, motion data was segmented into different stages, but then recombined with a different weighting given to the data associated with each segment. In both studies, a demanding Leave One Subject Out Cross Validation (LOSO–CV) test was 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 is that not all stages of walking equally contribute 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 of the segment of the leg could be weighted more heavily in classification.

Xu and Sakazawa [54] assumed that body movements such as gestures had 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 the 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 [55] also showed the benefit of giving weightings to different segments of motion data in emotion recognition. They examined a collection of knocking performances by thirty 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 using 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 gait and posture to identify emotions from walking styles.

B. Global Features

Global features represent the overall characteristics of an image rather than the properties of certain key points in the image. Sanghvi et al. [56] 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 showed 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) [57] which is extensively used in activity recognition systems but has potential for more use in affect recognition.

LMA has four major components: body, effort, shape and space. Hachimura et al. [58] deployed the LMA method but considered only the Effort and Shape components. Effort was broken down even further into weight, time, space and flow factors and Shape was broken down into shaping and shape flow.

a) Motion Capture

Zacharatos et al. [59] applied Laban movement analysis to classify the emotions of candidates playing exergames. Thirteen players played sports 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 agreed with the observers and were consequently utilised. For the analysis, the study only considered the space and the time motion factors. Concentration, meditation, excitement and frustration were recognised with an overall classification accuracy of 85.27% deploying ten-fold cross validation. 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 could have been misclassified by the system.

Fourati et al. [60] deployed a combination of local, semi global and global features on 11 subjects who displayed eight emotions whilst walking and performing basic actions. The subjects were trained by an actor and their motion data was recorded using an XSens inertial motion capture suit. A Random forest classifier was applied to the data on various movements (including walking), achieving an average recognition accuracy of 84.8% whilst the subject was walking.

b) Kinect

In their study, Woo Hyun et al. [61] proposed using an LMA to distinguish between emotions. In their experiment, they used Microsoft Kinect to study twenty points on the body and considered 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. [62] 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 could respond and interact appropriately. Body posture and movements in a seated position was examined to determine the emotion. 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 participants. Eight elderly individuals were recorded at two meal times resulting in 16 recordings. The authors utilised nine different learning techniques to compare their effectiveness, benchmarking them against the median value of twenty-one human observers with ten-fold cross validation. 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.

McColl et al. [63] studied social robotics contexts to determine the level of accessibility based on the nonverbal interaction and states analysis (NISA) scale. One expert in the scale was used to code a comparison truth. They deployed a Kinect system to generate a 3D ellipsoid model of a person's static pose to determine the trunk and arm orientation towards the robot. WEKA [64] data mining software was utilised with tenfold cross validation. Naïve Bayes, Logistic Regression, Random Forest, k-Nearest Neighbour, Adaboost with Naïve Bayes, Multilayer Perceptron, Support Vector Machine classifiers were deployed on 300 static poses from eleven different individuals. Here, the Adaboost algorithm together with the Naïve Bayes base classifier achieved the highest accuracy of 99.3%.

Piana and Staglian [65] deployed a Kinect sensor to recognise six emotions. They extracted global features such as kinetic energy, contraction, symmetry as well as raw local features and ran it through a linear SVM classifier with a random split. The authors were able to obtain an accuracy of 68.5% compared to 62.3% obtained when utilising data from a Qualisys motion capture system.

Senecal et al. [66] deployed LMA for emotion recognition in theatre performances recorded by a Kinect camera. Ten actors performed eight different emotions, resulting in 80 performances. They extracted 28 Features, which when combined, correspond to the four different LMA features; and then fed them into a neural network. Each feature contained the minimum, maximum, standard deviation and average value. The neural network contained 86 inputs and two outputs, which corresponded to the (x,y) coordinates on a Russel's emotion space, as shown in Figure 6. Since the authors used a continuity of emotion rather than discrete, their results are best displayed graphically in Figure 6.

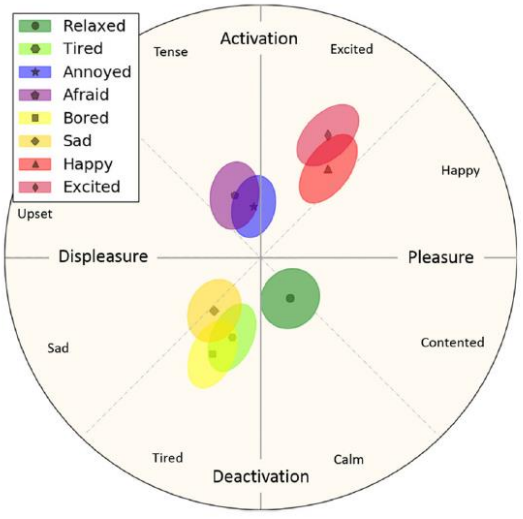

Figure 6 - Results from Senecal et al. [66] on Russell's emotion diagram

Kaza et al [67] used a deep learning classifier to

recognise five emotions in a dataset recorded with Kinect sensor. The authors used a neural network with stacked Restricted Boltzmann Machines (RBM). They deployed a range of global features which were broken up into the six groups of. Each of these six feature sets were fed into a different RBM and the output was then fed into a seventh RBM. The stacked RBM outperformed the RBM, MLP, SVM and Naïve Bayes algorithms, achieving the highest accuracy of 93%.

c) Video Analysis

Arunnehru and Geetha [68] recognised the three emotions happy, angry and fearful, in a surveillance video against a static black background. By extracting the global features of orientation, elongation, solidity and rectangularity, they recognised the emotions of the subjects whilst walking, siting and jumping. The authors applied the following algorithms in their classification system: SVM, Naïve Bayes and Dynamic Time Warping (DTW), which compares two sequences that may vary in time. The DTW classification achieved an overall accuracy of 93.39%.

Park et al. [69] 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 features such as the number of dominant points on the boundary, the coordinates of centroid, the aspect ratio and the coordinates of rectangle, as well as 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 multilayer perceptron. Recordings from three dancers were deployed as training and one dancer was utilised for testing data. They classified the emotions with an average accuracy of 73%.

Sanghvi et al. [56] also used global features in affective recognition in 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 were recorded, with a total of 44 recordings being utilised. Because of their age, the participants were unable to accurately identify their own levels of engagement. Instead the study utilised three coders to manually label the different sections of video as either engaged, not engaged and 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 global features quantity of motion and a contraction index, combined with the local features body lean angle and slouch factor. A variety of classifiers were tested with ADTree and OneR classifiers achieving the highest accuracy of 82% using ten-fold cross validation.

Barakova and Lourens [70] detected movements that expressed emotions. They examined the Laban sections of Weight, Time and Flow, then translated combinations of these into sadness, joy, fear and anger. Fifteen, twenty second recordings of waving patterns that demonstrated happiness, anger, sadness and nervousness were captured. A Neural Network classifier was deployed and 42 children were used to determine the ground truth in each case. Here, Barakova and Lourens achieved an overall accuracy of 63.8%.

Lourens et al. [71] studied subjects waving in an angry, happy, sad and polite emotion and discovered these states were associated with distinct acceleration profiles. A combination of skin colour tracking and motion analysis was deployed to view the movement of hand arm and head. It was shown that these emotions occupied distinct regions of weight, time and flow areas within the LMA.

A summary of the studies using video analysis is presented in Table 5.

C. *Using 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 for 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 has the potential to reduce a high-dimensional data to a lower-dimensional subspace [72].

Venture et al. [73] proposed the use of vector analysis and Principal Component Analysis (PCA) decomposition to detect emotions from gait. Four professional actors displayed four basic emotions walking in a straight line, whilst being recorded via a VICON motion capture system. The affective

states of neutral, joy, anger, sadness and fear were repeated five times by each actor, totalling 100 movements. They examined the features of position, velocity and acceleration of the markers, as well as the angle, velocity and acceleration of the joints. To determine the accuracy, a comparison was made between the detected emotion and one identified by twenty human observers viewing animations. Vector analysis, as well the animations produced from the performed emotions, indicated that the lower torso, waist rotations and head movements were the most important features in affect perception as the leg and arm data could bias the recognition process.

The authors subsequently utilised a similarity index computation to test similarity between test data trial 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 the 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% when using ¾ of data for training and ¼ for testing. A global database was developed from a combination of data from all the 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 this study, only a relatively small number of subjects were used and it was a possibile that deployment of more subjects could produce a different result. Both male and female actors were used in the study with no difference in recognition rates. The false negative classification seemed to be for neutral states rather than other emotions.

Kar et al. [74] applied quantity of motion as a dimensionality reduction tool based on a hypothesis that most movement comes from the most relevant body parts. A Kinect system recorded ten subjects, each performing five emotions. They extracted displacement features from the joints with the highest quantity of motion and combined these with expansion features. The authors then constructed Gaussian curves, extracting the peaks and variances of each feature, with the classified emotion being determined by the maximum value. An accuracy of 94.4 % was obtained using this Fuzzy system.

Samadani et al. [75] 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 movements was calculated after the training through Hidden Markov Models (MMM). 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 based on 13 demonstrators, recorded by an eight-camera motion capture system. The hand and arm dataset was collected independently from the full body data to prevent any confusion between them. 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 did not incorporate any high-level motion analysis.

Samadani et al. [72] 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 thirteen 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 movements. The algorithm did not perform as well on full body motion data with the highest recognition accuracy of 53.6% produced by FSPCA-GRBF with the leave-one-out cross validation method.

Karg et al. detected emotions using human gait and compared different component analysis

techniques and classifiers [9]. The Technische Universität München (TU München) gait database was utilised, which contained motion capture recordings of thirteen male non-professional actors demonstrating neutral, happy, sad and angry emotions. This contained a total of 520 strides. Initially, the motion capture data was applied to an animated puppet to determine the accuracy in determining human emotions purely from the gait, without any influence of facial expressions or physique. Human observers identified the emotions portrayed by the gait of the puppet 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. The feature space was then 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% utilising leave one out cross validation. 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%. The authors concluded that it would be useful to use a multimodal system with face and/or voice recognition combined with gait to improve accuracy. Following the same approach, they 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. The same gait database was deployed which contained a total of 780 strides for the affective dimensions. Using the same SVM on data from all joint angles, the system produced an accuracy of 88% for pleasure, 97% for arousal and 96% for dominance. However, there was no reported attempt to use PAD recognition models for classifying data into different emotions.

Authors	Emotions Studied	Dataset	Classifier	Features Deployed	Truth Comparison	Success Rate	Sensors
Xu and Sakazawa [54]	Neutral, Happy, Angry, Sad	60 Demonstrators Total of 2500 recordings	SVM with weighted segments	Entropy of each segment of movement	Actor's Intended Emotion	77% using leave one subject out cross validation	Motion Capture
Bernhardt and Robinson [55]	Neutral, Happy, Angry, Sad	30 Demonstrators hand knocking Total of 1200 recordings	SVM with polynomial kernels with weighting of limb speeds	Max dist. hand from body; Avg hand speed, acceleration and jerk	Actor's Intended Emotion	50% without weighting 81% with weighting utilising leave one subject out cross validation	Motion Capture
Karg et al. [9]	Neutral, Happy, Angry, Sad Displeased, Content, Bored, Excited and Obedient	13 Actors Total of 520 Recordings for discrete emotions Total of 780 Recordings for discrete emotions	SVM	velocity, stride length and cadence, as well as the minimum, maximum and mean joint angles	Intended emotion	using leave one out cross validation 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. [59]	Concentration, Meditation, Excitement & Frustration	13 Actors Total of 197 recordings	WEKA – MLP	LMA components Space and Time	4 Human Observers	85.27% using Ten Fold Cross Validation	Motion Capture
Fourati et al. [60]	Joy, Anger, Panic Fear, Anxiety, Sadness, Shame, Pride and Neutral	11 Subjects, total of 1025 recordings for Walking	Random Forest	Local features, semi-global (such as Feet arm and hands relationships, and symmetry), Global features (sagittal, vertical and horizontal directions of bounding box)	Intended emotion	84.8%	XSens Motion Capture
Venture et al. [73]	Neutral, Joy, Anger, Sadness, Fear	4 Professional Actors Total of 100 recordings	Similarity index	Coordinates of position, velocity and acceleration; Joint angle, velocity and acceleration	20 Human Observers 90% Agreement except Joy	78% for an individual, 69% for the group ¾ Training, ¼ test data	VICON Motion Capture
Samadani et al. [75]	Sadness, Happiness, Fear and Anger	13 Demonstrators. Total of 183 movements	HMM to calculate FS representatio ns, which are used in k-NN	Multivariate Times series movement sequence vector	Actor's Intended Emotion	77% using leave one out cross validation	Motion Capture
Samadani et al. [72]	Sadness, Happiness, Fear and Anger	13 Demonstrators Total of 183 movements	FSCPA-GRBF	Multivariate Times series movement sequence vector	Actor's Intended Emotion	53.6% using leave one out cross validation	Motion Capture

Table 3– Studies Deploying Motion Capture with Processed Data

Authors	Emotions Studied	Dataset	Classifier	Features Deployed	Truth Comparison	Success Rate	Sensors
Woo Hyan et al. [61]	Rejoicing & Lamenting	1 Participant total of 2 recordings	N/A	LMA components Space, Weight and Time	N/A	Two graphs of Space, Weight and Time were easily distinguishable for entire frames	Kinect
McColl et al. [62]	Valence & Arousal	8 elderly individuals, Total of 16 recordings	WEKA toolbox using various classifiers, best individual performances were: RBFN, Adaptive Boosting with Naïve Bayes	Bowing/Stretching of Trunk, Opening/ Closing of arms; vertical motion, speed and expansiveness of the body	Human Observer	Ten Fold Cross validation deployed V - 77.9%, A – 91.4% V – 70.0% A 93%	Kinect
[63] McColl et al	Level of Accessibility	Eighteen participants interacting with robot.	Naïve bayes, logistic regression, random forest, k-nearest neighbour, adaboost with naïve bayes (best), multilayer perceptron, support vector machine classifier was tested	trunk and arm orientation towards the robot	One Expert in NISA scale	99.3%. using Ten-fold cross validation	Kinect
Piana and Staglian [65]	happiness, fear, sadness and anger (surprise and disgust)	12 Participants totalling 100 videos	Linear SVM	kinetic energy, contraction, symmetry with raw local features	Intended emotion verified by human observers	82% for four emotions; 68.5% for six emotions with LOSO	Kinect
Senecal et al. [66]	happy, excited, afraid, annoyed, sad, bored, tired, and relaxed	10 Actors, totalling 80 performances	Neural Network	28 Features, which when combined, correspond to the four different LMA features	Intended emotion	Mostly distinct areas on Rusel Space Diagram (see Figure 6)	Kinect
Kaza et al [67]	anger, happiness, fear, sadness and surprise)	14 subjects	Stacked RBM performing best. Compared with SVM, RBM and MLP and Naïve Bayes	6 groups of features: kinematic related, spatial extent related, smoothness related, symmetry related, leaning related and distances related	Intended emotion	93%	Kinect
Kar et al. [74]	Happy, Anger, Fear, Disgust and Surprise	10 subjects, totalling 50 recordings	Maximum displacement on Gaussian curve	extracted displacement features from the joints with the highest quantity of motion and combined these with expansion features	Intended emotion	94.4%	Kinect

Table 4 - Studies deploying Kinect with processed data

Authors	Emotions Studied	Dataset	Classifier	Features Deployed	Truth Comparison	Success Rate	Sensors
Arunnehru and Geetha [68]	happy, angry and fearful	10 Subjects	SVM, Naïve Bayes and Dynamic Time Warping	orientation, elongation, solidity and rectangularity	Intended emotion	93.39%. 5 People used for training and 5 for testing	Video Camera
Park et al. [69]	happiness, surprise, anger and sadness	Four professional dancers Total of 16 recordings	Time Delayed MultiLayer Perceptron	number of dominant points on the boundary, the coordinates of centroid, the aspect ratio and the coordinates of rectangle, velocity and acceleration of each feature	Intended Emotion	71.5% 3 Dancers utilized for training data, 1 dancer for testing data	Video Camera
Sanghvi et al. [56]	Engaged, Not Engaged	Five eight-year-old subjects playing two chess exercises at different levels of difficulty Total of 44 recordings	variety of classifiers were tested with ADTree and OneR best performing	quantity of motion and a contraction index, combined with the local features body lean angle and slouch factor	3 manual coders	82% using ten-fold cross validation	two cameras; one looking at the child in a lateral view and one in a frontal view
[70] Barakova and Lourens	sadness, joy, fear and angry	Total of 15 recordings of waving patterns	Neural network	Laban sections of weight, time and flow	42 Children	63.8%	Video Camera
[71] Lourens et al.	angry, happy, sad and polite emotion	Total of 15 recordings of waving patterns	N/A	Laban sections of weight, time and flow	N/A	Demonstrate distinct regions of weight, time and flow sections of LMA	Video Camera

Table 5 - Studies deploying Video Camera

D. *Multiple Modality Fusion*

D'mello and Kory [76] performed a meta-analysis of the studies undertaken between 2003 and 2013. The accuracy of 90 affect recognition studies was examined, including unimodal and multimodal approaches. The multimodal systems used information from the face, voice, text, physiology, and body, mostly using a combination of two or more modalities. D'mello and Kory found that a multimodal approach to affective recognition consistently performed better than a unimodal system by an average of 9.8%.

In their work, Gunes and Picardi [77] 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. For example, 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 posture was used as an auxiliary mode in their system combined with facial recognition. Facial recognition and body posture recognition were first trained separately and then trained together. A variety of classifiers tested with BayesNet produced the best results for face and C4.5 providing the best results for body posture. The authors increased the recognition rate using facial information from 72.83% to 89.8%. They repeated the results with Adaboost and recognised 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 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 [78] also examined 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 appeared 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, 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). Three different actors were employed 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 the situation. The actors' responses were recorded by two cameras; one for the face and one 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 for 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 fusion 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 criterion was deployed to provide the eventual bimodal affect recognition. Although Gunes and Picardi expected the face to be the primary modality, experiments proved 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 looking at emotions generated with one or two hands, head, shoulders or combinations of these. For the unimodal approaches, they only obtained a success rate of 35.22% for facial expressions and 76.87% for body gestures. With combined modalities, they achieved an accuracy of 82.65% for feature level fusion and 78% for decision level fusion.

Shan et al. [79] also used the FABO database to study 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 deploying a combination of facial expression and body posture, the recognition rate increased to 88.5% compared to 79.2% from facial recognition alone.

Chen et al. [80] also considered fusing together information from both facial expressions and body cues with a temporal relationship. An alternative method was proposed to compensate for the complicated real time processing. A Motion History Image (MHI), a Histogram of Oriented Gradients (HOG) and an image-HOG was produced. Instead of only using the apex frame, they utilised data from the onset through the apex to the offset frames. After extracting MHI-HOG and Image-HOG, PCA was performed to reduce the feature dimension in each frame. Each frame was assigned 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 deployed the FABO database [78]. The approach of this study achieved an accuracy of 73% for combined facial

expressions and body gestures, using two thirds of the data as training and one third for testing. Although this was a lower accuracy than that recorded by Gunes and Picardi, Chen et al. believed 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 by themselves.

Chen and Tian [81] then proposed an alternative method of fusing together facial and body gesture information. They proposed 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, a one vs one classifier was trained using the optimally combined kernel and evaluated on the FABO database [78]. The facial features image-HOG and MHI-hog were extracted as well as the body gesture features of location, motion area, image-HOG and MHI-HOG. As applied in [80], each expression was then segmented into onset, apex, offset and neutral phases, and then a temporal normalisation procedure was undertaken. After this, the MCMKL method was used. 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. [82] combined multiple modalities into an emotion recognition system. They utilised 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). These eight emotions were chosen as they were equally distributed within the valence and arousal space. Two cameras were used, one for facial recognition and the other for body gestures, and a microphone on the participant's shirt recorded the voice. Kessous et al.'s system measured facial animation parameters (FAPs) tracking points and compared the deformation against a neutral frame. These FAPs, along with their calculated confidence levels were examined to provide the facial expression estimation. For body gestures, Kessous et al. used the EyesWeb [83] 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 feature extraction, a set of features based on intensity, pitch, Mel frequency cepstral coefficient, Bark spectral bands, voice segmented characteristics and pause length were deployed.

BayesNet from the WEKA toolbox, was deployed on all classifications 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 by initially determining whether there was an agreement in emotions between more than one of the modalities before reverting to the highest probability. When operating as a unimodal 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 was 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 unimodal approach with an accuracy of 62.5% for speech and face modalities and 75% for speech and gesture modalities.

A summary of the studies deploying multimodal recognition is presented in Table 6.

Authors	Emotions Studied	Dataset	Classifier	Body Language Deployed	Truth Comparison	Success Rate	Sensors
Gunes & Picardi [78]	anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness	Ten Subjects, Total of 170 recordings	Feature level fusion: Adaboost with Random forest of ten trees Decision Level Fusion: Face - Adaboost with C4.5 Body – Random forest of ten trees	general change within the feature, texture/motion, optical flow.	Intended Emotion	Feature level fusion – 82.65% Decision level fusion – 78% Ten-Fold Cross Validation deployed	Two video cameras, Face & Body
Shan et al. [79]	anger, anxiety, boredom, disgust, joy, puzzle and surprise	23 Actors total of 262 recordings	SVM Combined with CCA	Spatial-temporal features	Intended Emotion	Face -79.2% Body -72.6% Combined – 88.5% using 5 fold cross validation	Two video cameras, Face & Body
Chen et al. [81]	anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness	FABO database using 284 videos	SVM with RBF kernel	location features, motion area features, Image-HOG features, and MHI-HOG features	Intended Emotion	Combined – 73% using 3 fold cross validation	1 Camera on face & body
Chen & Tian [81]	anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness	FABO database using 255 videos	One vs one	location features, motion area features, Image-HOG features, and MHI-HOG features	Intended Emotion	Combined - 77.3% using 5 fold cross validation	1 Camera on face & body
Kessous et al. [82]	anger, despair, interest, pleasure, sadness, irritation, joy and pride	Ten non-actor subjects total of 240 recordings	Bayes Net (WEKA)	Quantity of Motion (QoM) and Contraction Index (CI) of the body, Velocity (VEL), Acceleration (ACC) and Fluidity (FL) of the hand's barycentre.	Intended Emotion	Facial – 48.3% Body – 67.1% Voice – 57.1% Combined – 74.6% using 10 fold cross validation	Two video cameras, Face & Body, microphone on shirt

Table 6 - Studies Deploying Multimodal Recognition

V. Conclusions and Future Direction

It is not difficult to conclude from the review conducted in this paper that affect recognition from gait and posture is at an early stage of its development. While the number of studies reported in the literature is not high, a thorough and systematic comparison between them is rather difficult due to the major differences among them in the type and set up of their experimental work, as well as the datasets and classification methods they deploy.

A number of studies use role play to act the emotions studied, though professional actors are not consistently used. In a role play scenario, the intended emotions should be correctly communicated. This, however, is not the case particularly when the subject is not a professional actor, resulting in poor performance and inconsistency of data across different studies. Some studies use a story to evoke an emotion in the observers, others rely on the actors recalling their own memories, while some leave the display of emotion to the imagination of the actor.

The style of emotions and the number of emotions deployed also significantly vary in different studies. For example Calvo and D'Mellow [84] 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. The validity of labelling emotions is questioned by some as they argue that emotions form a continuous spectrum [85], particularly with the challenges that arise from discrete emotion labelling [86]. In the context of affect recognition, this question needs further research.

In the literature examined, some studies deployed only two categories of emotions such as Valence and Arousal [62], or Rejoicing and Lamenting [61]. While some others have a larger set of emotions ranging from four including Neutral, Happy, Angry, Sad [54], [55], [9] up to 11 emotions of anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, uncertainty, puzzlement, and sadness [78], [81]. The larger the number of emotions the more difficult becomes the classification process as the emotions become less distinct.

Both the number of actors and observers used in the databases and datasets associated with affect recognition based on gait and posture is quite small compared to what currently available in facial expression databases. This small number decreases the reliability of the result obtained, as any outlier of the performance or opinion of the actors and observers will have a more significant effect on the overall results. Inconsistency across the human observers is highlighted in the reported agreement rates. In the study conducted by Venture et al. [17] the agreement on most emotions is said to be at least 90%, but joy only had an agreement rate of 65%.

This low agreement between observers, and between the observers and the intended emotion, highlight the problem in defining the true emotion or a ground truth for comparison. Training a classifier to recognize affect requires the training data to be tagged with specific emotions. Studies such as [73], [42], [46], [47], [43] use human observers to determine the emotion conveyed. This means that the classification is more likely to resemble how humans interpret emotions. However, there seems to be disagreement between human observers. Not only has there been differences in identification across cultures, but also age groups from the one culture can identify different emotions from the same body expression [87]. Alternatively, [48], [55], [54], [72], [75] use the emotion intended by actors as the intended emotion. This, however, relies on the expertise of actors, which is not often reliable and actors tend to only use the extremes of each emotion which can create an unwanted bias.

The low agreement between emotions also raises the question of whether people are able to recognise emotions through body language alone, or whether body language represents only one piece of the puzzle, particularly when dealing with real world emotions and not just the acted extremes. Better affect recognition can be produced through consideration of a range of features such as facial expression, environmental context, surrounding people and vocal expression. The results produced so far by the studies reviewed in this paper supports the conclusion that using gait alone may not produce results which are as accurate as those obtained when multimodal information is used. The work conducted by Gunes and Picardi [77] seems to be an exception as the deployment of multiple modalities decreases the overall success due to poor accuracy when facial recognition is combined with posture. This decrease in accuracy can be overcome through the

use of confidence ratings as shown in the follow up work by Gunes and Picardi [78]. All of the studies thus far, when used with multiple modalities, only use posture as a still picture, rather than dynamic body motion as a time series. This is an area that requires further study. Regardless, the approach taken by Gunes and Picardi [78] shows promise. They performed feature level fusion to combine face and body features. The authors were able to recognise 11 different emotions with an accuracy of 82.65% deploying ten-fold cross validation. This is a higher level of accuracy than reported in other studies in spite of considering a higher number of emotions in the analysis and using no body mounted sensors in collecting the data.

Perhaps an alternative way of approaching affect recognition is to assign 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. Currently most of the studies use actors who can display extremes of emotions but the intensity of such extremes varies 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 of the emotion is communicated in our gait.

In order to further examine the accuracy of an approach, classification can be applied to a dataset compromising more emotions than the classifier was trained on. For example, the system could be trained on emotions neutral, happy and sad, but then tested against emotions of neutral, happy, sad, angry and fearful. The classifier can estimate a confidence level for each category of emotions identified in the dataset, including the "unknown" category for emotions not classified.

The cited works use different databases and datasets which increases the complexity of a comprehensive comparison between them. Several studies deploy the FABO database but they only use a limited selection from the database rather than the whole set.

In some studies, the emphasis is on real time analysis of acquired data without requiring the user to wear any special equipment, whereas other studies use wearable sensors or motion suits, multiple cameras that require intensive computational analysis of data, not possible in real time. 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.

According to the literature, motion capture suits are the most popular method of acquiring gait and posture data in affect perception. Motion capture suits are unable to be used in real world scenarios due to the requirement of the subjects wearing specialised sensors or suits, but they capture more detailed and accurate data than other methods of collection. This indicates that utilising body language in affect recognition is still at an early stage of development. In Multimodal studies, video is used for data collection as the approach can be easily combined with facial recognition, which is extensively used in affect recognition.

Since each study uses its own dataset and data detection method, it is difficult to compare the analysis and classification methods. Studies that use a common data set and detection method need to be undertaken to enable a comparison of the various processing options (including raw data) to determine their comparative effectiveness. Current literature, however, appears to only report on the performance of the classifiers with the highest accuracy. The success of a classifier can depend on a number of factors including the size of the training data, the 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. Comparing classifiers within the same dataset and processing options should be considered to determine the more effective classifiers.

Rather than detecting acted emotions, some studies examine emotions portrayed through natural movements. This includes subjects playing Wii [46], [59]; subjects playing chess [56]; and interacting with a social robot [62], [63]. Natural emotions, are potentially less exaggerated and at the same time less consistent.

Interaction with robots and machines is a driving force behind further development and acceptance of methods and tools to perceive the emotion expressed by the user and response to it. One strong development in this direction is social robotics ([56], [62, 63]).

Zacheratos et al. [59] demonstrated the most successful application of classifying natural emotions. The authors recognised 197 recordings of the four emotions of concentration, meditation, excitement and frustration, with an accuracy of 85.27%. This is an impressive result among all of the single modality studies examined in this survey paper. Kapur et al. [42] produced the other significantly high level of accuracy within the

papers examined, but their analysis was based upon acted emotions.

The use of different ways to process the raw data seems to have shown its benefit over raw data alone. However, because of the variability of datasets and methods that are used, there is no simple way to compare the different methods. Examining the results of studies utilising raw data found in Table 1 and those deploying processed data in Table 3, shows that deploying discriminant analysis has a positive effect on the accuracy of classification. One of the more innovative and successful approaches taken in processing data is by Bernhardt and Robinson [55]. They demonstrated that by breaking up the motion into different segments, then recombining them with a weighting, the accuracy of the classification increases from 50% to 81%.

No study, however, has looked at using several different processing techniques at the same time. The use of different processing techniques at the same time may lead to further improvement. A future study could compare the effect of applying dimensional reduction, segmentation and a combination of dimensional reduction and segmentation.

The use of global features, including Laban movement analysis, in affective recognition shows some success, but there appears to be few studies of its use in recognising emotions from posture of the full body. The work conducted by Zacheratos et al. [59] is one of the few studies in which global features are applied to the whole body, which, as mentioned previously, achieved one of the highest accuracies with four different emotions observed. A combination of local and global features is used in object recognition [88] and this method is also used in action recognition [89] and, more recently, in facial expression recognition with encouraging results [90]. To date, however, this approach has only had limited application to automatic affect recognition from gait and posture.

ACKNOWLEDGEMENTS:

This research has been conducted with the support of the Australian Government Research Training Program Scholarship.

COMPLIANCE WITH ETHICAL STANDARDS:

The authors declare that they have no conflict of interest.

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

[2] 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, pp. 3475-3484, 2009.

[3] A. Kale, A. Sundaresan, A. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, *et al.*, "Identification of humans using gait," *Image Processing, IEEE Transactions on*, vol. 13, pp. 1163-1173, 2004.

[4] 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.

[5] 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.

[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, pp. 510-510, 1988.

[7] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey," *Affective Computing, IEEE Transactions on*, vol. 4, pp. 15-33, 2013.

[8] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, pp. 88-106, 1969.

[9] 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, pp. 1050-1061, 2010.

[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, pp. 39-58, 2009.

[11] M. Karg, A. A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation," *Affective Computing, IEEE*

Transactions on, vol. 4, pp. 341-359, 2013.

[12] 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, pp. 35-45, 2014.

[13] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI," *Journal of Intelligent & Robotic Systems*, vol. 82, pp. 101-133, 2016.

[14] L. T. Kozlowski and J. E. Cutting, "Recognizing the sex of a walker from a dynamic point-light display," *Perception & Psychophysics*, vol. 21, pp. 575-580, 1977.

[15] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the psychonomic society*, vol. 9, pp. 353-356, 1977.

[16] 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, p. 411, 1997.

[17] M. Demeijer, "THE CONTRIBUTION OF GENERAL FEATURES OF BODY MOVEMENT TO THE ATTRIBUTION OF EMOTIONS," *Journal of Nonverbal Behavior*, vol. 13, pp. 247-268, Win 1989.

[18] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, pp. 879-896, 1998.

[19] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, pp. 117-139, 2004.

[20] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, pp. B51-B61, 2001.

[21] 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, pp. 727-738, 1996.

[22] B. de Gelder, J. Van den Stock, H. K. Meeren, C. B. Sinke, M. E. Kret, and M. Tamietto, "Standing up for the body. Recent progress in uncovering the networks involved in the perception of bodies and bodily expressions," *Neurosci Biobehav Rev*, vol. 34, pp. 513-27, Mar 2010.

[23] S. Schneider, A. Christensen, F. B. Haussinger, A. J. Fallgatter, M. A. Giese, and A. C. Ehlis, "Show me how you walk and I tell you how you feel - a functional near-infrared spectroscopy study on emotion perception based on human gait," *Neuroimage*, vol. 85 Pt 1, pp. 380-90, Jan 15 2014.

[24] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, pp. 124-129, 1971.

[25] C. Crivelli, S. Jarillo, J. A. Russell, and J. M. Fernandez-Dols, "Reading emotions from faces in two indigenous societies," *J Exp Psychol Gen*, vol. 145, pp. 830-43, Jul 2016.

[26] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers*, vol. 18, pp. 1371-1389, 2006.

[27] H. A. Elfenbein, "In-Group Advantage and Other-Group Bias in Facial Emotion Recognition," in *Understanding Facial Expressions in Communication: Cross-Cultural and Multidisciplinary Perspectives*, ed, 2015, pp. 57-71.

[28] M. A. Quiros-Ramirez, "Considering cross-cultural context in the automatic recognition of emotions," *International journal of machine learning and cybernetics*, vol. 6, pp. 119-127, 2015.

[29] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning Personalized Models for Facial Expression Analysis and Gesture Recognition," *IEEE Transactions on Multimedia*, vol. 18, pp. 775-788, 2016.

[30] P. A. Wilson and B. Lewandowska-Tomaszczyk, "Affective Robotics: Modelling and Testing Cultural Prototypes," *Cognitive Computation*, vol. 6, pp. 814-840, 2014.

[31] 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," *Perception*, vol. 33, pp. 717-746, 2004.

[32] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Effort-Shape and kinematic assessment of bodily expression of emotion during gait," *Human Movement Science*, vol. 31, pp. 202-221, 2012.

[33] 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.

[34] M. Lankes, R. Bernhaupt, and M. Tscheligi, "Evaluating User Experience Factors Using Experiments: Expressive Artificial Faces Embedded in Contexts," in *Evaluating User Experience in Games: Concepts and Methods*, R. Bernhaupt, Ed., ed London: Springer London, 2010, pp. 165-183.

[35] S. Buisine, M. Courgeon, A. Charles, C. Clavel, J.-C. Martin, N. Tan, *et al.*, "The Role of Body Postures in the Recognition of Emotions in Contextually Rich Scenarios," *International Journal of Human–Computer Interaction*, vol. 30, pp. 52-62, 2014/01/02 2014.

[36] M. L. Willis, R. Palermo, and D. Burke, "Judging approachability on the face of it: the influence of face and body expressions on the perception of approachability," *Emotion*, vol. 11, pp. 514-23, Jun 2011.

[37] M. E. Kret and B. de Gelder, "Social context influences recognition of bodily expressions," *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, vol. 203, pp. 169-180, 04/17

[38] J. Van den Stock, M. Vandenbulcke, C. B. Sinke, and B. de Gelder, "Affective scenes influence fear perception of individual body expressions," *Hum Brain Mapp*, vol. 35, pp. 492-502, Feb 2014.

[39] M. Kret, K. Roelofs, J. Stekelenburg, and B. de Gelder, "Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size," *Frontiers in Human Neuroscience*, vol. 7, 2013-December-18 2013.

[40] A. Tajadura-Jiménez, M. Basia, O. Deroy, M. Fairhurst, N. Marquardt, and N. Bianchi-Berthouze, "As light as your footsteps: altering walking sounds to change perceived body weight, emotional state and gait," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2943-2952.

[41] P. M. Muller, S. Amin, P. Verma, M. Andriluka, and A. Bulling, "Emotion recognition from embedded bodily expressions and speech during dyadic interactions," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 2015, pp. 663-669.

[42] 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., ed Berlin: Springer-Verlag Berlin, 2005, pp. 1-7.

[43] A. Lim and H. G. Okuno, "The MEI Robot: Towards Using Motherese to Develop Multimodal Emotional Intelligence," *Ieee Transactions on Autonomous Mental Development*, vol. 6, pp. 126-138, Jun 2014.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[45] N. Bianchi-Berthouze and A. Kleinsmith, "A categorical approach to affective gesture recognition," *Connection science*, vol. 15, pp. 259-269, 2003.

[46] 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.

[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, pp. 1027-1038, 2011.

[48] 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," *Journal of Nonverbal Behavior*, vol. 32, pp. 79-92, Jun 2008.

[49] B. Fawver, G. F. Beatty, K. M. Naugle, C. J. Hass, and C. M. Janelle, "Emotional state impacts center of pressure displacement before forward gait initiation," *Journal of Applied Biomechanics*, vol. 31, pp. 35-40, 2015.

[50] T. Giraud, D. A. G. Jáuregui, J. Hua, B. Isableu, E. Filaire, C. L. Scanff, *et al.*, "Assessing postural control for affect recognition using video and force plates," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 109-115.

[51] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, *et al.*, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, pp. 116-124, 2013.

[52] 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.

[53] S. Li, L. Cui, C. Zhu, B. Li, N. Zhao, and T. Zhu, "Emotion recognition using Kinect motion capture data of human gaits," *PeerJ*, vol. 4, p. e2364, 2016.

[54] 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.

[55] 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.

[56] 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.

[57] R. Laban, *Principles of Dance and Movement Notation*. New York: Macdonald & Evans, 1956.

[58] 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.

[59] 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.

[60] N. Fourati and C. Pelachaud, "Multi-level classification of emotional body expression," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015.

[61] 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.

[62] D. McColl, G. Nejat, and Ieee, "Determining the Affective Body Language of Older Adults during Socially Assistive HRI," *2014 Ieee/Rsj International Conference on Intelligent Robots and Systems (Iros 2014)*, pp. 2633-2638, 2014.

[63] 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, pp. 1-15, 2016.

[64] 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, pp. 10-18, 2009.

[65] S. Piana, A. Staglian, #242, F. Odone, and A. Camurri, "Adaptive Body Gesture Representation for Automatic Emotion Recognition," *ACM Trans. Interact. Intell. Syst.*, vol. 6, pp. 1-31, 2016.

[66] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, "Continuous body emotion recognition system during theater performances," *Computer Animation and Virtual Worlds*, vol. 27, pp. 311-320, 2016.

[67] K. Kaza, A. Psaltis, K. Stefanidis, K. C. Apostolakis, S. Thermos, K. Dimitropoulos, *et al.*, "Body Motion Analysis for Emotion Recognition in Serious Games," vol. 9738, pp. 33-42, 2016.

[68] J. Arunnehru and M. Kalaiselvi Geetha, "Automatic Human Emotion Recognition in Surveillance Video," vol. 660, pp. 321-342, 2017.

[69] 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.

[70] E. I. Barakova and T. Lourens, "Expressing and interpreting emotional movements in social games with robots," *Personal and Ubiquitous Computing*, vol. 14, pp. 457-467, 2010.

[71] 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, pp. 1256-1265, 12/31/ 2010.

[72] A. A. Samadani, A. Ghodsi, and D. Kulic, "Discriminative functional analysis of

human movements," *Pattern Recognition Letters*, vol. 34, pp. 1829-1839, Nov 2013.

[73] G. Venture, H. Kadone, T. X. Zhang, J. Grezes, A. Berthoz, and H. Hicheur, "Recognizing Emotions Conveyed by Human Gait," *International Journal of Social Robotics*, vol. 6, pp. 621-632, Nov 2014.

[74] R. Kar, A. Chakraborty, A. Konar, and R. Janarthanan, "Emotion recognition system by gesture analysis using fuzzy sets," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 8298 LNCS, ed, 2013, pp. 354-363.

[75] 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, pp. 454-467, 2014.

[76] S. K. D'mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems," *ACM Comput. Surv.*, vol. 47, pp. 1-36, 2015.

[77] 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.

[78] 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, pp. 64-84, 2009.

[79] C. Shan, S. Gong, and P. W. McOwan, "Beyond Facial Expressions: Learning Human Emotion from Body Gestures," in *BMVC*, 2007, pp. 1-10.

[80] 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.

[81] 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.

[82] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, pp. 33-48, 2010.

[83] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, *et al.*, "Toward real-time multimodal processing: EyesWeb 4.0," in *Proc. Artificial Intelligence and the Simulation of Behaviour (AISB) 2004 Convention: Motion, Emotion and Cognition*, 2004, pp. 22-26.

[84] 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, pp. 18-37, 2010.

[85] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychol Rev*, vol. 110, pp. 145-72, Jan 2003.

[86] M. Lewis and L. Cañamero, "Are discrete emotions useful in human-robot interaction? Feedback from motion capture analysis," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 97-102.

[87] A. K. Matthias Rehm, Nicolaj Segato, "Perception of Affective Body Movements in HRI across Age Groups: Comparison between Results from Denmark and Japan," 2015, pp. 25-32.

[88] 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.

[89] 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.

[90] 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.

Benjamin Stephens-Fripp received his B.E. degree in Mechatronic Engineering from the University of Sydney, NSW in 2001 and completed his Master in Philosophy degree in 2017 with the school of Electrical, Computer and Telecommunications Engineering at University of Wollongong, NSW. He is currently undertaking a PhD with the school of Mechanical, Materials, Mechatronic and Biomedical Engineering at the University of Wollongong, NSW. His current

research interest is in developing non-invasive sensory feedback for a soft robotic prosthetic hand.

Fazel Naghdy has a demonstrated track record and leadership in research, teaching, and management. His research has had its focus on machine intelligence and control particularly in embedded mechatronics and robotics systems. He has more than 320 publications in international journals and conferences and as book chapters. He is also contributing reviewer to IEEE Transactions on Mechatronics Engineering, and International Journal of Intelligent Automation and Soft Computing, and many others. He has served on a large number of International scientific committee of various international conferences. He is the Director of Centre for Intelligent Mechatronics Research. His current research interests include haptic rendered virtual manipulation of clinical and mechanical systems, intelligent control and learning in non-linear and non-structured systems. He is currently a Professor of Robotics and Intelligent Systems at University of Wollongong. Fazel Naghdy was born in Tehran, Iran. He received his first degree from Tehran University in 1976, MSc and PhD from the Postgraduate School of Control Engineering, University of Bradford, England, in 1979, 1982 respectively.

David Stirling received his B.E. degree in electronic engineering from the Tasmanian College of Advanced Education, Hobart, Tasmania, in 1976, an M.Sc. degree in digital techniques from Heriot-Watt University, Edinburgh, U.K., in 1980, and his Ph.D. degree in computer science and machine learning from the University of Sydney, NSW in 1995. He is currently a Senior Lecturer in the School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong, NSW. His research is essentially interdisciplinary in nature and particularly allied with data mining and innovative applications of machine learning to challenging, real world problems.

Golshah Naghdy is an Associate Professor at the School of Electrical, Computer and Telecommunication Engineering, University of Wollongong. She received her BSc in Electrical Engineering and Electronic Engineering from Sharif (Aryamehr) University, Tehran. She did her Post-graduate studies in England where she received an MPhil in Control Engineering and PhD in Electrical and Electronic Engineering. Golshah was a Senior Lecturer at Portsmouth University before joining Wollongong University in 1989. Her research interests include biological and machine vision, in particular, a generic vision system based on “wavelet neurons” and its application in the development of artificial retina implants, medical image processing, content based image retrieval, and robotics.