AUTOMATIC EMOTION RECOGNITION FROM BODY MOVEMENTS

THEOCHARIS ZACHARATOS

University of Cyprus, 2021

Humans are emotional beings and their feelings influence the way they perform and interact with computers. Emotion recognition is important in a variety of applications such as intelligent tutoring systems, computer games, robotics, medical and human computer interaction. The most expressive modalities for humans is body posture and movement and it has lately received attention from researchers in the use for emotion recognition. Despite these advances, there is a significant gap in low recognition rates on emotion derived from body modality.

This thesis addresses shortcomings in emotion recognition and it proposes various methods towards an automatic emotion recognition system. Firstly a method using body postures to detect emotions in a game environment is presented. Even though there was no temporal information available it resulted in satisfactory levels of recognition. Secondly a method adding temporal information with high level notational systems is presented inspired by several temporal techniques and the theory of Laban [137], we have created a model that achieved an 89% recognition rates, showing that temporal information along with high level notation systems can increase the results. Thirdly a method of automatic segmentation of body movements is presented, proposing symmetry for automatic segmentation, which can be used as an input for fully automated systems of recognising emotions. Finally a method to classify emotions using deep CNN models on the 3D raw data is presented resulting to a high recognition rates of 81% for binary classification that shows a very promising path for future research.

AUTOMATIC EMOTION RECOGNITION FROM BODY MOVEMENTS

THEOCHARIS ZACHARATOS

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Cyprus

Recommended for Acceptance

by the Department of Computer Science

August, 2021

2021

VALIDATION PAGE

Doctoral Candidate: Theocharis Zacharatos

Doctoral Dissertation Title: Automatic Emotion Recognition from body movements

*The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the **Department of Computer Science** and was approved on the **August 27, 2021** by the members of the **Examination Committee**.*

Examination Committee:

Research Supervisor: _____
Dr. Chrysanthou Yiorgos

Committee Member: _____
Dr. Constantinos S. Pattichis

Committee Member: _____
Dr. Pelechano Nuria

Committee Member: _____
Dr. Panayiotou Georgia

Committee Member: _____
Dr. Aristidou Andreas

ACKNOWLEDGEMENTS

The completion of this endeavour would not have been possible without a few people that i consider mentors and friends. First and foremost I am extremely grateful to my supervisor, Prof. Yiorgos Chrysanthou who believed in me, his invaluable advice and continuous support.

I am also grateful to Dr. Christos Gatzoulis who stood by me in every step of this journey, pushing me always to the limits.

My gratitude extends to distinct colleagues from the computer science department of University of Cyprus, for their advices and support, Dr Georgiou, Dr Pallis, Prof Pattichis, Dr Aristidou, Dr Charalambous, Dr Stavrakis.

Finally i would like to dedicate this work to my family, my wife Maria for her huge patience, and my two kids Paulina and George.

CREDITS

As a result of the work carried out for this thesis a number of papers have been published as depict below:

1. Zacharatos, H., Gatzoulis, C., Charalambous, P. and Chrysanthou, Y. *Deep CNNs for Emotion Recognition from 3D Motion Capture Data*. In proceedings of 3rd IEEE Conference on Games, 17-20 Aug, 2021, IT University of Copenhagen, Denmark.

2. Zacharatos, H., Gatzoulis, C., and Chrysanthou, Y. *Automatic Emotion Recognition Based on Body Movement Analysis: A Survey*, in IEEE Computer Graphics and Applications, vol. 34, no. 6, pp. 35-45, Nov.-Dec. 2014, doi: 10.1109/MCG.2014.106.

3. Zacharatos, H., Gatzoulis, C., and Chrysanthou, Y. and Aristidou, A. *Emotion Recognition for Exergames using Laban Movement Analysis*. In Proceedings of Motion on Games (MIG '13), 2014. Association for Computing Machinery, New York, NY, USA, 61-66.
DOI: https://doi.org/10.1145/2522628.2522651

4. Zacharatos, H., Gatzoulis, C., Chatzitofis, A. and Chrysanthou, Y. *Recognizing Emotional Expressiveness in Raw 3D Body Motion Data*. Poster paper publication, Motion in games, May 7th-8th, 2016, Lisbon, Portugal.

5. Zacharatos, H., Gatzoulis, C., and Chrysanthou, Y. *Affect Recognition during Active Game Playing based on Posture Skeleton Data*. In GRAPP 2013 IVAPP 2013 - Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications (pp. 419-422), Barcelona, Spain, 2013.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

Scientific findings indicate that emotions play an essential role in decision making, perception and learning, hence influence the mechanisms of rational thinking. An imbalance of emotion can impair decision-making. According to Rosalind Picard [186], if we want computers to be genuinely intelligent and to interact naturally with us, we must give them the ability to recognize, understand and express emotions. Despite the significance of user affect in computing, technologists have largely ignored emotion, resulting in often-frustrating experiences for people. Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affect. Research in this field combines engineering and computer science with psychology, cognitive science, neuroscience, sociology, education, psychophysiology, value-centered design, ethics, and more. Detecting emotional information begins with active or passive sensors that capture data of the user's physical state or behavior. The data gathered can depict emotions from different input sources called modalities. For example, a video camera might capture facial expressions, body posture and gestures, while a microphone captures speech. Other sensors detect emotional cues by measuring physiological data, such as skin temperature and galvanic resistance. Even though some of the above modalities have been successfully explored,

new modalities have started to arise. One very promising and emerging modality that has lately started to receive attention from researchers is body posture and movement.

Various application areas exists for emotion recognition. In intelligent **tutoring** systems, emotion recognition can be used to adapt the presentation style when a learner is bored, interested, frustrated or to detect student motivation [15] [223]. This can be done by providing an estimation of the level of interest and engagement derived from the Theory of Mind [16], which emphasizes the principles and techniques that humans deploy in order to understand, predict, and manipulate the behavior of other humans. In online education, there is a lack of direct, timely, and effective communication and feedback between teachers and students. By enabling real-time emotion detection in online lessons ensures that feedback expressed by facial expressions can be provided to teachers real time which will enable them to adjust the teaching program and ultimately improve the quality and efficiency of online education [230].

In **games**, the affective state of players during game play has a significant effect on their motivation and engagement. Often players lose interest and stop playing a game due to negative emotions such as frustration or anger. On the other hand, players who experience positive emotions during game playing are more likely to continue playing. A system that recognizes player's emotions during game playing can be a useful tool for game designers, allowing them to employ artificial intelligence behaviors for the characters and the game in response to the user's state. New advances in non-intrusive user interfaces that allow human gestures as input have resulted in the high popularity of a new game genre called *exergaming*. *Exergames* go beyond the passive gameplay activity that traditional controllers such as gamepads, keyboard and mouse offer, and require players to become physically active using their body movements for interacting with the game. They are often used to promote a healthy lifestyle for both casual gamers that use such interfaces at home but also for special categories of users who need to advance their physical activity in order to improve specific health conditions [86] [124] [114] [113].

A fascinating challenge in the field of **robotics** and human-robot interaction is the possibility to endow robots with emotional intelligence in order to make the interaction more intuitive, genuine, and natural. To achieve this, a critical point is the capability of the robot to infer and interpret human emotions. For example in healthcare, we have interactive assistive robots designed to interact socially with humans and non-interactive robots for surgical, rehabilitation and medication delivery [80]. In Human Robot Interaction (HRI), emotions have been considered from the following main points: a) Formalization of the robots own emotional state, in order to improve their effectiveness and enhance their believability [110], b) emotion expression, the ability of robots to exhibit recognizable emotional expressions for social interaction [163] and c)Ability to infer the human emotional state in order to be more effective while interacting with people [41].

Driving is an essential activity that people do and research suggests that is affected by people's emotions [117]. The inability to manage one's emotions while driving is often identified as one of the major causes for accidents. A system can use physiological sensors to collect and analyze data in order to recognize the driver's affective state, by interpreting both the mental and physiological components of the particular emotion experienced, and respond accordingly [24] [105]. In this field companies are utilizing research results to create multimillion global operations, such us Affectiva, which analyse human states while driving, to save lives by improving road safety with in-cabin sensing [2].

Emotion recognition is also being used for **medical** applications and **healthcare**, to detect depression for patients suffering with dementia [133], or detect emotion in patients diagnosed with schizophrenia and autism [102]. Emotion recognition systems are particularly suited for the study of autism spectrum disorder (ASD), where patients with sever symptoms have developmental and long-term difficulties in evaluating facial emotions [153].

It is highly anticipated that the advent of 5G technology will affect a huge amount of information processing and numerous applications in internet of things. However, it should be a serious threat if

AI algorithms are utilized to attack a wide range of applications and promising systems including 5G-enabled autonomous cars, smart drones, AI-driven facilities, smart buildings, manufacturing machinery, and healthcare in advanced smart cities. Due to this threat a recent and promising emotion recognition area has emerged, the ability to distinguished humans from machines and things [125].

When it comes to reading and understanding emotions, humans are still way ahead of machines. It is foreseen that with further scientific progress, the emotional intelligence of machines will become more accurate, in order to become an increasingly integral part of our communication. My personal motivation is to contribute to this endeavour by creating beyond the state of the art reseach in emotion recognition and more specifically from body movements.

1.2 Problem Statement/Contributions

Over the past few years, many experiments with different approaches towards emotion recognition have been carried out. These approaches differ in various aspects, such as the modalities they utilize, the way they model the emotions, the emotion sets they tackle, and the techniques they deploy to achieve the recognition. The most widely used modality in emotion recognition is facial expressions [152] [157] [218]. A recent successful example using this modality was the FERC project [167] which deployed Convolutional Neural Networks in two phases. In the first phase to remove background information and in the second to extract a facial feature vector. Trained from a custom database of 10,000 images captured from 154 different persons, the developed model achieves a recognition rate of 96% accuracy for a set of 5 emotions. A similar approach [108] utilized a set of two sub-networks, each one utilizing a different CNN architecture. Each sub-network has a different task. The one network recognizes eye and mouth features, while the other focuses on macrostructure. After it was tested on public databases, the model achieved accuracy rates between 62.11% and 96.44%. While the face modality can potentially offer high recognition rates, test results indicate, that cannot generalize in all cases. Voice expressions

have also been used in the form of lexical cues [59] and a combination of acoustic and linguistic features [148] [143] but such techniques cannot generalize the verbal content of emotional speech. Brain and physiological signals such as Electroencephalography [154] and pulse/heart rate have been analyzed to recognize emotions with signal processing and traditional classification algorithms. Such techniques are still costly and experimental and do not yet provide recognition rates higher than facial features.

Using only motion to derive the affective state of the user is a challenging task, due to the fact that it involves precise motion segmentation, evaluation using human characteristics which are based on people origin, tradition or customs, and all these need to be performed in real time for useful results. As part of this thesis automatic emotion recognition methods from body movements are proposed, during a human computer interaction process.

Over the past years a set of approaches have been proposed that use expressions derived from body postures and movements [28] [50] [122] [128] [228]. A recent technique [200] deployed deep neural networks that receive low level sequences of body joints position and orientation to recognize seven different emotions from motion data collected with Microsoft Kinect. The results of experiments with different sets of emotions show that this technique can identify emotion above chance level (accuracies that vary between 33% and 89% for different emotions and different experiments). Another interesting statistic reported is that human performance for the given tasks can be as high as 63%, showing the subjectivity of perceiving emotions even when acted, leading clearly to a need of emotion recognition from one modality such us body motion or to a recommendation of incorporating the techniques as part of a multimodal system.

Another recent technique[208] proposed a CNN that models the skeleton data as a static graph, but adds additional connections among joints, adding supplementary information. They used a custom dataset of 5492 samples divided in four emotional classes and achieved 68.4% accuracy in emotion

recognition. Another technique[53] used the Riemannian center of mass for each motion classification on a set of five emotion classes, with acted data captured constrained in moving along a given path with average accuracy of 71.2%. The human evaluator results of this study reported an average of 74.2%. In addition to the above techniques mentioned there was another [3] that recognized emotions among five classes using a first step of ANOVA to remove irrelevant features, and a binary-chromosome genetic algorithms to further reduce the feature set. This system achieved accuracy of 86.6% using a proprietary data set. Part of the feature set used in this work was based on Laban Movement Analysis, which was used in an earlier experiment as part of the current thesis.

This thesis addresses some of the issues arise from of low recognition rates, on emotion derived from body modality. More specifically a method using body postures to detect emotions in a game environment is presented. Even though there was no temporal information available, and with a limited training data sets, it resulted to satisfactory levels of recognition. In addition a method adding temporal information with high level notational systems is presented. Adding more features by inserting 3D data as an input with temporal information and Inspired by several temporal techniques and the theory of Laban [137], we have created a model that achieved an 89% recognition rates, showing that temporal information along with high level notation systems , such us the Laban, can increase the results. Further more a method of automatic segmentation of body movements is presented, which can be used as an input for fully automated systems of recognising emotions. Proposing symmetry that can be used as a feature to automatic segmentation expressive clips, we have created a symmetry detection and segmentation algorithm, that produced satisfactory results. Finally a method to classify emotions using deep CNN models on the 3D raw data is presented resulting to a high recognition rates of 81% for binary classification that illustrates a very promising path for future research. A short description of the contributions over traditional approaches is presented over the next few paragraphs.

1.2.1 Emotion Recognition from Body Postures

Body postures do not contain temporal data but can still embed affective information. Previous studies have used postures captured from professional motion capture software to build and test emotion recognition models that perform at comparable to human base rates in which humans used only an avatar representation model with no facial expressions to rate postures [130]. A major difference of the current research was the use of a less accurate sensor for body skeleton (Microsoft Kinect) since professional motion capture systems are not available to home gaming systems, making results not applicable in short-term. This project investigated the comparison of an automatic recognition system based on postures information but provided the observers who performed the annotation with a double modality information, both skeleton data and video snapshot. This resulted in increased agreement levels and raised the human base rate. The system was trained based on the labelling of the observer with higher agreement levels. The classification problem included three classes of emotions concentrated, frustrated and triumph, therefore the chance level recognition probability was 33.33%. The overall agreement of observers measured before training was 72% showing that emotion recognition based on posture is not always a deterministic task. Despite this, the trained model achieved 61.22% recognition rate, using back-propagation, well above chance level and close to the observers' agreement. Since the training set that was used was limited (included only 147 postures) and unbalanced (half of postures belonged to the frustrated class), the proposed technique has potential of improvement, since there are now several data augmentation techniques [43] [100] [71] [136] [132] [196] [209] [218] that can enlarge and improve the training data set, anticipating higher performance levels for the model. Overall, postures may lack the temporal dimension, however a model can utilize a set of key postures (such as key frames) to extend its functionality in a temporal space, while maintaining its simplicity of expressive postures as input for the model.

1.2.2 Emotion Recognition Using Motion

The use of raw 3D skeleton data has proven successful in many experiments [19] [127] [162] [161] and researchers have made efforts to determine more sophisticated feature sets such as velocity, acceleration, fluidity amplitude and so on [40] [201] [202]. The current project utilized observer agreement to achieve annotations on non-acted motion clips for the emotions of excitement, frustration, concentration, and meditation. For the classification problem novel features derived from the Space and Time motion factors of the Effort component from the Laban Movement Analysis (LMA) theory [137] were calculated. Previous studies that deployed Laban's theories in examining emotion either did not perform automatic emotion recognition [161] [162] or did so with low recognition rates [35] [37] and on other domains such as dance. The current project focused on motion data captured from computer games sessions (more particular exergames, i.e. using the body as the input modality). Using the Space motion factor of LMA allowed the feature set to contain elements of human attention to surroundings. The proposed implementation of the Space component included percentages of narrowing down the body, and the orientation of the face vector with the four extremities. Further more implementing the Time motion factor of LMA enables to encode decision and intuition in the feature set. Sustained movements convey calmness, while sudden movements convey excitedness. Velocity, acceleration and jerk for the four extremities of the body were used to implement the Time factor. Data analysis revealed that the selected feature set allows distinguishing the four emotion classes. The overall results show high recognition rates for all four emotion classes (83-89%) showcasing the strength of LMA in engineering features for automatic emotion recognition. The current project success on deploying two LMA factors, highlighted that further engineering on Laban-based features can provide additional intuition towards emotion expressiveness via the body motion modality. Compared to the previous experiment with postures, this experiment deployed more complex raw data that was transformed based on the LMA theory and used

a bigger data set of 309 non-acted labelled motion clip samples. Nevertheless as mentioned, data engineering, augmentation or deployment of the latest deep learning techniques may further improve the recognition accuracy, however given the already high recognition rates this might not necessarily be the case.

1.2.3 Segmentation and Recognition of Emotional Expressiveness in Raw 3D Body Motion Data

The previous experiments showcased emotion recognition based on 3D posture and motion data recorded during gaming, however this was based on pre-segmented clips and human annotation. In an application scenario, recognition needs to happen in raw data produced in a time series format. In this case, the system should be able to determine emotion on clips that are not if given or predetermined size. The third experiment of this thesis provides a solution in this direction as well as a different approach towards recognizing emotional expressiveness. Raw data from 13 players who played sports games for 30 minutes was collected. The raw data is parsed by a function that calculates rotational kinetic energy at each frame, based on angular velocity. Based on observation of the raw data, energy increases during energetic motion. The energy signal is smoothed with the Savitzky Golay filter and clip segmentation takes place between local minima. Following the segmentation, a symmetry calculation based on hand and leg joints allows the analysis of the movements. Observation yielded that seeking above than average body symmetry combined with reduced lower body energy, appears to isolate a large percentage of the expressive clips and removes the vast majority of the non-expressive clips. While this novel approach has satisfactory preliminary results, its contribution can be further enhanced with further research in the given direction, such as the combination of this technique with either supervised learning where the LMA features are calculated after the segmentation, or with outlier detection considering emotions as outliers of a large cluster of non-expressive clips.

1.2.4 Deep CNNs for Emotion Recognition on Image transformation of 3D Motion Data

The final experiment of this thesis, focused on the classification of emotions from 3D body movements using modern Deep Learning techniques. Previous techniques focused on multimodal feature set that combined body motion with other modalities [150] [17]. Another technique [200] deployed sequential model of low level 3D joint data. The proposed technique utilizes a model that was used for action recognition [216] but this time on the context of emotion recognition. Each 3D motion skeleton data clip is converted to an RGB image with each row comprising a frame, and each column the temporal value series of a given feature. Features include joint to joint distances and their orientations with the orthogonal axes. The images are then used to apply transfer learning to train a version of the Inception V3 model and use it to solve a binary classification emotion recognition of happiness and sadness. The experiment used an acted emotional body movement data set [90]. The dataset contained scenarios of typical and natural expressions, captured by a motion capture system. Selected scenarios were carried out to include an equal number of men and women actors. Overall, 208 happiness and 194 sadness different input clips were deployed for training and testing the model. The results of binary classification between happiness and sadness exhibit a 81% recognition rate, showing that combining posture and subsequent frame motion dynamics in an image that uses rows as a temporal dimension and columns as dynamic features can capture affective information.

1.3 Document Structure

Chapter 2 summarises related affective computing literature around body movements. It provides an overview of emotion theories and the models of emotion recognition concentrating on body movements. It illustrates recent approaches on emotion recognition from body movements and summarises experimental methodologies to capture, validate such emotions.

Chapter 3 describes an initial contribution using active game playing data with postures to recognise three distinct emotions.

Chapter 4 describes the contribution of a real time emotion recognition system for four distinct emotions, using Laban Movement Analysis techniques to generate features used by the machine learning algorithms.

Chapter 5 Portrays the contribution of real time body symmetry detection and how it is related to emotion expressiveness from raw 3D body motion data.

Chapter 6 Covers our latest contribution utilizing deep neural networks for emotion recognition, using image-based transformations of 3D skeleton motion data.

Chapter 7 Summarizes the conclusions of the current thesis as well as the future directions from the research perspective.

Chapter 2

Related Work

2.1 Models of emotion

Recognition and analysis of human emotions have attracted a lot of interest over the past two decades and extensive research has been carried out in neuroscience, psychology, cognitive sciences, and computer sciences. There are various ways of representing emotions, either by using **distinct emotions** like happiness, sadness, fear, anger, surprise, disgust or by measuring and contextualizing emotions according to some **dimensional space** as illustrated in Figure 32, where emotions are represented in two dimensions of *valence and arousal* and each emotion can be viewed as point in the space defined by these dimensions.

Similarly, emotions are represented on dimensions containing *activation and evaluation*. Activation is understood as the tendency of the person to execute an action according to his emotion and evaluation reflects the global appraisal of the positive or negative feeling [187] [50].

Popular models have seen to be those dealing with psychological models of nonverbal communication like the *Pleasure-Arousal-Dominant model (PAD)*. In the PAD model, the Pleasure-Displeasure scale measures how pleasant an emotion may be, the Arousal-Nonarousal scale measures the intensity of the emotion, and the Dominance-Submissiveness scale represents the controlling and dominant nature

of the emotion [168].

Another model widely used in emotion research is the *OCC model* [178] that is applied in a large number of studies to generate emotions for embodied characters [66] [131]. This model states that the strength of a given emotion primarily depends on the events, agents, or objects in the environment of the agent exhibiting the emotion. The model specifies 22 emotion categories and consists of five processes that define the complete system that characters follow from the initial categorization of an event to the resulting behavior of the character. These processes are a) classifying the event, action or object encountered, b) quantifying the intensity of affected emotions, c) interaction of the newly generated emotion with existing emotions, d) mapping the emotional state to an emotional expression and e) expressing the emotional state.

An alternative approach for modeling affect is the *Appraisal theory*. Appraisal theory is the idea that emotions are extracted from our evaluations (appraisals) of events that cause specific reactions in different people. Essentially, our appraisal of a situation causes an emotional response that is going to be based on that appraisal. Appraisal theories offer a more flexible framework than discrete and dimensional models, being able to account for individual differences and variations of responses to the same stimulus by the same individual at two different moments in time [197].

All the above definition of emotions have been used in studies before. Distinct emotions are preferred in studies with a need for specific emotions to be recognized such as in a healthcare for recognition of depression or concentration. By using models of emotion in a dimensional way, you can combine distinct emotions, resulting in a dimension recognition, thus resulting in higher recognition rates. Dimensional approaches can be used in applications like games where a dimension of valence or arousal is enough for recognizing excitement vs boredom.

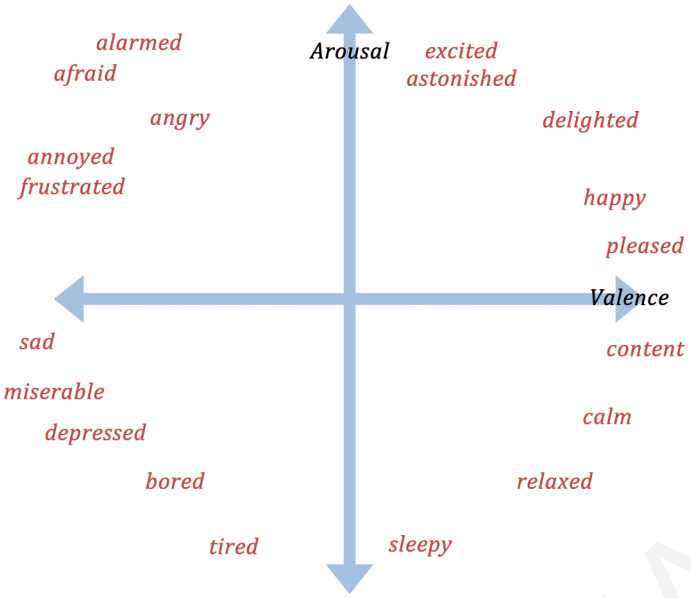


Figure 1: The Valence-Arousal space

2.2 Emotion recognition from various modalities

There has been extensive research in the field of emotion recognition using various modalities. The most widely used techniques are through **facial expressions** [157] [180], where recognition of actions of individual muscles called action units is performed, rather than emotions through the detection of differences between a spontaneous and an exaggerated data. Such studies incorporate temporal dynamics of facial actions and parameters like speed, intensity, duration, and the co-occurrence of facial muscle activation, in order to classify facial behavior presented as either deliberate or spontaneous [220] [152].

Another widely used modality is **voice** expressions. Through voice expressions [238]. Methods propose interpretation of speech signals in terms of certain application-dependent affective states. Lexical cues were used with better performance than paralinguistic cues to detect relief, anger, fear, and sadness in human-human medical conversations [59]. Some studies introduce the idea of using a combination of acoustic and linguistic features to improve vocal affect recognition [148] [143]. Although this can show an improvement on vocal recognition, in many cases the automatic extraction of these related features can be a difficult problem. Existing systems based on this idea cannot reliably recognize the verbal

content of emotional speech. Furthermore, extracting semantic discourse information is even more challenging. Most of these features are either extracted manually or by transcripts.

It is worth mentioning that over the past two decades several studies have focused on emotion recognition using **brain-machine interfaces**. They are based on the neural activity of the users [151] and the most used technique is electroencephalography(EEG) signal analysis. Neuroscientists develop new techniques in brain imaging that help us to map the neural circuitry that underlies emotional experience. An example of the contribution of neuroscience in understanding and recognizing emotion is the evidence provided to the discussions on dimensional models, where valence and arousal might be supported by distinct neural pathways. One of the most popular methods that neuroscientists use is the Functional Magnetic Resonance Imaging [142] by using evidence from brain damaged subjects to prove that emotions are very important in decision making [112]. Scientists observed that patients with a lesion in a particular section of the frontal lobe showed normal logical reasoning but yet they couldn't see the consequences of their actions and were unable to learn from their mistakes. From this observation scientists conclud that emotion related processes are required for learning, even in areas that had previously been attributed to cognition. The cost, time resolution, and the complexity of setting up protocols that can be used in real world activities are still problematic issues that put the application development with use of these techniques to a hold. Nevertheless, signal processing [18] and classification algorithms [154] for EEG have been developed in the context of building Brain Computer Interfaces.

Physiological signals can also be used to detect a user's emotional state by monitoring and analyzing their physiological signs such as pulse and heart rate (Blood Volume pulse), Galvanic Skin Response, and minute contractions of the facial muscles (Facial Electromyography). A person's Blood Volume Pulse (BVP) can be measured by a process called photoplethysmography, which produces a graph indicating blood flow through the extremities [186]. Facial Electromyography is a technique used to measure the electrical activity of the facial muscles by amplifying the tiny electrical impulses that are generated

by muscle fibers when they contract [139]. Nevertheless, this area is gaining momentum and there are now real products which implement the techniques. Galvanic Skin Response (GSR) is a measure of skin conductivity, which is dependent on how moist the skin is. As the sweat glands produce this moisture and the glands are controlled by the bodyś nervous system, there is a correlation between GSR and the arousal state of the body. The more aroused a person is, the greater the skin conductivity and GSR reading [186].

With the emergence of the affective computing field, various studies have been carried out to create systems that can recognize the affective states of their users by analyzing their **body expressions** in order to recognize, understand and model human emotion [228] [49] [128] [122]. For example, fear makes the body contract as an attempt to appear as small as possible, surprise causes orientating towards the object capturing attention, and joy may lead to movements of openness and acceleration of forearms upwards [28]. Bodily expressions have been recognized as important for nonverbal communication and changes in a persons affective state are also reflected by changes in their body posture. There are two separate pathways in the brain for recognizing biological information: one for form information and the other for the motion information [87].

2.3 Emotion recognition from body

Darwin was the first to describe the association between body language and posture with emotions in humans and animals [54]. State of the art emotion detection systems have overlooked the importance of body posture compared to facial expressions and voice recognition. Posture can offer information that is unavailable from conventional nonverbal measures such as the face and speech. For example, the affective state of a person can be decoded over long distances with posture, whereas recognition at the same distance from facial features is difficult or unreliable [227]. Meijer [169] defined some dimensions and qualities such as trunk movement: stretching - bowing; arm movement: opening - closing; vertical

direction: upward - downward; sagittal direction: forward - backward; force: strong - light; velocity: fast - slow; directness: direct - indirect. Those dimensions and qualities can be found in different combinations for different emotions. For instance, a joyful feeling could be characterized by a strong force, a fast velocity and a direct trajectory but it could instead have a light force or be an indirect movement.

2.4 Capturing body movements

Various ways are used to capture body movements. Motor data can be captured by physiological instruments to the body such as an accelerometer or electromyograph. These systems provide very accurate information on motor movement [48] but can be used only in controlled laboratory experiments where freedom of expression is limited. Motion capture has been found to be an accurate measurement technique for measuring body movements like gait [177] and arm movements [188], but requires a controlled environment with cameras and sensors as illustrated in Figure 2. Two dimensional video cameras are using less obstructive ways of measuring body movement, such as silhouette extraction [89] but they don't provide accurate description of movement for a specific body part. More complex features can be computed using 3D depth cameras like the Kinect sensor [171] although it is still not as accurate as traditional motion capture systems.

2.5 Motion Datasets for Emotion recognition

One of the firsts datasets made accessible was the FABO database that was created by Gunes and Piccardi [97]. It is a bimodal database that combines face and body expressions recorded simultaneously. The videos were obtained in a lab setting with artificial light and the emotions were posed. Subjects were placed in front of a plain blue background to facilitate background subtraction. 23 subjects were filmed,

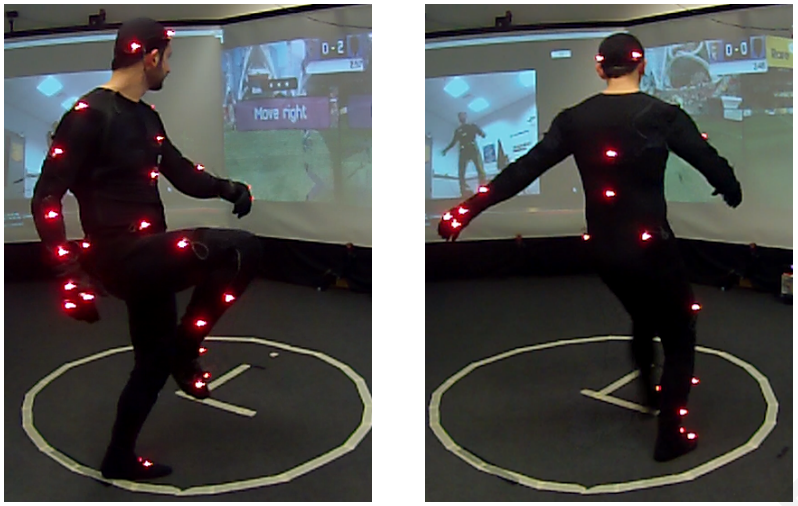


Figure 2: Motion Capture Process

with ages ranging from 18 to 50. The subjects were instructed to begin in a neutral expression and body position, and then perform the supposed emotion based on scenarios they were presented with. The expressions recorded were: neutral, uncertainty, anger, surprise, fear, anxiety, happiness, disgust, boredom, and sadness.

The GEneva Multimodal Emotion Portrayals (GEMEP) database contains audiovisual files that include 18 different emotions displays. Twelve of the emotion classes are categorized by two emotional dimensions: valence and arousal. The subjects that performed the emotions were French theatre actors between 25 and 57 years old. The videos were recorded in a controlled setting in a studio. The expressions were recorded with three digital cameras, one zoomed in on the face, the other zoomed out displaying the body and posture from a frontal view, and the other one from a side view.

2.6 Notational systems for body movements

Movement notational systems are required to access the expressive content of movement. They have been proposed for observing, describing, notating and interpreting human motion [213]. Many of the systems used from dance choreographers since 17th century. Some of them can describe only foot positions and symbols for each step like rhythm [67], while others can describe how emotions,

attitude and personality are conveyed in dynamic body postures and gestures [160]. Notation systems have been used contain information about body and limp positions of the extremities [20], orientation of the different body limbs expressed in a spherical coordinate system [69] and spatio-temporal coding [78] [103] that describes position and movement on the three Cartesian axes (sagittal, vertical and transverse).

Psychology coding systems describe orientation of postures and actions during seated interactions to check on attitudes and emotions like boredom/interest, agreement/disagreement [32]. Tracy and Robins [217] proposed a coding system to assess pride in static upper body postures. Ekman and Friesen [64] proposed a coding system for hand actions that has been used by various gesture-speech systems [62] [165]. Wallbott proposed a reliable correlation between different parts of body, emotion and movement qualities like activity, expansiveness and dynamics [228]. Birdwhistell [25] founded the theory of Kinesics, which assumes that nonverbal behavior is used in everyday communication systematically and can be studied in a similar way to language. A minimal part distinguished in kinesics is a kineme which is the smallest meaningful set of body movement, for example raising eyebrows or moving the eyes upward. Birdwhistell developed a complex system of kinegraphs to annotate kinemes for the research on body language.

A widely used movement notational system is *Laban Movement Analysis (LMA)* [137]. It was originally developed by dance artist and theorist Rudolf Laban in the early 20th century. Laban method focuses on the relationships between internal state, intention and attention and their effects on all human motions. One of the strong points of LMA is the ability to describe expressive content of movements, which makes it appropriate for emotion and behavior analysis. Many researchers have been trying to create a computational form of LMA for motion analysis [13] [240] [241]. Laban movement analysis has also been used to reproduce expressive movements in robots that could be interpreted as emotions by human observers and map low-level features to LMA parameters [173]. The main limitation it is tied

to the embodiment of the robot which has limited number of degrees of freedom. A system has been proposed that implements LMA using probability calculus and Bayesian theory using Bayesian models for computational LMA in form of Bayesian nets and joint distributions [191]. The results indicated that there is a set of low-level features that can be used as evidence for the Laban parameters and that the classifier is able to make online predictions, thus giving the system a sense of anticipation.

Another notational system called *Body Action and Posture* (BAP) is best suited for coding nonverbal emotion expression was proposed [52]. In this system a distinction is made between body posture units and body action units [103]. A posture unit represents the general alignment of one part, or set of parts of the body (head, trunk, arms) to a resting place, which shows periodic changes known as posture shifts e.g. arms crossed. An action unit is a path of one part or set of parts of the body (mostly the arms) outside the resting place with a distinct start point, a short duration and a distinct end point where the path returns to a resting place e.g. head shake, pointing arm gesture. All body skeleton movements are categorized as an action or posture with the exception of legs. BAP separates its behavior variables in 12 categories of action and posture: head orientation, trunk orientation, arms, whole body posture, gaze and other. Even though the coding system attempts to be as objective as possible, it offers challenges to its automatic implementation on how the data is segmented, containing noise and synchronization due to different sample rates for the different sensors. Moreover the same segment may have different labels depending on the previous and next segment. These faults are removed from a more recent approach called *AutoBAP* [221] which is a system that automates the annotation process of BAP. By preprocessing the sensor data it synchronizes the different sample rates and merges the data so it removes the synchronization issues.

2.7 Emotion recognition using notational systems

Even though notation systems are using motion analysis for expressive movements, only some of them have been used for emotion recognition. As previously described, the motion analysis is performed by using components from Laban's theory [137] [44] [176] [199]. Laban's theory is divided into four categories called body, space, effort and shape. Most of the work in emotion recognition focuses on the effort component that deals with the expressiveness and describes the dynamic qualities of the movement and the inner attitude towards using energy. By selecting a set of suitable features from the trajectories described by hands, foot and head, the effort component can be used as one descriptor for expressive movements. Laban saw effort as the inner impulse, a movement sensation, a thought, a feeling or emotion from which movement originates; it constitutes the interface between mental and physical components of movement.

Lourens et al. [155] extracted low level features from video and used Labanotation experts to classify the video clips to four emotional states manually. Samadani et al. [198] has used Laban effort components for hand arm movements and an approach of quantifying shape direction based on the average trajectory curvature. The results show a high correlation between Laban certified movement analyst and shape directions for the hand-arm movements dataset.

Another study used Laban features like whole-body movement, inclination of the body and area, to extract four emotions (pleasure, anger, sadness and relaxation) from a robot that has limited ways of movement [162]. Although they used observers to classify the robot movements to emotions, they have not used automatic recognition techniques for classification. They used empirical estimation of correlation between Laban features and emotional set.

Dael et al. [52] adopted the Body Action and Posture (BAP) system to examine the types and patterns of body movement from actors expressing twelve different emotions. They used Principal component analysis to reduce the 49 behavior variables of BAP, resulting to 16 extracted components, that applied to

Name	*Study*	*Description*	*Evaluation*
Ellis and March	[67]	Foot positions and symbols for each step like rhythm	Movement
Marsella et al.	[160]	How emotions, attitude and personality are conveyed in dynamic body postures and gestures	Emotion , Movement
Benesh	[20]	Information about body and limp positions of the extremities	Movement
Eshkol and Wachmann	[69]	Orientation of the different body limbs expressed in a spherical coordinate system	Movement
Frey and Pool	[78] [103]	Spatio-temporal coding that describes position and movement on the three Cartesian axes (sagittal, vertical and transverse)	Movement
Bull	[32]	Orientation of postures and actions during seated interactions to check on attitudes and emotions like boredom/interest, agreement/disagreement	Emotion
Tracy and Robins	[217]	Assess pride in static upper body postures	Emotion, Movement
Ekman and Friesen	[64]	Hand actions	Movement
De Silva and N. Berthouze	[56]	Description of body expressions in terms of anatomical body part configuration measured as joint rotations and joint distances	Emotion
Wallbott	[228]	Correlation between different parts of body, emotion and movement qualities like activity, expresiveness and dynamics	Emotion, Movement
Birdwhistell	[25]	Theory of Kinesics, which assumes that nonverbal behavior is used in everyday communication systematically	Movement
Laban	[137]	Laban movement analysis method focuses on the relationships between internal state, intention and attention and their effects on all human motions.	Emotion, Movement
Dael et al.	[52]	Body Action and Posture (BAP) describes of body movement on an anatomical level (different articulations of body parts), a form level (direction and orientation of movement), and a functional level (communicative and self-regulatory functions)	Emotion

Table 1: Notational systems for emotions and movements

a two-step clustering algorithm [45] to reveal natural groupings. This clustering algorithm was useful for handling large datasets with both categorial and continues variables. Even though some emotions were characterized by a specific behavior pattern, most emotions were encoded by a combination of clusters that also grouped other emotions resulting in a large amount of expressive variability and overlapping response profiles.

2.8 Establishing the ground truth

A major issue with emotion recognition is the subjectivity of the area and the difficulty in establishing a benchmark on when the recognition is successful or not. Most early automatic recognition systems relied on corpora that had been acted [10] [94] [169] [194]. More recent studies are using non-acted data [127] [129] [237] with body posture and movements, validating the results by using human observers. In order for a result to be valid, it should be compared to a database of valid postures or movements set as the ground truth. Ground truth of the expressed emotional state is a critical aspect, as even self-assessment can be manipulated by lying. Believability and authenticity of emotional expressions is usually increased if all modalities express the same state. Incongruence is correlated with lying. In this case, expressions of body movements seem to be more reliable than facial expressions because people do less bother to censor their body movement or physiology in daily life than facial expressions [63] [96].

To set the ground truth, after human annotation on postures and motion clips, a validation of the agreement level of all annotators is required. Various techniques have been used for the validation of agreement of the annotators like Cohen's Kappa [22] which works for two annotators, and Fleiss Kappa [75] that works for any fixed number of annotators. Correlation coefficients [193] measure pairwise correlation among annotators using a scale that is ordered and intra-class correlation coefficient [70] is defined as the proportion of variance of an observation due to between-subject variability in the true

scores. Another approach to agreement is limits of agreement with Bland-Altman plot [26] which is useful when there are only two annotators and the scale is continuous. Krippendorff Alpha [134] uses several specialized agreement coefficients by accepting any number of annotators. It is applicable to nominal, ordinal, interval, and ratio levels of measurement, being able to handle missing data, and being corrected for small sample sizes. Table 2 describes various statistical techniques for validating annotation agreement when ground truth is constructed.

Study	*Name*	*Use*
Kleinsmith et al. [129]	Cross Validation	Is a technique to assessing how the results of a statistical analysis will generalize to an independent data set. Cross-validation involves partitioning a sample of data into subsets, performing the analysis on one subset (training set), and validating the analysis on the other subset (testing set).
Carletta J.[39]	Cohen's Kappa	Is a statistical measure for assessing the agreement between two annotators
Fleiss J. L.[75]	Fleiss Kappa	Is a statistical measure for assessing the agreement between a fixed number of annotators
Pearson K.[183]	Correlation coefficients	measure pairwise correlation among annotators using a scale that is ordered
Koch G.[130]	Intra-class correlation coefficient	Is a statistical measure that can be used when quantitative measurements are made on units that are organized into groups. Is the proportion of variance of an observation due to between-subject variability in the true scores
Bland, J. and Altman, D.[26]	Bland-Altman plot	Is useful when there are only two annotators and the scale is continuous
Krippendorff K.[134]	Krippendorff Alpha	Uses several specialized agreement coefficients by accepting any number of annotators. It is applicable to nominal, ordinal, interval, and ratio levels of measurement, being able to handle missing data, and being corrected for small sample sizes.

Table 2: Statistical Agreement Techniques

2.9 Motion Segmentation for Emotion recognition

Humans identify emotion from body motion or posture based on visual parameters called gestures and our brain annotates them to a specific emotion. Gestures are also user specific, and each human has a personal vocabulary of gestures [1]. Efron [51] proved that humans use visual cues to segment motion sequences into gestures and he showed that gestures are used as the building blocks of complex motions. The methodology in automatic body emotion recognition uses the same principle to segment a motion or a posture. Segmentation enables the identification of each motion unit and its representation through a set of values of relevant parameters. In order to create quality motion capture data efficiently, captured sessions typically produce long streams of motion capture data. The solution is to preprocess the long motion capture data stream by breaking it up into short segments that are appropriate for an analysis tool. This process is often done manually, but it is a very laborious and time consuming. Moreover manual segmentation is subjective and it depends on the human perception on movement, and very often to cultural elements of the person. Segmentation done by experts instead of non experts on specific motion data, like LMA experts in dance, gives a better quality of segmented clips. Most of the emotion recognition studies [129] [201] [237] [121] use manual segmentation of postures or motion, by using human observers to classify expressive posture and movements and separate them. A more accurate solution is to create tools that automate the segmentation process.

An automatic segmentation program will produce the same segmentation given the same input motion capture. On the other hand, different people will produce different segmentations, given the same motion capture data. In addition, a person will often produce different segmentations of identical motion capture data. Different approaches have been proposed for automatic motion segmentation. The direct approach where temporal segmentation precedes recognition like Kang et al. [120], which first computes low-level motion parameters such as velocity, acceleration and trajectory curvature, and the approach like Kahol et al. [119] that uses motion parameters such as human body activity and then looks

for changes in those parameters to identify candidate gesture boundaries. Although both approaches are able to detect gesture boundaries with high accuracy, the method works only if each gesture is preceded and followed by non-gesturing intervals, a requirement not satisfied for continuous gesturing.

Various indirect methods detect gesture boundaries by finding intervals that give good recognition scores when matched with one of the gesture classes and detect gesture endpoint by comparing the recognition likelihood score to a threshold [6] [145] [175] [138] [190] [7]. These kinematic methods are extremely efficient, however they produce simple low-level segmentations. Different approaches produce higher level segmentations. Barbic et al. [14] implemented a method where the projection error resulting from Principal Component Analysis (PCA) increases on larger segments of motion capture data. They also proposed a segmentation method by tracking changes in the distance when data containing the frames that precede the segment fits to a Gaussian distribution model. They use expectation minimization clustering to estimate the Gaussian Mixture Model (GMM) while Lee and Elgammal [144] used k-means to estimate the GMM. These time series methods produce higher level segmentation than the kinematic methods, but they do not utilize semantic content of the motion.

Kahol et al [119] use a Naive Bayes approach, in order to find segmentation profiles of dance motion capture sequences. Starner and Pentland [212] implemented implicit segmentation using Hidden Markov Models (HMM). Both showed that supervised learning approach can be used to capture some of the complexity of decision making present in manual motion capture segmentation. Supervised learning based segmentation methods are difficult to implement for general motion, due to the enormous number of training data that needed in order to create a true general classifier. A different solution will be to use a classifier that works well on smaller set of classes but is effective on general motions. Bounchard and Badler [29] used Effort component of Laban Movement Analysis, that deals with the expressiveness and describes the dynamic qualities of the movement and the inner attitude towards using energy, in order to represent such a classifier. The inner impulse is expressed by usage of motion factors. Every human

movement including thought has the potential to engage the four motion factors: space, weight, time and flow.

A different approach yielding towards a computational emotion recognition system is by segmenting expressive motion automatically. Cammuri et al. [36] used quantity of motion for segmentation which is the amount of detected movement and its evolution in time. Quantity of motion has been seen as a sequence of bell-shaped curves in order to perform segmentation between pause and motion phases. In order to segment motion, a list of curves and their features has been computed by using empirical thresholds [36] [35]. Another approach [21] [185] used the concept of motion energy for segmenting motion primitives [76] for emotion recognition. Body's motion energy can be computed as a weighted sum of the rotational limb speeds, which will be large for periods of energetic motion and will remain small during periods of low motion energy. Table 3 shows various studies that used automatic segmentation techniques.

Study	*Segmentation Feature*	*Boundary*	*Movement*	*Evaluation*
Camuri et al.	Quantity of motion	Threshold	Body	Emotion
Fod et al.	Motion Energy - Sum of squares of angular velocity	Threshold	Body	Motion
Bernhadt et al.	Motion Energy	Threshold	Body	Emotion
Kahol et al.	Activity of Segment	Local minima	Body	Motion
Pianna et al.	Motion Energy	Threshold	Body	Emotion
Alon et al.	Continues Dynamic Programming	Gesture Spotting	Hand	Motion

Table 3: Automatic Segmentation for body movement

2.10 Emotion recognition using low level features

Existing research that attempts to recognize emotions using human motion data does not achieve sufficient recognition rates, is based on training the system with low level feature data that is very vague (such as rotation of a given joint on a given axes etc) and is selected without firm justification from movement analysis theories. Some recent approaches in robotics do achieve good quality recognition [161] [162] [19] showing that body language can be successfully used by humanoid robots to express emotions, however their task is more simplified since robots perform mechanic and predetermined movements while expressive human movement is more complex and non-deterministic. Different approaches have been used in order to get higher recognition rates for some basic emotions like uses of movement qualities such as amplitude, speed and fluidity of movement to infer emotions [40]. Similar to this it has been tried to recognize emotions from animation by using low level features such as angular velocity, acceleration for the body's arm, hand and right forearm and body directionality for spine and head [201]. In this approach, the recognition rate is average for individual emotions and higher when categorization is been done for high and low intensity of emotions.

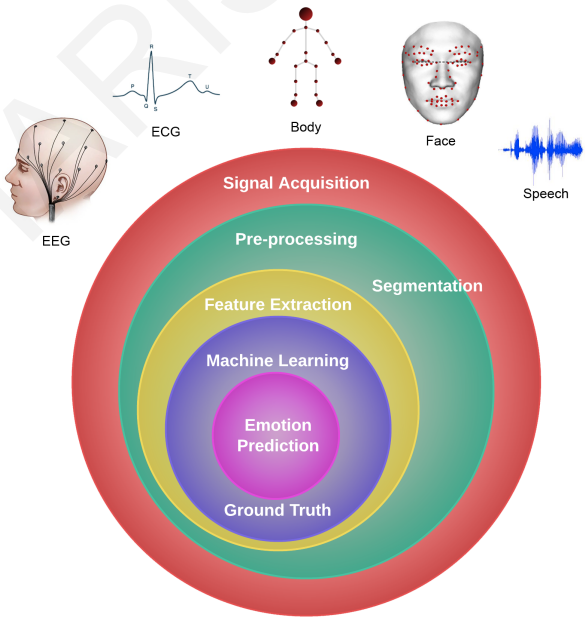


Figure 3: A Multimodal Automatic Emotion Recognition System

Another approach used balance postural control variables for automatic recognition. Human balance was assessed by analyzing center of gravity or center of pressure displacements [61]. Results showed that balance variables such as center of gravity and displacement variability were correlated to negative emotions and situation appraisals [88]. Low level 3D postural features were used also together with high level kinematic and geometrical features (body movement activity power, symmetry and bending) fed to a random forest classifier and achieved a high recognition rate [229]. In another study [47] various features to emotion classification were analyzed in order to measure their performance with respect to predict affective state of an input motion. Posture features: end effector positions, end effector orientations, bounding box. Dynamic features: velocity, acceleration, Jerk. Frequency-based features: output of fast Fourier transform for each position trajectory of the end effectors. The overall success rate was very high and the feature vector was combined by three feature sets, however actors were used instead on non-acted people. Table 4 shows various studies that used low level movement features for emotion recognition.

A more recent study [73] computed all the above features and categorised them into ten unique groups based on the type of movements, using a filter-based feature selection algorithm Analysis of Variance (ANOVA) [85] to select relevant features from each of the movement feature groups. Several top features from each feature group were used as inputs to the second layer of the framework. The number of features considered from each group was derived using normalised Multivariate Analysis of Variance (MANOVA) [215] score computed for each group separately. The number of relevant features selected from each group was based on the normalised MANOVA score computed for each motion feature group. A binary chromosome based genetic algorithm was utilised to extract a feature subset maximising the emotion recognition rate as displayed in Figure 4.

Study	*Feature used*	*Description*
[35]	Kinetic Energy	The overall energy spent during movement, estimated as the total amount of displacement in all of the points.
[137]	Construction Index	The Contraction Index is a measure, ranging from 0 to 1, of how the body occupies the space surrounding it and is the bounding volume of the minimum of a box surrounding the body.
[119]	Body Movement Activity and power	The body movement activity and power composed of three parameters, force, kinetic energy and momentum calculated hierarchically.
[205]	Body Spatial Extension	The Bounding Box calculated from the positions of joints in each frame and three spatial extents in x,y,z axis.
[195]	Symmetry	Spatial symmetric indexes are considered for the two hands in three coordinates. The three partial indexes are then combined in a normalized index that expresses the overall estimated symmetry.
[205]	Body Bending	Body bending forward or backward is measured by the velocity of the joint's displacement along its depth respective to the body position and orientation.
[228]	Smoothness	Hand trajectories curvature computation.
[228]	Fluidity	High curvature of the speed's trajectory in time means low fluidity, while low curvature means high fluidity. Fluidity = tangential velocity of the joint.
[169]	Directness	Movement Directness Index is computed from a trajectory by a joint as the ratio between the Euclidean distances, calculated between the starting and the ending point of the trajectory, and the trajectory's actual length The directness index tends to assume values close to 1 if a movement is direct and low values (close to 0) otherwise.
[207]	Periodicity	The periodicity transform decomposes sequences into a sum of periodic sequences by projecting onto a set of periodic subspaces. The Periodicity transform looks for the best periodic characterization of the length N sequence x.
[184]	Impulsiveness	A temporal perturbation of a regime motion.
[30]	Quantity of motion	The Quantity of Motion is computed as the weighted sum of the area of a layered silhouette: a wide movement will have a higher Quantity of Motion than a small one. An extended version is computing the direction of the motion by computing the gradients of motion history images.
[228]	Barycenter	The motion of the barycenter of a silhouette over time gives a good idea of the way a person is moving. In excitement the barycenter spans a bigger area and its movements are wider, while in sadness the movements are more quiet.

Table 4: Movement features for emotion recognition

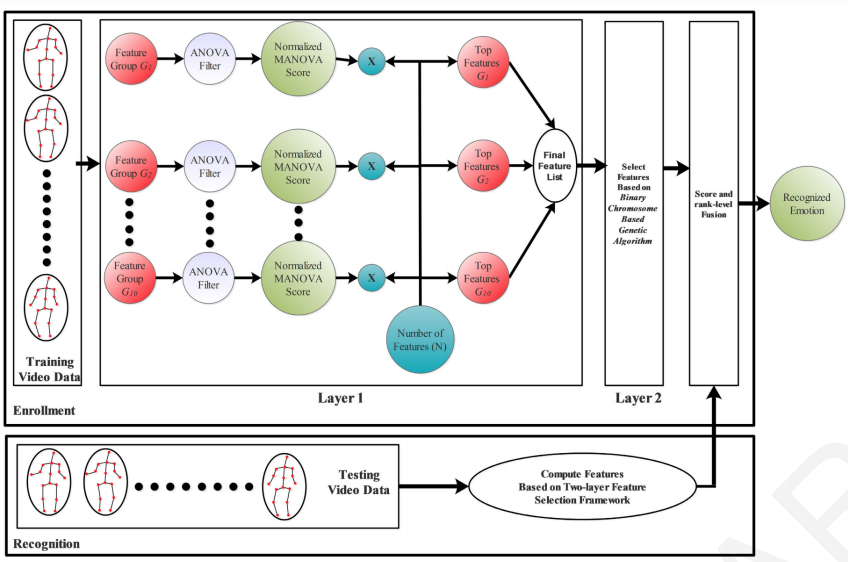


Figure 4: Framework for emotion recognition from fused features of body motion

2.11 Multimodal emotion recognition approach with body

As facial expressions and speech expressions dominate during face-to-face interaction, they cannot cover by themselves the full range of emotional states which can be recognized by observing a single modality [96]. This motivates the development of multi-modal emotion recognition systems. Most approaches for automatic recognition investigate the categorical emotions, in particular the basic emotions or a subset of them. Fewer studies refer to emotional dimensions. Nevertheless, recognition systems based on a dimensional model for emotions seem to be more appropriate for multi-modal emotion recognition because modalities differ in their intensity and distinctness to express various emotions.

Several multimodal approaches have been proposed, using facial, speech and body feature level fusion [84], using bi-modal of face and upper-body gestures [97] and by using face and body to automatically detect temporal segments or phases [98]. Results on multimodal approaches shows that explicit detection of the temporal phases can improve the accuracy of affect recognition. They also shows that the recognition from fused face and body modalities performs better than that from the face or the body

Study	*Type*
Joshi et al. [118]	upper body expressions and gestures
Griffin et al. [93]	laughter states (hilarious, social, awkward, fake, and non-laughter)
Kleinsmith et al. [129]	body postures
Scherer et al. [204]	head pose, eye gaze, facial expressions (smiles), hands and legs
Aung et al. [11]	body movements, EMG for paraspinal and trapezius muscles
Savva et al. [202]	body movements

Table 5: Studies on naturalistic body datasets

modality alone and synchronized feature-level fusion achieves better performance than decision-level fusion.

In cognitive and affective neuroscience, research on emotional body language is rapidly emerging as a new field. The involvement of the amygdala in emotional behavior has been known for some time [231]. A study measured dance-like body movements and biological movement patterns [27]. The patterns which were experienced as pleasant activated subcortical structures including the amygdala. Visual perception of biological motion activates two areas in occipital and fusiform cortex [95]. This indicates that areas that are known for processing faces are also involved in processing larger properties associated with human bodies. Hadjikhani and Gelder [99] used high-field fMRI and showed that exposure to body expressions of fear, as opposed to neutral body postures, activates the fusiform gyrus and the amygdala. The fact that these two areas have previously been associated with facial expressions suggest synergies between facial and body emotional expressions. Recent multimodal deep learning techniques has been proposed as described in the next section.

2.12 Bodily Emotion recognition using Deep Learning techniques

Current efforts as described in previous chapters, despite some improvement of accuracy, have relied on handcrafted features and classification techniques. The use of deep learning techniques to automatically extract effective features from multimodal information and classifications are new directions currently actively pursued by researchers, but several challenges remain in realising an end-to-end deep learning system. Deep learning is defined as a class of neural network based machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation or pattern analysis and classification [57]. With the availability of large datasets, deep learning has become a state-of-the art solution to problems such as emotion recognition. In [126] the authors use a CNN-based model for a hierarchical feature representation in the audio-visual domain to recognise spontaneous emotions. They showed that improvement of recognition accuracy is achieved when hierarchical features and multimodal information are adopted. In another effort, models are constructed from multiple physiological signals collected from sensors placed on the human body by adopting multimodal deep learning approach so as to improve their performance and reduce the cost of acquiring physiological signals for real world applications [150]. To classify spontaneous multimodal emotional expressions as positive or negative, researchers proposed a cross channel convolutional neural network (CCCNN) having the capability of learning and extracting general and specific features of emotions relying on body motion and face expression [17].These features were further passed through to cross-convolution channels to build the cross-modal feature representation. The Convolutional Neural Network (CNN) is a type of deep learning that is especially used in the processing of images, proposed by Lecun et al. [140] It is based on the foundation of conventional neural networks inspired by biological understanding of visual cortex. Figure 5 represents architecture of LeNet-5, one of the first initial architectures of CNN.

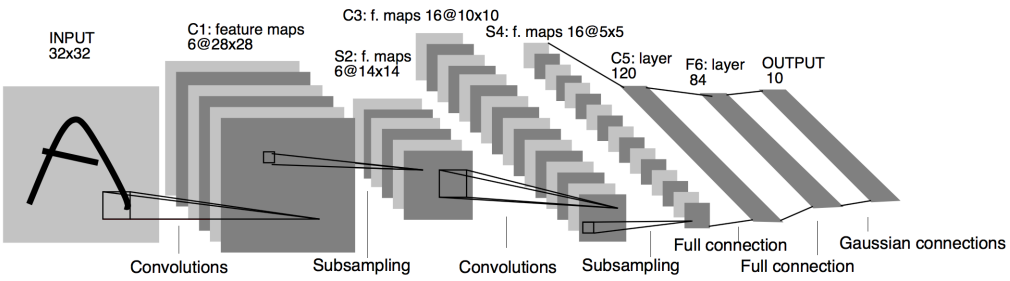


Figure 5: Architecture of LeNet-5, one of the first initial architectures of CNN [140]

The network as shown in Fig 5 applied convolution and sub-sampling alternatively to the input data, in the convolutional layers and sub-sampling layers. After two stages of this computation, the data is fed to a fully connected conventional neural networks, to complete the classification problem. Deep leaning based algorithms can be used for feature extraction and classification. With the use of CNNs the work spent on the pre-processing of the images is greatly reduced since the algorithm is already capable of detecting the best features needed to classify the images.

Because CNN based methods cannot reflect temporal variations, recently researchers have combined CNN, for the spatial features of single frames, with RNN networks that allow operation directly on time sequences. They are successfully applied to tasks involving temporal data such us speech recognition, language modelling, translation and gesture analysis. In RNNs, the output of the previous sequence time step is taken into consideration when calculating the result of the next one. However, a standard RNN does not handle long term dependencies well, due to the vanishing gradient problem. [107]. The Long Short Term Memory network (RNN-LSTM) is an extension for RNN, which works much better than the standard version. In RNN-LSTM architecture, RNN uses gateway units in addition to the common activation function, which extend its memory [12]. Such an architecture allows the network to learn and remember dependencies over more time steps, linking causes and effects remotely [106]. As seen in Figure 6 the above networks were used to identify gestures emotion recognition based on low level features inferred from the spacial location and orientation of joints within a track skeleton. [200]

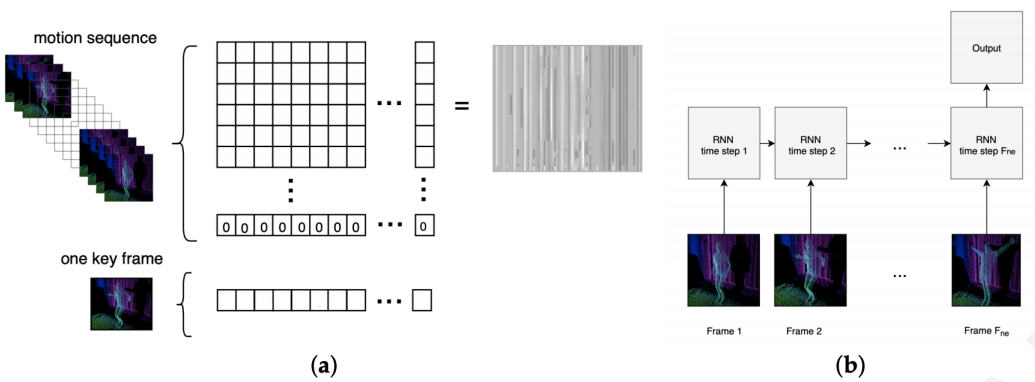


Figure 6: (a) The process of using a Convolutional Neural Network (CNN) for gestures-based emotion recognition shows the process of creating an matrices based on motion sequence. (b) The process of using a Recurrent Neural Network (RNN) for motion sequence analysis; each step of the motion sequence is evaluated by a RNN

Hierarchical Learning was used [65] for learning actions from body motion patterns of acted video recordings, using Grow When Required (GWR) networks. The separate processing of pose and motion features and their subsequent integration has been shown to improve the topological formation of visual representations in a hierarchical learning scheme. Due to an increased dimensionality of the neural weights along the hierarchy and concatenations of neural activation from previous layers, a recurrent variant called Gamma-GWR [182] was used that equips each neurone in the network with a temporal context as seen in Figure 7. Neural learning architecture with a hierarchy of self-organizing networks. The first layer processes separately pose and motion features from individual frames, whereas in the second a recurrent network learns the spatio-temporal structure of the joint pose-motion representations. For all the above deep learning approaches, a vast amount of data is needed to perform the training and learning. Unfortunately, the annotation of the motion data sets is a very laborious procedure, due to the fact that annotators need to perform a detailed analysis in every frame of a motion clip, to identify the exact frames that they emotion is depict. A solution to this problem are data augmentation techniques.

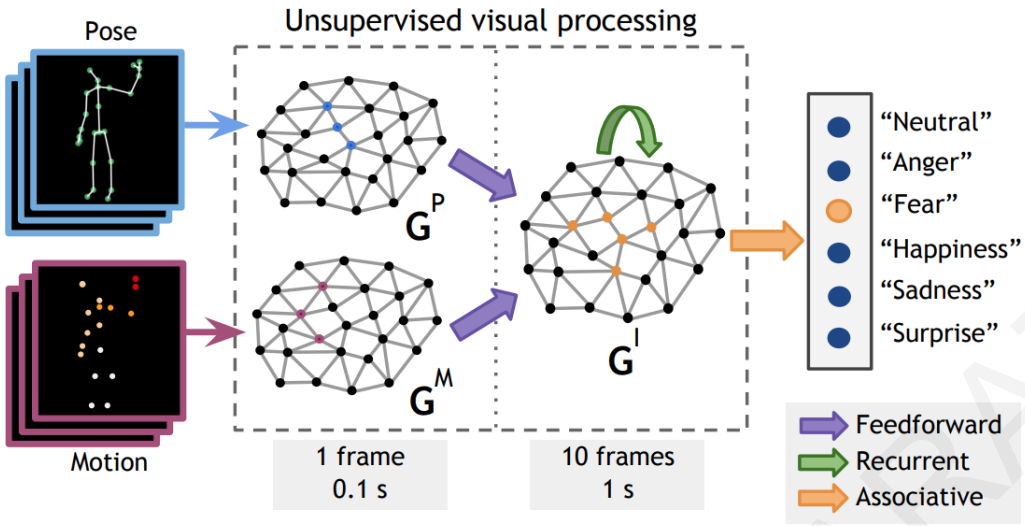


Figure 7: Neural learning architecture with a hierarchy of self-organizing networks

2.12.1 Data Augmentation

Obtaining an appropriate data set for training a desired Machine Learning model is a challenge by itself. Developers often have access to limited samples of data that are not sufficient to train and optimize a Deep Learning model, resulting in most of the time poor performance in terms of accuracy, precision, and recall. Moreover, the available training data for classification problems may also be imbalanced, causing issues to the model. Data Augmentation is a process that encompasses a suite of techniques which enhance the size and quality of training datasets leading to significant improvements in the performance parameters of Machine Learning models as well as providing developers with the opportunity to further iterate their model implementation. Data Augmentation has been extensively used with imaging data and computer vision tasks, involving image augmentation algorithms such us geometric transformations, color space augmentations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, generative adversarial networks, neural style transfer, and meta-learning [209]. While data augmentation is an established technique with image data sets employed widely in industry settings, there are numerous examples in the literature on possible ways to

also augment skeleton motion data for Deep Learning training purposes. One way to approach this is through the concept of time series data and the ways to perform data augmentation on this format [232]. In an example that uses time series 3D body action recognition, Dawar et al. [55] in a project that fuses depth images with signals captured by an inertia wearable sensor, proposed the simulation of various orientations of the depth camera and the sensor placement. Molchanov et al. [172] employed the popular methods of augmentation of video data by means of affine image transformations (rotation, scaling, mirroring). Additionally, they exploited the techniques of changing the sequence order of individual video frames (reverse ordering, mirroring, etc.). Another data augmentation technique proposed downsampling based on local averaging combined with data shuffling to allow further variations and avoid overfitting in wearable inertial measurement unit sensors [71]. A biomechanical-based approach to data augmentation for gesture recognition proposed populating human-like examples from a set of naturalistic features extracted from a single gesture sample while preserving human traits such as visual saliency and smooth transition [158]. Research on action recognition from motion and posture data has also benefited from data augmentation techniques. A Long Short-Term Memory Auto-Encoder was used to achieve effective spatial-temporal data augmentation for skeleton-based human action recognition. This technique outperformed traditional methods of scale and rotation [218]. In a study that involved static human skeleton data (postures) an image-based synthesis engine that uses annotated pictures and 3d skeleton postures to generate synthetic images with similar postures depicting the same action [196]. Regarding postures, another method augments the input data with rotation augmentation, and use pose estimation method multiple times for every frame, and select the most consistent pose, followed by a motion reconstruction for smoothing [132]. In a wider setting but inclusive of skeletal action recognition, new time-series classes can be obtained in warped space between sub optimally aligned input examples of different lengths to enrich the training dataset [136]. Another technique applies Dynamic Time Warping to preserve the relationships between neighboring elements in the warping. The results

are further improved by using the most discriminative sample in a batch as a teacher (reference sample) [116]. Oversampling minority classes is a common data augmentation technique and can range from changing the weight of the minority class, randomly oversampling (duplicating) the samples, to more sophisticated techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) [43] or Adaptive Synthetic (ADASYN) [100]. SMOTE interpolates between samples to generate new ones, while ADASYN is similar but applies an additional limitation to the number of new samples based on the number of nearby opposite samples [43], [136], [100]. A similar technique proposed averaging motion trajectories to generate new N!(N-1)!/2 artificial gestures from a database of size N [203]. The comparison results show significant improvement following the proposed augmentation technique.

A different approach is been used by Li et al [146], which the skeletal structure recorded by depth cameras, was converted to represent pseudo images. Subsequently, they performed image augmentation (standard affine transformations). Figure 8 displays a three dimensional action recognition approach using deep convolutional neural networks and data augmentation technique from images [111]. Given a skeleton sequence (a) extracted pose features (b-1) are transformed into a color image I (b-2) by an encoding technique, called PoF2I. The action image I is refined by a mechanism of adding or eliminating randomly selected skeleton frames. Manifold action images J (b-3) are generated from I for training set augmentation. These action images J are finally fed into deep CNNs (c) for action recognition.

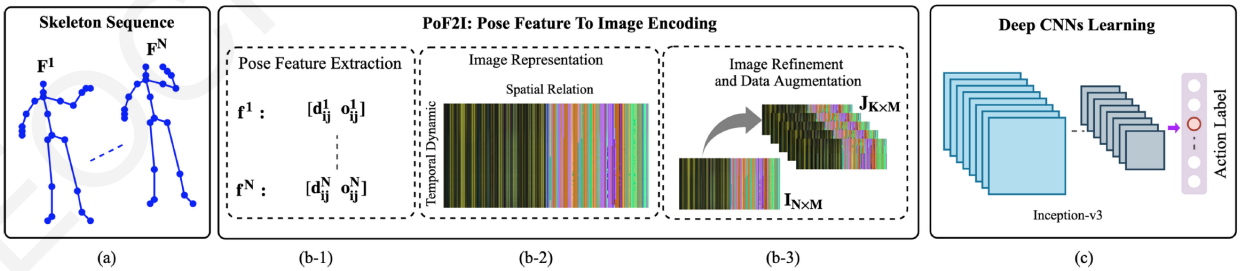


Figure 8: Schematic overview of a method of 3-D action recognition.

Generative adversarial networks (GANs) can be used to generate images from an adversarial training [92]. The generator attempts to produce a realistic image to fool the discriminator, which tries to distinguish whether it's input image is from the training set or the generated set. Generative adversarial nets are now widely used in many image tasks such as single image super-resolution [141], image manipulation [243], synthesis [58] and image to image translation [115]. Zhu et. al [242] proposed CycleGAN a network that can do image to image transition between two unpaired image domain. Researchers used CycleGAN network to generate labeled emotion images and show that these images are helpful in final image classification task. In Figure 9 shows a framework of GAN-based data augmentation. Both reference images and target images are collected from the original data and flow into the CycleGAN as domains R and T, respectively. G and F are two generators, transferring R & T and T & R, respectively. Supplementary data is generated through generator G. A CNN classifier is trained using original data and supplementary data as input.

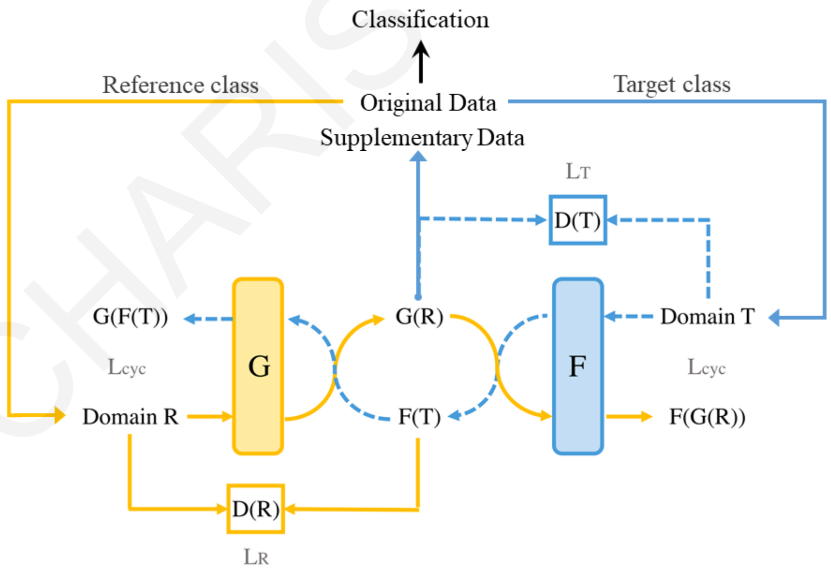


Figure 9: An illustration of the CycleGAN framework for data augmentation and classification using a CNN classifier.

2.12.2 Transfer Learning

Data mining and machine learning technologies have already achieved significant success in many knowledge engineering areas including classification, regression and clustering. However, many machine learning methods work well only under a common assumption: the training and test data are drawn from the same feature space and the same distribution. When the distribution changes, most statistical models need to be rebuilt from scratch using newly collected training data. In many real world applications, it is expensive or impossible to re-collect the needed training data and rebuild the models. It would be nice to reduce the need and effort to re-collect the training data. In such cases, *knowledge transfer* or *transfer learning* between task domains would be desirable. Transfer Learning partially resolves the limitations of the isolated learning paradigm: "The current dominant paradigm for ML is to run an ML algorithm on a given dataset to generate a model. The model is then applied in real-life tasks. We call this paradigm isolated learning because it does not take into account any other related information or any of the knowledge learned in the past [149]. Transfer Learning gives us the ability to share learned features across different learning tasks. Emerging transfer learning methods can leverage the knowledge from one emotion-related domain to another. The main premise behind such techniques is that people may share similar characteristics when expressing a given emotion. For example, anger may result in increased speech loudness and more intense facial expressions [226]. Fear is usually expressed with reduced speech volume and may produce increased heart rate [225]. These emotion-specific characteristics might be commonly met among people, contributing to the similarity among the various emotional datasets. Therefore, transfer learning approaches can learn common emotion-specific patterns and can be applied across domains for recognizing emotions in datasets with scarce or non-labeled samples. Such techniques can further result in generalizable systems, which can detect emotion for unseen data. Transfer learning in automatic emotion recognition has been used in speech based systems, and image or video based systems using facial expressions in which the state-of-art transfer learning approach to

the video-based emotion recognition includes obtaining high-level features using mainly a convolutional neural network (CNN) trained on large sources of data [8] [123] [174] [211] or transfering the knowledge from higher-quality auxiliary image datasets [234].

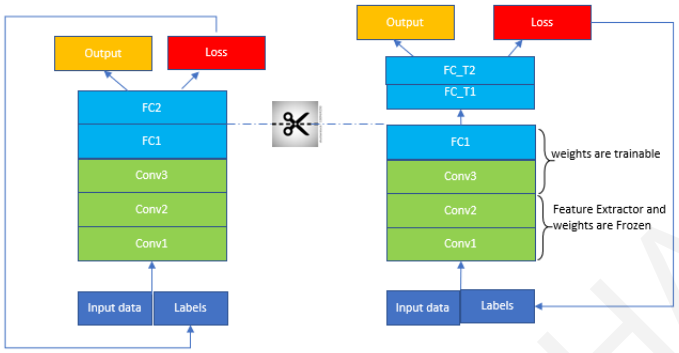


Figure 10: Transfer Learning of ResNet50 model

Very Deep Convolutional Networks for Large-Scale Image Recognition(VGG16) pre-train model [211] is a convolutional neural network trained on 1.2 million images to classify 1000 different categories. Its pre-trained architecture can detect generic visual features present on our emotion dataset.

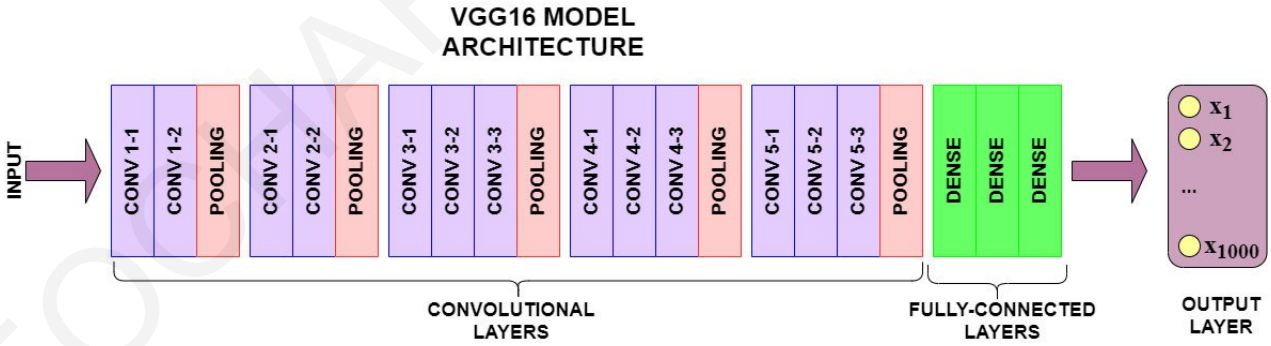


Figure 11: The VGG16 Model has 16 Convolutional and Max Pooling layers, 3 Dense layers for the Fully-Connected layer, and an output layer of 1,000 nodes

Another relevant model is the ResNet model. The main motivation behind this model was to avoid poor accuracy as the model went on to become deeper. Additionally the ResNet model is aimed to tackle the Vanishing Gradient issue. Other available deep learning models are made available alognside pre-trained weights. These models can be used for prediction, feature extraction and fine tuning. The available models are depict in the following table:

Model	Parameters	Accuracy	Time per inference step (GPU)
Xception	22,910,480	0.945	8.06ms
VGG16	138,357,544	0.901	4.16ms
VGG19	143,667,240	0.900	4.38ms
ResNet50	25,636,712	0.921	4.55ms
ResNet101	44,707,176	0.928	5.19ms
ResNet152	60,419,944	0.931	6.54ms
InceptionV3	23,851,784	0.937	6.86ms

Figure 12: Pre-train models for transfer learning

Chapter 3

Emotion recognition on Postures during active game playing

3.1 Introduction

The affective state of a player during game playing has a significant effect on his/her motivation and engagement. Often players lose their interest and stop playing a game due to negative emotions such as frustration, anger, and many more. On the other hand, players who experience positive emotions during game playing are more likely to continue playing the specific game. Therefore, a system that recognises player's emotions during game playing can be a useful tool for designing games that can receive affective feedback from the player and respond to this through sophisticated artificial intelligence behaviours that can be applied to the game characters and the game itself.

A system was designed to recognise emotions using posture data. The objectives for the research were to (a) investigate how to construct a database of postures labelled with emotions. (b) Record data from both single and multiplayer competitive games to capture the rich expressiveness of both game scenarios (c) Utilise state of the art non-intrusive interfaces such as Microsoft Kinect [171] to capture data and provide a system that can be used in today 's games. This is particularly important as commercial systems are increasingly adopting such interfaces in contrast to traditional motion capture systems that are expensive, and have unrealistic space requirements so cannot be supported. Taking into

consideration the limitation of movements and postures that an interface like Kinect can recognise due to its simple setup, it is interesting to see whether it performs similarly to other motion capture systems in terms of recognition accuracy.

As previously illustrated in table 5 there has been research in the field of emotion recognition using various body modalities. One approach uses cyclic arm movements to recognise fundamental emotions [21] and another investigates recognition accuracy, confusions, and viewpoint difference while attributing emotions to postures [49]. In another study, automatic recognition models grounded on low-level posture descriptions were built and tested for their ability to generalize to new observers and postures [127] [129]. The automatic models achieve recognition percentages comparable to the human base rates. A different approach uses movement qualities such as amplitude, speed and fluidity of movement to infer emotions [40]. Similar to this, researchers tried to recognise emotion from animation rather than posture [201]. In this approach, human recognition of emotion is taken through observation of skeletal information only. Our approach used, both video and skeleton data is given to the human observers during labelling, as during experimentation, a number of postures could not be predicted using skeleton data only. Moreover, all the above approaches use body posture or animation data captured from traditional motion capture equipment which is not as ubiquitous as Kinect. However, as Kinect is not as accurate as traditional motion capture devices, it is uncertain if it is practical to capture training data for the recognition system using Kinect.

3.2 Methodology

3.2.1 Data Collection

The Kinect SDK skeleton data was used to capture and store the posture of players in each frame. Six male players were asked to play games with the Xbox integrated with a Kinect. A second Kinect

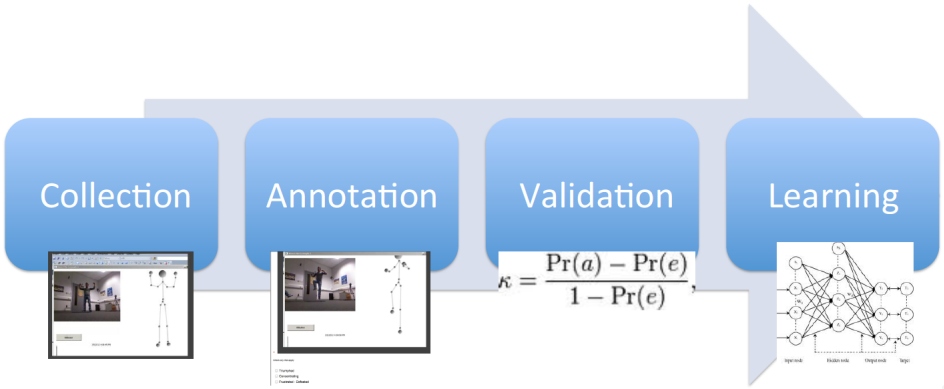


Figure 13: Experiment Overview

was connected to a PC and was used separately to record the motion data of the players. The software used for the motion capture also had the capacity to capture and display both video and skeleton data in real-time. The PC screen was captured to provide replay capacity for the extraction of the apex poses. In contrast with the previous research [127] skeleton and emotion postures were recorded during both actual game play sessions and replay sessions, as it may not be sufficient for an emotion recognition system to be trained with different data than the one it is aimed to recognise.

3.2.2 Data Analysis

Two different male students replayed the captured video to determine and select affective postures for four emotional states: Triumph, Concentrating, Frustrated and Defeated. The selection of these emotional states has been decided from previous research [127], describing some basic emotions that are valuable for game-play scenarios. Although we have used the same emotions in order to be able also to compare and test against prior research, the system can be used with any emotional representations, that ground truth has been collected.

The students asseed the videos and expressed a concern on selecting a posture between frustrated and defeated, as they considered these emotional states similar in many circumstances. For this reason, the emotion label set was reduced to three (triumph, concentrating, frustrated/defeated). The students

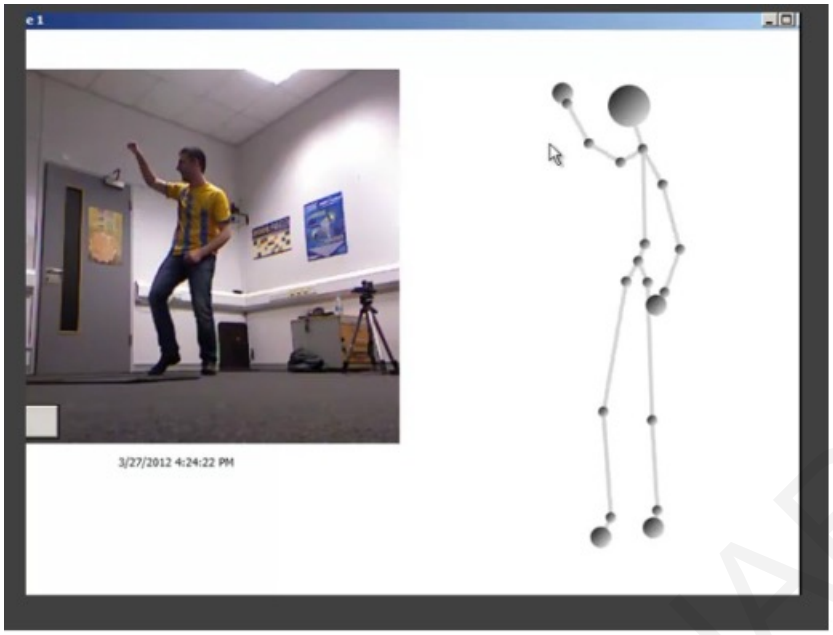


Figure 14: Example on skeleton capturing using human observers

were given the option to identify emotional states based on both video and skeleton data. Both students agreed that the actual video made it easier for them to decide on the emotional state in many instances where the skeleton data was not helpful. A total of 147 postures were extracted using this technique. For all the postures that were extracted, a screenshot of the corresponding frame of the software was taken for each posture. A digital questionnaire was created and four different observers, three male and one female, were asked to annotate the screenshots 14 with an emotion label of one of the three above mentioned emotional states.

Table 6 presents observer agreement for all possible pairs among the four observers, measured based on Cohen's Kappa value (Berry 1998). As can be seen, agreement strength is above or equal to 'good'at all cases, demonstrating that the labelling quality is satisfactory. Observer 2 clearly performs better on all paired cases. The average Cohen's Kappa value for observer 2 on all three pairs is 0.812.

Observer pair	*Kappa*	*SE of Kappa*	*95% confidence interval*	*agreement strength*
1-2	0.806	0.043	0.721 - 0.890	Very good
1-3	0.739	0.049	0.643 - 0.835	Good
1-4	0.749	0.047	0.656 - 0.842	Good
2-3	0.782	0.045	0.694 - 0.871	Good
2-4	0.849	0.038	0.775 - 0.923	Very good
3-4	0.686	0.052	0.585 - 0.787	Good

Table 6: Paired observer agreement strength

Table 7 lists the actual number of agreements and percentage of agreement for all possible pairs that include observer 2. It also compares against chance level agreement. The score is significantly higher than the chance level agreement and above 86% on all pairs, making observer 2 a suitable candidate for the labelling of postures. Considering this and the Kappa value, the labelling from observer 2 was used for posture data annotation for the different emotion recognition tests.

Observer IDs	*Observer agreements*	*% of agreement*	*Chance Agreements*	*% of expected agreement*
1 and 2	129	87.76	54.41	37.04
2 and 3	127	86.39	55.1	37.5
2 and 4	133	90.48	54	36.76

Table 7: Agreement score for Observer 2

Finally, the overall agreement of all observers was far above chance level at 72.3%, agreeing on 107 out of the 147 postures. This is used as the benchmark for evaluating the performance of the collected skeleton data postures as input for the emotion recognition system. Figure 15 presents the distribution of labels across the 147 postures according to observer 2.

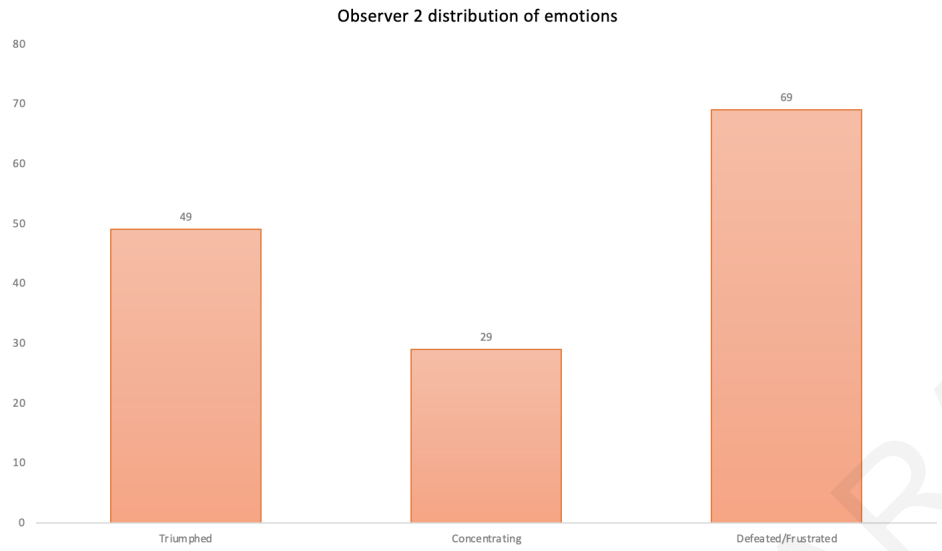


Figure 15: Distribution of emotion labels for the 147 postures, according to observer 2

3.2.3 Emotion Recognition Method and Results

After the observer performance was completed, the skeleton data for each labelled posture was extracted and annotated with the corresponding emotion label taken from the annotation set by observer 2. The skeleton data comprises standard 3d rotational information for each of the joints of the body. Automatic emotion recognition was tested based on the capacity to recognise new postures. The labelled data was used in WEKA [101] to build and test a model. Using supervised learning, back-propagation algorithm with 10 fold cross-validation we have tried to generalised based on new postures and based on new players.

The model successfully recognized 90 out of the 147 postures, resulting in overall recognition rate of 61.22%, which is approximately double of what might be expected from chance. Figure 16 presents the confusion matrix for the conducted test. The recognition rates are balanced across all three emotional labels, with defeated/frustrated slightly higher than the other two labels. It can be seen that as the number of postures increases, the recognition rate improves. It is possible that a more balanced database of sample can result into more balanced and improved recognition rates.

Back Propagation Algorithm		Predicted Concentrated	Defeated/ Frustrated	Triumph
	Concentrated	16 (55%)	8	5
Actual	Defeated/ Frustrated	6	45(65%)	18
	Triumph	5	15	29(59%)

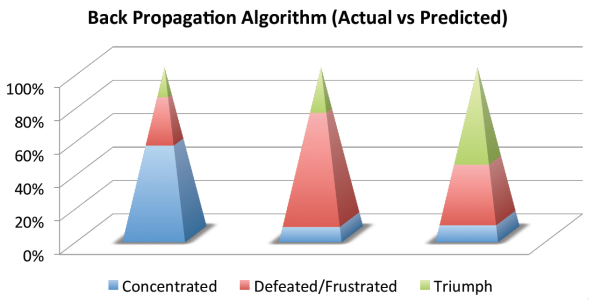


Figure 16: Recognition rates for new postures using the back-propagation algorithm. Recognized successfully 90 out of 147 postures, 61.22% recognition rate

Another recognition test evaluated the capacity to recognise emotions on new players. The data was separate into two sets (a) the first set comprises 108 postures that was captured by four players (b) the second set includes 39 postures captured by two separate players. The first set was used to train the back-propagation algorithm (Phansalkar, 1994), while the second one was used as test data. Figure 17 presents the confusion matrix for the conducted test. The correctly classified postures were 22 out of the 39, resulting in 56.4% overall recognition rate. Recognition rate is almost the same for triumph and defeated/frustrated labels, but is significantly smaller for concentrating (42%). However, this is expected due to the small amount of training data for the new concentrating label, which was reduced to 17, as 12 of the 29 are now part of the test data. In general, it appears that the database again performs above chance level and can generalise to new players, although the training data sample is small.

Back Propagation Algorithm		Predicted Concentrated	Defeated/ Frustrated	Triumph
	Concentrated	5 (42%)	6	1
Actual	Defeated/ Frustrated	1	10(62%)	5
	Triumph	0	4	7(64%)

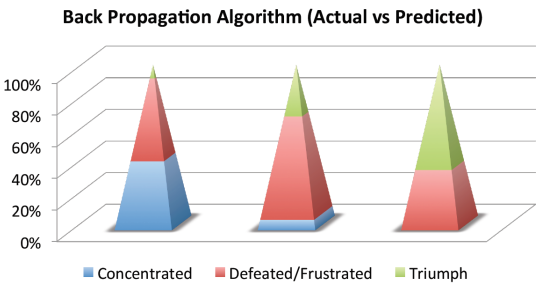


Figure 17: Recognition rates for new players using the back-propagation algorithm. Recognized successfully 22 out of 39 postures, 56.4% recognition rate

The above results in recognising emotions have based on posture data captured using Microsoft's Kinect and that time there was no previous research that used non-acted data captured from Kinect to recognise emotions. The results are similar to those in existing literature that use traditional motion capture equipment. This suggests that Kinect provides sufficiently valid data to construct such a model, which can be used in today's games. Improved recognition of concentrating labels compared to researchers [127] above and can be attributed to smaller emotion set (three instead of four) or to better observer annotation due to the provision of camera input apart from skeleton posture data. Compared to the benchmark rate taken from the observers in the current study, the overall recognition was 11.08% lower. This again, is probably due to the fact that observers had the advantage of camera input, while the algorithm uses only skeleton data. It would be interesting to test whether observer?s agreement is influenced by this, as in preivous research [127] agreement is lower than in our experiment. This could test and indicate the superiority that a multimodal system may have against single modalities such as posture or animation data.

3.3 Conclusion

The recognition system described above yielded satisfactory results showing that images with postures can be used as a tool for automatic recognition where skeleton tracking is possible. However, i believe that it can be further optimised by using mirror postures. Moreover, the training data size can be adjusted to improve performance since the current data was collected using specific sports games. It would be interesting to investigate if the constructed database can recognise emotions captured during game playing of different game genres and to compare the benchmark of acted, non-acted and hybrid databases.

Chapter 4

Emotion Recognition for Exergames using Laban Movement Analysis

4.1 Introduction

New advances in non-intrusive user interfaces that use natural human gestures as input have resulted in high popularity of a new game genre called Exergaming. Exergames go beyond the passive gameplay activity that traditional controllers such as gamepads, keyboard and mouse offer, and require game players to become physically active. Through this, exergames are often used to promote a healthy lifestyle for both casual gamers that use such interfaces at home but also for special categories of users who need to advance their physical activity in order to improve specific health conditions [86] [181]. Further to this, exergames provide a novel and livelier game experience that can also augment the fun factor, however research in this area is still in the early stages [124] [114] [113].

A major issue of the available exergames is that they do not have the capacity to detect whether the players are really enjoying the game-playing. The games are not intelligent enough to detect significant emotional states and adapt according to them in order to offer a better user experience for the players. While facial and audio information have been used successfully to detect emotions on users of desktop applications [91] [34] [72] [222], exergame players express their emotions using their bodies as these modalities are more active and energetic during exergaming. Existing research that attempts to recognize

emotions using human motion data does not achieve sufficient recognition rates, and is based on training the system with low level feature data that is very vague (such as rotation of a given joint on a given axes etc) and is selected without firm justification from movement analysis theories. Some recent studies in robotics do achieve good quality recognition [161] [162], however their task is more simplified since robots perform mechanic and predetermined movements while expressive human movement is more complex and non-deterministic. Therefore it is not clear how applicable these methods are to real game playing situations.

This contribution is a step towards overcoming the above limitations, by providing a novel method that achieves high recognition rates using real human motion data, captured during genuine game playing. It presents an emotion recognition model that makes use of human motion data dynamics derived from the widely accepted and applied movement analysis theories of Laban [137]. The features that are used to describe the emotional state vector are derived from the theories of Laban on Effort movement qualities. Four different game-playing related emotional states (excitement, frustration, meditation and concentration) are studied and training features extracted so that they can distinguish either single emotions or subsets of the above mentioned emotion.

4.2 Methodology

Many studies have been carried out for motion analysis by using Laban theory's [44], [176] and [199]. Camurri et.al examined emotion in dance [36]. The results ranged between 31% and 46% for recognizing four emotions, far less than the observer recognition rate of 56%. Lourens [155] extracted low level features from video and used Labanotation experts to classify the video clips to four emotional states manually. Another study used Laban features like whole-body movement, inclination of the body and area, to extract four emotions, pleasure, anger, sadness and relaxation from a robot that has limited ways of movement [162]. Although they used observers to classify the robot movements to emotions,

they have not used automatic recognition techniques for classification. They used empirical estimation of correlation between Laban features and emotional set. In my approach i used human motion capture data to extract some of the Laban features for the body's extremity parts such as arms, legs and head and then perform automatic recognition techniques on those features to classify our emotional set: excitement, frustration, concentration and meditation. For my experiment, 'meditation'is the mental state during which users ignore the environment and focus on themselves. An example of this, a breathing moment, stretching and any other movement which draws the user's attention to his own body.

4.2.1 Laban Movement Analysis

Laban Movement Analysis (LMA) is a theory for observing, describing, notating and interpreting human motion. It was originally developed by dance artist and theorist Rudolf Laban in the early 20th century. The method focuses on the relationships between internal state, intention and attention and their effects on all human motions. One of the strong points of LMA is the ability to describe expressive content of movements, which makes it excellent for emotion and behavior analysis. Many researchers have been trying to create a computational form of LMA for motion analysis [13] [241] [240]. Nakata [173], reproduced expressive movements in a robot that could be interpreted as emotions by a human observer.

Theory divides LMA in four components shown in Figure 18. The experiment focuses on the Effort component that deals with the expressiveness and describes the dynamic qualities of the movement and the inner attitude towards using energy. By selecting a set of suitable features from the trajectories described by hands, foot and head, the effort component can be used as one descriptor for expressive movements. Laban sees Effort as the inner impulse-a movement sensation, a thought, a feeling or emotion- from which movement originates; it constitutes the interface between mental and physical components of movement. The inner impulse is expressed by way of Motion Factors. Every human

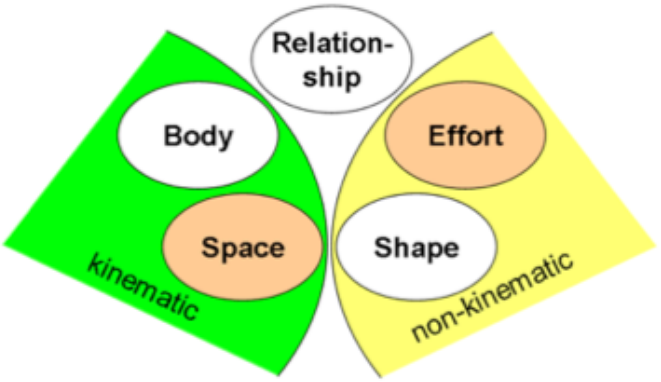


Figure 18: Four major Components of Laban Movement analysis. Adopted by ZHAO 2002

movement including thought has potential to engage the four motion factors: Space, Weight, Time and Flow. Table 8 shows the motion factors, the underlying cognitive process associated with and the bipolar quality between two extremes of Effort component.

Motion Factor	*Cognitive process*	*Extremes*
Space	Attention-Thinking	Indirect-Direct
Weight	Intention-Sensing	Light-Strong
Time	Decision-Intuiting	Sustained-Sudden
Flow	Progression-Feeling	Free-Bound

Table 8: Effort motion factors

Space Motion Factor

As observed by Maletic [156], motion factors have correlation with cognitive processes. The emphasis on attitudes toward Space can be associated with the cognitive capacities of orienting, attending and organizing. It addresses the quality of active attention to the surroundings. The two Extremes are Direct (Concentrate, Focused, pinpointing, narrowing down) and Indirect (multi-focused, with all-round attention).

Weight Motion Factor

The predominance of Weight qualities may indicate sensing or sensibility for assuming light or firm Intention towards an action. It senses the physical mass and its relationship with gravity. The two Extremes are Light (Accepting or Adjusting to gravity, delicate, lesser muscular tension) and Strong (resisting the pull of gravity, firm, forcefull).

Time Motion Factor

A great frequency of Time qualities may indicate an intuitive readiness for Decision making. Its mastery gives a calm or alert approach to thought or movement actions. The two Extremes are Sustained (Calm, slow tempo of movement) and Sudden (Excited, immediate, unexpected).

Flow Motion Factor

The emphasis on Flow can be associated with the emergence of feelings that depending on the interaction with self or others, free or bind the continuity of movement and give either a controlled and careful or exuberant and outgoing Progression. The two Extremes are Free (Accepting the continuity of movement, go with the flow) and Bound (Resisting the flux of movement, controlled, restrained).

4.2.2 Data collection and processing

Thirteen players (ten male and three female) were asked to play sports games for 30 minutes each on the Xbox integrated with the Microsoft Kinect [171]. The motion data was collected using a PhaseSpace Impulse X2 motion tracking system with 8 cameras. A camera was also used to record all the sessions on video, in order to aid at a later stage the annotation of emotional states. The data annotation was done in a two-step process. First motion clips (of size no longer than 2 seconds) that potentially exhibit one of the 4 investigated emotions were extracted manually. Special care was given not to include frames at the beginning and end of the clip that are not significantly expressive in order to reduce noise in the learning process. A total of 309 clips were extracted. In the second step, four different observers through

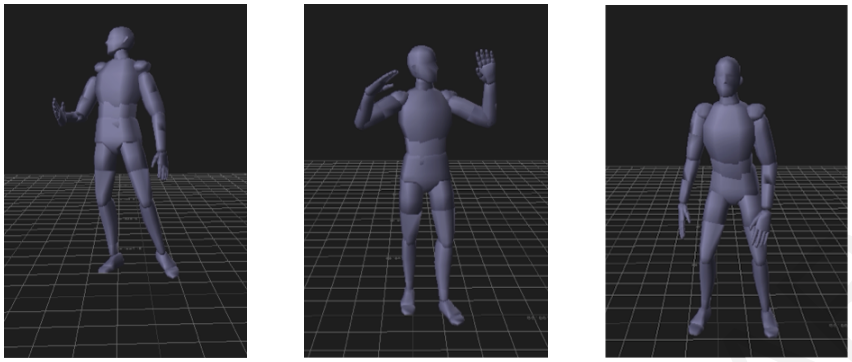
a multiple-choice questionnaire annotated each of the extracted clips, resulting on an agreement on 197 clips that became the ground truth for our system.

4.2.3 Feature Analysis

In the current implementation, the Space and Time motion factors of the Effort component were implemented. According to Laban [137], the Space motion factor represents the person's attention to the surroundings. It is related to attention and thinking. Indirect Space is multi-focused with all-around attention, while direct is focused with a tendency to align joints and bend. Laban states that concentrated behavior has direct space quality. Through observation of motion data it is easy to see that excitement and frustration are not focused movements, while meditation is a state of focusing on one's whole body rather on a single point. In the current study, Space motion factor is used to try to recognize concentrate emotional states from the other three emotional states. Space motion factor is implemented similar to Masuda [161], but taking into consideration the above theories of Laban.

Through experimentation we have used and discarded multiple features like Quaternion Velocity and Acceleration, Torsion, Corner Curvature, Angular Displacement, Angular Velocity and Acceleration, Swivel Angles, Sternum Height [240]. These features have been tested and received lower recognition rates than the features proposed below. In particular quaternions are very good in representing rotations as compared to Euler angle or matrix representations and eliminating gimbal lock. In our case recognition using quaternion velocity and acceleration has resulted in lower rates probably due to the need for a global coordinate system, thus positional information was used. Further more, other features such as Torsion and Corner Curvature, that are independent of the way the trajectory is traversed, showed that are higher in strong motions versus light motions. However, path curvatures of light motions are larger than in strong motions. These motions does not give much value to emotion recognition correlation as an individual feature and that is why was excluded from our experiments. On the contrary, swivel angles

show that Laban's Indirect and Free movements tend to be driven by the elbow, so swivel angle changes significantly during a movement.



(a) Frustration (b) Excitement (c) Concentration

Figure 19: Emotion recognition using Laban's features on the body's extremity parts

Although that we have tested various features derived from swivel angles, as discussed in the literature from Laban movements like a) swivel angle changing rate (velocity), b) total sum of swivel angle velocities, c) number of zero-crossings of the second derivative, d) total pendulum distances (swivel angle changes between all the neighbouring zero-crossings) and e) the difference between maximum and minimum swivel angles, for emotion recognition did not provide any higher recognition rates so we excluded them to reduce the input information. Furthermore by analysing the feet of the player we excluded left foot which was not giving any higher recognition rates versus using both feet. This maybe was to the type of the game (football) where most of the participants where kicking the ball with the right foot. This shows also that in a generalization of the algorithm, we need to take into consideration left and right handed/feet people separately into the training data.

Sternum height feature on the other hand is measured as the distance between the lowest and the highest point in a movement. Motion capture systems cannot directly measure muscular tension, so sternum height is used as an indirect indicator of muscular tension. Sternum height has been used for

Discarded features	
Quaternion Velocity	Angular Velocity
Quaternion Acceleration	Angular Acceleration
Torsion	Angular Displacement
Swivel Angles	Corner Curvature
Swivel angle changing rate (velocity).	Sum of swivel angle velocities
Swivel angle number of zero-crossings of the second derivative	Total pendulum distances (swivel angle changes between all the neighbouring zero-crossings)
The difference between maximum and minimum swivel angles	

Figure 20: Discarded features

discrimination of motion factors that are almost the same, such us sudden time and strong weight that are simultaneusouly active in the same movement. For this reason it can not be used as a stand alone feature for recognition. Nevertheless, we have inspired from Sternum height to create a new feature called percentage of narrowing down P_{ND} as depict below:

Percentage of narrowing down P_{ND} in the clip, is calculated as the difference of the initial Y position of the head minus the average head Y position of the clip, divided by the initial Y position. Through observation, it is easy to see that in concentration clips, the player tends to bend resulting in significantly lower head positions throughout the clip frames.

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

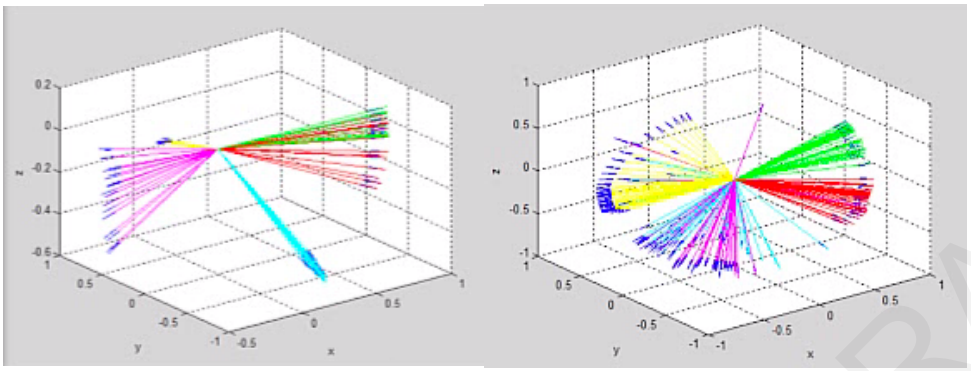
Further to the above feature, to highlight a prospective focus of the movement to a given point (direct behaviour) the face direction $\overrightarrow{F}$ and the unit movement vectors of the four extremity points of the skeleton are used.

$$S = \{\overrightarrow{L}_{hand}, \overrightarrow{R}_{hand}, \overrightarrow{L}_{foot}, \overrightarrow{R}_{foot}\}$$

The dot product of the face vector with each of the four extremity movement vectors are calculated at each frame of the clip.

$$\forall x \in S, F \cdot x \tag{2}$$

Their signs are tested to see if the angle of each pair of vectors (hands or foot) is above or below 90° (resulting in indirect or direct movement), as can be seen in Figure 21.



(a) Concentration (b) Excitement-Frustration-Mediation

Figure 21: Face Vector with extremity points

For each extremity point, the average of the dot product values of all the direct frames and the indirect frames are calculated as two separate features. At the end eight features are calculated for all four extremity points. Together with the P_{ND} feature, they form the Space feature vector, as seen in Table 9.

Feature	*Description*
P_{ND}	Percentage of narrow down
$DotL_{hand}Direct$	$(\overrightarrow{F} \cdot \overrightarrow{L}_{hand})$ for direct frames
$DotR_{hand}Direct$	$(\overrightarrow{F} \cdot \overrightarrow{R}_{hand})$ for direct frames
$DotL_{foot}Direct$	$(\overrightarrow{F} \cdot \overrightarrow{L}_{foot})$ for direct frames
$DotR_{foot}Direct$	$(\overrightarrow{F} \cdot \overrightarrow{R}_{foot})$ for direct frames
$DotL_{hand}inDirect$	$(\overrightarrow{F} \cdot \overrightarrow{L}_{hand})$ for indirect frames
$DotR_{hand}inDirect$	$(\overrightarrow{F} \cdot \overrightarrow{R}_{hand})$ for indirect frames
$DotL_{foot}inDirect$	$(\overrightarrow{F} \cdot \overrightarrow{L}_{foot})$ for indirect frames
$DotR_{foot}inDirect$	$(\overrightarrow{F} \cdot \overrightarrow{R}_{foot})$ for indirect frames

Table 9: The Space feature vector

The Time component represents the speed of the movement. According to Laban, it has to do with decision and intuition. Sustained movements are calm, with slow tempo, while sudden movements, are immediate, excited, unexpected and with fast tempo. Laban's theory about Time and emotions correlates:

(a) {meditation,concentration} $\in$ Sustained

(b) {frustration,excitement} $\in$ Sudden

Time is implemented using the positional velocity(υ), acceleration(α) and jerk(j) (acceleration derivative) for the extremities of the body, both hands and foot. In Figures 22,23 and 24 we have shown the average velocity, acceleration and jerk of each extremity joint across all the clips. It shows that for the left foot the variation is small and thus does not contribute much and can be omitted from the feature set.

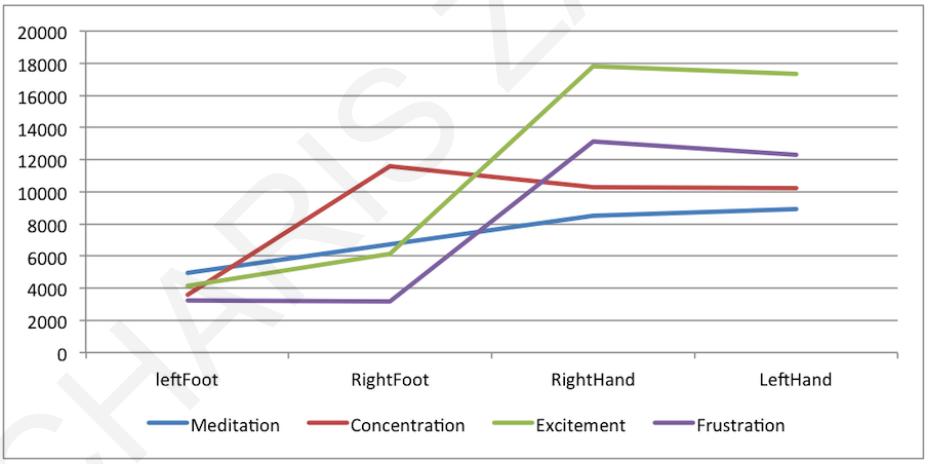


Figure 22: Average Velocity

The final feature set for the Time component comprises nine features, positional velocity, acceleration and jerk for left hand, right hand and right foot respectively as seen in Table 10 below.

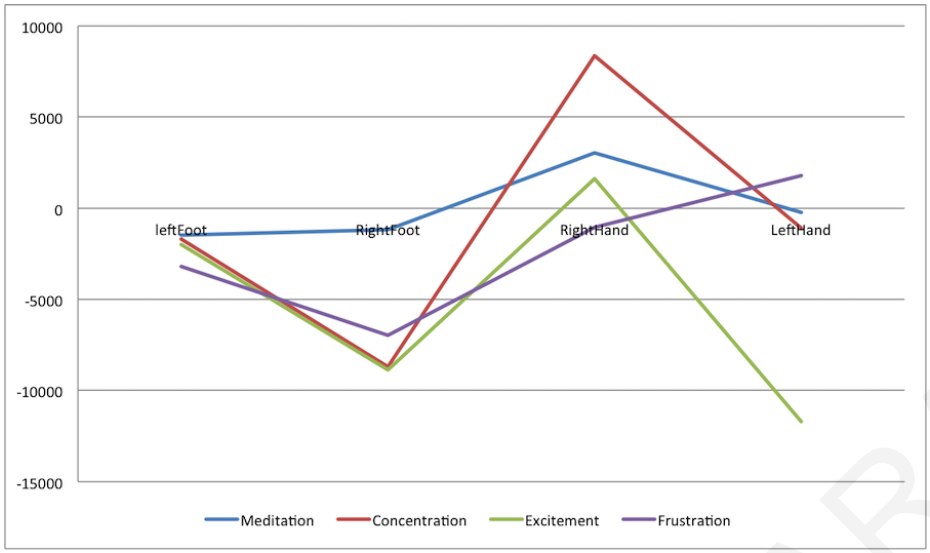


Figure 23: Average Acceleration

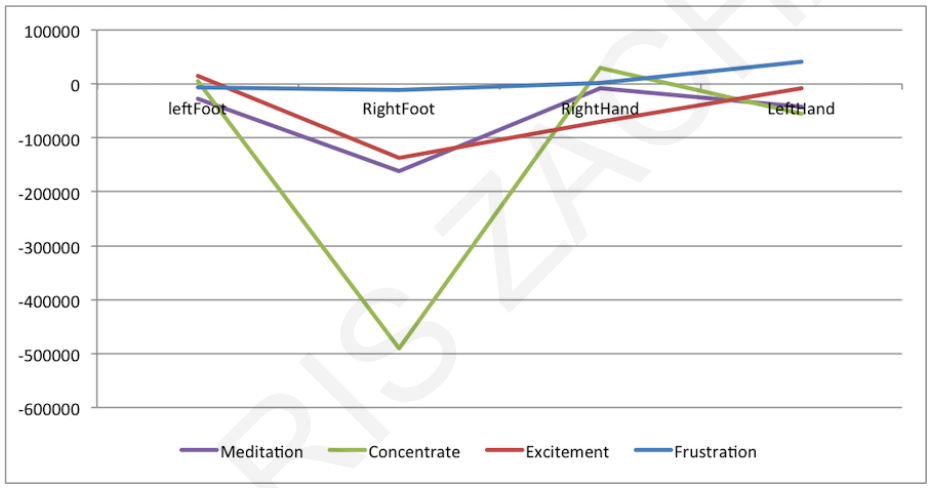


Figure 24: Average Jerk

4.2.4 Machine Learning Approach

To assess the validity of the selected feature sets for the Space and Time motion factors, and to measure the success in recognition of the targeted emotional states, the following tests were conducted:

Test1: Annotate all non-concentrate clips as one category and test to see if Space factor can distinguish between concentrate and non-concentrate clips. This can be used during automatic measurements of concentration on exergames in which acute cognitive benefits such as temporal improvements in concentration are being evaluated [83].

Feature	*Description*
$L_{hand}V$	Velocity(υ) for right hand
$R_{hand}V$	Velocity(υ) for left hand
$R_{foot}V$	Velocity(υ) for right foot
$L_{hand}A$	Acceleration(α) for right hand
$R_{hand}A$	Acceleration(α) for left hand
$R_{foot}A$	Acceleration(α) for right foot
$L_{hand}J$	Jerk(j) for right hand
$R_{hand}J$	Jerk(j) for left hand
$R_{foot}J$	Jerk(j) for right foot

Table 10: The Time feature vector

Test2: Annotate all excitement and frustration clips as the one category and all meditation and concentration clips as another and attempt to see how well the Time factor can recognize between the two categories. This can be used in a scenario where the valence of the emotion state of the user is required to be measured.

Test3: Attempt to recognize all four emotions against all others using a combined feature set.

I have used WEKA [101] to distinguished all 4 emotions against all others using a combined feature set. The whole data set was divided to 10 folds and each fold was used once as a testing set, while the rest acted as training sets. All the three tests have computed by Multi Layer Perceptron Classification algorithm. The results presented in this paper are the averages of the 10 trials.

4.2.5 Results

The results showed an overall 92,38% recognition rate for the binary set of concentrate emotion or other. As seen in Table 4, 36 from the 44 clips were recognized as concentrate, and 146 from 153 clips as other.

Concentration	*Other*	
$\mathbf{36}(\mathbf{82}\%)$	8	Concentration
7	$\mathbf{146}(\mathbf{95}\%)$	Other

Table 11: Concentrate or not classification using the Space factor

For Test 2 i have defined a binary set of emotional states, when the clip is sustained or sudden. It showed a 91,87% recognition rate for the binary set of emotion, with 86 out of 96 clips were recognized as Concentrate or Meditation and 95 out of 101 clips recognized as Excitement or Frustration. The confusion matrix can be seen on Table 5.

Concentrate-Meditation	*Excitement-Frustration*	
$\mathbf{86}(\mathbf{90}\%)$	10	Con.-Med.
6	$\mathbf{95}(\mathbf{94}\%)$	Exc.-Fru.

Table 12: Concentrate-Meditation vs Excitement-Frustration classification using the Time factor

For Test 3, this time with all the four emotional states available i have combined space features and time features in one set, with overall classification of 85.27%, with Kappa statistic 0.8031. The Confusion matrix can be seen on Table 6.

Meditation	*Concentrate*	*Excitement*	*Frustration*	
45(**87**%)	2	0	5	Meditation
5	**39**(**89**%)	0	0	Concentrate
1	1	**39**(**83**%)	6	Excitement
3	1	5	**45**(**83**%)	Frustration

Table 13: All four emotions classification using the combined Space and Time feature set

4.3 Conclusion

The results shows the conclusion that Laban Movement Analysis is a valid and promising approach for emotion recognition from body movements due to the abstract level of Laban's technique. Specifically i have shown that two of Effort's component motion factors, Time and Space can result to high emotion recognition rates. The implementation of the rest of the Laban motion factors and components is one of current goals and part of my future work. It is anticipated that this will further improve the recognition rates. This is very important as emotion recognition systems must be very accurate before they can be used within games to adapt game behavior, as any emotion recognition mistakes can have the opposite effect on the player's experience. It would also be interesting to integrate the method to an automatic emotion recognition system capable to be used by Exergames.

Chapter 5

Recognizing Emotional Expressiveness in Raw 3D Body Motion Data

5.1 Introduction

As the gaming industry evolves, more sophisticated and natural user interfaces are being introduced, gradually replacing traditional controllers. Such interfaces make extensive use of modalities such as body motion gestures and voice commands, becoming popular nowadays in gaming and virtual reality applications. The former, motion data, has the potential to achieve an added level of immersion through the physical embodiment of the player character in real time [114]. This makes it a very strong candidate for an emotion recognition modality, and has already been identified and is being explored by researchers [40] [47].

While recognizing specific emotions is a very interesting and challenging task, being able to detect moments of emotional expressiveness for a game player could offer multiple benefits. Aside from the fact that this is a first step towards specific emotion recognition, emotional expressiveness detection can be used to automate part of the manual segmentation of 3d motion data, requiring users to watch long clips and segment portions as emotionally-labeled candidates. This would allow the quick creation of large databases of emotionally expressive clips, that many institutions construct manually and use for many purposes such as qualitative observation or training sets for predictive models [33] [224] [77].

Moreover, it would also be useful from a marketing perspective to be able to extract expressive motion clips from raw data, similar to the technique used to extract candid portraits from videos [74]. For example, some exergames that use Microsoft Kinect sensor extract screenshots from the gameplay sessions at the end of the games.

The proposed method uses automatic segmentation based on motion primitives in order to determine which clips to segment. Using symmetry detection in the segmented clips and energy of upper and lower body we can determine expressive clips that can benefit emotion recognition and emotion database creation experiments.

5.2 Methodology

The proposed method uses automatic segmentation of motion data, in order to separate high from low energy sequences. The input of the methodology is motion capture continues data sequences that was given as a raw signal in the algorithm. Then an automatic segmentation algorithm is used for segmenting high energy clips, by calculating the rotational kinematic energy of the upper and lower body, and segment sections based on a predetermined threshold. Furthermore a symmetry detection algorithm is applied in order to detect the symmetrical percentage for hands and legs, derived from our observations that when expression occurs then we have symmetry in some relevant body bones. A comparison of the warping motions is done through a multidimensional Dynamic Time Warping algorithm, that shows the symmetry in percentage for each frame and the average total symmetry. Due to the fact that our segmented clips mostly comprised by non-expressive gameplay clips, i have used a set of features that empirically appear to exhibit possible correlation with the expressiveness of the motion in the given gameplay context The set of features includes hand and leg symmetry and kinetic energy that allows the isolation of a pool of expressive clips. Figure 25 depicts the overall methodology and process.

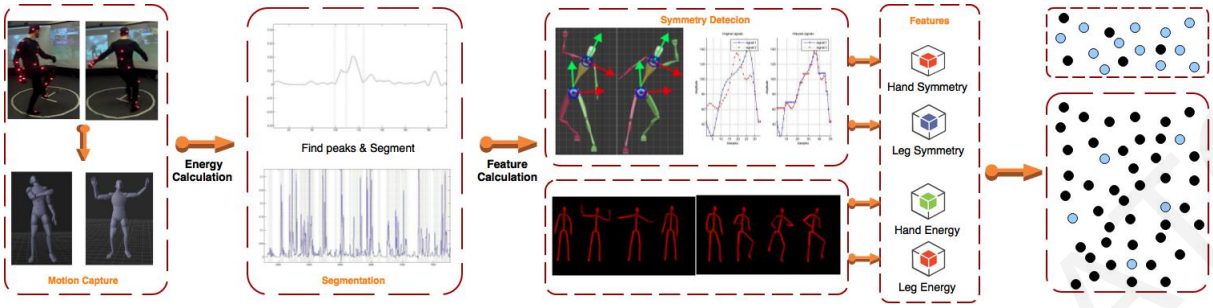


Figure 25: Methodology and process

5.2.1 Data Collection and Segmentation

Players were asked to play sports games for 30 minutes each on the Xbox integrated with the Microsoft Kinect. The motion data was collected using a PhaseSpace Impulse X2 motion tracking system with 8 cameras. The calculations and processing time was reduced to 30 frames per second. We have created an animation session representation in Unity3D, in order to aid the annotation of emotional states and create the ground truth. An automatic segmentation system has been implemented in order to parse high-dimensional body movements into a sequence of more basic primitives [76]. The segmentation is based on function E(f) which measures the rotational kinematic energy [206] at frame f by using the angular velocity of different joints. Let $\theta_{f,k}$ represent the angular speed of the k-th rotational degree of freedom at frame f, assuming that I is equal for all joints. Then, we can define the body's rotational kinematic energy as the weighted sum of the rotational joint velocities. Energy will be high when energetic motion occurs.

$$E(f) = \sum_{k=1}^{n} w_k ||\theta||^2_{f,k} \tag{3}$$

$$\sum_{k=1}^{n} w_k = 1 \tag{4}$$

Where w_k is the weight for the k-th rotational degree of freedom. After the calculation of the rotational kinematic energy of each frame, we smooth the signal with Savitzky Golay filter and segment the clips with the previous and next local minima that are created. Initially we have defined an empirically-deterimend threshold for the level of Energy, and observed that some times there are continues movements that should be combined to one, since they represent the same overall movement with different energy. At first to solve this, we used a Gaussian function to smooth the signal, resulting to a smoother signal but loosing many small expressive movements. Finally we have used another approach to define the initial and end cutting points, by when the percentage of the Energy falls below 90% comparing with the previous minima, thus showing that the movement has come to an end. This approach resulted in combining the small movements, but also keeping the valuable expressive information.

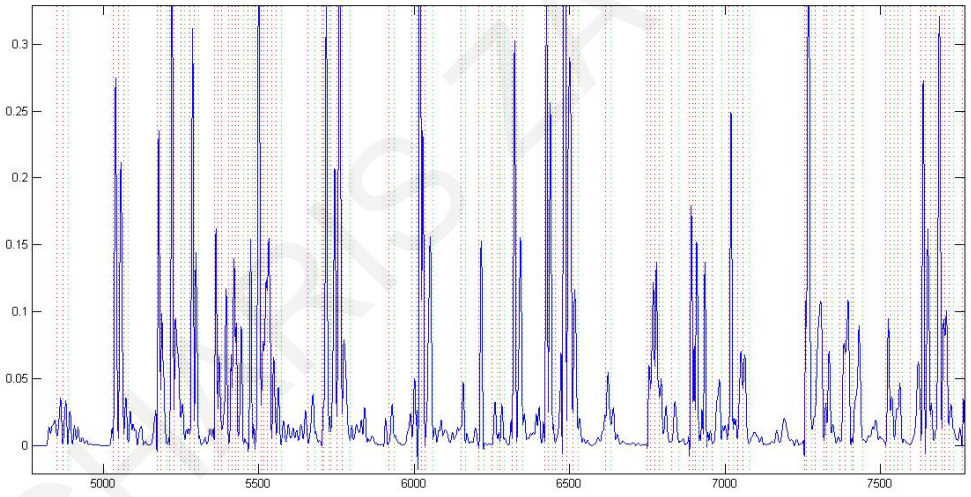


Figure 26: Motion primitives segmentation

As seen in figure 27 and 28, segmentation sessions contains both gameplay , and expressive clips, thus an evaluation of movement is needed to differentiate.

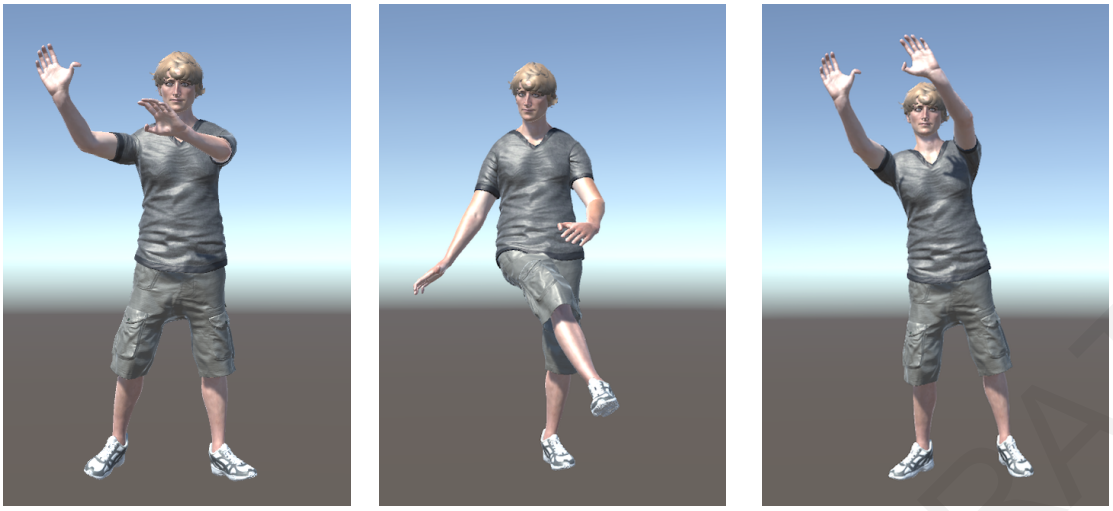


Figure 27: GamePlay Clips

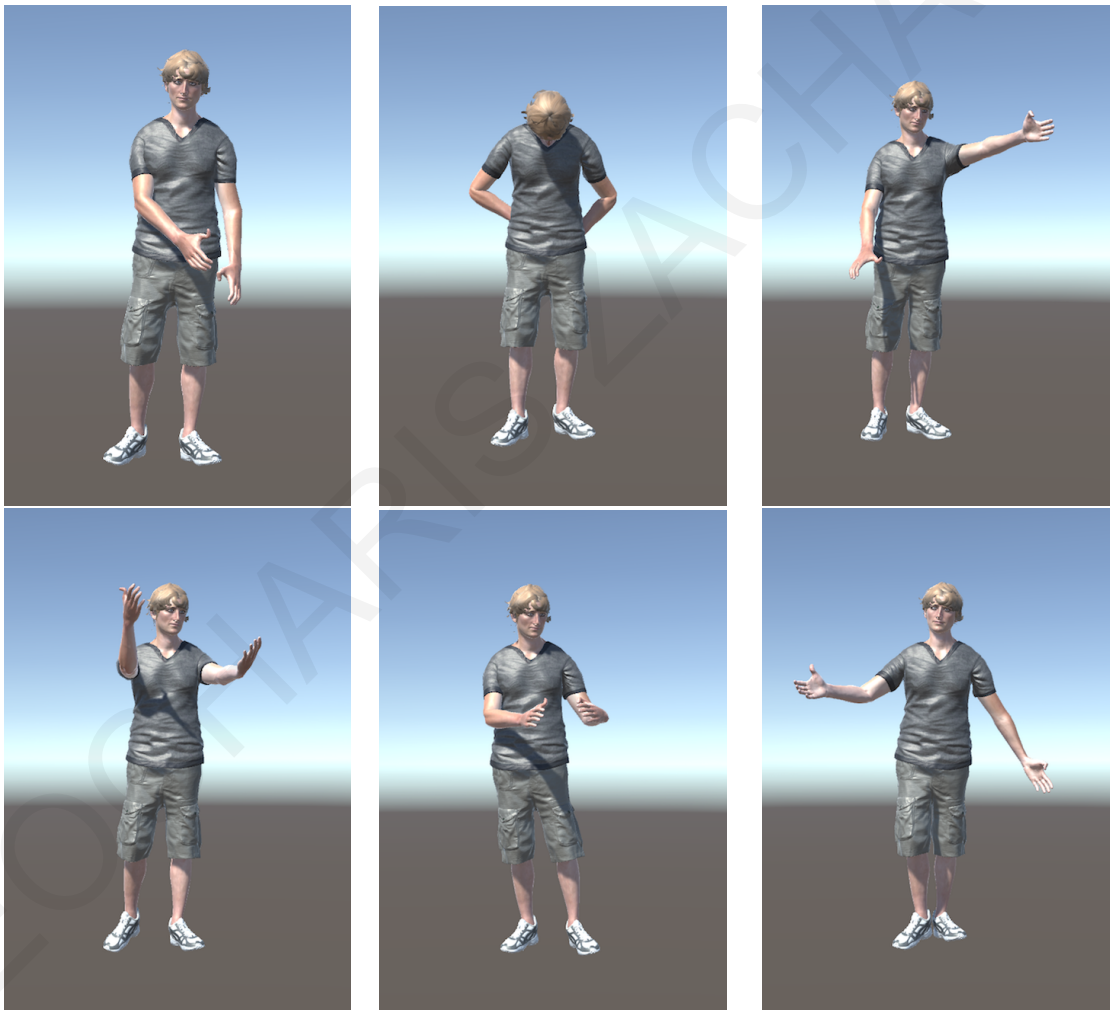


Figure 28: Expressive Clips

5.2.1.1 Symmetry Detection

Symmetric Motion Extraction

The human body is analyzed partially, the upper and the lower body, and the symmetrical motion of each segment is extracted separately in order to produce the final symmetric motion. For the upper body, taking into account the positions of the spinemid $\mathbf{p}_s$, the neck $\mathbf{p}_n$ and the left shoulder $\mathbf{p}_{ls}$, and, for the lower body, the positions of the spinebase $\mathbf{p}_{sb}$, the spinemid $\mathbf{p}_s$ and the left hip $\mathbf{p}_{lh}$, the orientations $\mathbf{R}_{ub}$ and $\mathbf{R}_{lb}$ are extracted for the upper and lower body respectively. In particular, let $\mathbf{v'}_{x1}$ be the normalized vector from the neck to the left shoulder position and $\mathbf{v}_{y1}$ be the normalized vector from the neck to the spinemid position. The cross product between $\mathbf{v}_{y1}$ and $\mathbf{v'}_{x1}$ gives $\mathbf{v}_{z1}$. Therefore, the cross product between $\mathbf{v}_{z1}$ and $\mathbf{v}_{y1}$ gives $\mathbf{v}_{x1}$. As a result, the upper body rotation matrix $\mathbf{R}_{ub}$ is defined by the normalized vectors $\mathbf{v}_x$ $\mathbf{v}_y$ $\mathbf{v}_z$. The sequence of the equations is presented below.

$$\mathbf{v'}_{x1} = \mathbf{p}_{ls} - \mathbf{p}_n \tag{5}$$

$$\mathbf{v}_{y1} = \mathbf{p}_s - \mathbf{p}_n \tag{6}$$

$$\mathbf{v}_{z1} = \mathbf{v}_{y1} \times \mathbf{v'}_{x1} \tag{7}$$

$$\mathbf{v}_{x1} = \mathbf{v}_{z1} \times \mathbf{v}_{y1} \tag{8}$$

$$\mathbf{R}_{ub} = [\mathbf{v}_{x1}\mathbf{v}_{y1}\mathbf{v}_{z1}]^T \tag{9}$$

Similarly, for the lower body, using $\mathbf{p}_{sb}$, $\mathbf{p}_s$ and $\mathbf{p}_{lh}$ we have:

$$\mathbf{v'}_{x2} = \mathbf{p}_{lh} - \mathbf{p}_{sb} \tag{10}$$

$$\mathbf{v}_{y2} = \mathbf{p}_{sb} - \mathbf{p}_{s} \tag{11}$$

$$\mathbf{v}_{z2} = \mathbf{v}_{y2} \times \mathbf{v'}_{x2} \tag{12}$$

$$\mathbf{v}_{x2} = \mathbf{v}_{z2} \times \mathbf{v}_{y2} \tag{13}$$

$$\mathbf{R}_{lb} = [\mathbf{v}_{x2}\mathbf{v}_{y2}\mathbf{v}_{z2}]^{T} \tag{14}$$

Since the orientations $\mathbf{R}_{ub}$ and $\mathbf{R}_{lb}$ are known in each body frame of the motion, the joint positions that belong to the body segment can be used to extract the symmetric motion. The symmetric motion of a human body joint is defined as the saggital symmetry of its position in each frame. The symmetric position of the joint *j*, can be calculated using the orientation $\mathbf{R}_b$ of the a body segment. Let **p** be the position of the joint and $\mathbf{p}_{sym}$ corresponding symmetric one. Using eq. (15), the point **p** is aligned to the global coordinate system and gives **p'**, the eq. (16) gives the saggital symmetric point $\mathbf{p'}_{sym}$ and finally, eq. (17) gives the symmetric point $\mathbf{p}_{sym}$.

$$\mathbf{p'} = \mathbf{R}^{-1}\mathbf{p} \tag{15}$$

$$\mathbf{p'}_{sym} = [-\mathrm{p'}_{x}\mathrm{p'}_{y}\mathrm{p'}_{z}] \tag{16}$$

$$\mathbf{p}_{sym} = \mathbf{R}\mathbf{p'}_{sym} \tag{17}$$

$$\tag{18}$$

Thus, applying the equations above for each joint position using the orientation of the body segment that the joint belongs to, we create the symmetric position of each joint, thus the symmetric motion is extracted. It is worth to be mentioned that the extraction of the symmetric and not the mirrored motion is desired, so the symmetric positions are assigned to the symmetric joints and not to the same one (e.g. the symmetric motion of the right wrist is the motion of the left one wrist in the symmetric motion).

Since the symmetric motion is extracted, in case there exists symmetry in a human motion, the symmetric and the original motions must be quite similar. In other words, the symmetric motion is extracted in order to be compared with the original one and the outcome will give the symmetry. To achieve that, the proposed comparison between the motions must be defined. In such an approach, the "technique" of the motion is important, not the speed or the synchronization of the performed action. For example, a jumping jack is a symmetric motion but also the performance of two sequent punches, a left-handed and a right handed straight ones, is a symmetric motion too. Thus, a warping between the motions is necessary in order to compare the warped frames of the motion. For this purpose, multiple multidimensional Dynamic Time Warping algorithm (DTW) is applied [81]. The joints that are analyzed to compare the motions are the wrists, the elbows, the ankles and the knees. For each of these human body joints, motion features as the position and the velocity are extracted in each frame. Thus, applying the DTW algorithm on the 3d vectors of the joint positions and the joint velocities, the warping between the symmetric and the original motions is achieved. Then, the euclidean distance between the 3d vectors,

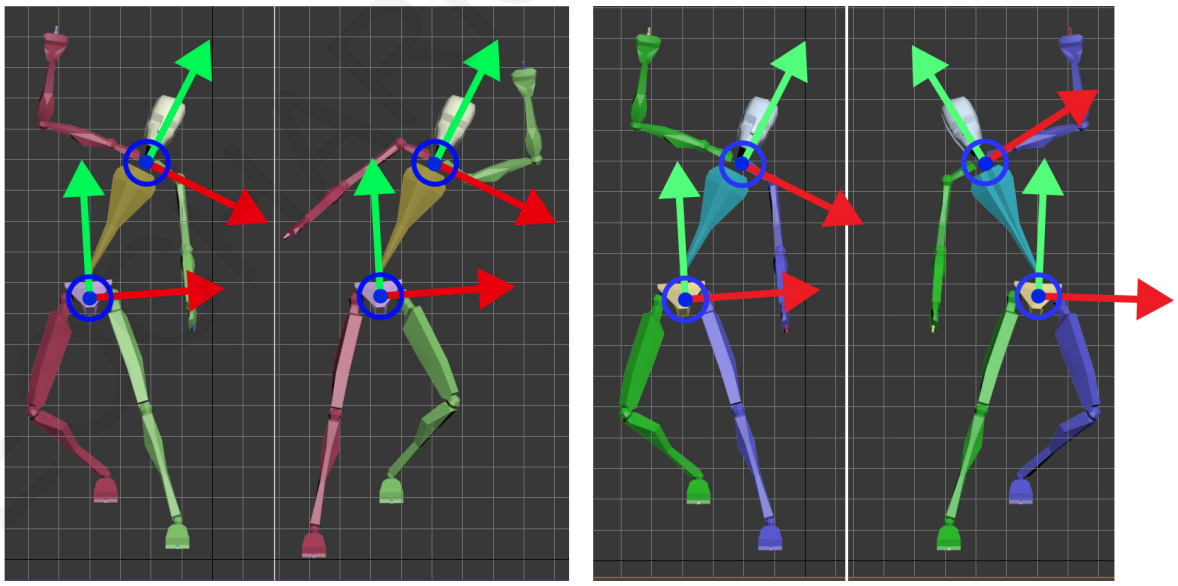


Figure 29: Symmetry detection

normalized with a maximum distance value that declares the maximum error, represents the error for each feature. Finally, a function $S_i(e_1, e_2)$ gives the symmetry in percentage for frame i, using weights, and the average of the motion frames symmetry gives the total symmetry. The equations are below:

$$w_2 = \frac{||\mathbf{u}_{or}||}{u_{max}} \tag{19}$$

$$w_1 = 1 - w_2 \tag{20}$$

$$e_1 = \frac{||\mathbf{p}_{or} - \mathbf{p}_{sym}||}{d_{max}} \tag{21}$$

$$e_2 = \frac{||\mathbf{u}_{or} - \mathbf{u}_{sym}||}{u_{max}} \tag{22}$$

$$S(e_1, e_2) = w_1 e_1 + w_2 e_2 \tag{23}$$

$$FS_{total} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{24}$$

where $\mathbf{u}_{or}$ and $\mathbf{u}_{sym}$ are the 3d vectors of the linear velocity of the original and symmetric motion respectively, u_{max} is the maximum value of the euclidean distance between 3d vectors of linear velocity, w_1 and w_2 are the weights of the error based on positions and velocities respectively, $\mathbf{p}_{or}$ and $\mathbf{p}_{sym}$ are the 3d vectors of the relative positions of the original and symmetric motion respectively and FS_{total} is the final evaluation of the symmetry in percentage.

5.2.2 Feature Space

Our overall set of segmented clips comprise mostly non-expressive gameplay clips. In order to successfully separate the majority of the expressive clips from the rest, i have used a set of features that empirically appear to exhibit possible correlation with the expressiveness of the motion in the given gameplay context. The set of features includes hand symmetry and leg symmetry (based on the previous section) as well as hand energy and legs energy. Energy is calculated in the form of kinetic energy based on the velocity of the corresponding body joints. The feature space is seen below:

Space: [Hand Symmetry, Leg Symmetry, Hand Kinetic Energy, Leg Kinetic Energy]

5.3 Results

The method presented has been tested using Microsoft Kinect sports game. Without requiring any heavy pre-processing from the motion capture, only some skeleton adjustments, the skeleton data are collected and the necessary info is extracted. In order to reduce the computational cost, motion capture data can be collected with less frames per second, or use only the end-effectors (head, hands, legs) for training and not all the skeleton joints. The stream of data is segmented automatically using a threshold on the total energy of hands and feet at every frame. Each segmented clip is individually evaluated using the criteria specified in previous section. For evaluation only purposes, the data was also annotated, before being segmented, by 2 human observers.

For an example player that plays a football game, the total interaction consisted of 10600 frames as depict Figure 30. The human observers identified in this 16 distinct expressions while the rest was marked as game-play. The system segmented the game into 292 clips. It is worth noting that due to the rather naive segmentation method used, a movement or expression is sometimes split into 2 or more clips. The aim is to be able to pick up at least one clip from each expression performed by the user. Using the formula described in previous paragraphs, with the thresholds for the hand and feet symmetries set to 50 and 80 respectively, the methods marks 18 clips as expression. 13 of these fall on the 10 out of the 16 expressions while the other 5 fall on 4 different game-play movements.

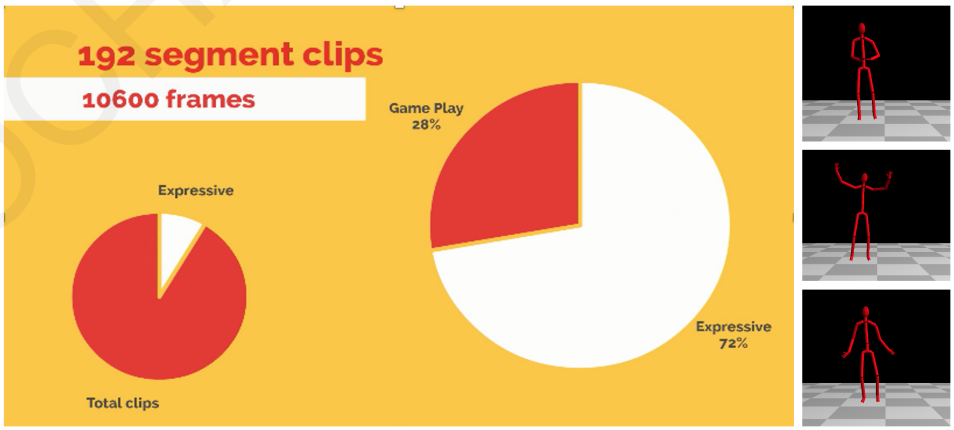


Figure 30: Results showed that 72% of the selected clips were expressive

5.4 Conclusion

In this experiment i have presented the preliminary results towards recognising emotional expressiveness from raw 3d body motion data.Through observation and testing, it appears that expressiveness in motion is linked to motion symmetry and minima or maxima of the energy of body extremities, depending on the gameplay context. For example, from experiments conducted using a football exergame, observation revealed that seeking above than average body symmetry and reduced lower body energy, isolates a large percentage of the expressive clips and removes the majority of the non expressive clips. Future work in this direction includes testing more features and building the automatic thresholding in all phases, starting from segmentation. Moreover, to conduct an agreement level test, among real users and the system, to determine a more reliable and realistic success metric, as the subject of emotion recognition is subjective, especially with non-acted data.

Chapter 6

Deep CNNs for Emotion Recognition based on Image transformation of 3D Skeleton Motion Data

6.1 Introduction

As the gaming industry evolves, more sophisticated and natural user interfaces are being introduced, gradually replacing traditional controllers. Such interfaces make extensive use of modalities such as body motion gestures and voice which are nowdays becoming popular in gaming and virtual reality applications. The former, motion data, has the potential to achieve an added level of immersion through the physical embodiment of the player character in real time [114]. This makes it a very strong candidate for an emotion recognition modality, and has already been identified and is being explored by researchers [40] [47] [9]. Deep Learning (DL) has been deployed in computer vision applications offering significantly improved results compared to traditional machine learning techniques. Particularly for human action recognition from motion data, Convolutional Neural Networks (CNNs) have been used extensively due to their high performance success on images or videos tasks [219]. This experiment focuses on the classification of emotions from 3D body movements, which are transformed to 2D images, that encode posture and motion dynamics in pixel values. Those images are used as input to train the last layers of a pre-trained Deep CNN applying the popular methodology of transfer learning.

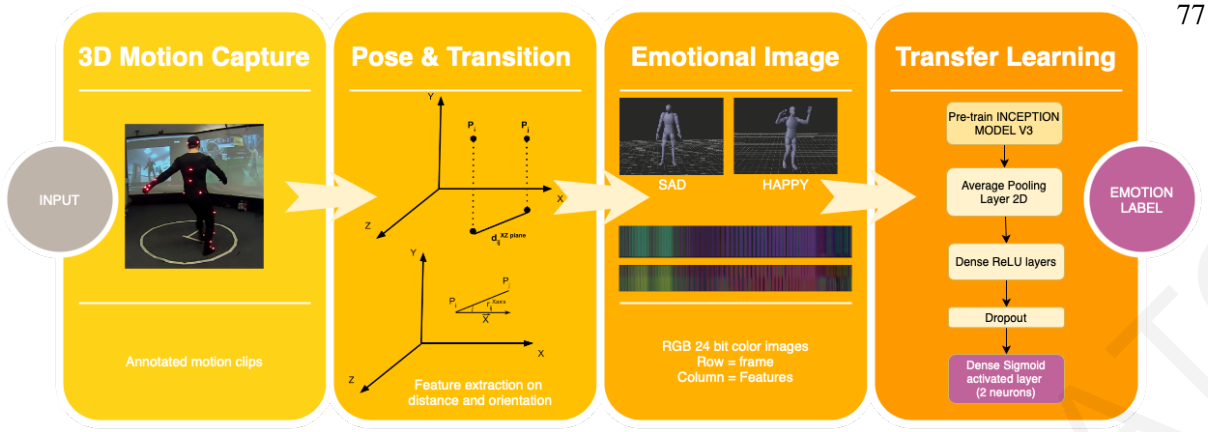


Figure 31: The overall Architecture

Traditional machine learning techniques have been improving in terms of accuracy but rely on hand-crafted features [40] [47] [9]. The use of deep learning techniques to automatically extract effective features from multimodal information and classifications are new directions currently actively pursued by researchers, but several challenges remain in realising an end-to-end deep learning system. With the availability of large datasets, deep learning has become a state-of-the art solution to problems such as emotion recognition. Kim et al. for example propose a CNN-based model for a hierarchical feature representation in the audio-visual domain to recognise spontaneous emotions [126]. Results showed that improvement of recognition accuracy is achieved when hierarchical features and multimodal information are adopted. In another effort, models are constructed from multiple physiological signals collected from sensors placed on the human body by adopting a multimodal deep learning approach so as to improve their performance and reduce the cost of acquiring physiological signals for real world applications [150]. To classify spontaneous multimodal emotional expressions as positive or negative, researchers proposed a cross channel convolutional neural network (CCCNN) having the capability of learning and extracting general and specific features of emotions relying on body motion and face expression [17].These features were further passed to cross-convolution channels to build the cross-modal feature representation.

The CNN is a type of deep learning that is especially used in the processing of images, proposed by Lecun et al. [140] It is based on the foundation of conventional neural networks inspired by biological understanding of the visual cortex. In this work the CNN applies convolution and sub-sampling alternatively to the input data, in the convolutional layers and sub-sampling layers. After two stages of this computation, the data is fed to a fully connected conventional neural networks, to complete the classification problem. Deep leaning-based algorithms can be used for feature extraction and classification. With the use of CNNs the work spent on the pre-processing of the images is greatly reduced since the algorithm is already capable of detecting the best features needed to classify the images.

Because CNN-based methods cannot reflect temporal variations, recently researchers have combined CNN, for the spatial features of single frames, with Recurrent Neural Networks (RNNs) that allow operation directly on time sequences. They are successfully applied to tasks involving temporal data such us speech recognition, language modelling, translation and gesture analysis. In a RNN, the output of the previous sequence time step is taken into consideration when calculating the result of the next one. However, a standard RNN does not handle long term dependencies well, due to the vanishing gradient problem. [107]. The RNN Long Short Term Memory Network (RNN-LSTM) is an extension for RNN, which works much better than the standard version. In the RNN-LSTM architecture, RNN uses gateway units in addition to the common activation function, which extend its memory [12]. Such an architecture allows the network to learn and remember dependencies over more time steps, linking causes and effects remotely [106]. In recent research, an RNN-LSTM was used to identify gestures emotion recognition based on low level features inferred from the spacial location and orientation of joints within a track skeleton. [200]. For all the above deep learning approaches, a vast amount of data is needed to perform the training and learning. Moreover, encoding raw skeleton data to images and then recognise emotions faces the limitation of a frame by frame representation of emotions. My method creates features related to time from raw skeleton data and converts them to images.

The proposed technique is inspired by recent research on action recognition methods that depict skeleton information into image-based representations and create features from 3D skeleton sequences [216]. The feature matrix that is created contains pose and transition dynamics using distance and orientation features.

For the *pose distance feature* within any given frame, the joint-to-joint Euclidean distance for all the joint pairs combinations was calculated by projecting the 3D joint coordinates to the three planes perpendicular to the axes x, y, z in a global coordinate system. The pose distance feature between two joints i and j for a given frame t is given by the below equation:

$$\mathbf{D}_{ij}^{t} = [d_{ij}^{XYplane,t}, d_{ij}^{YZplane,t}, d_{ij}^{XZplane,t}] \tag{25}$$

where:

$$d_{ij}^{XYplane,t} = ||P(i_x, i_y)^t - P(j_x, j_y)^t|| \tag{26}$$

In the above equation, P is the 2D point created from the projection of joint i or j on the XY plane for a given frame t.

In a similar way, the *transition feature* calculates the joint-to-joint Euclidean distance for all possible joint pairs combinations but within two consecutive frames:

$$\mathbf{C}_{ij}^{t} = [c_{ij}^{XYplane,t}, c_{ij}^{YZplane,t}, c_{ij}^{XZplane,t}] \tag{27}$$

where:

$$c_{ij}^{XYplane,t} = ||P(i_x, i_y)^t - P(j_x, j_y)^{t-1}|| \tag{28}$$

observe the difference from calculating distance for frames t and t-1. Two additional features are calculated based on joint-to-joint orientations with respect to the horizontal axes X, Y, Z. Calculating the

dot product of each joint-to-joint orientation with each of the 3 axis vectors allows the extraction of the orientation angle from the inverse cosine function. An example is given below for a joint-to-joint vector $\overrightarrow{ij}$ and the X axis vector $\overrightarrow{X}$

$$r_{ij}^{Xaxis,t} = cos^{-1}\left(\frac{\overrightarrow{ij^t} \cdot \overrightarrow{X}}{||\overrightarrow{ij^t}|| \times ||\overrightarrow{X}||}\right) \tag{29}$$

And below is the vector from all 3 axes for a single pair of joints i and j.

$$\mathbf{R}_{ij}^t = [r_{ij}^{Xaxis,t}, r_{ij}^{Yaxis,t}, r_{ij}^{Zaxis,t}] \tag{30}$$

In a similar way, a transition of orientation is calculated across two consecutive frames, with the same formula but now vector $\overrightarrow{ij}$ is calculated with joint i from frame t and joint j from frame t-1:

$$\mathbf{G}_{ij}^t = [g_{ij}^{Xaxis,t}, g_{ij}^{Yaxis,t}, g_{ij}^{Zaxis,t}] \tag{31}$$

The four features are calculated for all applicable joint pairs and are normalized using min and maximum values to (0,1). They are then concatenated in a row to form a feature set for a given frame. The same process is repeated for each frame starting from frame number 2 and moving further taking into consideration the dynamics with the previous frame 3D joint data. Given this configuration, at the end i had a 2D matrix with every row being the data for each frame and every column representing a feature for a particular pair of joints. This data is then converted to a 2D RGB image.

6.2.1 Emotion image generation

There are various ways of representing emotions, either by using **distinct emotions** like happiness, sadness, fear, anger, surprise, disgust or by measuring and contextualizing emotions according to a **dimensional space** as illustrated in Figure 32, where emotions are represented in two dimensions of *valence in x axis and arousal in y axis* and each emotion can be viewed as point in the space defined by these dimensions.

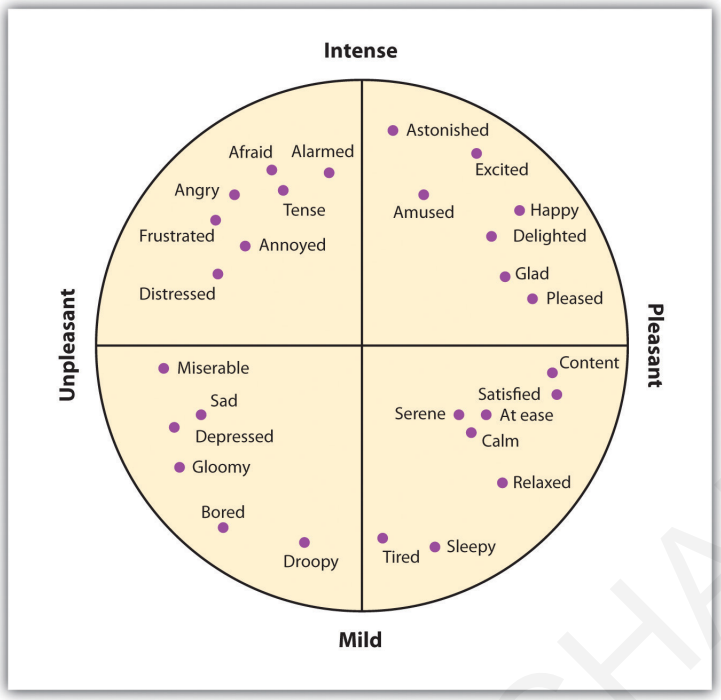


Figure 32: The Valence-Arousal space

Starting from the hypothesis that motion data can represent emotion information, to prepare motion clips for use in a CNN, i have proposed the transformation of 3d data to pixel data in the form of normalized posture and motion dynamics using an approached that has proven to be successful for action recognition [216]. The posture and motion features are encoded to RGB 24-bit color images. Each row of the image represents a single frame of the clip and each column a different posture or motion dynamics feature as seen in Figure 33.

All clips depict a single skeleton therefore the number of features is the same in all clips, making the width dimension of the image common for all of them. However, since each input motion clip can have a different length in terms of number of frames, the generated images have different sizes with respect to the image's height. To prepare the data for input for the selected pre-trained CNN, images needed to be converted to a standard size. This was achieved by determining the maximum height of all images, witch have variable frames in length and i have padded zeros to the remaining images extended pixels.

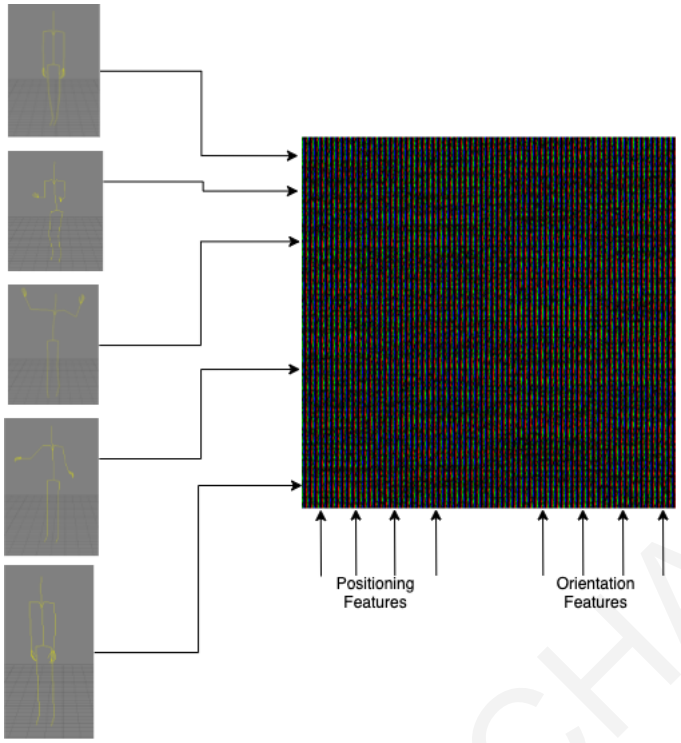


Figure 33: Image representing a series of postures (rows) with features (columns)

6.2.2 Transfer Learning

The rapid developments in Computer Vision has been further accelerated by the advent of Transfer Learning. Transfer learning allows us to use a pre-existing model, trained on a huge dataset, for our own tasks. Consequently reducing the cost of training new deep learning models and since the datasets have been vetted, we can be assured of the quality. To address our given classification problem, i have tested and compared different pre-train models that used on millions of images as described in chapter 2. Among these models we choose Inception V3 model which gave very good results in prior work.

Inception V3 [214] is an image recognition model that has been shown to attain greater accuracy on the ImageNet dataset. The parameters of the Inception module are 24 Million as can be seen in Figure 34. I have removed the last layers of the model adding my own layers, to accommodate my architecture with the total parameters reaching 24.5 million out of which 2.6 million are trainable.

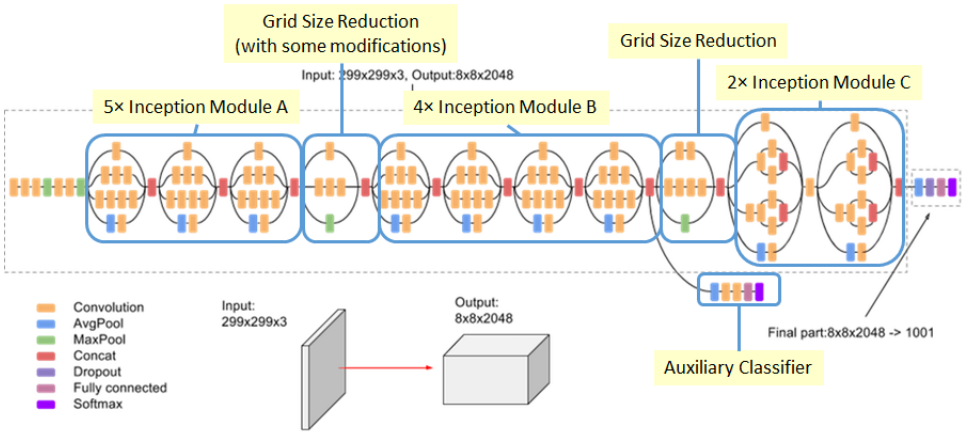


Figure 34: Inception-V3 model

I had used binary cross-entropy as the loss metric as i had 2 target classes (happy and sad). I have added new trainable layers as seen in Figure 35. The new layers contain a global average 2D pooling, then multiple dense RELU activation layers, and then dropout of 0.3, ending on two neurons for prediction of the targeted two classes.

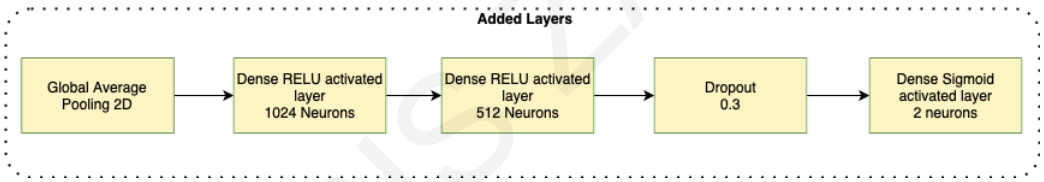


Figure 35: Added Layers on pre-trained Inception-V3 model

The model learned to convert the existing features into predictions on the new dataset. The summary of the model can be seen in Figure 36. The overall architecture of my emotion recognition method is showed in Figure 31.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, 83, 322, 2048)     21802784
_________________________________________________________________
global_average_pooling2d (Gl (None, 2048)              0
_________________________________________________________________
dense (Dense)                (None, 1024)              2098176
_________________________________________________________________
dense_1 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout (Dropout)            (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 513
=================================================================
Total params: 24,426,273
Trainable params: 2,623,489
Non-trainable params: 21,802,784
_________________________________________________________________
```

Figure 36: Our Transfer learning model architecture

6.2.3 Data Capture

I have used an acted emotional body movement dataset [90] in order to execute a pilot test with 2 emotions that differ in both dimensions of the Valence-Arousal space. The dataset contained scenarios to perform a typical and natural expression, captured by a motion capture system Axis Neuron. I have selected scenarios of equal man and women actors and in total i have used 208 happiness and 194 sadness different inputs. All the data were setup using 17 body joints with both positional and rotational data; i have only considered the positional data.

6.2.4 Results

The Inception model network was trained for 30 epochs using a learning rate of 10^{-3}. I have used 80% of the input clips for training and 20% for validation. All experiments are implemented on an Intel i9-07920x CPU @ 2.9Ghz, with one NVIDIA GeForce RTX 2080 Ti card. The training model was tested with an un-seen dataset of 16 motion clips (8 happiness, 8 sadness), which resulted in an average of 81% recognition rate as can be seen in table 14

Hapiness	*Sadness*	
$\mathbf{7}(\mathbf{88}\%)$	1	*Hapiness*
2	$\mathbf{6}(\mathbf{75}\%)$	*Sadness*

Table 14: Happiness or Sadness classification using Transfer Learning

6.3 Conclusion

Previous studies [200] showcased that movement dynamics can be used for emotion recognition. Up to now we have not seen research contributions in the Affective computing domain, that utilise image representations of pose and movement dynamics from 3D skeleton motion data. This technique has been used with success previously for action recognition [216] and in the current project i attempted to apply it in the context of emotion recognition.The proposed technique utilizes both posture and motion

dynamics to construct image representations of the motion clips. The images are then annotated with the emotion class of their source clips. The current technique shows that combining posture and subsequent frame motion dynamics in an image that uses rows as a temporal dimension and columns as dynamic features can capture affective information. While the initial results are promising, the study needs to be extended to a larger set of emotion classes, to determine how descriptive is the encoding of affect into the produced images. Moreover, new representations of images should be tested, such as those derived from other sets of motion dynamics, for example Laban Movement Analysis features. Further to this, the training data can be enriched with standard data augmentation techniques to potentially improve the classification accuracy. The data augmentation can take place either directly to the skeleton data before the creation of images (noise on joint properties, time warping, autoencoder-based among others) or to the resulting images with traditional image-based data augmentation techniques. Finally, while the current work deployed the Inception V3 model, there are other successful pre-trained CNN models that should be tested and compared in terms of performance.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This section presents the conclusions to the current thesis, as well as a discussion on limitations and advancements on the elements of emotion recognition from body movements.

7.1.1 Is there really an Emotion? Improving the Ground Truth

The ongoing research presented in this thesis and other studies [127] [129] [237] highlighted the importance of establishing ground truth for recognizing emotions. The task of emotion recognition relates to human traits and behavior which is not a well-defined task such as human action recognition (e.g., walking, running, picking up an object) object detection (detecting cars, animals and so on) where a corpus of training data can be created with high certainty for its validity. For example, in the first experiment of the current work, inspection of the agreement levels among observers, gives evidence that postures cannot represent expression of emotions deterministically. This is because humans express emotions in different ways, based on personality, culture, and perhaps other factors. Similarly, humans apply their individual experiences and characteristics when trying to recognize emotions expressed by others. Achieving a high percentage ground truth is already a challenging task and requires research

focus as a topic itself. Having to train a model with data that is not deterministic limits the opportunity for high accuracy rates. The uncertainty can be partially mitigated by utilizing acted postures or motion clips [10] [94] [169] [194], however this has also limitations. Creation of acted data is expensive and time consuming and it is impossible to generate a large corpus of data equivalent to the amounts of data that is used in other tasks to effectively train modern deep learning architectures. Data augmentation can enhance and improve the training data to an extent but based on the fundamental principles of acquiring training data, many actors that represent a wide set of different groups need first to be engaged. Extensive research in this field is likely to highlight a need for separate models to be trained and deployed for different type and/or demographics of players. It would also be interesting to see the agreement observers on the acted clips as this can further establish the ground truth.

Since the creation of a large corpus of acted clips is difficult, the training data can be designed to include selected motion clips from raw data captured during gameplay. Several trained observers can then identify, segment, and annotate the data accordingly. As the selection of training data can affect the performance of an algorithm, selection of the data to be used must include high agreement levels. Data samples case where agreement levels are not high, can also create opportunities for identification of reasons why agreement is not achieved, as this can also help towards the feature and data engineering process. Observed annotations can also be correlated with information collected by other modalities such as face, heart-rate monitor signals, electroencephalography, and self-reporting, to strengthen the ground truth. It is worth noting that in older and recent studies [60] [4], correlations among such modalities have been detected but as anticipated they are not strong due to human parameter measurements. When deploying observers and self-reporting, there are also other parameters to consider such as deception and self-deception, cognitive load, and demographics [235].

7.1.2 Are the selected features descriptive? Data Engineering

The different techniques that were deployed in this thesis and in similar studies [32] [137] [159] [217] [227] to engineer the feature set that will be used to train and test the model, indicate that there are different ways to try to encapsulate emotion in data. The first experiment showcased that using raw 3D skeleton transformation data as a feature set can be successfully used to recognize emotion from single postures. However, when working with motion clips, using the whole set of 3D transformation data enlarges the feature space and makes it hard to train a model. Transforming the raw data using LMA [137] and other similar techniques [56] [78] [103] proved to be successful in capturing affect, due to the elements that such methodologies attempt to embed. The results of the second experiment of this thesis showed that using features (such as percentage of narrow down, orientation features, velocity, acceleration and its derivative) derived from the Space and Time motion factors of the Effort component yielded high performance results on non-acted clips that were manually segmented and annotated by observers. In a side experiment that is not documented in this thesis, an attempt was made to train an autoencoder using raw 3D transformation motion clip data. While the compressed feature set captured and decoded the motion clips with high visual detail, it did not prove to capture and encode emotional information, as tests indicated. It would be interesting to see if transforming the raw data to variations of LMA representations first, and then compressing those through the encoder would encapsulate affect. The last experiment showcased that transforming raw data to posture and motion dynamics and converting those to images has the potential to represent emotions and allow recognition of some of those using Deep CNNs. The latter technique can also be combined with LMA as an alternative way to capture motion dynamics. It is apparent that there are many more feature set technique combinations that may further enhance the current results and offer insight to the problem domain. It is also possible to examine separately particular body segments only, and derived feature set for those, as indicated in the second experiment of this thesis.

7.1.3 Which model? Machine Learning and other Techniques

Recent advances in Machine Learning gave access to specialized new architectures that can be utilized to solve classification and pattern analysis problems. Deep Learning neural networks allow the construction and use of automatically transformed features that are capable to encode and capture information that relates to given problems. This thesis demonstrated how architectures such as Multilayer Perceptron, CNNs as well as ad-hoc approaches such as symmetry detection can be applied to recognize affect or the presence of emotional expressiveness. Regarding the CNN, this thesis applied Transfer learning by using the Inception v3 model. While the Inception model has been transferred successfully to experiments that address different machine problems, there is room for more experimentation by testing other readily available CNN trained models such as the VGGNet [211], REsNet [104], AlexNet [135], among others. Additionally, different architectures such as RNN-LSTM [12] [71], Autoencoders [5], and Transformers [46] provides a way to treat the problem at its original form when it comes to raw data, more specifically, a time-series domain. Considering the difficulty in segmenting the clip before processing it, these seem techniques that can offer a way to deal with the motion data within frame windows, the size of which can be chosen empirically after experimentation. More recently, Reinforcement Learning (RL) has been used to achieve emotion detection and the gradual emotional changes within conversations [147]. Similar principle may have potential to skeleton data, with the action of the RL agent to be set selecting the emotion label. Using pre-segmented data allows more traditional techniques, but the time dimension becomes more significant as the focus shifts towards application that use emotion recognition.

7.1.4 3D Postures for Emotion Recognition

Manual examination of body postures provides a way to examine, determine and report emotional expressiveness. Observers that are given sets of postures are capable to recognize the expressed emotion in most of the cases. It is possible to train machine learning models with training sets that comprise postures annotated by observers, and such systems are capable to generalize to new postures and new users with results comparable to humans. However, as mentioned earlier in the section about improving the ground truth, providing observers with postures only snapshots, does not yield high agreement levels, which reduces the ground truth and the expected accuracy. Using postures from acted clips can eliminate this issue, but the bias of selecting those posture remains. Moreover, postures do not hold temporal information and are utilized isolated from the actual movements that embed the selected postures. Perhaps an emotion recognition that uses posture representation as a feature set need to be enhanced by ss different mechanism that operates on a raw signal, recognizing emotion frame by frame, but apply a temporal mechanism to determine emotion based on the time-series of postures. After all, the 3D representation of a posture itself cannot embed information on what action caused this posture and what type of movement has led to it. A particular posture can be the result of a large number of 3D movements and can lead to an equally large number of subsequent postures, depending on the movement of the player/actor. Modern models of Recurrent Neural Networks [12] [106] [200] can offer some ideas for experimentation towards the use of postures in emotion recognition.

7.1.5 3D Body Motion for Emotion Recognition

The second experiment showed that 3D body motion data can be used to calculate new features such as those derived from Laban Movement Analysis that are distinctive for some emotion categories. This makes Laban features suitable for training machine learning models to recognize emotions with high success. An interesting question arises on whether implementing more features that are derived from

LMA or other body motion theories can be used with modern Deep Learning architectures to further improve the system. The current implementation in this thesis, calculated Laban features as per clip, using average values. As described above, an interesting venue to move ahead would be to calculate LMA features per frame and establish them as a time-series training data. Another interesting question is whether those can be used either directly as input to an Autoencoder, or similarly to the last experiment of the thesis by embedding LMA features as pixels of images and applying CNN techniques, that proved promising during the last experiment. However, it is not very clear that images created this way present descriptive image data for all classes of emotions that will allow classification of more than two emotion labels. Overall, 3D motion data offer the additional dimension information that postures of 2D motion data from videos does not have, but theories of motion analysis require to handle the complexity of the additional dimension.

7.1.6 3D motion data in Multimodal systems

While individual modalities have proven successful towards recognition of affective information, the prospect of multimodal systems can combine the strength of each modality by fusing them into a feature set of by combining the output of different modalities before reporting the emotional prediction [84] [97] [98]. The current thesis did not explore this opportunity, however given the limitations of each modality and the ambiguity that is caused by the temporal duration of the emotional expressions, this approach needs to be further explored and exploited. Accuracy rates can potentially be improved and the temporal effect of emotions can be examined.

7.1.7 Final remarks

This thesis presented a set of different techniques that have potential in recognizing specific emotions or emotional expressiveness from 3D body posture and motion data. This thesis is concentrated on methods and algorithms for emotion recognition which are not binded on a specific emotion, but more generalized. The topic of emotion recognition has been explored by many researchers in the past years and despite the promising achieved findings, there are numerous areas of improvement and explorations. The thesis contributed with results across a number of different facets of the problem and has also presented a number of opportunities to further contribute towards this non-deterministic and open area of research.

7.2 Future Work

This section presents different but interconnected directions for future work, in the field of automatic emotion recognition using body movement analysis.

7.2.1 Automatic emotion recognition using high level movement notational systems

Even though there have been some approaches that used movement notational systems, the complexity of the task allows further investigation. Most approaches so far have focused on specific aspects or subcomponents of a notational system. Even though in many cases results are promising, all methods are ad-hoc and are far from applicable in real life situations. They also fail to generalize in not only broader context, but also in minor adjustments of some parameters.

Another common problem of existing movement notational systems is that they are designed for purposes that do not necessarily regard emotional states. Laban Movement Analysis [137], Beauchamp [67], Benesh [20], Eshkol and Wachmann [69] movement notational systems were created to describe

dance movement based on some movement qualities or properties. Specifically Laban Movement Analysis has been tested in its suitability for emotion recognition (described earlier in 4.2). Although encouraging results have been produced regarding recognition of emotion in pre-segmented clips under specific context, the presented Laban Movement Analysis computational models have several limitations in their experimental data, as there is no evidence that the success of the recognition can occur in other applications and different context. In other words, the rules that encode the different Laban Movement Analysis categories (Body, Effort, Shape, Space) into a computational model have not been successfully tested across different databases. It would be interesting to see how LMA can be studied to offer a more comprehensive and holistic system that can be used to detect emotion of users in different context. Such an approach would require that a set of emotions and their bodily manifestations are studied and observed across different context cases and correlations are derived from the results of the observations. Context and segmentation are significant components of such a system if the purpose is to achieve automatic emotion recognition in real time for different context (e.g. different types of exergames).

Other notational systems [228] [217] [32], which have been designed for purposes that are related to emotions have not been extended and applied to wider cases and are only proof of concepts. Perhaps the most common limitation of these approaches and any other applications of notational systems described earlier in 4.2, is the fact that all movement modalities provide raw and continuous signals which make it very difficult to segment or to recognize emotions within this raw signal rather than from between a predetermined set of annotated clips. The concept of segmentation for emotion is a different area of future work itself and it is described in the next paragraph.

Furthermore, the area of automatic emotion recognition requires not only a segmentation process, but also some form of recognition to take place during this segmentation. Perhaps the closest notational system to achieve the goal of emotion recognition is the BAP [52]. BAP encodes a list of behaviors as variables, all of them annotated with a short description that can help develop a computational model

from BAP that can be used to study emotion expression. A significant limitation of this approach is that leg movement and whole body posture cannot be studied in detail due to the camera settings. This limitation could be tackled by utilizing different notational systems (such as LMA) to evaluate whole body posture or lower limb movement. Also, as the authors of BAP mention [52], categorical coding systems that focus on specific hand shapes, orientations, positions and movement trajectories have been established in the fields of sign language [189] and linguistics and gesture studies [31]. Such information can be integrated into BAP or into another layer of a holistic system that uses BAP. Another limitation of BAP is the manual annotation of the movement clips with the labels/behaviors of the system. This was partially solved by an extension called AutoBAP [221] that loads continuous body motion and outputs a labeled XML file with a 62% support of the behaviors of BAP and with good agreement level with a manual annotator. This does not contribute yet to emotion recognition, but as BAP evolves as a coding system, AutoBAP can support the use of BAP by replacing the manual annotation. AutoBAP has some limitations that can be improved in future research. The experimental results need to be validated by the introduction of more natural datasets in the system. The current approach uses scripted datasets. Moreover, as the authors of AutoBAP state, BAP was designed for situations of standing characters with upper body information available. This was already highlighted earlier as a limitation of BAP and will need to be addressed in the future. Finally, BAP and AutoBAP are recently established and will need to be tested and verified. It is likely that through evaluation, some behaviors of BAP may prove more significant for emotion recognition. Systems similar to AutoBAP may focus on those behaviors instead. Different approaches for automatic annotation and emotion recognition could be derived, based on other notational systems. This is discussed further in the next section.

7.2.2 Automatic emotion segmentation

A very challenging component of an automatic emotion recognition system is the ability to segment continuous body motion signals so that recognition can occur using the segmented motion clips. The majority of current emotion recognition approaches described earlier in this paper [129] [201] [237] [121] were based on manual segmentation and annotation of postures. Automating this process can yield significant benefits such as achieving automatic emotion recognition itself, saving expensive human intervention, avoiding ambiguity due to the subjectivity of human perception and decision and so on. A rational approach would be to try and establish automatic emotion recognition based on automatic gesture segmentation.

Currently, gesture-based motion segmentation is an unresolved topic under investigation. Techniques that use continuous signals of motion parameters and look for sudden changes to those parameters have been used [119] but with limitations to signals with gestures detected among non-gesture signals. However, that assumption is not valid for most of the time for real application signals such as in exergaming. To address this problem, research effort is needed to initially determine a dynamic separation of gesture and non-gesture signals using a technique like the one described in [119], and then test a set of different techniques that can provide emotion recognition as performed with manually segmented clips or with the use of context short memory mechanisms and continuous frame windows for recognition. Also it would be important for a system to be able to separate different context periods within a signal. For example, in the case of games, it would be good if the signal could detect movements that are actual gameplay and movements in which the player does not interact with gameplay. It would then require a different technique to be used for emotion recognition for the above two cases, as emotion detection during gameplay could look into properties like arousal and energy of movement for gestures, while for non-gameplay the recognition of more standard gesture properties could provide insight. Another possible extension of existing research would be to compare the success of different

emotion recognition techniques in high-level against low-level segmentation techniques [6] [14]. For example, LMA could be tested on both high and low level segmentations due to the correlations found between LMA qualities and kinematic features. Moreover, it would be worth examining the use of fuzzy logic to determine membership of emotions for each clip. This kind of membership can be used both for emotion recognition, but also for studying the richness of motion with observations and annotations of experts in emotion expression. Furthermore, an unsupervised learning technique could detect outliers in frame windows to try and identify either context from non-context signal segments, or expressive from non-expressive clips. On the contrary, supervised techniques can also provide solutions for this topic. In Fiss et al. [74] describe how to select frames from a video of a human face that effectively communicate the moment and work well as candid portraits. To automate this task, they have collected a large dataset of human ratings, and trained a predictive model to select those frames that are most or least effective as candid portraits. This technique can be used in selecting motion clips instead of frames for two different approaches. The first would be to use it for an automatic selection and extraction of expressive clips without the need to determine a definite emotion for annotating it. Those clips can be then given for annotation by observers, as done in techniques described in section 4.4. The second approach would be to establish automatic emotion recognition using a set of different features as in the current emotion recognition literature [36] [76] [185]. Section 4.5 provides more information on this topic that can give more ideas for future exploration.

7.2.3 Context knowledge of environment

Context knowledge describes information about the surrounding environment and more specifically interaction items such as objects or other humans/avatars. Analyzing the context of environment when a movement is performed can lead to more accurate emotion recognition. Today's body movement-driven computational models do not take into account the context knowledge of environment into affect

recognition results. However, successful recognition of emotional states in HCI is linked to individual human characteristics and behavior types. Therefore, an emotion recognition system has to take into account the context of the interaction, as well as the user types of anyone involved in such an interaction.

Caridakis et al. [38] presented a neural network architecture that highlights the need for adaptation and has the capacity to adapt to the user. They used a combination of different sensors to achieve a multimodal fusion input based on fuzzy logic that included facial expressions and hand/body gestures (upper body). However, this is limited to recognition of the active quadrant of Whissel's wheel activation/valence representation [233] and not to given emotional states. Moreover, it can adapt to the user, rather than the overall interaction context. This technique can be further improved with the introduction of affective user models [166] [236] [109] which can be really important for: (a) design adaptive interactive applications for personalized experiences, and (b) globalize our systems by transferring a user's model among different applications. Metallinou et al. [170] showed how adaptation and learning can be enhanced by context-sensitive frameworks for emotion classification. In this approach, context is defined at temporal level as the emotional context of past and future observations. They presented a context-sensitive Hidden Markov Model (HMM) equipped with this type of memory mechanism that outperforms a context-free HMM in terms of emotion classification performance for valence and valence-activation space into clusters. This approach tested audio and visual cues with the former being more significant. However, this definition of context is more of a memory system and would need to be broader to include knowledge and understanding of more general type as described earlier in this paragraph.

Earlier in time, researchers used domain ontologies as components to define context-aware emotions [42]. This ontology uses different modules to organize fundamental affective computing concepts. The ontology allows the use of a formal description of abstracted emotions that can be personalized by users, language and culture. Another similar approach uses a set of ontologies to use context and culture to

improve the recognition process [164], however the testing results are limited to facial cues. Ontologies have also recently been used to apply context and user profiles in the recognition of emotional states using EEG signals [239]. Such ontologies can be tested on other modalities such as body motion data and movement notational systems to allow emotion recognition based on context information. A recent approach, which is of particular interest to this survey, incorporated context in the bodily reaction and the cognitive input to demonstrate how context influences the way humans determine and interpret emotions felt by others [23]. As humans are the standard benchmarking technique to test emotion recognition for non-acted data in the literature, the results of this study show that such ontology can offer another step towards more realistic and successful emotion recognition based on body motion data.

Context is not only important during the actual recognition process, but also during the creation of databases that contain emotion information and are used to train systems. In that particular case and for the utilization of non-acted data, context may influence not only the emotion expression style, but also the opinion of the person who annotates a given set of data with an emotion label. In such scenarios the expressions are uncontrolled and annotation becomes a subjective decision with varying agreement levels being the major benchmark measurement. Nevertheless, experimentation in emotion recognition shows that invalid annotation can lead to poor results. Moreover, expressions themselves may vary based on context, resulting in poor results when attempting to generalize by transferring emotion captured during a given context as training data for different contexts. Siegert et al. [210] investigated how context information influences the assignment of labels with realistic data. They used video and audio channels to present data to the annotators. Their experimental results show that contextual information may be more important in specific emotions, however their experimental results are limited and do not provide a definite set of rules as a result. It would be interesting to investigate how to determine a set of features or an adaptation of a notational movement that allow data sets that can generalize to a set of emotions under different contexts. Another alternative approach would be to assemble a large

collection of training data sets captured from different contexts, using different feature sets and different combinations of modalities, organized appropriately in graphs or hierarchies, to allow the database to perform under different context or even attempt to determine and recognize the context itself as a reverse engineering task.

7.2.4 Brain and body emotion recognition

Emotional body language consist of an emotion expressed in the whole body, comprising coordinate movements and meaningful actions which so far have been investigated in isolation and not related to the perception of emotional body language. Darwin [54] has shown the close relationship between emotions and adaptive behaviour. He has also described in detail the body expressions associated with emotions in animals and humans with a particular emphasis on the link between emotion and action [179] [79]. As described earlier [27] [95] [99] fusiform gyrus and the amygdala of the brain is responsible for processing properties related to emotional body language. The amygdala decodes the affective relevance of sensory inputs and initiates adaptive behaviours via its connections to the motor systems [68]. More recently, the discovery of neurones that encode complex movements and actions (mirror neurones) [192] [82], provide the neurobiological bases for all emotional and social cognition skills. This has changed the role that motor areas have in perception of body movement and emotional body language.

Brain activity directly associated with exposure to emotional body language is relatively unexplored and emotion researchers are trying to get closer to the link between emotion and behaviour, in order to shed some light on disorders that combine motor and emotional components such as autism, schizophrenia and Huntington's disease. With respect to the perception of bodily expressed emotions, humans rely on a mixture of visual form and motion cues. Attempts to emulate human emotion recognition in machines will require detailed knowledge not only of how all the different subsystems of the brain operate but also of how they interact, which is currently a focus of research in cognitive neuroscience. Another

interesting use of brain technology on body-based emotion recognition would be to compare the success of an emotion recognition system trained from clips annotated by brain signals with a system that is trained using manual annotation. Moreover, as technology now allows neuron headsets to act as input devices to robotic controllers and virtual simulations, it would be interesting to see how given body motions and gestures are triggered from specific emotions and determine correlations between the two of them.

7.2.5 Datasets for emotion recognition based on movements

As described in sub-section 2.5, only a few datasets exist for emotion recognition from body movements such us the FABO database and the GEMEP database [97], to be used from researchers as ground truth to test their experiments. Throughout this thesis, we have created datasets from motion capture and from Kinect devices, that have been used successfully in the experiments described in previous sections. One of our future goals is to publish and open this library of videos with the respective annotations, to be used freely in the community as open-source. For this work further annotation needs to be made from different people, from different genders, ages and ethnic backgrounds in order to generalize better the human annotation.

7.2.6 Variational auto-encoders

Lately, deep learning based generative models have gained more interest due to improvements in the machine learning field. Relying on huge amounts of data and well-designed network architectures, deep generative models have shown an incredible ability to produce highly realistic content of images, text, sounds etc. Among these deep generative models are also the Variational Autoencoders (VAEs). VAEs are meant to compress the input information into a constrained multivariate latent distribution (encoding) to reconstruct it as accurately as possible (decoding).

A VAE encodes the data into a latent space R, as seen below in Figure 37.

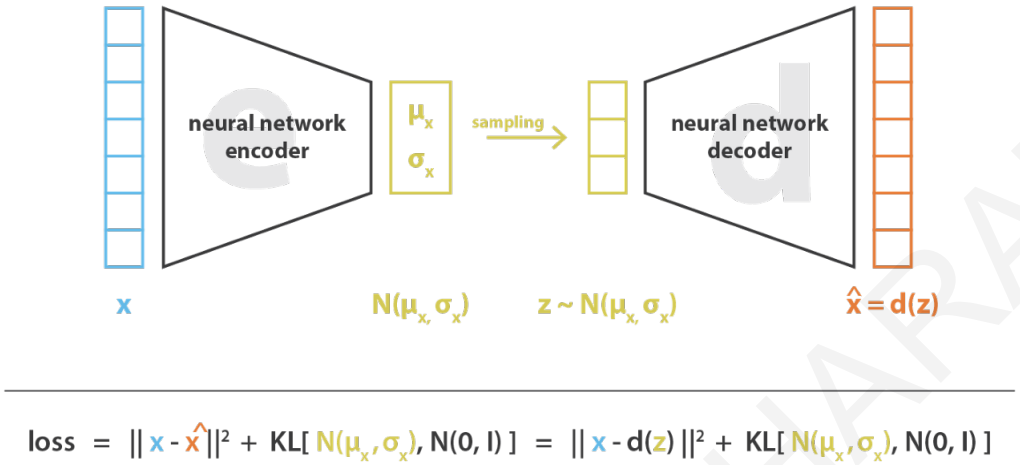


Figure 37: Variational Auto-Encoder

VAEs should be selected if precised control is needed over latent representations and what it is needed to represent. However if auto-encoders suffices for the feature representation its preferable to use auto-encoders due to their simple and uncomplicated structure. As a future research we will try to use the VAEs latent space, in a continuous motion capture data, to learn distributions, that can be used as an extra feature space for emotion recognition or generate new expressive datasets.

Bibliography

[1] KENDON A. How gestures can become like words. *Cross cultural perspectives in nonverbal communication.*, pages 131–141, 1988.

[2] Affectiva. http://www.affectiva.com.

[3] Ferdous Ahmed, A. S. M. Hossain Bari, and Marina L. Gavrilova. Emotion recognition from body movement. *IEEE Access*, 8:11761–11781, 2020.

[4] A. Ali and C. Gatzoulis. An exploratory pilot study on human emotions during horror game playing. *IEEE Bahrain and IET International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, pages 1–6, 2020.

[5] Sevegni Odilon Clement Allognon, Alessandro L. Koerich, and Alceu de Souza Britto Jr. Continuous emotion recognition via deep convolutional autoencoder and support vector regressor. *CoRR*, abs/2001.11976, 2020.

[6] J ALON, V. ATHITSOS, and S. SCLAROFF. Accurate and efficient gesture spotting via pruning and subgesture reasoning. *IEEE ICCV Workshop on Human Computer Interaction*, pages 189–198, 2005.

[7] J ALON, V. ATHITSOS, Q. YUAN, and S. SCLAROFF. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1685–1699, 2009.

[8] S. F. Aly and A. L. Abbott. Facial emotion recognition with varying poses and/or partial occlusion using multi-stage progressive transfer learning. *Scandinavian Conference on Image Analysis (Norrköping: Springer)*, pages 101–112, 2019.

[9] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. Emotion analysis and classification: Understanding the performers' emotions using the lma entities. *Computer Graphics Forum*, 34(6):262–276.

[10] A.P. ATKINSON, W.H. DITTRICH, A.J. GEMMEL, and A.W. YOUNG. Evidence for distinct contibutions of form and motion information to the recognition of emotions from body gestures. *Cognition*, 104:59–72, 2007.

[11] MSH AUNG, N. BIANCHI-BERTHOUZE, P. WATSON, and C. WILLIAMS. Automatic recognition of fear-avoidance behaviour in chronic pain physical rehabilitation. In *Pervasive Computing Technologies for Healthcare*. ACM, 2014.

[12] D. Avola, M. Bernardi, L. Cinque, G.L. Foresti, and C. Massaroni. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Transactions Multimedia*, 21:234?245, 2018.

[13] N. BADLER, B. WEBBER, and C PHILLIPS. Simulating humans. *Computer Graphics, Animation and Control, Oxford Univ Press*, 1993.

[14] J. BARBIC, A. SAFONOVA, J.Y. PAN, C. FALOUTSOS, J.K. HODGINS, and N.S. POLLARD. Segmenting motion capture data into distinct behaviors. *Graphics Interface Conference*, pages 185–194, 2004.

[15] W. BARLESON. *Affective learning companions: Strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation and perseverance.* PhD thesis, Massachusetts Institute of Technology, 2006.

[16] S. BARON-COHEN. Mindblindness. *MIT press*, 1995.

[17] P. Barros, C. Weber, and S. Wermter. Emotional expression recognition with a cross channel convolutional neural network for human-robot interaction. *IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 582–587, 2016.

[18] A. BASHASHATI, M. FATOURECHI, R.K. WARD, and G.E. BIRCH. A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. *J. Neural Eng.*, 4:R32–R57, 2007.

[19] A. BECK, L. CANAMERO, and K.A. BARD. Towards an affect space for robots to display emotional body language. *IEEE internation symposium on Robot and Human Interactive Communication*, pages 464–469, 2010.

[20] R. BENESH and J. BENESH. Reading dance: The birth of choreology. *McGraw-Hill Book Company Ltd*, 1983.

[21] D. BERNHARDT and P. ROBINSON. Detecting affect from non-stylised body motions. *Proceedings of ACII07*, 2007.

[22] K. BERRY and P. MIELKE. A generalization of cohen kappa agreement measure to interval measurement and multiple raters. *Education and Psychological Measurement*, page 48, 1988.

[23] F. BERTHELON and P. SANDER. Emotion ontology for context awareness. *Cognitive Infocommunications*, pages 59–64, 2013.

[24] L. BIANCHI-BERTHOUZE, N.and LISETTI. Modeling multimodal expression on user's affective subjective experience. *User modeling and User-adapted interaction*, 12(1):49–84, 2002.

[25] R.L. BIRDWHISTELL. *Introduction to Kinesics:An annotation sysytem for analysis of body motion and gesture.* PhD thesis, University of Kentucky Press, 1952.

[26] J.M. BLAND and D.G. ALTMAN. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, pages 307–310, 1986.

[27] E. BONDA, M. PETRIDES, D. OSTRY, and A. EVANS. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *The Journal of Neuroscience*, 16(11):3737–3744, 1996.

[28] T.R BOONE and G. CUNNINGHAM. Children's decoding of emotion in expressive body movement: The development of cue attunement. *Development physchology*, 34:1007–1016, 1998.

[29] D. BOUNCHARD and N. BADLER. Semantic segmentation of motion capture using laban movement analysis. *Intelligent Virtual Agents*, pages 37–44, 2007.

[30] G. R. BRADSKI and J.W. DAVIS. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002.

[31] J. BRESSEM. Notating gestures - proposal for a form based notation system of coverbal gestures. unpublished.

[32] P. BULL. *Posture and Gesture*. Elsevier Science and Technology Books, 1987.

[33] C. BUSSO, M. BULUT, C.-C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J.N. CHANG, S. LEE, and S.S. NARAYANAN. Iemocap:interactive emotional dyadic motion capture database. *J. Language Resources and Evaluation*, 42(4):335–359, 2008.

[34] C. BUSSO, Z. DENG, S. YILDIRIM, and M. BULUT. Analysis of emotion recognition using facial expressions, speech and multimodal information. *6th international conference on Multimodal interfaces (ICMI '04).ACM, New York, NY, USA,*, pages 205–211, 2004.

[35] A. CAMURRI, I. LAGERLOF, and G. VOLPE. Recognising emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal Human Computer Studies*, 2003.

[36] A. CAMURRI, B. MAZZARINO, M. RICCHETTI, R. TIMMERS, and G. VOLPE. Multimodal analysis of expressive gesture in music and dance performances. *Gesture-Based Communication Human-Computer Interactions*, pages 59–70, 2004.

[37] A. CAMURRI, B. MAZZARINO, and G. VOLPE. Expressive interfaces. *Cognition, Technology and Work*, 6:15–22, 2004.

[38] G. CARIDAKIS, K. KARPOUZIS, and S. KOLLIAS. User and context adaptive neural networks for emotion recognition. *Neurocomputing, Elsevier*, 71(13-15):2553–2562, 2008.

[39] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[40] G. CASTELLANO, S. VILLALBA, and A. CAMURRI. Recognising human emotions from body movement and gesture dynamics. *Affective Computing and Intelligence Interaction*, (LNCS 4738):71–82, 2007.

[41] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P Dario. Emotion modelling for social robotics applications: a review. *Journal of Bionic Engineering*, 15:185–203, 2018.

[42] I. CEARRETA, J.M. LOPEZ, and N. GARAY-VITORIA. Modeling multimodla context-aware affective interaction. *Proceedings of the Doctoral Consortium ACII07*, 2007.

[43] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority oversampling technique. *Journal Of Artificial Intelligence Research*, 16:321–357, 2002.

[44] J. CHEN, W. LIN, K. TSAI, and S. DAI. Analysis and evaluation of human movement based on laban movement analysis. *Tamkang Journal of Science and Engineering*, 13(3):255–264, 2011.

[45] T. CHIU, D. FANG, J. CHEN, Y. WANG, and C. JERIS. A robust and scalable clustering algorithm for mixed type attributes in large database environment. *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–268, 2001.

[46] Sangwoo Cho, Muhammad Hasan Maqbool, Fei Liu, and Hassan Foroosh. Self-attention network for skeleton-based human action recognition. *CoRR*, abs/1912.08435, 2019.

[47] G. CIMEN, H. ILHAN, and T. CAPIN. Classification of human motion based on affective state descriptors. *computer animation and virtual worlds*, 24(3-4):355–363, 2013.

[48] S. A. COOMBES, J. H. CAURAUGH, and C. M JANELLE. Dissociating motivational direction and affective valence. *Psychological Science*, 18:938–942, 2007.

[49] M COULSON. Attributing emotion to static body postures: Recognition accuracy, confusions and viewpoint difference. *Journal of Nonverbal Behaviour*, pages 28,117–139, 2004.

[50] R. COWIE and E. DOUGLAS. Feeltrace: An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop(ITRW) on speech and emotion.*, pages 5–7, 2000.

[51] EFRON D. Gesture and environments. *King Crown Press*, 1941.

[52] N. DAEL, M. MORTILLARO, and K. Scherer. The body action and posture coding system (bap):development and reliability. *Nonverbal behaviour*, 36:97–121, 2012.

[53] M. DAOUDI, S. BERRETI, P. PALA, Y. DELEVOYE, and A. Del Bimbo. Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices. *Image Analysis and Processing - ICIAP*, Lecture Notes in Computer Science, vol 10484. Springer, 2017.

[54] C. DARWIN. The expression of the emotions in man and animals, 1872.

[55] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz. Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. *EEE Sensors Letters*, 3:1–4, Jan 2019.

[56] R. DE SILVA and N. BIANCHI-BERTHOUZE. Modeling human affective postures:an information theoretic characteriation of posture features. *computer animation and virtual worlds*, 15(3-4):269–276, 2004.

[57] L. Deng and D. Yu. Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387, 2013.

[58] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *MIT Press*, 1:1486–1494, 2015.

[59] L. DEVILLERS and I VASILESCU. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. *International conference on spoker language processing*, 2006.

[60] A. Drachen, L. E. Nacke, G. N. Yannakakis, and A. L. Pedersen. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. *ACM SIGGRAPH*, (6):49–54, 2010.

[61] M. Duarte and S. M. S. F. Freitas. "revision of posturography based on force plate for balance evaluation. *Revista Brasileira de Fisioterapia*, 14(183-192), 2010.

[62] D. DUNCAN, S.and McNEILL and K. McCULLOUGH. How to transcribe the invisible and what we see. *special issue of code*, pages 75–94, 1995.

[63] P. EKMAN. Darwin, perception, and facial expression. *Analysis of the New York Academy of Sciences*, pages 105–221, 2003.

[64] P. EKMAN and W. V. FRIESEN. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98, 1969.

[65] N. Elfaramawy, P. Barros, G. Parisi, and S. Wermter. Emotion recognition from body expressions with a neural network architecture. *Proceedings of the 5th International Conference on Human Agent Interaction*, 17:143–149, October 2017.

[66] C.D. ELLIOTT. *The Affective reasoner: A process model of emotions in a multi-agent system*. PhD thesis, The institute for the learning sciences, Northwestern University, Evanston Illinois, 1992.

[67] M. ELLIS and C. MARCH. La danse noble, and inventroty of dances and sources. *Broude Brothers Ltd*, 1992.

[68] N.J. EMERY and D. G. AMARAL. *In Cognitive Neuroscience of Emotion*. Oxford University Press, 2000.

[69] N. ESHKOL and A. WACHMANN. Movement notation. *Weidenfield and Nicholson*, 1958.

[70] B. EVERITT. *Making Sense of Statistics in Psychology*. Oxford University Press, 1996.

[71] Steven Eyobu and Han DS. Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. *Sensors*, 18:2892, 2018.

[72] B. FASELA and J. LUETTIN. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[73] A. Ferdous, B. Hossain, and M.L. Gavrilova. Emotion recognition from body movement. *IEEE Access*, 8:11761–11781, 2019.

[74] J. FISS, A. AGARWALA, and B. CURLESS. Candid portrait selection from video. *SIGGRAPH Asia Conference*, 30(6):128:1–128:8, 2011.

[75] J. L. FLEISS. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[76] A. FOD, M. MATARIC, and O.C. JENKINS. Automated derivation of primitives for movement classification. *Autonomus Robots*, 12(1):39–54, 2002.

[77] N. FOURATI and C. PELACHAUD. Emilya: Emotional body expression in daily actions database. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation, ELRA*, 2014.

[78] S. FREY and J. POOL. A new approach to the analyses of visible behavior. *Bern: Forschungsberichte aus dem Psychologischen Institut*, 1976.

[79] N.H FRIJDA. The emotions. *Cambridge university press*, 1986.

[80] NEJAT. G., Y. SUN, and M. NIES. Assistive robots in health care settings. *Home health Care managment and practice*, 21(3):177–187, 2009.

[81] E.A. Hendriksa G.A. ten Holta, b M.J.T. Reindersa. Multi-dimensional dynamic time warping for gesture recognition. *Thirteenth annual conference of the Advanced School for Computing and Imaging*.

[82] V. GALLESE, C. KEYSERS, and G. RIZZOLATTI. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8(9):396–403, 2004.

[83] Y. GAO and R.L. MANDRYK. The acute cognitive benefits of casual exergame play. *In CHI '12: Proceedings of the 30th international conference on Human factors in computing systems., Austin, Texas, USA.*, pages 1863–1872, 2012.

[84] G. Garidakis, G. CASTELLANO, L. KESSOUS, A. RAOUZAIOU, L. MALATESTA, S. ASTERIADIS, and K. KARPOUZIS. Multimodal emotion recognition from expressive faces body gestures and speech. *International conference on Artificial Intelligence and applications*, (7):375–388, 2007.

[85] A. Gelman. Analysis of variance: Why it is more important than ever. *Annals of Statistics*, 33(1):1–53, 2005.

[86] K. GERLING, J. SCHILD, and M. MASUCH. Exergame design for elderly users: the case study of silverbalance. *7th International Conference on Advances in Computer Entertainment Technology (ACE '10). ACM, New York, NY, USA,*, pages 66–69, 2010.

[87] M.A. GIESE and T. POGGIO. Neural mechanisms for the recognition of biological movements. *Neuroscience*, 4:179–191, 2003.

[88] T. GIRAUD, O. LIMSI, T. GIRAUD, G. JAUREGUI, J. HUA, B. ISABLEU, E. FILAIRE, C. Le SCANFF, and J.C. MARTIN. Assessing postural control for affect recognition using video and force plates. *Affective Computing and Intelligent Interaction (ACII)*, pages 109–115, 2013.

[89] D. GLOWINSKI, N. DAEL, A. CAMURRI, G. VOLPE, M. MORTILLARO, and K. R. SCHERER. Towards a minimal representation of affective gestures. *IEEE Transactions on Affective Computing*, 2(2):106–118, 2011.

[90] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, and H. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):215–220, 2000.

[91] A. GOLDMAN and C. SRIPADA. Simulationist models of face-based emotion recognition. *Cognition*, 94(3):193–213, 2005.

[92] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. pages 2672–2680, 2014.

[93] HJ. GRIFFIN, MSH. AUNG, P. ROMERA, C. McLOUGHLIN, G. McKEOWN, W. CURRAN, and N. BIANCHI-BERTHOUZE. Laugher type recognition from whole body motion. *Proceedings of Affective Computing and Intelligent Interaction*, 2013.

[94] M.M GROSS, E.A. CRANE, and B.L. FREDRICKSON. Methodology for assesing bodily expression of emotion. *Journal of Nonverbal Behaviour*, 34:223–248, 2010.

[95] E.D. GROSSMAN and R. BLAKE. Brain areas active during visual perception of biological motion. *Neuron*, pages 1167–1175, 2002.

[96] H. GUNES and M. PANTIC. Automatic dimensional and continues emotion recognition. *International Journal of Synthetic Emotions*, 1(1):69–99, 2010.

[97] H. GUNES and M. PICCARDI. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30:1334–1345, 2007.

[98] H. GUNES and M. PICCARDI. Automatic temporal segment detection and affect recognition from face and body gestures. *IEEE Transactions on Systems, Man and Cybernetics*, 39(1):64–84, 2009.

[99] N. HADJIKHANI and B. GELDER. Seeing fearful body expressions activates the fusiform cortex and amygdala. *Current Biology*, 13(24):2201–2205, 2005.

[100] E. A. Garcia Haibo He, Yang Bai and Shutao L. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence*, pages 1322–1328, 2008.

[101] M. HALL, E. FRANK, and G. HOLMES. The weka data mining software an update. *SIGKDD Explorations*, 11(1), 2009.

[102] M. HARMS, A. MARTIN, and G. WALLACE. Facial emotion recognition in autism spectrum disorters: A review of behavioral and neuroimaging studies. *Neuropsychology*, (20):290–322, 2010.

[103] J. A. HARRIGAN, R. ROSENTHAL, and K. SCHERER. Proxemics, kinesics, and gaze. *The new handbook of methods in nonverbal behavior research*, pages 137–198, 2005.

[104] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[105] J. HEALEY and R. Picard. Detecting stress during real-world driving tasks. *IEEE Transactions on Inteligence Transportation systems*, 6(2):156–166, 2005.

[106] M. Hermans and B. Schrauwen. Training and analysing deep recurrent neural networks. *Advances in Neural Information Processing Systems*, 1:190–198, 2013.

[107] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6:107?116, 1998.

[108] W. Hua, F. Dai, L. Huang, J. Xiong, and G. Gui. Hero: Human emotions recognition for realizing intelligent internet of things. *IEEE Access*, 7:24321–24332, 2019.

[109] E. HUDLICKA. Affective game engines: motivation and requirements. *4th International Conference on Foundations of Digital Games*, pages 299–306, 2009.

[110] E. HUDLICKA. Guidelines for designing computational models of emotions. *International Journal of Synthetic Emotions*, pages 26–79, 2011.

[111] T. Huynh-The, C. Hua, and D. Kim. Encoding pose features to images with data augmentation for 3-d action recognition. *IEEE Transactions on Industrial Informatics*, 16:3100–3111, May 2020.

[112] M.H. IMMORDINO-YANG and A. DAMASIO. We feel, therefore we learn: The rel- evance of affective and social neuroscience to education. *Mind, Brain, and Education*, 1(3-10), 2007.

[113] K. ISBISTER, M. KARLESKY, J. FRYE, and R. RAO. Scoop!: using movement to reduce math anxiety and affect confidence. *the International Conference on the Foundations of Digital Games ACM, New York, NY, USA*, pages 228–230, 2012.

[114] K. ISBISTER, R. RAO, U. SCHWEKENDIEK, E. HAYWARD, and J. LIDASAN. Is more movement better?: a controlled comparison of movement-based games. *6th International Conference on Foundations of Digital Games ACM, New York, NY, USA*, pages 331–333, 2011.

[115] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[116] Brian Kenji Iwana and Seiichi Uchida. Time series data augmentation for neural networks by time warping with a discriminative teacher. *25th International Conference on Pattern Recognition*, 2021.

[117] L. JAMES. *Road Rage and Aggressive Driving*. Prometheus Books, 2000.

[118] J. JOSHI, R. GOECKE, G. PARKER, and M. BREAKSPEAR. Can body expressions contribute to automatic depression analysis? *Proceedings IEEE, International Conference Automatic Face and Gesture Recognition*, 2013.

[119] K. KAHOL, P. TRIPATHI, and S. PANCHANATHAN. Automated gesture segmentation from dance sequences. *Automatic Face and Gesture Recognition*, pages 883–888, 2004.

[120] H. KANG, C. LEE, and K. JUNG. Recognition based gesture spotting in video games. *Pattern Recognition Letters*, 25(15):1701–1714, November 2004.

[121] A. KAPUR, A. KAPUR, N. VIRGI-BABUL, G. TZANETAKIS, and P. DRIESSEN. Gesture-based affective computing in motion capture data. *ACII 2005, Beijing, China*, pages 1–17, 2005.

[122] M. KARG, ALI-AKBAR SAMADANI, R. GORBET, K. KUHNLENZ, J. HOEY, and D. KULIC. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4):341–359, 2013.

[123] H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Visual Computing*, 65:66–75, 2017.

[124] K. KIILI and S. MERILAMPI. Developing engaging exergames with simple motion detection. *14th International Academic MindTrek Conference: Envisioning Future Media Environments ACM, New York, NY, USA*, pages 103–110, 2010.

[125] Hyunbum Kim, Jalel Ben-Othman, Lynda Mokdad, Junggab Son, and Chunguo Li. Research challenges and security threats to ai-driven 5g virtual emotion applications using autonomous vehicles, drones, and smart devices. *IEEE Network*, 34(6):288–294, 2020.

[126] Y. Kim, H. Lee, and E. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3687–3691, May 2013.

[127] A. KLEINSMITH and N. BIANCHI-BERTHOUZE. Modeling non-acted affective posture in a video game scenario. *International conference Kansei Eng. Emotion*, 2010.

[128] A. KLEINSMITH and N. BIANCHI-BERTHOUZE. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2012.

[129] A. KLEINSMITH, N. BIANCHI-BERTHOUZE, and N. STEED. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics*, 41(4):1027–1038, 2011.

[130] G. KOCH. Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences*, pages 213–217, 1982.

[131] T. KODA. *Agents with Faces: A study on the effect of personification of software agents.* PhD thesis, MIT media lab, 1996.

[132] Michinari Kono Kohei Toyoda and Jun Rekimoto. Post-data augmentation to improve deep pose estimation of extreme and wild motions. *IEEE VR 2019 Workshop on Human Augmentation and its Applications, Osaka, Japan*, 2019.

[133] S. KOMLOSI, G. CSUKLY, G. STEFANICS, and I. GRIGLER. Fearful face recognition in schizophrenia: an electrophysiological study. *Schizophrenia*, (149):135–140, 2013.

[134] K. KRIPPENDORFF. Content analysis: An introduction to its methodology. *Thousand Oaks*, pages 221–250, 2013.

[135] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[136] Kamycki Krzysztof, Kapuscinski Tomasz, and Oszust Mariusz. Data augmentation with suboptimal warping for time-series classification. *Sensors*, page 3116–3124, 2020.

[137] R. LABAN. *The Language of Movement: A Guidebook to Choreutics*. Kalmbach Publishing Company, Plays, Boston, Mass, USA, 1974.

[138] J. LAFFERTY, A. MCCALLUM, and F. PEREIRA. Conditional random 17 fields: Probabilistic models for segmenting and labeling sequence data. *18th International Conference on Machine Learning*, pages 282–289, 2001.

[139] J.T. LARSEN, C.J. NORRIS, and J.T. CACIOPPO. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, 40(5):76–85, 2003.

[140] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[141] C. Ledig, L. Theis, F. Husz?ar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang. Photo-realistic single image super-resolution using a generative adversarial network. *Cornell University, Arxiv.org*, 2016.

[142] J. LEDOUX. The emotional brain: The mysterious underpinnings of emotional life. *Simon and Schuster*, 1998.

[143] C.M. LEE and S.S. NARAYANAN. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Processing*, 13(2):293–303, 2005.

[144] C.S. LEE and A. ELGAMMAL. Human motion synthesis by motion manifold learning and motion primitive segmentation. *Conference of Articulated Motion and Deformable Objects*, pages 464–473, 2006.

[145] H. LEE and J. KIM. An hmm-based threshold model approach for gesture recognition. *Pattern Analysis and Machine Intelligence*, 21(10):961–973, October 1999.

[146] Bo Li, Mingyi He, Xuelian Cheng, Yucheng Chen, and Yuchao Dai. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. *Cornell University, Arxiv.org*, 2017.

[147] Huadong Li and Hua Xu. Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems*, 204:106172, 2020.

[148] D. LITMAN and K. FORBES. Recognizing emotion from student speech in tutoring dialogues. *Proceedings of ASRU*, 2003.

[149] B Liu. Lifelong machine learning: a paradigm for continuous learning. *Frontiers of Computer Science*, 11:359–361, 2017.

[150] W. Liu, W. Zheng, and B. Lu. Multimodal emotion recognition using multimodal deep learning. *Cornell University - CoRR*, 2016.

[151] Y. LIU, O. SOURINA, and M. NGUYEN. Real-time eeg-based human emotion recognition and visualization. *International conference on Cyberworlds, Singapore*, (20-22):262–269, 2010.

[152] A. LONARE and S. JAIN. A survey on facial expression analysis for emotion recognition. *International journal on Advanced Research in Computer and Communication Engineering*, 2(12), December 2013.

[153] E. Loth, L. Garrido, and J. Ahmad. Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits? *Molecular Autism*, 9:7, 2018.

[154] F. LOTTE, M. CONGEDO, A. LE CUYER, F. LAMARCHE, and B. ARNALDI. A review of classification algorithms for eeg-based brain- computer interfaces. *Neural Eng.*, 4(R1-R13), 2007.

[155] T. LOURENS, R. BERKEL, and E. BARAKOVA. Communicating emotions and mental states to robots in a real time parallel framework using laban movement analysis. *Robot. Autonomus Systems*, 58(12):1256–1265, 2010.

[156] VERA MALETIC. Dance dynamics: Effort and phrasing, 2005.

[157] G. MANDLER. *Handbook of Psychology*, volume 1. John Wiley and Sons, 2003.

[158] Cabrera Maria, Eugenia and Wachs Juan, Pablo. Biomechanical based approach to data augmentation for one-shot gesture recognition. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[159] S. MARSELLA, S. CARNICKE, J. GRATCH, A. OKHMATOVSKAIA, and A. RIZZO. An exploration of delsarte's structural acting system. *6th International Conference, IVA*, 2006.

[160] Stacy C. Marsella, Sharon Marie Carnicke, Jonathan Gratch, Anna Okhmatovskaia, and Albert Rizzo. An exploration of delsarte's structural acting system. pages 80–92, 2006.

[161] M. MASUDA, S. KATO, and H. ITOH. Emotion detection from body motion of human form robot based on laban movement analysis. *Principles of Practice in Multi-Agent Systems, 12th International Conference, Nagoya, Japan*, 2009.

[162] M. MASUDA, S. KATO, and H. ITOH. Laban-based motion rendering for emotional expression of human form robots. *PKAW, LNAI*, (6232):49–60, 2010.

[163] N. Mavrides. A review of verbal and non-verbal human-robot interactive communication. *Journal of Robotics and Autonomous Systems*, 63:22–35, 2015.

[164] G. McINTYRE and R. GOECKE. Towards affective sensing. *HCII*, 2007.

[165] D. McNEILL. Gesture and thought. *Chicago: University of Chicago Press*, 2005.

[166] S. W. McQUIGGAN, B.W. MOTT, and J.C. LESTER. Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, 2008.

[167] N. Mehendale. Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2:446, 2020.

[168] A. MEHRABIAN. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[169] M MEIJER. The contribution of general features of body movement to the attribution of emotions. *Jornal of NonVerbal Behavior*, 13(4):247–268, 1989.

[170] A. METALLINOU, M. WOLLMER, A. KATSAMANIS, F. EYBEN, B. SCHULLER, and S. NARAYANAN. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012.

[171] MICROSOFT. Kinect controller for xbox360, http://www.xbox.com/en-us/kinect.

[172] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Hand gesture recognition with 3d convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–7, 2015.

[173] T. NAKATA, T. MORI, and T. SATO. Analysis of impression of robot bodily expression. *Journal of Robotics and Mechatronics*, (14):27–36, 2002.

[174] T. Q. Ngo and S. Yoon. Facial expression recognition on static images. *International Conference on Future Data and Security Engineering (Nha Trang: Springer)*, pages 640–647, 2019.

[175] R. OKA. Spotting method for classification of real world data. *The Computer Journal*, 41(8):559–565, July 1998.

[176] S. OKAJIMA, Y. WAKAYAMA, and Y. OKADA. Human motion retrieval system based on lma features using interactive evolutionary computation method. *Innovations in Intelligent Machines*, SCI(376):117–130, 2012.

[177] L. OMLOR and M.A. GIESE. Unsupervised learning of spatio-temporal primitives of emotional gait. *Perception and Interactive Technologies*, 4021:188–192, 2006.

[178] A. ORTONY, G. CLORE, and A. COLLINS. The cognitive structure of emotions. *Cambridge university press*, 1988.

[179] J. PANKSEPP. Affective neuroscience: The foundation of human and animal emotion. *Oxford University press*, 1998.

[180] M. PANTIC and L. J. M. ROTHKRANTZ. Automatic analysis of facial expressions:the state of the art. *IEEE Transaction on pattern analysis and machine intelligence*, 22:1424–1445, December 2000.

[181] M. PAPASTERGIOU. Exploring the potential of computer and video games for health and physical education: A literature review. *Computers and Education*, (53):603–622, 2009.

[182] G. Parisi and S. Wermter. Lifelong learning of action representations with deep neural self-organization. *AAAI Spring Symposium on Science of Intelligence: Computational Principles of Natural and Artificial Intelligence, Standford*, pages 608–612, 2017.

[183] K. PEARSON. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Sociaty of London*, (58):240–242, 1985.

[184] S. PIANNA, A. STAGLIANO, F. ODONE, and A. CAMURRI. A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. *IDGEI International Workshop*, 2013.

[185] S. PIANNA, A. STAGLIANO, F. ODONE, A. VERRI, and A. CAMURRI. Real-time automatic emotion recognition from body gestures. *Cornell University Library*, 2014.

[186] R. PICARD. Affective computing. *MIT press*, 1998.

[187] R. PLUTCHIK. Emotion, a psychoevolutionary synthesis. *Harper and Row*, 1980.

[188] F. E. POLLICK, H. M. PATERSON, A. BRUDERLIN, and A. J. SANFORD. Perceiving affect from arm movement. *Cognition*, 82(2):51–61, 2001.

[189] S. PRILLWITZ. Hamburg notation system for sign languages. an introductory guide. *International studies on sign language and communication of the deaf*, 5:46, 1989.

[190] A. QUATTONI, S. WANG, P. MORENCY, M. COLLINS, and T. DARREL. Hidden conditional random fields. *IEEE Transaction on pattern analysis and machine intelligence*, 29(10):1848–1852, 2007.

[191] J. RETT and J. DIAS. Computational laban movement analysis using probability calculus. *ROBOMAT*, 2007.

[192] G. RIZZOLATTI and L. CRAIGHERO. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

[193] J.L. RODGERS and W.A. NICEWANDER. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, February 1988.

[194] C. ROETHER, L. Omlor, A. CHRISTENSEN, and M.A. GIESE. Critical features for the perception of emotion from gait. *Journal of Vision*, 8(6:15):1–32, 2009.

[195] C. ROETHER, L. OMLOR, and M.A. GIESE. Lateral asymmetry of bodily emotion expression. *Current Biology*, 18(8):329–330, 2008.

[196] Gregory Rogez and Coredlia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3116–3124, 2016.

[197] I. ROSEMAN and A. EVDOKAS. Appraisals cause experienced emotions: Experimental evidence. *Cognition and Emotion*, (18):1–28, 2004.

[198] ALI-AKBAR SAMADANI, SARAHJANE BURTON, ROB GORBET, and DANA KULIC. Laban effort and shape analysis of affective hand and arm movements. *Affective Computing and Intelligent Interaction (ACII)*, pages 343–348, 2013.

[199] L. SANTOS, J. PRADO, and J. DIAS. Human robot interaction studies on laban human movement analysis and dynamic background segmentation. *IEEE international conference on robots and systems*, 2009.

[200] T. Sapinski, D. Kamiska, A. Pelikant, and G. Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21(7):646, 2019.

[201] N. SAVVA and N. BIANCHI-BERTHOUZE. Automatic recognition of affective body movement in a video game scenario. *International conference on intelligent technologies for interactive entertainment*, 78:149–158, 2012.

[202] N. SAVVA, A. SCARINZI, and N. BIANCHI-BERTHOUZE. Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(3), 2012.

[203] Aleksander Sawicki and Slawomir Zielinski. Augmentation of segmented motion capture data for improving generalization of deep neural networks. *Saeed K., Dvorsky J. (eds) Computer Information Systems and Industrial Management*, 12133, 2020.

[204] S. SCHERER, G. STRATOU, M. MAHMOUD, J. BOBERG, J. GRATCH, A. RIZZO, and LP. MORENCY. Automatic behaviour descriptors for psychological disorder analysis. *Proceedings IEEE, International Conference Automatic Face and Gesture Recognition*, 2013.

[205] S. J. SCHOUWSTRA and J. HOOGSTRATEN. Head position and spinal position as determinants of perceived emotional state. *Perceptual and motorl skills*, 81(2):673–674, 1995.

[206] A. Serway Raymond and John W. Jewett. *Physics for Scientists and Engineers*. Thomson-Brooks/Cole, 6th ed. belmont edition, 2004.

[207] W.A. SETHARES and T.W. STALEY. Periodicity transforms. *IEEE Transactions on Signal Processing*, 47(11):2953–2964, 1999.

[208] J. Shi, C. Liu, C.T. Ishi, and H. Ishiguro. Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network. *Sensors 2021*, 21:205, 2019.

[209] C. Shorten and T. M. Khoshgoftaaw. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 2019.

[210] I. SIEGERT, R. BOCK, and A. WENDEMUTH. The influence of context knowledge for multimodal affective annotation. *Human Computer Interaction*, V(8008):381–390, 2013.

[211] K. Simonyan and A. Zesserman. Very deep convolutional networks for large-scale image recognition. *Computer Vision and Pattern Recognition*, 2014.

[212] T. STARNER and A. PENTLAND. Visual recognition of american sign language using hidden markov models. Master's thesis, Massachusetts Institute of Technology, 1995.

[213] V. SUTTON. Dancewriting shorthand for modern and jazz dance. *Center Sutton Movement*, 1982.

[214] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z Wojna. Rethinking the inception architecture for computer vision. *arXiv, cs.CV*, 2015.

[215] B. G. Tabachnick, L. S. Fidell, and L. S. Ullman. Using multivariate statistics. *USA:Pearson*, 5, 2007.

[216] H. Thien, H. Cam-Hao, N. Trung-Thanh, and D. Dong-Seong. Image representation of pose-transition feature for 3d skeleton-based action recognition. *Information Sciences*, page 513, 2019.

[217] J. L. TRACY and R. W. ROBINS. The prototypical pride expression: Development of a nonverbal behavior coding system. *Emotion*, 7(4):789–801, 2007.

[218] H. Liu Tu, F. Meng, M. Liu, and R. Ding. Spatial-temporal data augmentation based on lstm autoencoder network for skeleton-based human action recognition. *25th IEEE International Conference on Image Processing (ICIP)*, pages 3478–3482, 2018.

[219] T. Yang V. Sze, Y. Chen and J. S. Emer. Efficient processing of deep neural networks: A tutorial and survey. *in Proceedings of the IEEE*, 105:2295–2329, 2017.

[220] MF. VALSTAR, H. GUNES, and M. PANTIC. How to distiguish posed from spontaneous smiles using geometric features. *ACM International Conference on multimodal interfaces (ICMI07)*, pages 38–45, 2007.

[221] E. VELLOSO, A. BULLING, and H. GELLERSEN. Autobap: Automatic coding of body action and posture units from wearable sensors. *5th Humaine Association Conference on Affective Computing*, 2013.

[222] D. VERVERIDIS and C. KOTROPOULOS. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(8):1162–1181, 2006.

[223] A VICENTE and H. PAIN. Informing the detection of students motivational state: an empirical study. *Proceedings of the 6th International Conference on Intelligent Tutoring Systems, Springer*, pages 933–943, 2002.

[224] E. VOLKOVA, S. DE LA ROSA, H.H. BULTHOFF, and B. MOHLER. The mpi emotional body expressions database for narrative scenarios. *PLoS ONE*, 9(12), 2014.

[225] Hodges W. and Spielberger C. The effects of threat of shock on heart rate for subjects who differ in manifest anxiety and fear of shock. *Psychophysiology*, pages 287–294, 1966.

[226] Siegman A. W. and Boyle S. Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear and anxiety and sadness and depression. *Journal of Abnormal. Psychology*, pages 102–430, 1993.

[227] R. WALK and K. WALTERS. Perception of the smile and other emotions of the body and face at different distances. *Bull. Psychonomic Soc.*, 26:510, 1998.

[228] H. WALLBOTT. Bodily expressions of emotion. *European journal of social psychology*, 28(879-896), 1998.

[229] W. WANG, V. ENESCU, and H. SAHLI. Towards real-time continues emotion recognition from body movements. *Human Behaviour Understanding*, 8212:235–245, 2013.

[230] W. Weiqing, X. Kunliang, N. Hongli, and M. Xiangrong. Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. *Complexity*, (4065207), 2020.

[231] L. WEISKRANTZ. Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *Journal of Comparative and Physiological Psychology*, 49(4):381–391, 1956.

[232] Qingsong Wen, Liang Sun, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv*, 2020.

[233] C WHISSEL. The dictionary of affect in language. *Emotion: Theory, Research and Experience: The Measurement of Emotions*, 4:113–131, 1989.

[234] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal. Video emotion recognition with transferred deep feature encodings. *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 15–22, 2016.

[235] Georgios N. Yannakakis and John Hallam. Ranking vs. preference: A comparative study of self-reporting. pages 437–446, 2011.

[236] G.N. YIANNAKAKIS, H. HALLAM, and H. LUND. Entertainment capture through heart rate activity in physical interactive playgrounds. *User Modeling and User-Adapted Interaction*, 18(1-2):207–243, 2008.

[237] H. ZACHARATOS, C. GATZOULIS, Y. CHRYSANTHOU, and A. ARISTIDOU. Emotion recognition for exergames using laban movement analysis. *ACM SIGGRAPH conference on Motion in Games*, pages 61–66, 2013.

[238] Z. ZENG, M. PANTIC, G. ROISMAN, and T.S. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Transaction on pattern analysis and machine intelligence*, 31:39–58, 2009.

[239] X. ZHANG, B. HU, J. CHEN, and P. MOORE. Ontology-based context modeling for emotion recognition in an intelligent web. *World Wide Web*, 16(4):497–513, 2013.

[240] L. ZHAO. *Synthesis and acquisition of laban movement analysis qualitive parameters for communicative gestures*. PhD thesis, University of Pennsylvania, 2002.

[241] L. ZHAO and N. BADLER. Acquiring and validating motion qualities from live limb gestures. *Graphical Models*, 67(1):1–16, 2005.

[242] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[243] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.