2022-11-23

Bi-Modal Deep Neural Network for Gait Emotion Recognition

Bhatia, Yajurv

UNIVERSITY OF CALGARY

Bi-Modal Deep Neural Network for Gait Emotion Recognition

by

Yajurv Bhatia

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA

NOVEMBER, 2022

Abstract

Emotion Recognition systems can be used for autonomous tasks such as video gaming experiences, medical diagnosis, adaptive education, and smart homes. Several biometric modalities, including face, hands, and voice have been successfully used for emotion recognition tasks. Gait Emotion Recognition (GER) is an emerging domain of research that is focused on identifying the emotional state of a person from gait biometric, which represents the person's manner of walking. In comparison to the other modalities, gait provides a non-intrusive method to collect data remotely without an expert's supervision. Moreover, unlike facial expression-based emotion recognition, it does not require high-resolution data for inference. Early works in GER produced limited feature sets and used classical machine learning methodologies to infer emotions, but could not achieve high performance. This thesis proposes powerful architectures based on deep-learning to accurately identify emotions from human gaits. The proposed Bi-Modal Deep Neural Network (BMDNN) architecture utilizes robust handcrafted features that are independent of dataset size and data distribution. The network is based on Long Short-Term Memory units and Multi-Layered Perceptrons to sequentially process raw gait sequences and facilitate feature fusion with the handcrafted features. Lastly, the proposed Bi-Modular Sequential Neural Network (BMSNN) has a low number of parameters and a low inference time, hence making it suitable for deployment in real world applications. The proposed methodologies were evaluated on the Edinburgh Locomotive MoCap Dataset and outperformed all recent state-of-the-art methods.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Marina Gavrilova, who guided me through the various challenges I faced. Her unwavering faith in me helped me achieve my goals and successfully conclude my work for my master's degree in Computer Science at the University of Calgary. I am extremely thankful that I had the chance to work under the mentorship of an accomplished, knowledgeable, talented, kind and understanding individual. I would also like to thank my colleague, ASM Hossain Bari, whose feedbacks, comments, and discussions helped shaped my work and publications. Next, I would like to thank my parents and my sister for believing in me when I failed to do so myself and for motivating me to strive. I would also like to extend my gratitude towards my friends who kept me focused and helped me through tough times. During my time in the Master of Science program I was surrounded by incredible people who continue to inspire me every day.

Table of Contents

List of Figures

List of Tables

List of Symbols, Abbreviations, and Nomenclature

Abbreviations	Definition
GER	Gait Emotion Recognition
BMDNN	Bi-Modal Deep Neural Network
BMSNN	Bi-Modular Sequential Neural Network
mAP	mean Average Precision
DL	Deep Learning
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
MLP	Multi Layer Perceptron
DNN	Deep Neural Networks
JRA	Joint Relative Angles
JRD	Joint Relative Distances
DWT	Discrete Wavelet Transform
EEG	Electroencephalogram
ECG	Electrocardiogram
EDA	Electrodermal Activity
EMG	Electromyogram
BVP	Blood Volume Pulse
DC	Direct Current
GSR	Galvanic Skin Response
SCR	Skin Conductance Response
PPG	Photoplethysmogram

EOG	Electrooculography
PCA	Principal Component Analysis
DOF	Degrees Of Freedom
ANOVA	Analysis Of Variants
MANOVA	Multivariate Analysis Of Variants
SVM	Support Vector Machine
NB	Naïve Bayes
RF	Random Forest
DT	Decision Tree
LR	Linear Regression
ANN	Artificial Neural Network
LMA	Laban Movement Analysis
GA	Genetic Algorithm
HOC	Higher Order Crossing
PSD	Power Spectral Density
FER	Facial Emotion Recognition
CNN	Convolutional Neural Network
FT	Fourier Transform
POS	Part Of Speech
SOM	Self Organizing Maps
GCNN	Graph Convolutional Neural Network
ADF	Affective and Deep Features
HAPAM	Hierarchical Attention Pooling and Affective Mapping
SGD	Stochastic Gradient Descent
ADAM	Adaptive Moment estimation
RMSprop	Root Mean Square propagation
STGCN	Spatial Temporal Graph Convolutional Network
STEP	STGCN for Emotion Perception

Chapter 1

Introduction

1.1 Motivation

Recognizing emotions of a person has fueled several human-related autonomous tasks [27]. For instance, emotion recognition from gait has been adopted in smart home design for fall prevention, in disaster management during evacuation [59] and can be applied in medicine for detection and treatment of Parkinson's disease [201]. Gait Emotion Recognition (GER) can be beneficial for security and privacy as well [60, 157]. It can be deployed to observe individuals for suspicious behavior at national or state borders [59] (Figure 1.1a). It can also be used for developing emotionally aware robots [59] (Figure 1.1b) for medicine or smart homes. Virtual reality and gaming experiences can potentially be curated according to the subject's emotional state [63].

(a) A camera on a lookout tower in Poland [42]

(b) A Honda ASIMO robot at the Miraikan museum of emerging science and innovation [119]

Figure 1.1: Areas of application for Gait Analysis and Emotion Recognition

Human emotions can be inferred through various biometric modalities, such as facial expressions [104],

gait [186, 4, 3], hand gestures [65], voice tones [114], and text [184]. The phenomenon of expressing emotions through human body movements has been observed from the early 1900s [31, 127, 41]. Although, using a person's facial expressions has been the prevalent emotion recognition method, gait-based emotion recognition systems have started gaining popularity [59]. The increase in the number of research works being conducted for gait emotion recognition can be attributed to the numerous benefits it provides over other modalities. In comparison to other modalities, gait provides a non-intrusive way to remotely collect data and to infer emotions from a distance.

1.2 Problem Statement

A person's walk can be described by the sequence of their body joint positions over a period of time. In the literature, this is referred to as the gait sequence of the person (see Figure 1.2). Every individual has a certain uniqueness associated with their gait [137] which allows it to be used for person identification (termed gait recognition). Gait also provides descriptive information about the emotional state of the subject [31]. This information can be recorded and processed to infer the emotion being experienced by the subject. This is called Gait Emotion Recognition (GER). While there have been many works in gait recognition, GER had not been studied till recently. GER works are applicable in a plethora of industries including video games, medicine, and education. Hence, it is imperative to develop a methodology for identifying emotions from gait effectively.

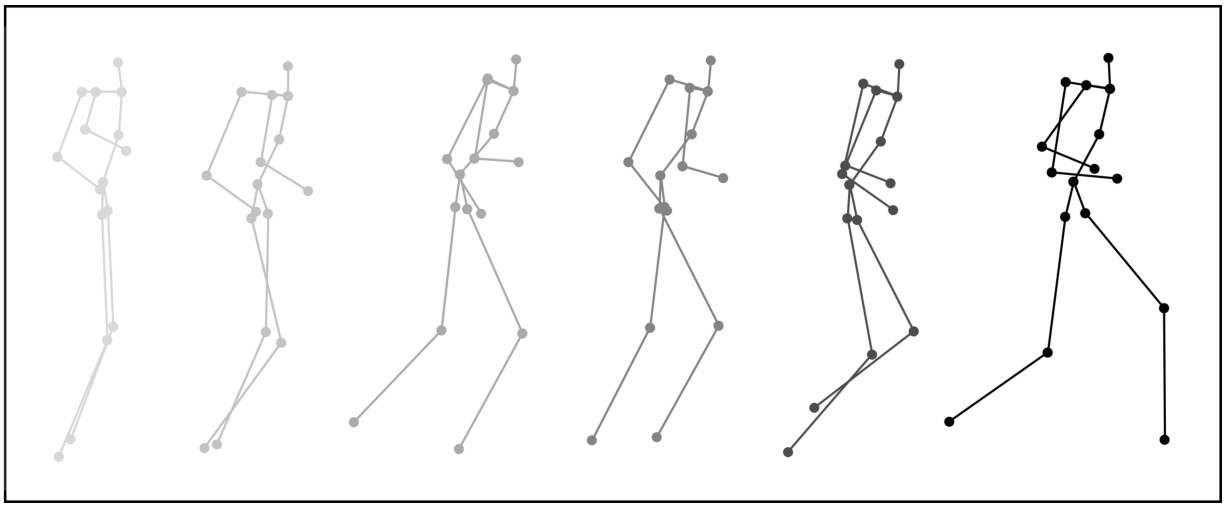


Figure 1.2: An example of human gait as sequence of body joint positions over time

The preliminary works of Gait Emotion Recognition (GER) relied on limited sets of features extracted manually from the gait sequences. These feature sets were usually processed by a classical machine learning

algorithm to infer the emotions [87, 176, 105]. The problem of exploring a limited set of features was addressed by the recent works in the domain by adopting Deep Learning (DL) approaches. The deep neural networks facilitated automatic feature extraction, selection, and classification. However, these models had a high number of network parameters which resulted in slow processing of the emotions from gait. Moreover, the previously proposed handcrafted features were discarded but a robust set of discriminating features was not extracted by the DL methods.

As research progressed, various DL methods were proposed to address the above drawbacks. Graph and pseudo-image-based methodologies benefited from inherent structural information of the human body but failed to explore low-level features between vertices placed far away from each other. Some works based on Recurrent Neural Networks (RNN) processed gaits as sequences to explore all dependencies but used ineffective training methodologies or a sub-optimal network. This thesis proposes powerful neural networks that process gaits sequentially and are trained on real gaits for emotion recognition. The following research questions are addressed in this work:

1. Can an attenuated hybrid deep learning architecture be devised to achieve low inference time for gait emotion recognition?

2. Can a deep learning architecture for gait-based emotion recognition be designed to identify distinctive sequential and temporal features extracted from body joints?

3. Can a light deep learning architecture combining a sequential neural network and multi-layered perceptrons be used to accurately recognize emotions from human gaits?

4. Can handcrafted features based on the geometric relationships between body joints be combined with the deep learning architecture to further improve recognition performance and to make the architecture resilient to class imbalance in the dataset?

5. Can domain-specific handcrafted features be fused with latent deep features to improve gait emotion recognition performance?

6. How do the Laban Movement Analysis feature groups affect the performance of the proposed network?

1.3 Objectives

This thesis aims to develop a DL-based neural network trained on real human gaits to accurately infer emotions from gait. Hence, the thesis aims to accomplish the following goals:

1. The deep learning architecture must be designed to identify distinctive sequential and temporal features extracted from body joints.

2. The deep learning architecture must incorporate dynamic handcrafted features based on the geometric relationships between body joints to improve recognition performance.

3. The deep learning architecture must have fewer number of parameters to ensure low inference times while ensuring high recognition performance.

4. The handcrafted features must be processed with latent deep features to improve gait emotion recognition performance.

5. The deep learning architecture must be resilient to class imbalance in the dataset.

To accomplish the above goals, a novel sequential neural network based on Long Short Term Memory (LSTM) and Multi Layered Perceptrons (MLP) will be introduced in this thesis.

1.4 Challenges and Limitations

The development of the desired GER system posed numerous challenges that were overcome during research:

1. **Lack of data:** One of the disadvantages of using DL methodologies is the high data requirement. Since the domain of gait emotion recognition is fairly new, there is a lack of large emotionally labelled gait datasets. This forced researchers to combine multiple datasets or even produce artificial gait sequences to obtain samples for model training [20, 129, 144, 21].

2. **Dataset Variance:** Factors like the number of body joints recorded in the dataset, the positioning of body markers in the motion capture suit, or the position of body joints detected by pose estimation algorithms from a video can introduce unwanted noise to the dataset. Consequently, training a model on such a dataset would result in a high variance model that learns features irrelevant to the task [124]. Hence the models would not perform well. The proposed model resolves this issue by eliminating the requirement of large datasets. This is done by ensuring that the number of trainable parameters is a fraction of what the recent state-of-the-art methods have. Additionally, the model utilizes robust handcrafted features which limits the reliance of the model on a large number of data samples.

3. **High Number of Parameters:** The system is required to consolidate the information extracted from the raw gait sequences and the handcrafted features. In Deep Neural Networks (DNN), higher number of parameters can be used in a layer to produce more features at that stage, and more layers

can be added to produce higher level features in later stages [132, 193]. However, the objective of this research is to build a neural network capable of extracting high-level features from gait sequences while keeping the number of parameters low. This issue was overcome by designing a tapered neural network. The attenuated design ensured that the new layers being added had equal or lesser number of units. The design also ensured that the information extracted by a layer in the network is represented using smaller feature vectors in the subsequent layers; eventually leading the network to develop a condensed high-level feature vector towards the end.

4. **Feature sets:** Previous works had explored handcrafted features to a limited extent [87, 175, 144, 20]. The features sets used were restricted to a handful of angle and distance measures using a few body joints. They did not include crucial metrics that are beneficial for emotional analysis. This work uses all Joint Relative Angles (JRA) and Joint Relative Distances (JRD) to ensure that all possible angle and distance measures are considered while inferring emotions.

5. **Class imbalance:** The reliance on purely DL methodologies caused models to become susceptible to distribution of the classes in the dataset. Models using data-driven approaches performed poorly for under represented classes. In this thesis, the problem was mitigated with the help of robust discriminating Laban Movement Analysis-based (LMA) handcrafted features. The handcrafted features are calculated from defined functions unaffected by data distribution and hence provided robust information about the subject's motion.

6. **Neural network development:** Developing an effective Long Short Term Memory (LSTM) and Multi Layered Perceptron (MLP) based neural network involved numerous iterations of designing, testing and modifying the structure of the network. To prevent overfitting without compromising on the high performance, various techniques such as kernel regularization and Batch Normalization were applied.

1.5 Contributions

The thesis contributions can be outlined as follows:

1. A novel hybrid deep learning architecture is presented that utilizes Long Short-Term Memory (LSTM) units followed by Multi-Layered Perceptrons (MLP) to extract a distinctive feature map from raw gaits and recognize four emotions, namely happy, angry, sad, and neutral. (Published in IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 2021 [17])

2. A novel Bi-Modular Sequential Neural Network (BMSNN) with significantly low number of parameters compared to prior research is introduced. The network achieves the lowest inference time among recent state-of-the-art deep learning-based GER methods. (Published in MDPI Sensors, 2022 [19])

3. The proposed BMSNN network utilizes geometric features Joint Relative Angle (JRA) and Joint Relative Distance (JRD) for Gait Emotion Recognition (GER). (Published in MDPI Sensors, 2022 [19])

4. A novel fusion of robust LMA-based domain-specific handcrafted features with latent features extracted from a deep neural network is proposed. (Published in the 17th International Symposium on Visual Computing (ISVC), 2022 [18])

5. A novel powerful Bi-Modal Deep Neural Network (BMDNN) to facilitate the combination and processing of the handcrafted features with latent deep features is introduced. (Published in the 17th International Symposium on Visual Computing (ISVC), 2022 [18])

6. An ablation study of the proposed BMDNN deep learning architecture to validate the performance on an imbalanced dataset of real gait samples and the importance of the introduced feature fusion is performed. A comprehensive analysis of GER performance with respect to different LMA-based feature groups is also conducted. (Published in the 17th International Symposium on Visual Computing (ISVC), 2022 [18])

The research presented in this thesis introduces methodologies that are superior to the previous works in the domain. The proposed BMSNN architecture exhibits the highest overall emotion recognition performance while inferring a gait sample within a fraction of the time taken by prior methods. Additionally, the proposed BMDNN architecture outperforms the state-of-the-art methods, including BMSNN, in an overall emotion recognition, and provides a robust method to infer emotions for each individual class, even if it is underrepresented in the dataset. Moreover, the experimental results presented in Chapter 4, describe the relation of Laban Movement Analysis with the model performance for each emotion class.

Hence, the thesis extends the work in the domain of Gait Emotion Recognition by providing a faster processing technique that could enable mobile deployment, and processing on less powerful systems. Secondly, it provides a more accurate method for identifying emotions to build reliable emotion recognition systems that needs to identify rare occurrences. Finally, the exploration of the effects of Laban Movement Analysis Features opens up the opportunities for the development of more informed emotionally aware robots, and improves understanding of exhibited emotions for Human Computer Interactions (HCI).

1.6 Thesis Outline

The remainder of the thesis is organized into the following sections. Chapter 2 provides an overview of the previous work in this domain. First, various types of biometrics are discussed in terms of emotion recognition systems. Next, the thesis describes the various emotion representation models in the literature and how the ground truth can be defined for emotional gaits. Finally, the methods previously developed for gait emotion recognition are explored. Chapter 3 describes the proposed methodologies and the intuitions behind the design of the proposed architectures. Chapter 4 presents the results of extensive experimentation, involving the development of the proposed deep neural network, its performance without crucial components, and its performance in comparison to state-of-the-art methods. In Chapter 5, a summary of the contributions of this thesis is provided along with the limitation and the future directions of this research work.

Chapter 2

Literature Review

This section will provide an overview of the key concepts and the previous research works in the domain of Gait Emotion Recognition (GER). Subsection 2.1 includes a discussion on the various types of biometrics and their usage for emotion recognition tasks. Next, emotion representation models are described in Subsection 2.2. Preliminary works in the domain relied on manually selected handcrafted feature sets and classical machine learning algorithms. Only recently, the methodology shifted towards deep learning methods to address the shortcomings of the classical machine learning methods. Subsection 2.3 summarizes previous approaches used for GER and highlights their advantages and disadvantages.

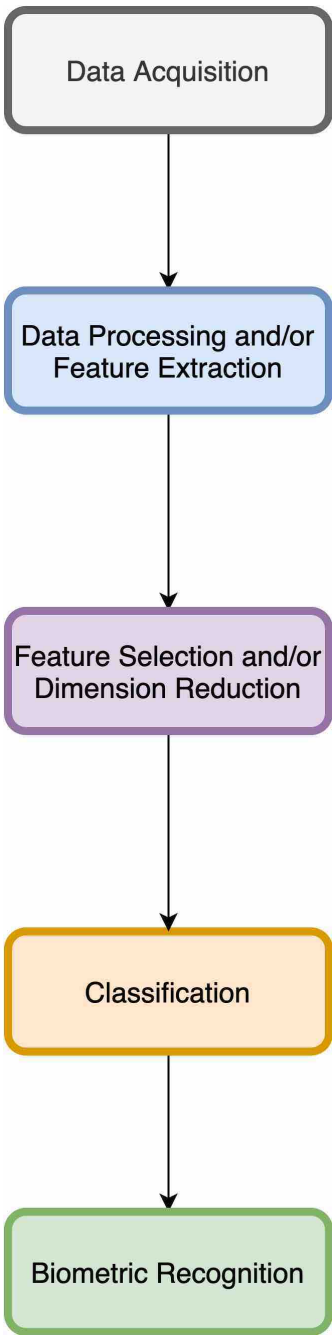
(a) Adaptive Rehabilitation [159]

(b) Smart Homes [153]

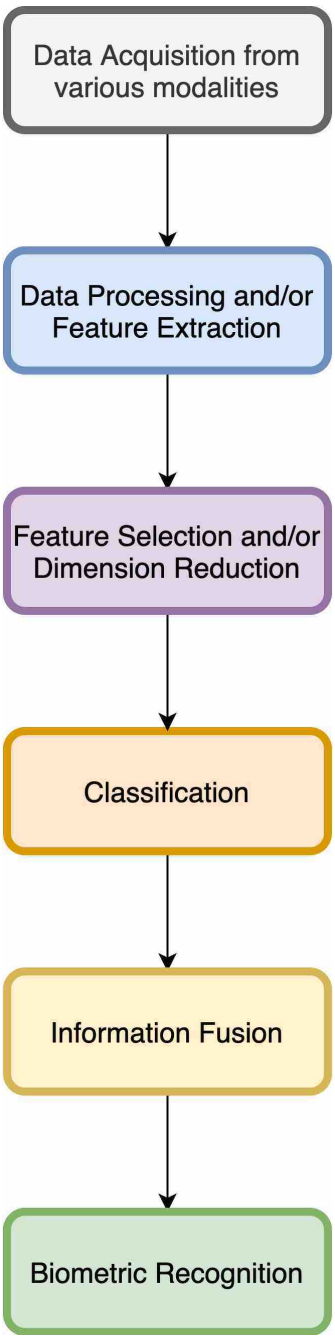
Figure 2.1: Areas of application for an Emotion Recognition system

2.1 Overview of Emotion Recognition Systems

Emotion recognition systems analyze and process data collected from one or more modalities to provide important information for the interpretation of a subject's emotional state [158]. This information is crucial



(a) A general framework for unimodal biometric systems



(b) A general framework for multi-modal biometric systems

Figure 2.2: General frameworks for biometric systems

for understanding human interactions with other humans or with computer systems. Moreover, numerous biometric modalities are suitable for emotion recognition and have applications in industries like gaming, education, advertisement, automotive industry, healthcare, and smart home design (Figure 2.1). Biometric data can be used to identify the emotional state of a person [76]. Some physiological biometrics provide a non-invasive way to collect data for identifying emotions. However, they can sometimes be intrusive and might elicit hesitance. On the other hand, behavioral biometrics are collected remotely, and provide a comfortable experience for the user. Social Behavioral Biometrics further ease the data acquisition by observing online user behavior. Biometrics have been predominantly researched for user authentication and identification systems, with very limited approaches for emotion recognition.

Based on the number of modalities utilized, a biometric system can be classified as one of two types: Unimodal or Multimodal. Most biometric systems have a general framework that consists of four main steps: collecting the required data, processing it to extract features, selecting all or a subset of the features, and using the final feature set to classify emotions (see Figure 2.2a). Multimodal biometric systems, as the name suggests, rely on two or more biometric modalities. To take advantage of different types of data, these systems either merge the information derived from the modalities, or combine the classifications resulting from each modality, or both. Hence, multimodal biometric systems contain additional fusion module (see Figure 2.2b). For most biometrics, the early research revolved around optimizing feature extraction and selection methods using classical machine learning. With the new trends, the works in biometric domain slowly shifted to adopt Deep Neural Network (DNN) based methodologies.

2.1.1 Physiological Biometrics

Physiological Biometric traits refer to the subset of biometrics that are intrinsic properties of the human body. Size and shape of the ear, print of the palm, shape and structure of the face, and fingerprints are some of the commonly used physiological biometrics that barely change during the lifetime of a person [8]. Although the task of emotion recognition is highly associated with the behavior of the subject, there are certain physiological biometrics that are capable of expressing emotion.

Electroencephalogram (EEG)

The electronic signals of a subject's brain can be recorded through electrodes placed on their scalp. These signals are representative of the brain activity of the person and hence can be used for emotion recognition. In the early 2000s, statistical features were proposed to encapsulate the information of the EEG signals [138, 169]. Later, the domain witnessed Discrete Wavelet Transform (DWT) based features [128], Higher

Order Crossing (HOC) based features [136], and asymmetry index and power spectral density (PSD) based features [108]. In 2011, the authors of [111] presented a fractal dimension based algorithm for recognizing emotion through EEG in real-time. Recent works in the field have been directed towards using more features than prior methods either through combining previously proposed features [106] or via Deep Neural Networks (DNN) [163].

Electrocardiogram (ECG)

ECG is the process of measuring electrical signals of the heart using electrodes attached to the outer skin of the subject's thorax. ECG can be processed to extract relevant features including heart rate which varies with emotions [49, 26]. Much like any other bio-signal, ECG can be used to extract basic statistical features [90], DWT features [84], frequency domain features [179], or a combination of them all [187]. While most studies focused on optimizing the feature extraction, works like [187, 188] proposed improvements to feature selection methods. In the 2010, the domain witnessed further advancement in feature extraction methods [1, 155, 32]. Research prior to 2017 utilized classical machine learning methods for final classifications. Deep learning methodologies have only recently been incorporated for recognizing emotions using ECG [91].

Electrodermal Activity (EDA)

Electrodermal activity, also known as galvanic skin response (GSR), is another commonly used physiological biometric for emotion recognition. This metric measures the conductivity of the skin of the subject which is induced by the arousal elicited by emotions. Consisting of two main components: the Direct Current (DC) level of the skin and the Skin Conductance Response (SCR), EDA metrics can be used to extract features using statistical analysis [138, 95, 94]. Furthermore, it is often combined with other signal based physiological biometrics like ECG and EMG [148, 68, 103, 130].

Electromyogram (EMG)

Measuring electrical activities from muscles can also be beneficial for emotion recognition. DWT-based features extracted from EMG signals have been previously used to identify emotions [34, 200]. Most of the research in emotion recognition from EMG typically uses it in combination with other metrics with classical machine learning techniques [95, 94, 145, 97]; however, there has been some attempts based on deep neural networks [190] that used only EMG.

Blood Volume Pulse (BVP)

Blood Volume Pulse (BVP) refers to the amount of blood flowing through a vessel at a given time. BVP is usually measured using a Photoplethysmogram (PPG). Like most other physiological traits, this biometric has been predominantly used to produce statistical features for classification using classical machine learning models [68, 93, 116].

Respiration

The variation in breathing can be indicative of the emotional state of a subject [74]. However, research using this modality has been limited. Studies using respiration for emotion recognition systems have relied on statistical metrics to extract features and on classical machine learning methods for classification [148, 66].

2.1.2 Behavioral Biometrics

Behavioral biometrics are defined as the characteristics, patterns, mannerisms and other properties of a human activity [147]. A person's voice tones, unique signature, banking activity, and visual auditory preferences fall under the category of behavioral biometrics. Behavioral biometrics such as gait, gestures, and voice have been utilized for user authentication in the past [167]. Behavioral biometrics have certain advantages over conventional physiological biometrics; they provide a non-obtrusive way of collecting biometric data, that is difficult to circumvent, more acceptable, and can be integrated with existing technologies. Moreover, behavioral biometrics can be used on a regular basis for continuous authentication instead of explicit authentication checks [156].

Facial Expressions

User identification through facial structures has been extensively researched and is used in the industry today. Although the shapes and sizes of various facial features are physiological metrics, the changes in those features are behavioral traits. Facial expressions have been shown to be a reliable and accurate source for emotion recognition [96]. However, it requires specially recorded data such as 3D infrared images to overcome pose and illumination based variations, or high frame-rate videos to capture micro emotions [96]. Figure 2.3 shows a few samples from the extended Cohn-Kanade Dataset for facial emotion recognition.

Facial Emotion Recognition (FER) systems usually rely on features derived from facial landmarks, which are important points on the subjects face (such as tip of the nose, ends of the mouth, etc.). The authors of [62] used geometric features consisting of the changes in the angles and distances between various facial landmarks. Works like [71] used global appearance-based features while some studies [61] focused on combin-

Figure 2.3: Samples from the extended Cohn-Kanade Dataset (CK+) [113])

ing individual regional facial features instead. Research combining geometric and appearance based features were also published in the field [52, 61].

Deep learning-based works in FER primarily rely on Convolutional Neural Networks (CNN). The authors of [25] studied a model trained to recognize facial emotions to establish that the domain could benefit from Deep Neural Networks. Subsequently, many studies using CNN for temporal and geometric feature extraction [86] and multi label training [198] were published. Sequential models were also used in combination with CNNs to capture the temporal information more efficiently [48, 92, 39, 72, 67, 80].

Speech

Speech signals have been used in unimodal [162] and multimodal [7] user recognition systems and is another biometric that facilitates identification of emotions in humans [24]. Features such as excitation signals [191, 14], vocal tract features [161, 22], and prosodic features [182, 195] are representative of the emotional state of the subject. Signal processing techniques such as Fourier Transform (FT) [161], and Mel-frequency cepstral coefficients [131] are prevalent in the domain. Works such as [75, 24] have also attempted emotion recognition using a combination of the features mentioned earlier. Figure 2.4 shows an example of a speech spectrogram, which can be used to extract features for emotion recognition.

Until the mid 2010s, most works in the domain utilized explicit feature extraction methods and traditional machine learning classifiers to identify emotions using those features [139, 131, 178]. However, in recent years, the field has witnessed numerous works employing deep neural networks. Combinations of CNNs and RNNs have been a popular approach [197, 171, 192]. Though not as popular, combinations of traditional machine learning and deep learning methods were also explored by researchers [53, 166].

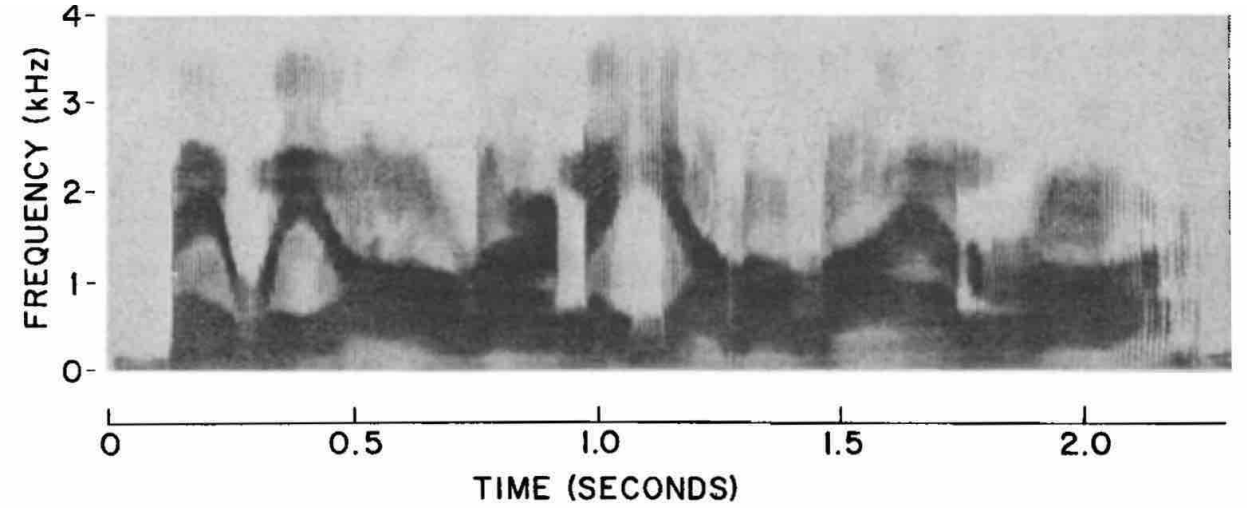


Figure 2.4: A spectrogram of a speech signal uttering "May we all learn a yellow lion roar" [12]

Eye

Various devices can acquire the data about the motions of the human eye [47, 40] (See Figure 2.5). This behavioral information can be processed to identify users [55, 88] or infer their emotional state [172, 181]. The authors analyzed various behavioral traits of the human eye for emotion recognition, such as fixation features [172, 110], pupillary responses [11, 6], eye motion [146], and features based on predefined models [10]. Certain physiological features derived from the eye, such as the pupil diameter [112, 134] and Electrooculography (EOG) [181, 135] are also beneficial for eye based emotion recognition.

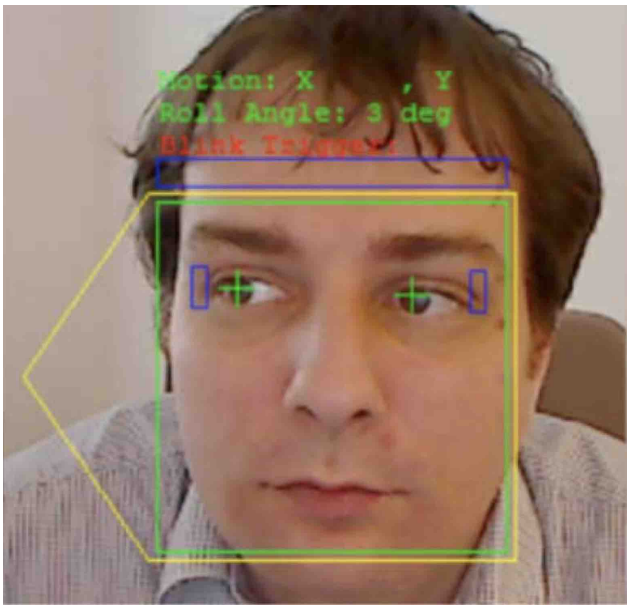


Figure 2.5: An example of real-time eye gaze detection using a webcam [40]

Gait

Gait based person identification systems use one of two approaches: appearance-based and model-based [107]. Appearance-based approaches disregarded the skeletal structure and processed the shape of the human body to extract features (Figure 2.6). In contrast, model-based methodologies represented human gaits as movements of the skeleton. The approaches used in gait-based user recognition system were similar to some works on Gait Emotion Recognition (GER) [87, 45]. However, the researchers also identified novel methods to extract or select emotion-relevant features. The authors of [133] identified the key features that facilitate reconstruction of emotional gaits. [176] derived their feature set using auto-correlation matrices of the degrees of freedom of body joints. In 2019, [3] proposed using Laban Movement Analysis for the task with a Analysis of Variance (ANOVA) based feature selection method. The emotion recognition systems based on gait exclusively relied on classical machine learning methods until recently.

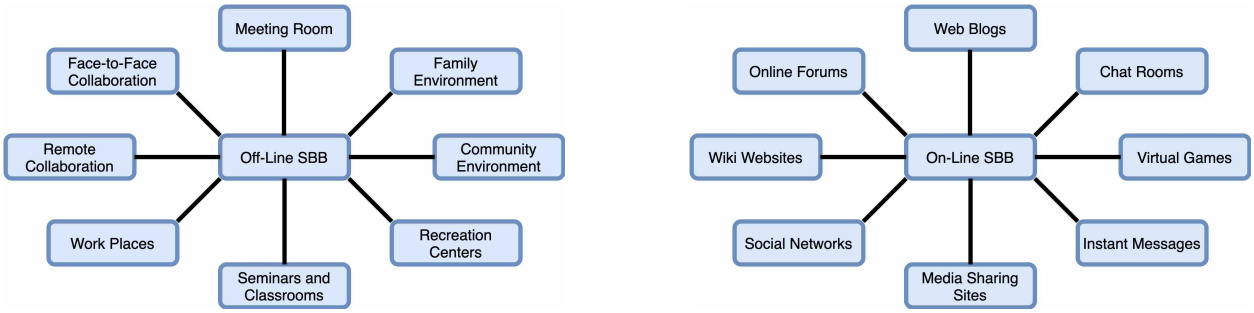
Figure 2.6: An example of an extracted silhouette from a video frame of a gait sequence in the Casia-B dataset [194]

Deep learning methodologies for GER typically use one of three types of networks on their own or in combination: Graph Neural Networks (GNNs), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). CNNs and GNNs model gaits are rigid structures to benefit from the structure of the human body and combine the information available in each node/pixel to extract spatial and/or temporal features [20, 129]. On the other hand, RNNs processes gaits sequentially to extract spatial and temporal features simultaneously [144, 21].

2.1.3 Social Behavioral Biometrics

Social behavioral biometrics is a branch of biometrics that is defined as a study of human behavior in a social context [168]. Furthermore, the social interactions where these behaviors are observed can be either

on-line or off-line, hence a behavioral biometric qualifies as a social-behavioral biometric if it is exhibited in a social setting. Figure 2.7 lists the various scenarios where off-line and on-line Social Behavioral Biometrics are observed. The field of social behavioral biometrics has recently gained popularity and has been proposed for person identification [160, 173]. Social network behavior, textual conversation or commenting patterns, and body movement in a social setting are some biometrics that fall in this domain. Online social behavioral biometrics can facilitate continuous authentication for a user without being intrusive. Additionally, these biometrics can be used for emotion recognition. For instance textual analysis, monitoring keystroke patterns, and even user aesthetic preferences can be indicative of the emotion being experienced by the user.



(a) Application areas for Off-line SBB

(b) Application areas for On-line SBB

Figure 2.7: Application areas of Social Behavioral Biometrics (SBB), adapted from [174]

Monitoring keystroke and mouse dynamics can be used for incorporating periodic user authentication in existing devices without any additional hardware. Features associated with key strokes such as dwell time, flight time, and other time difference metrics have been successfully used in identifying people [64, 33]. Similarly, features derived from mice such as pointer speed, and mouse clicks have also been employed for user authentication [73, 58].

These keyboard and mouse features also contain information about the emotional state of the subject [202, 51]. Most works used generalized key press features and classical machine learning methodologies, but some works incorporated the information about the specific keys being pressed as well [177, 99]. In addition to keystroke features, studies also collected data from mouse movements [202], touch screen parameters [99], and pressure [115] for enhancing the emotion recognition accuracy.

Perhaps one of the most commonly used authentication methods, text is used readily in a variety of applications [89]. However, it can also be analyzed for emotion recognition purposes. The process of inferring emotion from textual data is called sentiment analysis [121]. Features used by text based emotion classifiers can be broadly categorized as: word-level, phrase-level, and sentence-level features. The research focused on features at a word-level at first [9, 165] but was soon extended to include information derived from sentences [117, 126], the part-of-speech (POS) context [13], and characters and documents [35]. Some works also aimed

to improve the classification methodologies for textual emotion recognition [142]. Finally, recent works have eliminated feature extraction and selection steps by resorting to deep learning methodologies for sentiment analysis [37].

2.2 Emotion Representation Models

Once the data for an emotion recognition system is collected, it is imperative to choose a suitable representation for the emotions. Human emotions are complex and can be described by a myriad of systems or schemes. Over the years, several researchers have developed these systems that stem from careful analysis about what constitutes an emotion. These systems can be broadly classified into: Discrete models, Dimensional Models, and Componential Models.

2.2.1 Discrete Models

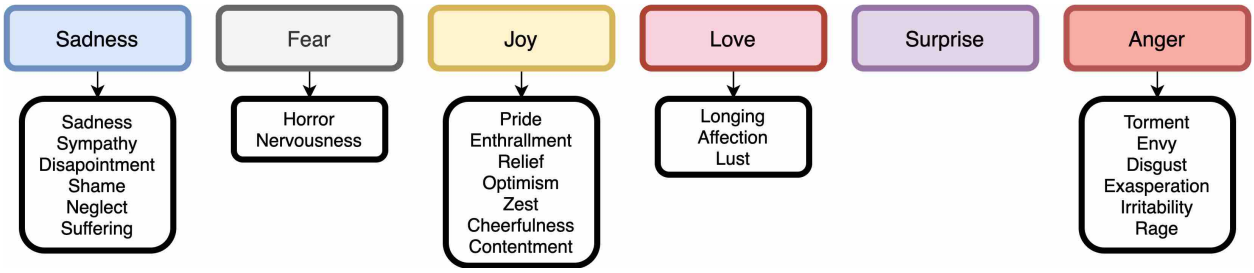
This class of emotion models is based on the idea of emotions existing as mutually exclusive categories. The emotions, defined by most models, in this class may be divided further into finer emotions, but are not proposed to be combined to form complex emotions. In [43], Darwin argued that humans developed emotions as a result of evolution. While Ekman agreed that emotions root from evolution, he believed that emotions can also be learned throughout the lifetime of a person [49]. Consequently, six emotions were defined as the basic emotions: anger, disgust, fear, happiness, sadness, and surprise (see Figure 2.8a). A similar ideology was used by Parrot, where he defined six emotion classes with further branches to capture more specific emotional states. He proposed a tree structure for emotion models which described a total of 100 emotions (refer to Figure 2.8b). Izard proposed 12 basic emotions that could not be further deconstructed, but could be combined with each other to form more complex emotions [78] (Figure 2.8c).

The emotions presented by Ekman's model were limited to emotions that can be expressed by facial expressions [28]. Furthermore, discrete emotion models in general exclude complex emotions that exist between two described emotion classes and emotional states with two or more emotions exhibited simultaneously. Despite these shortcomings, discrete emotion models are the most suitable and acceptable method of representing emotions for computer applications and hence are used in this thesis.

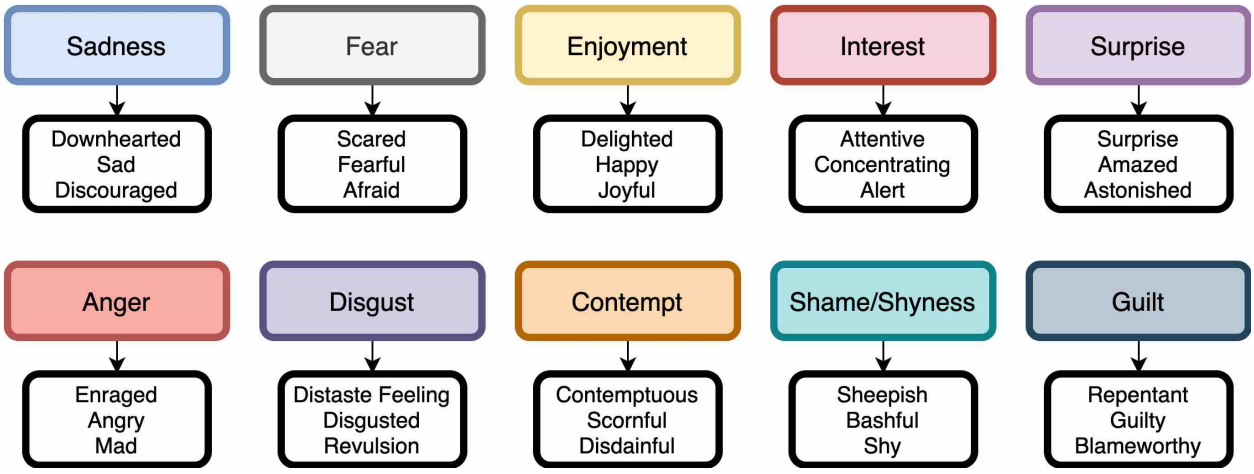
2.2.2 Dimensional Models

Dimensional models assume that every emotion is composed of certain fundamental variable properties. Russell proposed that these properties are Arousal and Valence, and that all emotions are the result of a

(a) Discrete categories of emotions proposed by Ekman



(b) Parrot's discrete emotion model

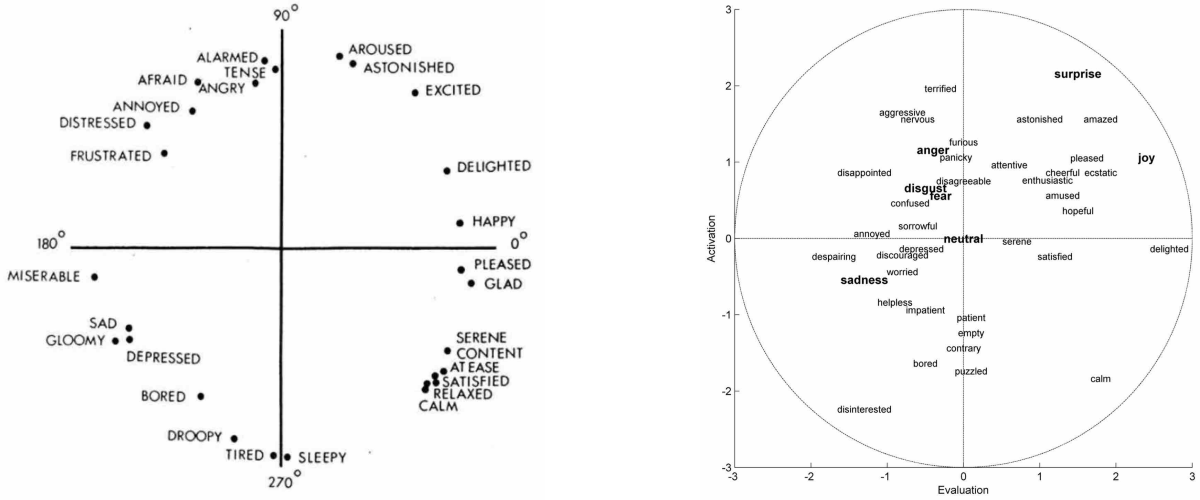


(c) The discrete emotions proposed by Izard

Figure 2.8: Discrete emotion representation models

varying degree of these two dimensions [149]. Arousal determines how intense an emotional stimulus is, while the Valence refers to the pleasantness of the emotion. The 150 emotions in this model were described as coordinates on the circumference of a circle defined by two axes representing arousal and valence (Figure 2.9a). This model was slightly modified by Scherer [152] to include coordinates within the described circle. A similar approach was adopted by Whissell [183], who described various emotions with Activation and Evaluation, synonymous with arousal and valence, as the two dimensions (Figure 2.9b). The two-dimensional approaches were extended by introducing a third dimension of dominance by Mehrabian [122]. The third

dimension of dominance, also referred to as potency or power, represents the amount of affect the emotion has on a person. In other terms, it defines how much control a person has over a particular emotion. In addition to potency, Fontaine et al. introduced a new dimension called unpredictability in 2007 [54].



(a) Russell's circumplex model

(b) Dimensional Emotion model proposed by Whissell

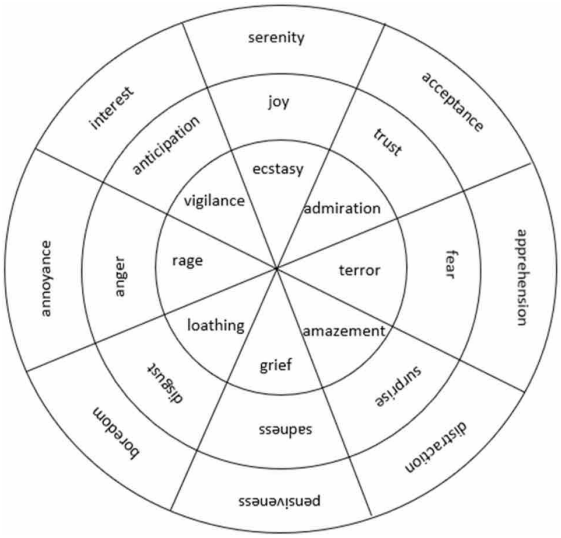
Figure 2.9: Dimensional Emotion Models [28]

Dimensional representation of emotion allows overlapping of emotions and provides a continuous spectrum that describes them. These models are fairly accurate representations of human emotions but unsuitable for use in computing applications due to the ambiguous nature of emotional descriptions.

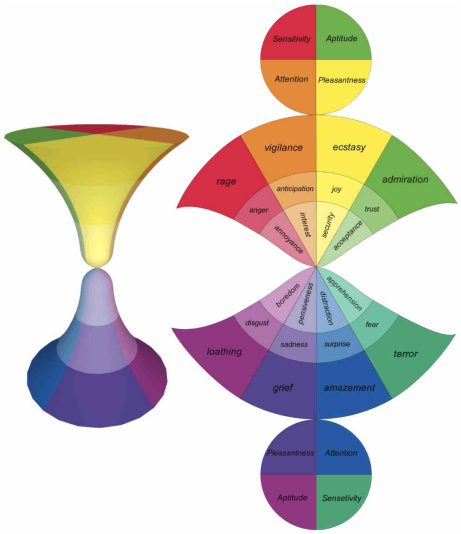
2.2.3 Componential Models

Emotion representation models that allow discrete emotions to be combined to form complex emotions are categorized as componential models. One such model was proposed by Plutchik [140]. This hybrid model considered eight discrete emotions, arranged in a color wheel to describe primary, secondary, tertiary, and opposite emotions. In this model similar emotions were placed adjacently, opposite emotions were placed at a 180° angle, and complex emotions were described as combinations of these basic emotions (Figure 2.10a). Plutchik's model inspired an hourglass model that was defined using four independent properties (Pleasantness, Attention, Sensitivity and Aptitude) of emotions [28]. This model defines emotions similar to color theory and proposes that complex emotions can be represented as a combination of two or more basic emotions (Figure 2.10b).

In comparison to dimensional representation models, discrete and componential representations have limited emotional categories but provide clear interpretation for emotion recognition tasks. Componential



(a) Plutchik's Emotion model



(b) The hourglass Emotion representation

Figure 2.10: Componential Emotion Models [28]

models only differ from discrete representations in terms of the composition of complex emotions, and hence are also compatible for computational tasks.

2.3 Overview of Gait Emotion Recognition Systems

2.3.1 Gait Biometric and Gait Emotion Recognition

As established earlier, gait analysis can be beneficial for numerous industries such as security, medicine, internet of things, robotics, disaster management, entertainment, advertising, education, virtual reality, computer vision, affective computing, and assisted living [59]. Gait biometric is a unique trait that can be used for user identification [125] in conditions where other biometric data is unavailable or unidentifiable, for instance, surveillance footage. It can also be used to prevent illegal activities by monitoring person behavior in public places. Additionally, it can be used in the healthcare industry in diagnosing chronic diseases [85, 57, 83, 81]. Furthermore, gait analysis can benefit rehabilitation [201] and assisted living [154]. Robots that can identify users [36], and are emotionally aware [129] can also be developed. Moreover, emotion recognition has already been used to develop adaptive education and entertainment experiences [63], and can be incorporated with virtual reality [118] applications. In addition to the variety of applications, gait, as a biometric trait, provides significant advantages over other modalities:

1. **Remote** – A significant part of any biometric system is the acquisition of data. The collection of most biometric traits such as face can only be recorded in close proximity. However, gait allows data to be

collected and processed from a distance [186].

2. **Non-intrusive** – The collection of biometric data can be overwhelming for users and induce hesitance towards such procedures [141]. Since gait can be recorded from a distance, it is less likely to cause stress to the users.

3. **No Expert Supervision required** – Biometric traits, like EEG, usually require an expert to set up the system for data collection, which might increase the time and cost required to collect data. In contrast, gait data can be collected using normal cameras or depth-based sensors which do not require significant expertise for operating.

4. **Universal** – Although gestures may have different interpretations depending on the observer [164], walking is a universal action that is common for all ethnic backgrounds. Gait is a robust source that has more consistent emotional interpretation [79].

5. **Robust to data quality** – Depending on the biometric data capturing techniques and conditions, the recorded data might be of low resolution. This is problematic for modalities like face, which relies heavily on the quality of the image, but gait emotion recognition can still be performed [186].

Gait based systems are compatible with various data collection methods. Platforms like Kistler force plates [82] and infrared light barriers [100] have been used in prior research for collecting gait velocity. Accelerometers in wrist bands can also be employed to record similar gait features [196, 143]. Additionally, depth map sensors, such as Microsoft-Kinect-V2, capture information about the distance of the object from the camera and provide a way of identifying body joint coordinates [102, 105]. However, the most accurate method of recording the subject's body structure is via motion capture systems [176, 45]. Systems like Vicon or Xsens Motion Capture, require markers to be placed on various body joints of a person and record the positions of these markers throughout the gait of that person. This thesis uses a dataset recorded using Motion Capture systems since they provide the most accurate readings. However, special data acquisition equipment is not a requirement for gait analysis. Video recordings of people walking can be simply analyzed using appearance-based methods or processed using pose-estimation algorithms to identify body joints for model-based approaches.

Once the data has been collected, it is processed to produce relevant features. Processing techniques for gait analysis can be broadly categorized into: appearance-based, and model-based. Appearance-based methodologies work on video data and derive visual features such as shape and area of the subject's body. These methods produced images, where each pixel captured the information about the silhouette [151], amount of motion, [23] and/or the general posture of the subject throughout the gait [70]. Since these

methodologies utilize visual features, they are sensitive to the subject's clothing and environmental factors. Hence, these approaches do not perform well under conditions where the subject wears baggy clothes or the clothes and the background have identical colors. Moreover, they require a side view and high-resolution video data to work.

On the other hand, model-based methods map the collected gait sequences to predefined skeletal structures. This representation is used to calculate gait parameters such as stride length and cadence [16], joint angle and distance trajectories [180, 170]. Unlike appearance-based techniques this approach facilitates gait analysis that is unaffected by view, pose, clothing, and most environmental changes. Preliminary works in gait based user recognition systems used appearance-based methods and shifted to model-based methods once the issues mentioned above were identified. However, since the advancements in gait emotion recognition have commenced recently, most works adopted model-based techniques. Consequently, these features can be processed to identify emotions by classical machine learning algorithms or deep neural networks, that are discussed in detail in the subsequent sub-sections.

2.3.2 Classical Machine Learning-Based Methods

Machine Learning can be defined as a system that emulates the human learning [50]. These work by analyzing data and determining patterns in that data to identify rules for decision making. Systems based on machine learning methods slightly modify themselves as they parse data. This process is referred to as the training stage and is intended to be an equivalent of the human learning process. Similar to how humans gain experience with practice, these systems tune themselves by repeatedly processing the input data. To ensure that these algorithms develop a general scheme of rules rather than a set of rules specific to the training data the systems' performance is measured on unseen data, this stage is called testing. Classical machine learning based methods refer to a subset of Machine Learning methods that require explicit feature extraction, which are then processed by various algorithms. Some popular machine learning methods that belong to this class are Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB), and Linear Regression [30].

Over the years, classical machine learning works in Gait Emotion Recognition (GER) have explored various data. One of the earliest works to automate identification of emotions from gaits used force platforms and infrared light barriers to measure reaction forces and walking speed [82]. This work utilized Self Organizing Maps (SOM) for distinguishing between neutral, joyous, sad, and angry gaits. With the advent of smart wearable devices, works like [196] and [143] developed methodologies for recognizing emotions from accelerometers found in smart watches/bracelets. These works were based on statistical features derived from acceleration and gyroscopic data. Furthermore, a mobile gait emotion recognition methodology was

proposed [38] that used a pose estimation algorithm to generate 2D gait skeletons from a video. These skeletons were used to produce statistical features and processed remotely on a server.

However, most methods relied on model-based gait depictions and focused on feature extraction and feature selection improvements. The domain witnessed a trend of using Fourier Transform (FT) and Principal Component Analysis (PCA) for generating frequency and phase features, and to reduce the dimension of the feature set. A 2010 paper, published by Karg et al. [87], generated two sets of features: one containing kinematic features, and one containing posture, frequency, and phase shift features. PCA and FT were used to generate the second set, while both sets underwent PCA for dimension reduction. Venture et al. [175] proposed feature sets based on degrees-of-freedom (DOF) of body joints instead, and used a similarity index on PCA components for emotion classification. The DOF feature set and the similarity index calculation were optimized in a subsequent publication [176]. Model-based research until 2014 was conducted using Vicon motion capture data. However, in 2016, Li et al. [105] used data collected from a depth sensor, Microsoft-Kinect-v2, to produce FT-based features for emotion classification. A similar work by the same authors [102] also proposed using FT to produce frequency-domain and time-domain features. Both works reduced the dimension of the final feature set using PCA.

Later works in the domain improved decision making modules using ensemble learning. In 2018, Ahmed et al. [4] proposed a methodology that utilized Laban Movement Analysis for describing the emotional movements of the subject. Unlike prior research work, this paper reduced the feature set using histogram values and applied score and rank level fusion on four classifiers. In 2019, the work was improved with the introduction of ten feature groups of emotion-relevant body movement features [3]. They also proposed Analysis of Variants (ANOVA), Multivariate ANOVA (MANOVA), and Genetic Algorithm (GA) for the feature selection module in this work.

Prior works on traditional machine learning methods applied commonly used machine learning classifiers. Support Vector Machines (SVM), Naïve Bayes (NB), Random Forests (RF), and Decision Trees (DT) were particularly popular in this domain. These methods highlighted the features important for recognizing emotional gaits. Furthermore, these studies provided insights on various approaches that can be taken while performing Gait Emotion Recognition (GER). However, these methods had some fundamental flaws. Classical machine learning methods were only compatible with the type of features they were designed for. For instance, methodologies that considered gaits as trajectories of body joints performed well with signal processing features but not with statistical features. Some works attempted combination of different types of feature, but were restricted by the abilities of classical machine learning models, which were accurate only if the feature sets were small. As a result, researchers resorted to discarding certain features. Additionally, these methods were not scalable and did not perform well with different data. Lastly, the methods worked

on a handful of known and easily describable features. They could not exploit latent features that could have resulted in a significant performance gain. These issues were mitigated with the transition to deep learning-based methodologies.

2.3.3 Deep Learning-Based Methods

Overview of Deep Learning

Deep Learning is a subdomain of Machine Learning that achieves learning through Artificial Neural Networks (ANN) [124]. These networks were originally inspired by the biological neuron-based processing in animals. However, artificially simulated neural networks and biological neural networks have various differences. These data structures allow flow of data in a specified manner and update the network's parameters accordingly to identify patterns more accurately. These networks typically contain several layers to facilitate this learning process, hence the name deep learning [124]. Furthermore, this technique allows the feature extraction and selection process to be automated. Each layer outputs a matrix of values calculated using non-linear functions called activation functions. Stacks of such layers allow the network to develop more complex functions as compared to the relatively simple functions learned by classical machine learning algorithms. Therefore, these networks are well suited for complex tasks such as identifying emotions from gaits. The recent boom in deep learning methodology can be accredited to the advancements in computation power, and the increased availability of data [46]. Since the inception of the idea of neural networks [120], many general neural network architectures have been developed catering to different processing needs. Multi-Layered Perceptrons, known as vanilla neural networks [44], were developed as multi-purpose artificial neural networks to process all kinds of data but are most well-suited for processing data as feature vectors similar to traditional machine learning classifiers. Images have certain repeating patterns that can be more efficiently processed by Convolutional Neural Networks (CNNs) using lesser number of parameters [5]. Similarly, Graph Neural Networks (GNN) are built to effectively process graphical data using message passing [185]. Finally, Recurrent Neural Networks (RNN) utilize the concept of a memory cell and exhibit high performance in sequential or signal data processing tasks [109].

Multi Layered Perceptron

Linear regression, a classical machine learning classifier, is capable of learning a simple linear functional relationship between various inputs (X) and an output variable (y) by estimating the values of θ and b, Equation (2.1). This operation can be treated as one unit and can be stacked to learn more complex non-linear functional dependencies by estimating the parameters (θ and b) for all units. Multi Layer Perceptrons

are such networks that have an input layer to accept the input variables, at least one hidden layer to add complexity, and an output layer to map the processed values to the output. The values of these parameters can be initialized using a distribution, randomly, or as constants, and can be tuned using an optimizer. Optimizers utilize loss functions that represent the poorness of the model performance using the current values of the parameters. Once the loss has been calculated, optimizers modify the model parameters to reduce the loss. This process is known as training and is repeated over several iterations (called epochs). MLPs are suitable for processing vectors of data and, with enough data, outperform all classical machine learning classifiers.

$$y = \theta X + b \tag{2.1}$$

Convolutional Neural Network-Based Methods

Convolutional Neural Networks were developed to process images effectively. Processing a high resolution image using Multi Layer Perceptron (MLP) would require a huge number of parameters that identify patterns between all pixels. However, images have certain repeating patterns than can be identified once and detected multiple times within the image. Some of these patterns can be simple curves, strokes and edges that occur more than once in an image. This approach require significantly lesser parameters to perform the same task in comparison to MLPs.

The mentioned idea is realized by using kernels (also called filters), which are learnable matrices that can identify common features in the image. CNNs use these grids of values to align with a portion of the image, and perform a dot product between the kernel and the selected portion of the image. This operation is performed iteratively for other portions of the image until the entire image has been processed. The output of the various dot product operations form a feature map. This map contains the information about the presence or the absence of the pattern described in the kernel. The entire process of sliding the kernel window and performing dot product operations to produce a feature map is called a convolution operation (see Figure 2.11). Usually, multiple kernels can be used on an image to detect several patterns, and the kernel values are tuned during training to identify optimal patterns. In addition to the convolution operation, CNNs support other operations such as pooling to downscale an image, and flattening to unravel a feature map into a feature vector.

Sometimes CNNs are applied on matrices formed from non-image data sources, these matrices are referred to as psuedo-image. In 2020, Narayanan et al. [129] proposed a representation of gait sequences as stacks of pseudo-images. The coordinate values from each body joint, was placed into a matrix to form one pseudo-image. They used a CNN architecture with 2D convolution layers and maxpooling layers to identify emotions

from emotional gaits. Due to the lack of large datasets in the domain, the authors of this paper used a data augmentation technique that calculated new gait sequences for various observation angles.

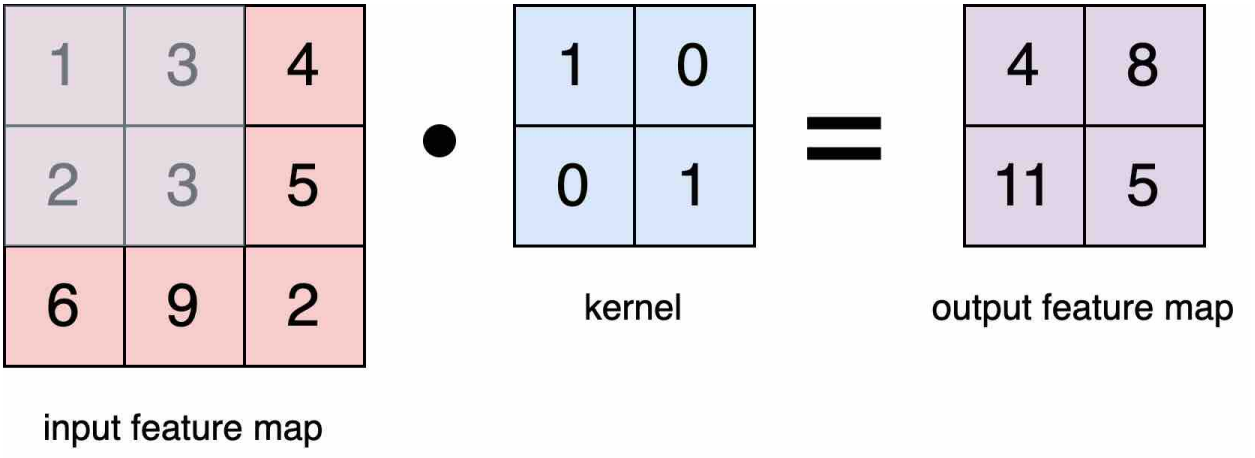


Figure 2.11: An example of a convolution operation

The architecture greatly benefited from the structured representation of gaits using psuedo-images, as the kernels learned patterns with respect to the position of body joint values in the pseudo-image. CNNs combine convolutionally extracted features hierarchically. Thus, low level features between coordinates that are placed far away from each other are not explored. However, certain dependencies between distant joints are crucial for GER, for instance, a slouched head and short steps might indicate sadness. Moreover, the ability of CNNs to detect and identify recurring patterns was not fully exploited in this work. The architecture processed pseudo gait images using 2D convolutional operations only. These operations were able to encode the spatial information present in each frame of the sequence however temporal information across frames was not extracted from the gaits. Hence, information corresponding to the change in a subject's behavior throughout the emotional gait was not captured efficiently. Lastly, data augmentation techniques results in the model learning information specific to the dataset used and might not perform well on other datasets.

Graph Neural Network-Based Methods

Graph Neural Networks (GNN), sometimes referred to as Graph Convolutional Neural Networks (GCNN) are adapted specifically to process graph data. In a conventional convolutional operation, the information in a locality (of pixels) is combined using a kernel. This idea is extended in GNNs, where the information in the locality of nodes is combined. Hence, the operations in a GNN are often compared to convolutional operations.

Depending on the application, a node-level, edge-level, or a graph-level prediction might be required from the inputted graphical structures. Each node and edge in a graph can contain descriptive feature vectors.

Graph neural networks combines the information from these individual node and edge feature vectors to form new representations for them. This can be done iteratively to combine information from nodes that are not immediate neighbors. The process that facilitates this incorporation of information from node and edge feature vectors is called message passing.

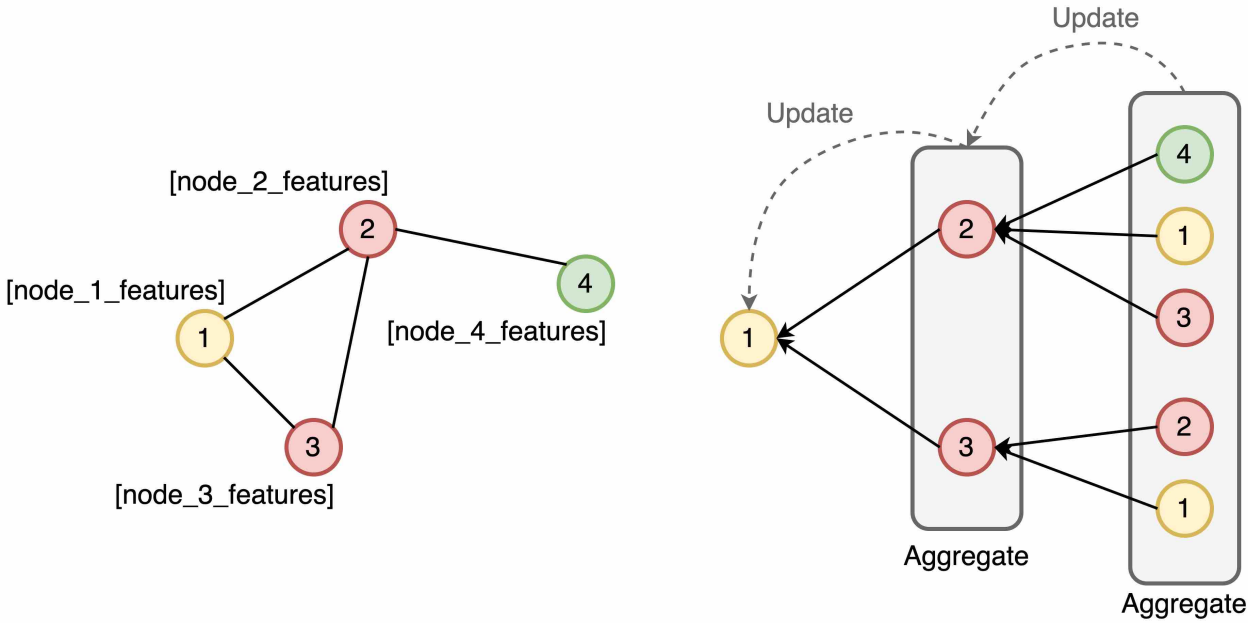


Figure 2.12: Message Passing computational graph (right) for the embedding of node 1 (h_u) for a sample graph input (left)

A message passing operation can be described in two steps: aggregating and updating, described in Equation (2.2) where $h_u^{<k>}$ is the embedding of a node u at the k^{th} step, and v represents a neighbouring node. During aggregation, features from immediate neighbours ($N(u)$) of a node (u) are combined. Subsequently, the node's feature vector (also called a node's embedding) is updated by combining its previous value ($h_u^{<k>}$) and the newly formed feature vector. The specific method for aggregation and updating depends on the exact implementation of the GNN. It can range from simple mathematical functions such as addition and multiplication to more sophisticated approaches like neural networks. The new embedding of a node ($h_u^{<k+1>}$) contains information about itself and its neighbours after one iteration of message passing (Figure 2.12). If more message passing layers are used (more message passing iterations), each node embedding in the graph contains information from farther nodes. The final embedding of the node can then be used for node-level prediction, or embeddings from all nodes can be combined for graph-level predictions. For simplicity, the message passing procedure for only node embeddings were described above, but the same process can be described for producing new edge embeddings with each layer instead.

$$h_u^{<k+1>} = Update^{<k>}(h_u^{<k>}, Aggregate^{<k>}(h_v^{<k>}, \forall v \in N(u))) \tag{2.2}$$

In the domain of Gait Emotion Recognition, many researchers view gait as a sequence of skeletal graph, or sometimes simply a graph. These approaches utilize GNNs for processing gait. In 2018, Yan et al. [189] proposed a Spatial Temporal Graph Convolutional Network (ST-GCN) that modelled the entire gait sequence as one connected graph. Each frame consisted of a skeletal representation of a pose and the body joints in this skeletal graph were connected to the the corresponding body joints in the skeletal graphs of the previous and the next frame. The authors proposed three partitioning schemes for the graph convolutional operations: unilabelling, distance partitioning, and spatial configuration. Although this method was introduced for action recognition, it was modified for gait emotion recognition purposes by the authors of [20] in 2020 to create the ST-GCN for Emotion Perception model (STEP). This paper used an encoder-decoder based networks, where the network was trained to re-produce the input. To construct the encoder, they added a 2D average pooling layer and two parallel 2D convolution layer to the ST-GCN network. Similarly, the decoder contained a de-convolution layer before the ST-GCN network. After the encoder-decoder network was trained to reconstruct emotional gaits, the encoder was used in addition to a handful of affective features to identify the emotional labels.

Similar to Convolutional Neural Networks (CNN), GNNs benefit from the inherent rigid structure of the human body for gait analysis. However, this approach has certain flaws that are detrimental to the network performance for GER tasks. The message passing operations in GNNs do not allow low level features to be formed between nodes that are placed far away in the graphical structure. Furthermore, GNNs allow the combination of node and edge embeddings, but the ST-GCN only considers node embeddings which ignores the relations between the bones and the joints in the body. This demerit propagates to the STEP network as well, which has additional shortcomings. To address the lack of data, the authors combined different datasets collected under varying conditions. Moreover, they generated synthetic gaits from these datasets to increase the number of samples. The amalgam of datasets was then used to train and test the network. Hence, the dataset that was used in this research had undesired variation in the samples and gaits that were not representative of real-world gaits. Moreover, since the train and test sets were derived from real and synthetic gaits generated from models trained on those real gaits, the test set had some representation of the training set. Hence, the success of this research was overestimated.

Recurrent Neural Network-Based Methods

Recurrent Neural Networks were designed for sequential data processing. Time-series or signal data pertaining to finance, language, music, etc. can be effectively processed by this technique. Similar to CNNs' idea of sharing the learning throughout the whole image, RNNs share their learnings across a sequence. For instance, the word "not" should have a similar contextual meaning irrespective of its position in the sentence.

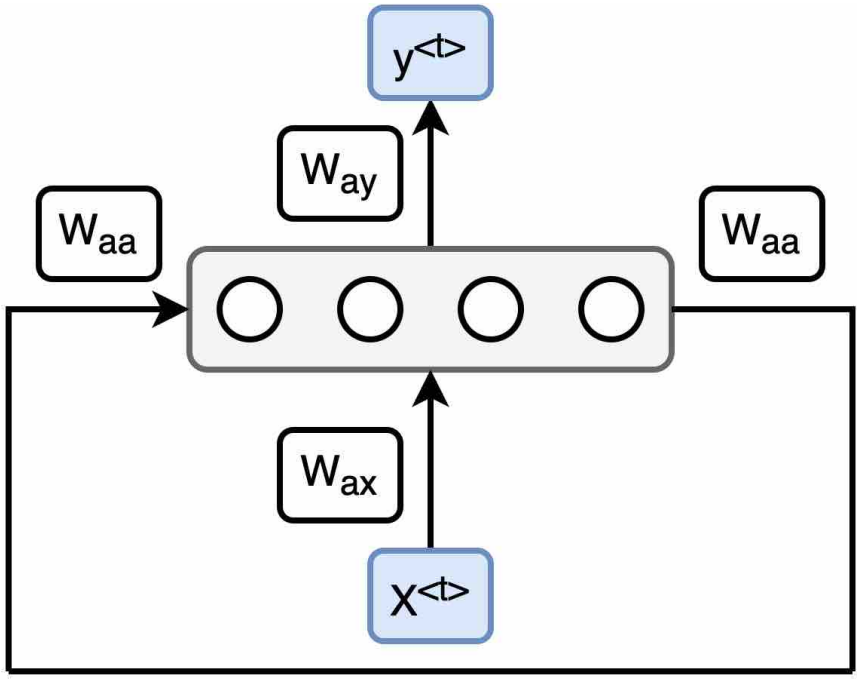


Figure 2.13: A flowchart describing the information processing in Recurrent Neural Networks

The sharing of information in Recurrent Neural Networks is achieved by processing data in time steps and accepting information from previous time-steps. Activations for the current time step are produced using previous time-steps' activations ($a^{<t-1>}$) and current information ($X^{<t>}$). This calculation is described in Equations (2.3) and (2.4), where g and g' are activation functions. The processing of information in RNNs is visualized in Figure 2.13. This method of processing data is effective for short sequences. There are multiple weight matrix (W_{aa}, W_{ax}, W_{ay}) products required for processing long sequences. These weight matrices usually have small values and repeated multiplications of such values cause the product to diminish over time steps. Hence, vanilla recurrent neural networks are ineffective for processing long term dependencies and are rarely used in literature.

$$a^{<t>} = g(W_{aa}a^{<t-1>} + W_{ax}X^{<t>} + b_a) \tag{2.3}$$

$$y^{<t>} = g'(W_{ay}a^{<t>} + b_y) \tag{2.4}$$

This issue is mitigated in Long-Short Term Memory (LSTM) units and Gated Recurrent Units (GRU), by using a dedicated memory cell to hold information. In LSTMs, this memory cell vector is controlled by an update gate (γ_u) and a forget gate (γ_f). At each time step the old value of the memory cell ($c^{<t-1>}$) and a newly computed candidate value of the memory cell ($c^{'<t>}$) are considered to calculated the new value ($c^{<t>}$). This calculation is described in Equation (2.9). Additionally, an output gate (γ_o) governs the activation of the unit, see Equation (2.10). The two gates to calculate a new memory cell value ($c^{<t>}$) along with an output gate to regulate the value passed on to the next step provide more control (Figure 2.14).

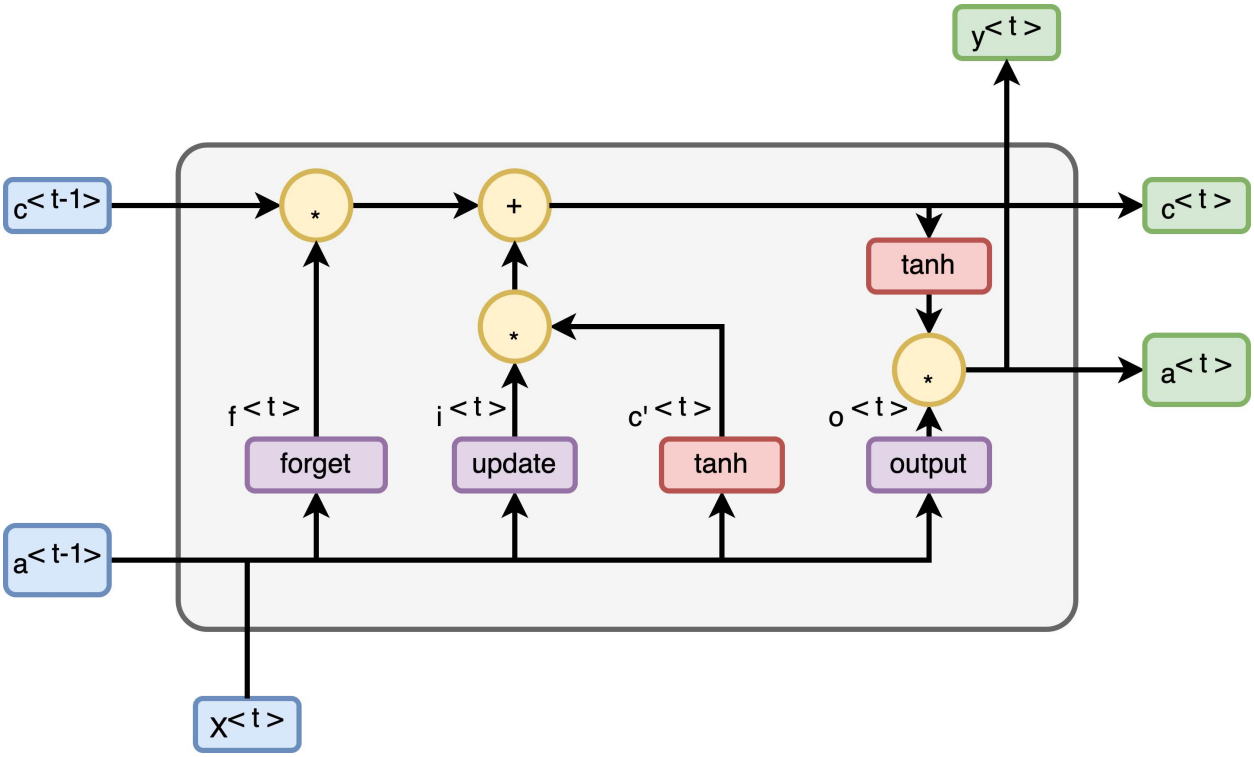


Figure 2.14: Visualized calculation of a general Long Short Term Memory unit

$$c'^{<t>} = tanh(W_c[a^{<t-1>}, X^{<t>}] + b_c) \tag{2.5}$$

$$\gamma_u = \sigma(W_u[a^{<t-1>}, X^{<t>}] + b_u) \tag{2.6}$$

$$\gamma_f = \sigma(W_f[a^{<t-1>}, X^{<t>}] + b_f) \tag{2.7}$$

$$\gamma_o = \sigma(W_o[a^{<t-1>}, X^{<t>}] + b_o) \tag{2.8}$$

$$c^{<t>} = \gamma_u * c'^{<t>} + \gamma_f * c^{<t-1>} \tag{2.9}$$

$$a^{<t>} = \gamma_o * tanh(c^{<t>}) \tag{2.10}$$

In contrast, GRU's have a relevance gate (γ_r) which is used to signify the importance of retaining the old value of the memory cell, see Equation (2.11). While the gate describes the weightage of $c'^{<t>}$ and $c^{<t-1>}$ during the calculation of the new value of the memory cell vector ($c^{<t>}$), the value essentially depends on only one update gate (γ_u). Moreover, GRU lacks the final activation facilitated by the output gate in LSTMs. Hence, GRUs offer less control over the final activation values, in comparison to LSTM units. The calculations of a typical Gated Recurrent Unit are visualized in Figure 2.15.

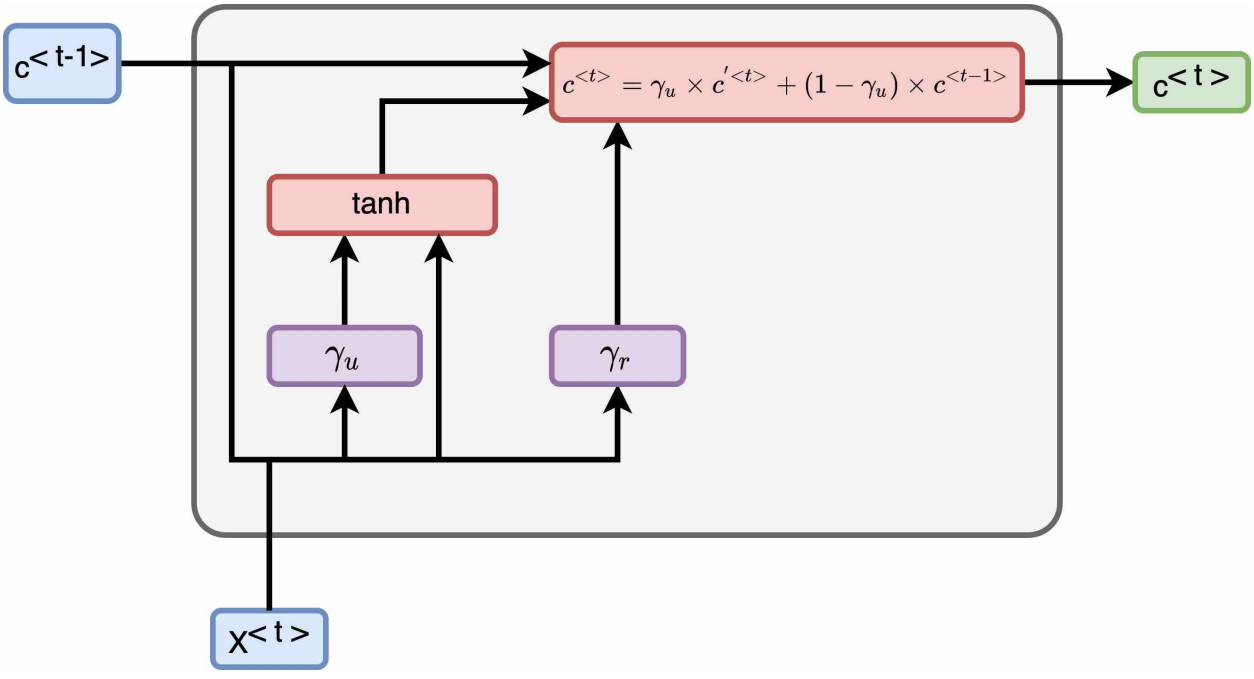


Figure 2.15: Visualized calculation of a Gated Recurrent Unit

$$c'^{<t>} = tanh(W_c[\gamma_r \times c^{<t-1>}, X^{<t>} + b_c]) \tag{2.11}$$

$$\gamma_u = \sigma(W_u[c^{<t-1>}, X^{<t>} + b_u]) \tag{2.12}$$

$$\gamma_r = \sigma(W_r[c^{<t-1>}, X^{<t>} + b_r]) \tag{2.13}$$

$$c^{<t>} = \gamma_u \times c^{'<t>} + (1 - \gamma_u) \times c^{<t-1>} \tag{2.14}$$

Since the features in RNNs are calculated by performing dot product of the input data ($X^{<t>}$) and a weight matrix (W_{ax}), all possible low-level features are explored. This mitigates the issues with models such as GNNs and CNNs, that use a rigid structure to represent gaits. Moreover, sequential models are better suited to extract temporal features from time-series data.

In 2019, Randhavane et al. [144] used a methodology similar to [20]. They used a LSTM based encoder-decoder trained to reconstruct human gaits. The trained encoder from this network was used to produce gait embeddings which were then aggregated with posture and movement features. The combined feature set was passed to a Random Forest classifier to recognize emotions. This work is referred to as Affective and Deep Features (ADF) in this thesis. In 2020, Bhattacharya et al. [21] used the encoder-decoder methodology as well, with a few differences. The network trained to recreate gait sequences used Hierarchical Attention Pooling and Affective Mapping (HAPAM) and was based on GRUs. The trajectories corresponding to each body joint's rotation throughout the gait were independently processed by two layers of GRU at first. Linear layers were then employed to produce feature representations of the body joint rotation trajectories. These representations were then combined using sum pooling layers to form segments of the human skeleton. For instance, feature vectors corresponding to the trajectories of left shoulder, left elbow and left hand were pooled to represent a feature vector for the entire left arm. These combined feature vectors were processed linearly to form an embedding of the entire gait sequence. The decoder followed a similar approach in reverse and used linear, GRU and unpooling layers to generate gait sequences. For classification, the authors produced gait embeddings using the encoder, as described above, combined it with affective features and used an MLP for mapping it to the output labels.

These works provided a base-line for recurrent neural network based methodologies for gait emotion recognition; however they failed to reap all the benefits of using sequential deep learning models. The LSTM network used in [144] was poorly designed and ineffective at producing gait embeddings. Additionally, like the affective feature set used in [20], the posture and movement based features contained only a few measurements. Lastly, the decision making module in this architecture was a Random Forest Classifier,

which can be outperformed using a basic Artificial Neural Network (ANN). On the other hand, the work published in [21] utilized an MLP classifier but had various other design flaws. One advantage RNNs offer over CNNs and GNNs is the ability to explore low-level dependencies between distant body joints, but the hierarchical approach adopted in this paper eliminates the possibility of capturing those spatial features. Moreover, the network employed GRUs which offer a limited control over the memory cell. Hence, the temporal features derived from the gaits are not optimal either. A demerit common with HAPAM [21] and STEP [20] methodologies is that these models were trained on a dataset containing a mixture of real and synthetic gaits made from those real gaits. Hence, most of the data used was not representative of the real world and also exaggerated the performance of the network, since the test set contained synthetic gaits built from a network trained on real gaits present in the training data. Both of the RNN based approaches in this domain were devised by the same group of authors as the STEP [20] paper and used the same encoder-decoder design. One major flaw with such an approach is that the network tunes its parameters to capture information relevant for re-constructing the image. This information is then re-purposed for emotion recognition. Hence, there is a misalignment between the objective of the research and the network design. Furthermore, the affective feature set used in these researches works contain only a small number of features. This characteristic was common with studies prior to 2010, since intensive computation tasks were not feasible with older hardware and traditional classical machine learning methods. However, deep learning networks are capable of handling larger feature sets on improved hardware. Hence, the affective feature sets were also not optimal.

Concluding Remarks

Gait Emotion Recognition witnessed only a handful of works based on deep learning. All of these methods relied on deep neural networks with a lot of parameters. Hence, they required large datasets to train on, which led researchers to mix different datasets and use synthetic gaits. The large networks also required more time to process a gait sequence. Furthermore, most recent methods proposed did not have architectures targeted towards emotion recognition, which resulted in sub-optimal performance. Additionally, the limited affective feature set used was introduced towards the end of their networks. The deep features and the affective features were never processed together before classification. Hence, the classifiers had to rely on information from the low level affective features. Moreover, the architectures did not fully exploit the advantages the respective type of neural network had to offer. Hence, this thesis introduces a model that addresses the shortcomings in the domain.

Chapter 3

Proposed Bi-Modal Gait Emotion Recognition Methodology

Since gait is a time-series data containing coordinates for each joint in the subject's gait skeleton, it is imperative to employ sequential data processing methods to extract temporal features. Furthermore, exploration of spatial features exhibited by the gait skeleton in each frame is also important for processing gait data. Hence, the deep learning architectures proposed in this thesis are made up of neural networks that are proficient at processing sequential data to extract spatial as well as temporal gait features.

3.1 BMSNN Architecture for Gait Emotion Recognition

The proposed Bi-Modular Sequential Neural Network (BMSNN) utilizes Long Short Term Memory (LSTM) units. These units explore low-level features in the temporal as well as the spatial dimension, by employing memory cells to calculate features across frames and a weight matrix to extract features within a frame, respectively. This addresses the fundamental issue with GNNs in gait recognition, where the low-level features are computed only from neighbouring joints, while low-level features dependant on body-joints/graph-nodes far away from one another (spatially or temporally) are unaccounted for.

Once the LSTM subnetwork produces the feature set, the features must be consolidated to obtain the classification results. As discussed in the previous section, due to their ability to automatically extract and select new discriminating features, Multi Layer Perceptrons (MLP) are a great choice for the selection of latent features and for the subsequent classification. Therefore, the functionality of the MLP subnetwork is twofold: feature extraction to produce a condensed feature set and mapping those features to the various

emotion classes.

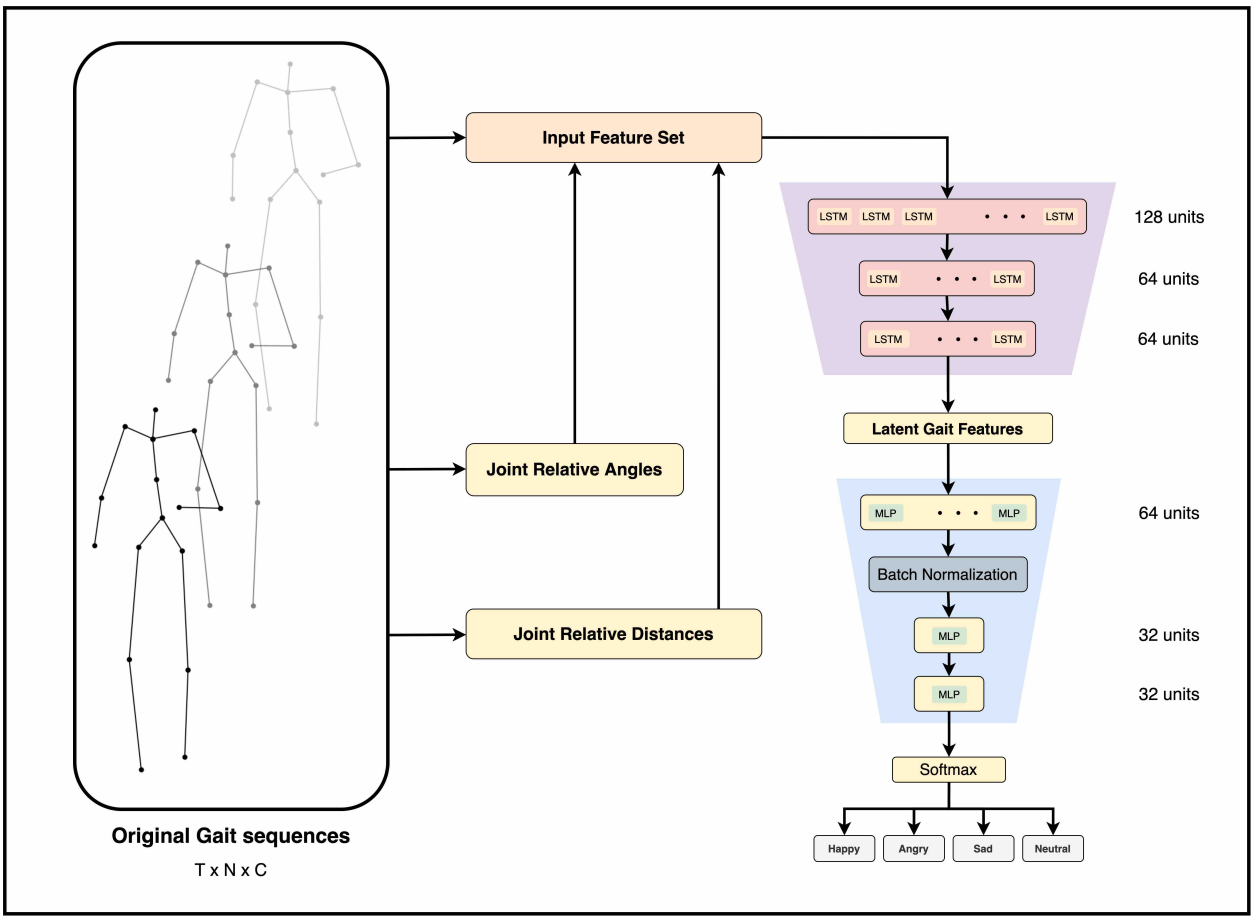


Figure 3.1: Architecture of the proposed Bi-Modular Sequential Neural Network

3.1.1 Deep Learning Architecture

The proposed Bi-Modular Sequential Neural Network (BMSNN) is comprised of two modules: the LSTM subnetwork and the MLP subnetwork, as described in Figure 3.1. To address the first research question: "Can an attenuated hybrid deep learning architecture be devised to achieve low inference time for gait emotion recognition?", the overall architecture of the network has a tapered design. This produces a condensed representation of the input by ensuring that the amount of information received is represented by a smaller set of features towards the end. This is achieved through an architecture where the next layer has either an equal or smaller number of units than the previous one. Such a design also ensures that the network has a low number of parameters, since the number of units in each layer decreases with the depth of the network. The second research question: "Can a deep learning architecture for gait-based emotion recognition be designed to identify distinctive sequential and temporal features extracted from body joints?", is answered

by implementing the first half of the network using sequential neural networks that can extract spatial and temporal features effectively, and the second half of the network that is adept at identifying distinctive features. Combined with the attenuated design requiring lesser parameters, the proposed network addresses the third research question: "Can a light deep learning architecture combining a sequential neural network and multi-layered perceptrons be used to accurately recognize emotions from human gaits?".

Unlike prior research, the network is trained with a categorical cross entropy loss for each of the emotion classes, which ensures that network weights are trained specifically to extract features relevant for the task of emotion recognition. The network accepts an input vector of size $[T, (N \times C) + F]$, where T is the number of the time steps in each gait sequence, F is the size of the handcrafted feature set, N is the number of body joints in the body skeleton, and C is the number coordinates for each body joint (thus, for an input of only raw gait sequences, $F = 0$). This sequential gait input is processed by a three-layered LSTM subnetwork containing 128, 64, and 64 LSTM units, respectively. All of these layers have a Hyperbolic Tangent (Tanh) activation function to ensure that the negative activation values are not ignored while introducing non-linearity. The resulting feature set of size 64 is passed onto the MLP subnetwork, which has three layers with decreasing number of units (64, 32, and 32). Additionally, the activations of the first MLP layer are batch normalized with a momentum of 0.1 that resulted in a robust performance across all classes of emotions as shown in Chapter 4. This subnetwork is responsible for producing a denser set of 32 features that is mapped to the four emotion classes, following which a Softmax activation is used to convert the scores into probabilities. Lastly, the class with the highest probability value is chosen.

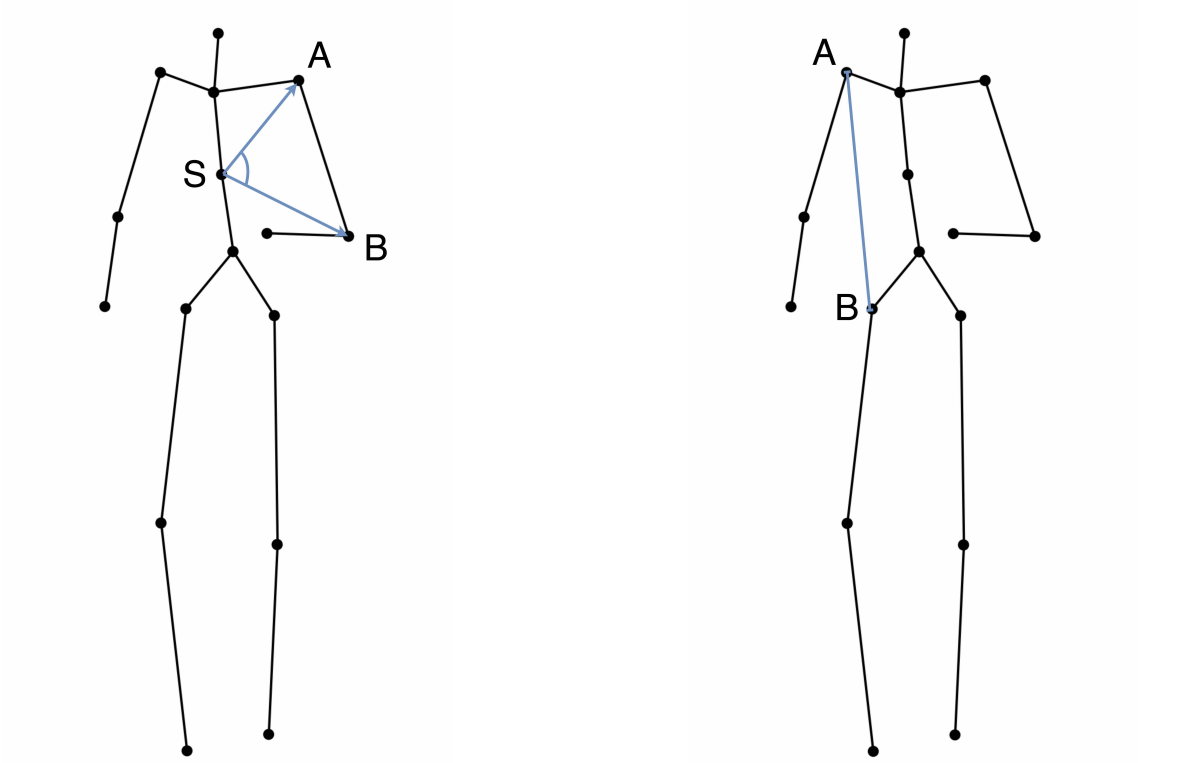
3.1.2 Leveraging Handcrafted Features

While using a purely deep learning approach has its advantages, in this thesis we demonstrate that it can be further empowered by the previous knowledge from the problem domain. The data driven approach of deep learning methodologies is highly influenced by the dataset used to train the deep learning network. This characteristic can cause poor performance for classes in the dataset which do not have as many samples as the other classes. The problem can be mitigated by avoiding relying solely on deep learning. Therefore, the proposed architecture incorporates robust handcrafted gait features as well. This component of the proposed system also addresses the fourth research question, "Can handcrafted features based on the geometric relationships between body joints be combined with the deep learning architecture to further improve recognition performance and to make the architecture resilient to class imbalance in the dataset?".

One way to make use of the sequential handcrafted features with a deep learning architecture is to concatenate them with the output of the LSTM subnetwork before feeding them to the MLP subnetwork.

This can be done by utilizing statistical values that hold information about the sequential data, for instance the mean or the highest value of the series. Unfortunately, using such values ignores the latent features present in that sequential data, and as such not practical. Therefore, the network is designed to accept a feature set as its input, which contains two types of handcrafted features in addition to the raw gait sequences. These features, along with the raw gait sequences, are used to produced higher level features in the first half of the network which are then refined and classified into emotions by the second half of the network.

To find the most suitable feature set for the model, combinations of four gait-specific handcrafted features that demonstrated best performance in prior research for processing gait data are considered [2, 15]. The features: Joint Relative Angles (JRAs), and Joint Relative Distances (JRDs), describe the geometric and directional motions of a subject's body joints [15]. All handcrafted features mentioned below are computed for each frame in a gait sequence. Furthermore, all possible relative angles and distances are considered in the proposed method to overcome the limitation of favoring only a few body joints.



(a) The Joint Relative Angle formed by the right shoulder joint and right elbow joint

(b) The Joint Relative Distance between the left shoulder joint and the left hip joint

Figure 3.2: Examples of the JRA and JRD Geometric Handcrafted Features

Joint Relative Angles

The JRA between two body joints, $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$, is the angle formed at the mid-Spine joint $S(x_0, y_0, z_0)$ between the vectors $\vec{SA}$ (vector from the mid-Spine joint to body joint A) and $\vec{SB}$ (vector from the mid-Spine joint to body joint B), (see Figure 3.2a). The angle is defined as the inverse Cosine of the dot product of $\vec{SA}$ and $\vec{SB}$ over the product of the magnitude of the two vectors. Equation (3.1) describes the calculation mathematically. The $N-1$ body joints (excluding the mid-Spine joint) result in $N-1$ vectors originating from the mid-Spine joint, which are used to calculate the $\binom{N-1}{2}$ angles formed between all possible pairs of vectors. Thus, the size of the JRA feature set is $(T, \binom{N-1}{2})$, where T is the number of time steps in the gait sequence. This feature is representative of relative angular motions of the various body joints. A stable joint that remains mostly stationary throughout the gait is required to be the relative joint. Hence, the mid-Spine is chosen as the relative joint for all the features [15].

$$JRA(A, B) = \cos^{-1}\left(\frac{\vec{SA} \cdot \vec{SB}}{\|\vec{SA}\|\|\vec{SB}\|}\right) \tag{3.1}$$

Joint Relative Distances

Additionally, the JRD considers the relative motion of various body joints in terms of distance [2]. The JRD between two points $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ is calculated as the euclidean distance between the two joints, (see Figure 3.2b). The mathematical formula for this calculation is mentioned in Equation (3.2). Similar to JRAs, $\binom{N}{2}$ JRDs are calculated for all possible combinations of two joints in the body skeleton for one frame of the gait sequence.

$$JRD(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{3.2}$$

Two more features, namely Joint Relative Triangle Areas (JRTAs) and Joint Relative Cosine Dissimilarities (JRCDs), introduced in [15] were considered as a part of the input feature set. JRTAs can be calculated for any three body joints $A(x_1, y_1, z_1)$, $B(x_2, y_2, z_2)$, and $C(x_0, y_0, z_0)$. It is described as half of the norm of the cross product of $\vec{AB}$ (the vector from point A to point B) and $\vec{BC}$ (the vector from point B to point C), as shown in Equation (3.3). Similarly, JRCDs can be described as the Cosine distance between two points $\vec{A}$ and $\vec{B}$, as shown in Equation (3.4).

$$Ar(\triangle ABC) = \frac{\|\vec{AB} \times \vec{BC}\|}{2} \tag{3.3}$$

$$\delta_{cosine} = 1 - \left(\frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \right) \tag{3.4}$$

However, on comparing the feature set of raw gait sequences with JRA and JRD and with JRTA and JRCD, it was found that JRTA and JRCD do not affect the model precision. Hence, the final feature set included raw gait sequences, JRAs and JRDs, of size $[T, (N \times C) + \binom{N-1}{2} + \binom{N}{2}]$, where T is number of time steps in a gait sequence, N is number of body joints for each time step, C is the number of coordinates for each joint in a time step, and $\binom{N-1}{2}$ and $\binom{N}{2}$ are the sizes of the handcrafted feature set of JRA and JRD, respectively.

This section first introduced the novel Bi-Modular Sequential Neural Network (BMSNN) architecture that consists of the LSTM subnetwork and the MLP subnetwork. These subnetworks facilitate efficient extraction and selection of sequential features that result in high emotion recognition performance, as will be shown in Chapter 4. The performance is further increased by incorporating gait-specific geometric handcrafted features (JRAs and JRDs). The architecture parameters such as the activation function, learning rate, batch size, etc. were optimized as will be discussed in Chapter 4. The results demonstrate that the proposed novel architecture surpasses all other recently developed deep-learning based methods for gait emotion recognition.

3.2 Improved BMDNN Architecture for Gait Emotion Recognition

While the previously described architecture proposes a novel fusion of deep features extracted using a sequential neural network and robusts handcrafted features, it can be further enhanced by incorporating a broader set of latent deep features as seen in Figure 3.3. The stability and recognition performance are further improved by incorporating domain-specific Laban Movement Analysis features to capture the dynamic structural properties of a subject's body while walking. Information-rich LMA-based features are fed to the MLP subnetwork to fuse with deep features which results in robustness and resilience to the imbalanced dataset. The proposed improved architecture achieves remarkable precision scores across all emotion classes and outperforms, in addition to previous architecture, all recent state-of-the-art methods. Hence, the architecture presented in this section answers the fifth and the sixth questions, namely: "Can domain-specific handcrafted features be fused with latent deep features to improve gait emotion recognition performance?" and "How do the Laban Movement Analysis feature groups affect the performance of the proposed network?".

The Bi-Modal Deep Neural Network (BMDNN) consists of two updated modules. The LSTM-based

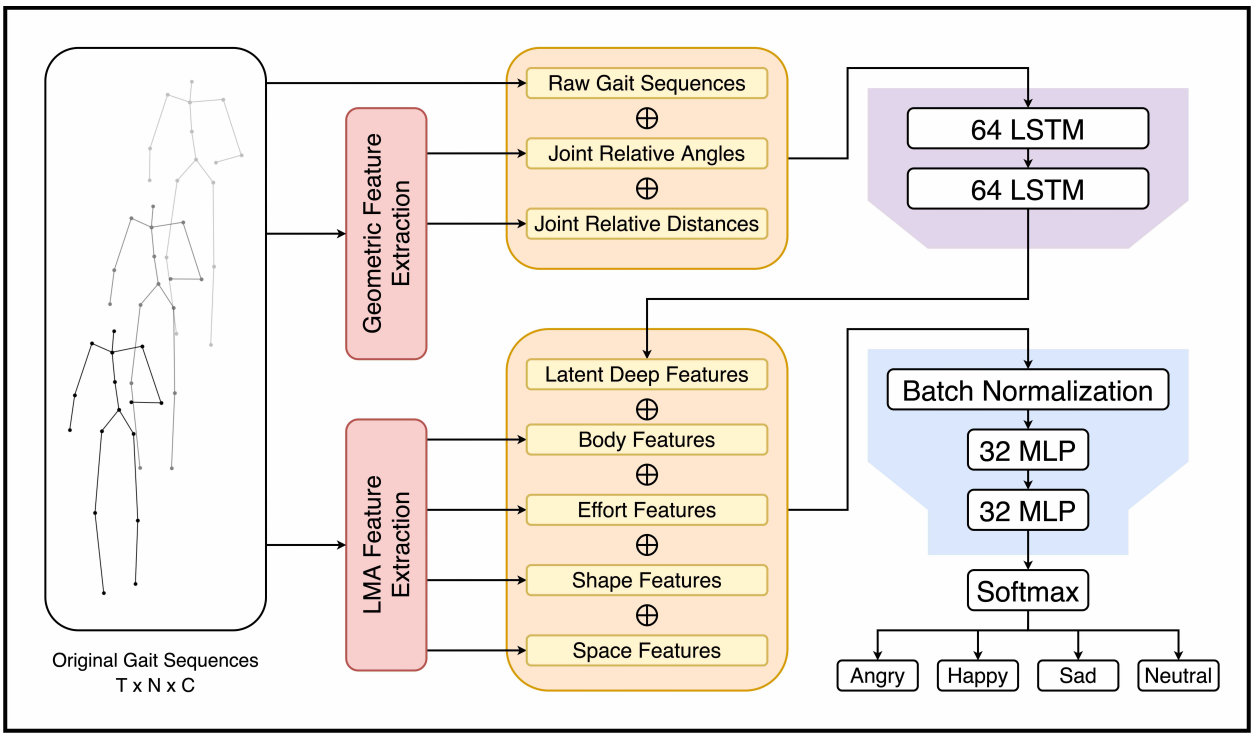


Figure 3.3: Architecture of the proposed Bi-Modal Deep Neural Network

feature extraction module sequentially processes gait data to produce a rich feature vector. The MLP-based decision module is responsible for combining and condensing the information extracted by the LSTM subnetwork and the features calculated using LMA, and for mapping to the four emotion classes. Moreover, the batch normalization layer in the MLP module ensures lower loss during training with fast and smooth parameter updates.

3.2.1 Deep Learning Architecture

The input to the first module is a concatenated vector of size $[T, (N \times C) + F]$, where T is the number of frames of each gait sequence, N is the number of body joints, C is the number of coordinates for each body joint, and F is the combined size of the angle and distance-based handcrafted features (JRAs and JRDs), described in Section 3.1.2. The input to the second module consists of the latent deep features of size 64, extracted from the first module, concatenated with the robust domain-specific LMA features. If the inputs are not normalized, optimization of the model is skewed and produces too large or too small gradient values, restricting optimal parameter updates. Moreover, since the input to the first module is gait data, normalizing it through conventional pre-processing methods would distort the structure of the gait sequences. Hence, the combined feature vector from both inputs is normalized by the network via the batch normalization layer in the beginning of the MLP module.

The first half of the network has two LSTM layers, each with 64 units, a Hyperbolic Tangent (Tanh) activation, and a L2 regularizer with a penalty of 0.01. The regularizers penalize high weights and biases, hence preventing overfitting. This improvement resulted in an increase in the system's performance, as shown in Chapter 4. The Tanh activation ensures that the negative values from the inputs are not ignored while producing the activations. The second half of the architecture contains a batch normalization layer before the two MLP layers, which improves the network's precision for under-represented classes, as shown in Chapter 4. Each layer has 32 units with Tanh activations to ensure an overall tapered design for refining features. The Tanh activation function was selected for both subnetworks via comparative experimentation with other activation functions (see Chapter 4, Section 3). The second subnetwork combines the information from the features extracted by the LSTM subnetwork with the LMA-based handcrafted features, to produce high-level features. These high-level features are mapped to the four emotion classes using a Softmax activation.

Out of the commonly used optimizers, the model exhibited the best loss convergence with RMSprop (see Chapter 4, Section 3). Similarly, 400 was found to be the optimal value for the number of training epochs (further information on these experiments is provided in Chapter 4). Hence, the model parameters are optimized using RMSprop optimizer with a momentum of 0.5, a rho of 0.3, and an epsilon of $1e-7$ for 400 epochs. Furthermore, the training is performed using an experimentally optimized batch size of 64 (refer to Chapter 4) and a categorical cross-entropy loss function. As mentioned earlier, DNN-based approaches are sensitive to the composition of the training data and might result in low performance for under-represented emotion classes. Therefore to introduce more robustness, this thesis proposes a novel hybrid architecture that integrates the domain-specific features.

3.2.2 Laban Movement Analysis Features

Laban Movement Analysis (LMA) [98] has been effective at discriminating emotions from body movements [101]. Prior works have successfully identified emotions from human motion using LMA based features [77, 29]. However, no previous work in gait emotion recognition combined the powerful LMA based handcrafted features with features extracted using a deep neural network. The domain has also not seen any work that processes handcrafted and deep features together to derive more information rich features. Additionally, the sensitivity of deep learning models toward the data distribution was not addressed previously.

To improve the performance of the model for emotion classes with low data representation, the proposed method employs statistically cumulated handcrafted features that are resilient to unbalanced datasets. These handcrafted features are based on the Laban Movement Analysis (LMA) [98] which provides a structural description of the movement of a subject's body using four groups: body, effort, shape, and space. These

groups comprise of 17 features calculated in the temporal domain, i.e. the features are calculated for each time frame of the gait sequence and have a combined size of $(17, T)$. The LMA features must be converted to a one-dimensional vector to make it compatible for the MLP subnetwork. This flattening of the features is performed by calculating histogram values on 100 bins, thus making the final feature set of shape $(1700, 1)$. The final feature set is concatenated with the latent deep features extracted from the LSTM subnetwork $(64, 1)$ to form the input vector for the MLP subnetwork.

Body Features

The *Body* feature group (Head Inclination Angle, Flex Angle, Abduction Angle, Knee Angle, Stride Angle, Knee Stride Length, Foot Stride Length) describes the physical and structural characteristics of the body using seven angle and distance measures. These features capture the information about the connections of the body as it moves [56], are beneficial for emotion recognition [2] and can be calculated using formulae using Equations (3.1) and (3.2) described in Subsection 3.1.2.

Effort Features

The *Effort* feature group (Kinetic Energy, Knee Average Velocity, Foot Average Velocity, Elbow Average Velocity, Wrist Average Velocity) encapsulate the subtle intent behind the motion of a body by measuring the energy/force put into the motion. This feature group describes the amount of expressiveness [101]. The velocity of a particular joint at a time frame i can be calculated as the difference between the joint's position at the i^{th} frame and the joint's position at the $i + 1^{th}$ frame, described in Equation (3.5). Additionally, the Kinetic Energy at a given time frame i is calculated using Equation (3.6), where m is the mass of the joint $(m = 1)$, v is the velocity of the k^{th} joint and N is the total number of the body joints.

$$V_A^i = A^{i+1} - A^i \tag{3.5}$$

$$KE^i = \frac{1}{2N} \sum_{k=1}^{N} m.v_k^{i^2} \tag{3.6}$$

Shape Feature

The next movement component, *Shape*, contains a single feature: Density Index, which captures the progression of the body's shape change with respect to time. The metric represents the variation of the body shape throughout the gait, which indicates the smoothness/unevenness of the movements, linked to the comfort level of the subject [150]. First, the centroid C of the body for each frame i is calculated according to

Equation (3.7), where J_k is the vector containing the x, y, and z coordinates of the k^{th} body joint, and N is the total number of body joints. Finally, the Density Index (DI) is calculated as described in (3.8), where J_{kx} is the x coordinate of the k^{th} body joint.

$$C^i = \frac{1}{N}\sum_{k=1}^{N} J_k \tag{3.7}$$

$$DI^i = \frac{1}{N}\sum_{k=1}^{N} \sqrt{(C_x^i - J_{kx}^i)^2 + (C_y^i - J_{ky}^i)^2 + (C_z^i - J_{kz}^i)^2} \tag{3.8}$$

Space Features

The fourth category in LMA is *Space* (Whole Body Bounding Volume, Upper Body Bounding Volume, Lower Body Bounding Volume, Spatial Symmetry Index), which delineates the way a subject makes use of the surrounding space during a gait. The Spatial Symmetry Index is indicative of relaxation [123]. The Bounding Volume (BV) is the product of d_x, d_y and d_z (Equation (3.12)), which are distances calculated in Equations (3.9), (3.10), and (3.11). Lastly, the Spatial Symmetry is calculated by computing the barycenter of the skeletal body for each frame, and then using it to calculate the Symmetric Indices for each axis. The Symmetric Index (SI) at a given time frame i for an axis w is defined in Equation (3.13), where LW, RW, and BC represent the coordinates of the left wrist joint, the right wrist joint and the barycenter of the body. Subsequently, the overall Symmetry Index is calculated according to Equation (3.14).

$$d_x = max_{k\in K} J_{kx} - min_{k\in K} J_{kx} \tag{3.9}$$

$$d_y = max_{k\in K} J_{ky} - min_{k\in K} J_{ky} \tag{3.10}$$

$$d_z = max_{k\in K} J_{kz} - min_{k\in K} J_{kz} \tag{3.11}$$

$$Bounding\ Volume = d_x \times d_y \times d_z \tag{3.12}$$

$$SI_w^i = \frac{(LW_w^i - BC_w^i) - (RW_w^i - BC_w^i)}{(LW_w^i - BC_w^i) + (RW_w^i - BC_w^i)} \tag{3.13}$$

$$SI^i = \sqrt{(SI_x^i)^2 + (SI_y^i)^2 + (SI_z^i)^2} \tag{3.14}$$

Table 3.1: Summarized description of the Laban Movement Analysis features

LMA Feature Group	**Feature Name**	**Description**
Body	Head Inclination Angle	The angle formed at the Neck joint, by the Head joint and the Base Spine joint
	Flex Angle	The average of the angles formed at the left and the right Elbow joints, by the corresponding Shoulder and Wrist joints
	Abduction Angle	The average of the angles formed at the left and the right Shoulder joints, by the corresponding Elbow joints and the Neck joint
	Knee Angle	The average of the angles formed at the left and the right Knee joints, by the corresponding Hip and Ankle joints
	Stride Angle	The angle formed at the Base Spine joint, by the left and right Ankle joints
	Knee Stride Length	The Euclidean distance between the left and right Knee joints
	Foot Stride Length	The Euclidean distance between the left and right Ankle joints
Effort	Kinetic Energy	The square of the velocities of all the body joints
	Knee Average Velocity	The average of the velocities of the left and the right Knee joints
	Foot Average Velocity	The average of the velocities of the left and the right Ankle joints
	Elbow Average Velocity	The average of the velocities of the left and the right Elbow joints
	Wrist Average Velocity	The average of the velocities of the left and the right Wrist joints
Shape	Density Index	The sum of the distance between the centroid and each body joints, divided by the number of body joints
Space	Whole Body Bounding Volume	The volume of the cuboid formed by the farthest body joint coordinates in all three axes
	Upper Body Bounding Volume	The volume of the cuboid formed by the upper body joint coordinates (Head, Neck, left and right Elbows, Wrists, and Shoulders) in all three axes
	Lower Body Bounding Volume	The volume of the cuboid formed by the upper body joint coordinates (Base Spine, left and right Hips, Knees, and Ankles) in all three axes
	Spatial Symmetry Index	The ratio of the difference and sum of the displacements of the left and right wrist joints from the barycenter of the body

This section introduced a Bi-Modal Deep Neural Network that improves on the BMSNN architecture by introducing domain-specific and emotionally relevant Laban Movement Analysis-based features (Summa-

rized in Table 3.1). This modification results in a remarkable improvement in the network precision that outperforms the recent state-of-the-art methods as well as the proposed BMSNN architecture, as shown in Chapter 4. The experiments presented in Chapter 4 justify the optimal values of the network and validate the importance of the key components of the network through an ablation study. Furthermore, the effects of various groups of the LMA features on the performance of the proposed BMDNN architecture are also studied.

Alternate Network Configurations for the BMDNN

While developing the proposed architecture, alternative variations were experimented with. However, it was found that when LMA features were processed sequentially in concatenation with JRAs, JRDs and raw gait sequences by the LSTM subnetwork, the overall performance of the model decreased while its sensitivity towards the dataset's class distribution increased. This was caused due to the tendency of neural networks to learn the biases in the dataset. A similar decrement in the overall performance was observed when the JRAs and JRDs were passed alongside the LMA features to the MLP subnetwork. The reasons behind it is threefold. Firstly, adding new features to the input of the MLP increased the vector size and hence the network complexity which caused the network to start overfitting. Secondly, low-level metrics contained in JRAs and JRDs were no longer being processed to form higher-level features. Lastly, the distinction in the processing of the angular and distance metrics present in the LMA feature group and the geometric features was lost, which led to similar feature being produced.

Inputting the JRAs and JRDs with the raw gait sequences allowed the network to sequentially produce high-level features. Simultaneously, passing the higher-level LMA features directly to the MLP subnetwork promoted resilience towards the class imbalance. Hence, the current configuration of the Bi-Modal Deep Neural Network was optimal for Gait Emotion Recognition problem. The detailed experimentation on the proposed network is presented in the next chapter.

3.3 Summary

The methodology proposes a novel bi-modal deep learning architecture to identify human emotion from gait. The proposed architectures employ LSTM units to allow low-level feature extraction from all possible combinations of body joint coordinate values in the spatial as well as the temporal domain. The developed BMSNN and BMDNN architectures are trained using categorical cross-entropy loss to effectively distinguish between various emotion classes. The BMSNN architecture presents a novel light design with fewer parameters that outperforms the prior works while ensuring inference times that are a fraction of the comparators, as will

be shown in the next chapter. The architectures proposed in this thesis incorporated, for the first time, powerful JRAs and JRDs geometric features with Laban Movement Analysis features to create light-weight architecture resilient to data imbalance for emotion recognition from gaits. The next section presents extensive experimentation that confirm that the proposed methodology outperforms all recent state-of-the-art deep learning methods. Additionally, a discussion of each Laban Movement Analysis feature group's affect on the proposed BMDNN architecture's performance is included.

Chapter 4

Experimental Results

Various experiments were performed in this research for optimization, validation, and comparison purposes. First, the hyper-parameter optimization was performed to fine-tune the model to ensure the highest emotion recognition performance. Next, an ablation study was performed to demonstrate that all the elements of the proposed architecture play a crucial role in recognizing emotion. Then, experiments on handcrafted features and LMA were conducted. Finally, the proposed architecture was compared to the recent state-of-the-art methods. This chapter describes the above-mentioned experiments and discusses the findings in detail.

4.1 Experimental Setup

To ensure that the proposed architecture learns optimal parameter values, both the precision and the loss metrics were monitored for the Bi-Modal Deep Neural Network's (BMDNN) training and validation stages. The BMDNN model was trained using a categorical cross-entropy loss function. The loss function, $L_{categoricalcross-entropy}$, is described in Equation (4.1), where C is the total number of classes, y_i^{true} is the true label value for a class i, and y_i^{pred} is the network's predicted value for class i.

$$L_{categoricalcross-entropy} = -\sum_{i=0}^{C} y_i^{true} \log(y_i^{pred}) \tag{4.1}$$

After the model was trained successfully, micro and macro mean Average Precision were recorded to measure the model's ability to classify emotions correctly. The macro mAP is computed as the mean of all individual class average precision scores (described in Equation ((4.2)), where AP_i is the average precision for class i, and C is the total number of classes). Hence, macro mAP is a good metric to measure the model's resilience towards the imbalance in the dataset. Micro mAP considers each sample as a unique class. This

metric is suitable for measuring the overall performance on an unbalanced datasets like the one used in this research, since it considers the varied representation of classes in the dataset. The micro mAP measure is defined as the global average of the Average Precision metrics. Hence, micro mAP measures the number of samples classified correctly irrespective of their classes and macro mAP represents the model's overall performance with respect to all the classes [199].

$$macro\ mean\ Average\ Precision = \frac{AP_1 + AP_2 + ... + AP_{\mathrm{C}}}{C} \tag{4.2}$$

4.2 Dataset

All the experiments were run with a data split of 80:10:10 for training, validation and testing using a stratified shuffling on a subset of the Edinburgh Locomotion MoCap Dataset (ELMD) [69]. This dataset was originally collected by the researchers from the University of Edinburgh and later annotated by Bhattacharya et al. [21]. The dataset was collected to understand and recreate human locomotion, and hence contains gait sequences with varying conditions: walks that are slow, fast, in different directions, non-linear, around obstacles, etc. The modified ELMD dataset consists of 1835 gait sequences recorded for 4 seconds at 60 Hz. Thus, each gait sequence in the dataset has 240 frames. Each gait sequence consists of 48 values which correspond to 3 coordinates of 16 body joints (see Figure 4.1 for details). While Gait Emotion Recognition (GER) can be performed using gait skeletons with fewer number of body joints and/or lower dimensional coordinates, having fewer amount of data points would be detrimental to the network's performance. These gait sequences were labelled by [21] into four categories of emotions: Angry, Happy, Sad, and Neutral. The labels were generated by various participants on a crowd sourced website that provided ratings to each sequence based on the emotions they perceived from that gait. This resulted in a dataset of 1835 gait sequences, with each sequence consisting of 240 frames, and each frame containing 16×3 values. These 1835 gait sequences consist of 1048 Angry gaits, 454 Happy gaits, 254 Sad gaits and 79 Neutral gaits.

ELMD is the benchmark dataset for emotion recognition research from gait. This dataset is very popular since it is an emotionally-labelled gait data, which is not artificially synthesized and has a sufficient number of samples for training a deep learning methodology. It has, however, an unequal number of class samples corresponding to different emotions. In all of the prior research, higher precision values were observed for the classes with more samples and vice-versa. However, a positive outcome of this attribute is that the dataset allows for a clear identification of methods that are too susceptible to the class sample distribution.

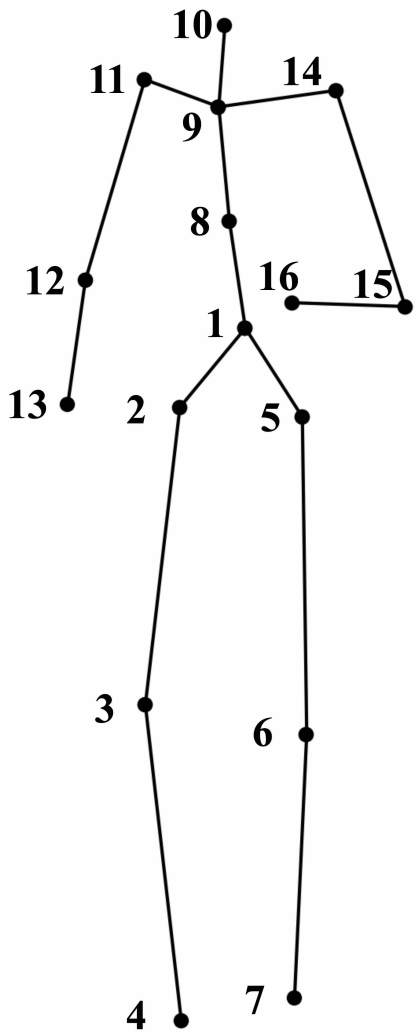
Gait Skeleton	Body Joints
	1. Base Spine 2. Left Hip 3. Left Knee 4. Left Heel 5. Right Hip 6. Right Knee 7. Right Heel 8. Mid Spine 9. Neck 10. Head 11. Left Shoulder 12. Left Elbow 13. Left Hand 14. Right Shoulder 15. Right Elbow 16. Right Hand

Figure 4.1: Modified gait skeleton joints from Edinburgh Locomotive MoCap Dataset

4.3 Hyperparameter Tuning Experiments

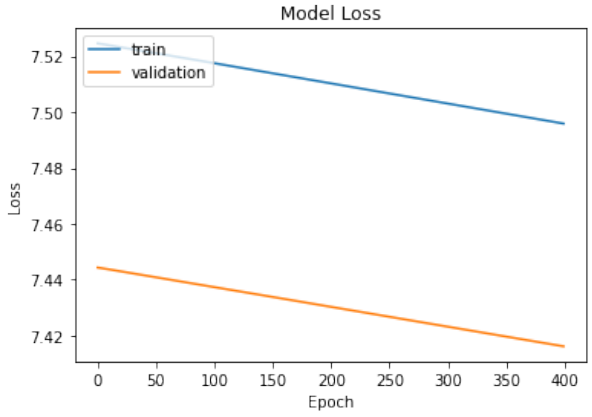
4.3.1 Optimizer Selection

In a nutshell, the optimizer of a neural network is responsible for updating the network's trainable parameters while ensuring that the overall network loss is reduced. There are a variety of optimizers available in the domain of Deep Learning, each one offering a different way to utilize the loss function value to change the network's parameter values. This research considered four commonly used neural network optimizers for the proposed BMDNN model: Adam, RMSprop, SGD, and AdaDelta.

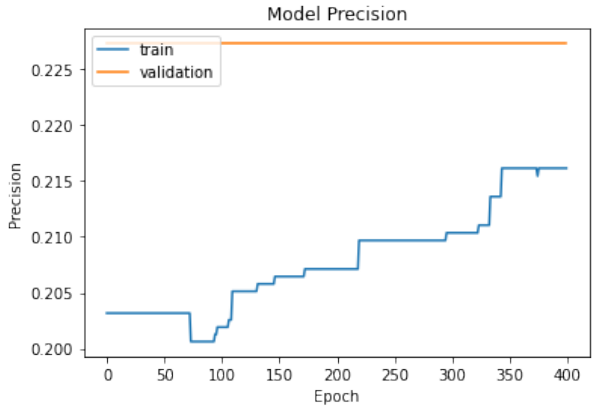
The network was unable to train properly with an AdaDelta optimizer as seen from the poor loss and precision graphs in Figure 4.2a and 4.2b respectively, and the low precision scores in Table 4.1. A slight improvement was seen with Stochastic Gradient Descent (SGD) optimizer, which provided a smooth learning curve for the network but was unable to achieve a good performance, see Figure 4.3. On the other hand, ADAptive Movement estimation (Adam) and Root Mean Square propagation (RMSprop) optimizers led to steep training loss graphs, which showed a significant decrease in the network loss (Figures 4.4a and 4.5a). Out of the two, RMSprop can be clearly identified to have had the best overall performance, followed by Adam optimizer (see Table 4.1). Hence, RMSprop optimizer was selected as the optimizer for the model.

Table 4.1: Performance comparison of the proposed model for different optimizers

Optimizers	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
AdaDelta	0.52	0.22	0.18	0.08	0.25	0.25
SGD	0.94	0.58	0.23	0.08	0.58	0.46
Adam	0.99	0.93	0.95	0.91	0.97	0.94
RMSprop	**0.99**	**0.95**	**0.97**	**0.91**	**0.98**	**0.96**



(a) The loss values for the training and validation of BMDNN with the AdaDelta optimizer

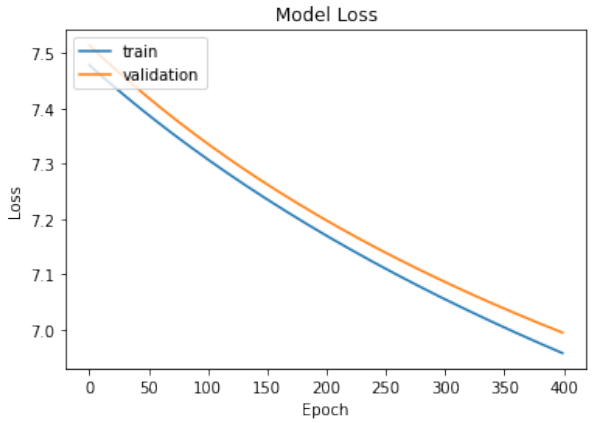


(b) The precision values for the training and validation of BMDNN with the AdaDelta optimizer

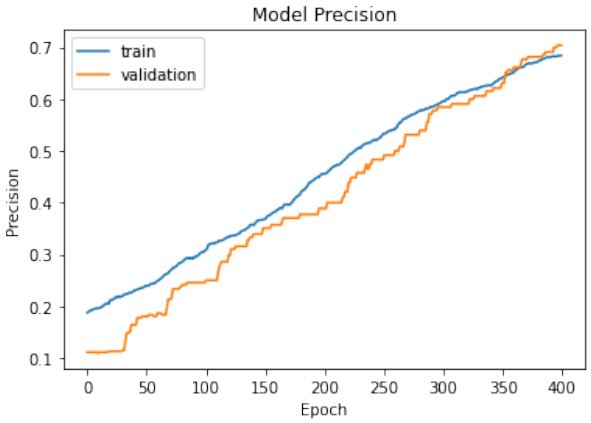
Figure 4.2: The proposed BMDNN's loss and precision graphs for the training and validation datasets with the AdaDelta optimizer

4.3.2 Learning Rate Selection

The learning rate of a model regulates the magnitude of the weight updates and therefore, how fast the model converges. A large learning rate would result in unstable parameter updates, and produce an untrained model. This is generally depicted by an oscillating (or a rugged) loss curve since the model's loss keeps varying without gradually decreasing. Simultaneously, if the chosen learning rate is too small, the model

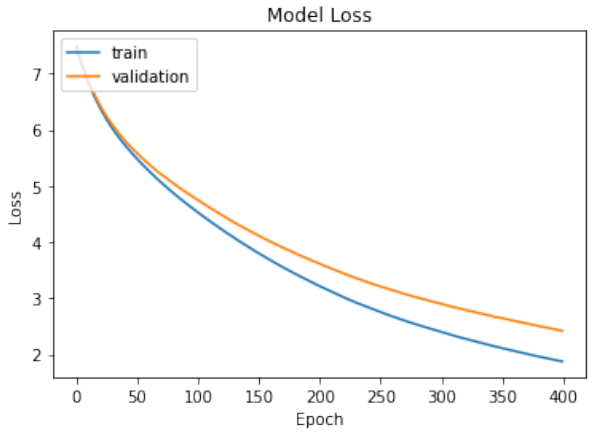


(a) The loss values for the training and validation of BMDNN with the SGD optimizer

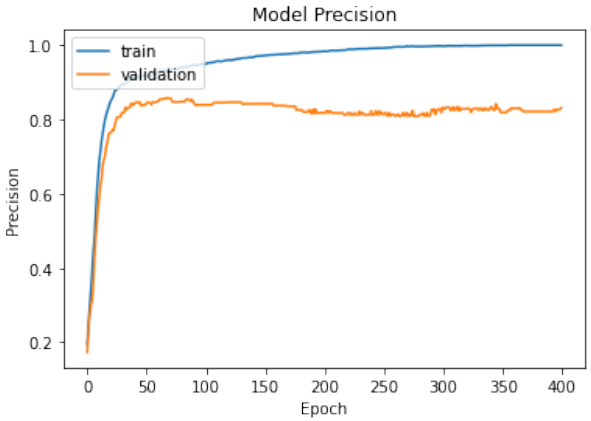


(b) The precision values for the training and validation of BMDNN with the SGD optimizer

Figure 4.3: The proposed BMDNN's loss and precision graphs for the training and validation datasets with the SGD optimizer



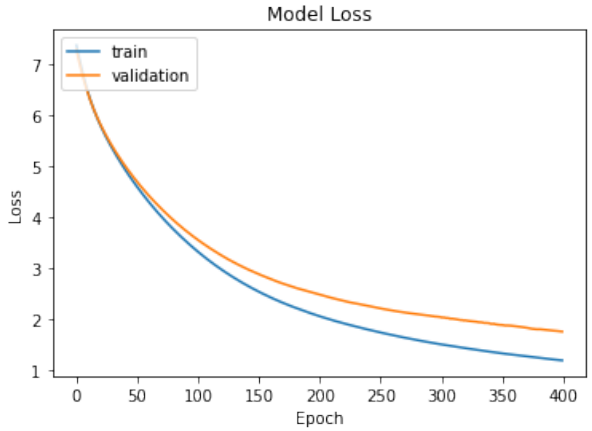
(a) The loss values for the training and validation of BMDNN with the Adam optimizer



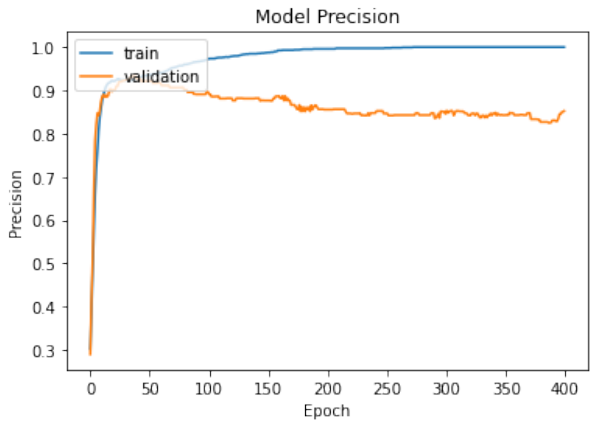
(b) The precision values for the training and validation of BMDNN with the Adam optimizer

Figure 4.4: The proposed BMDNN's loss and precision graphs for the training and validation datasets with the Adam optimizer

will make minuscule changes to the network parameters. This will result in smooth loss curves; however, the model will remain untrained after numerous epochs. Hence, an optimal learning rate ensures that the network parameter updates result in a consistent and a sufficient amount of reduction in the model loss. To determine an optimal value for this hyper parameter, the BMDNN model was trained on learning rates ranging from $1e-6$ to $1e-4$. As seen in Table 4.2, the learning rate of $1e-5$ results in the best precision of the model and hence was selected.



(a) The loss values for the training and validation of BMDNN with the RMSprop optimizer



(b) The precision values for the training and validation of BMDNN with the RMSprop optimizer

Figure 4.5: The proposed BMDNN's loss and precision graphs for the training and validation datasets with the RMSprop optimizer

Table 4.2: Performance comparison of the proposed model for different learning rates

Learning Rates	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
1e-6	0.99	0.84	0.80	0.39	0.94	0.76
2e-6	0.99	0.88	0.86	0.74	0.96	0.87
4e-6	0.99	0.93	0.95	0.87	0.98	0.94
8e-6	0.99	0.96	0.94	0.81	0.98	0.93
1e-5	**0.99**	**0.95**	**0.97**	**0.91**	**0.98**	**0.96**
2e-5	0.99	0.90	0.94	0.86	0.97	0.93
4e-5	0.99	0.93	0.95	0.81	0.97	0.92
8e-5	0.99	0.92	0.95	0.84	0.97	0.93
1e-4	0.99	0.90	0.90	0.81	0.95	0.90

4.3.3 Batch Size Selection

Batch Size refers to the number of samples processed by the model to calculate the loss before the parameters are updated. This hyper parameter dictates the regularization of the model. Too small batch sizes result in the model fixating on only a few samples at a time while tuning its parameters. Hence, it results in an unstable learning which translates to rough/rugged loss curves. In contrast, batch sizes that are too large produce smooth loss curves but result in fewer updates within one epoch (training cycle) of the network. Hence the number of updates required to properly train the model increases.

Various batch sizes ranging from 16 to 128 were used with the network to ensure that a sufficient number of examples are processed by the network before updating its weights. Table 4.3 displays the various batch sizes that were used to train the network and their effect on the model's precision scores. Consecutively, a batch size of 64 ensured that each class sample had appropriate representation in a batch and hence resulted

in the highest mean average precision values, while avoiding overfitting to the data. This behaviour can be observed as the high mAP value for a batch size of 64 in Table 4.3.

Table 4.3: Performance comparison of the proposed model for different batch sizes

Batch Sizes	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
16	0.99	0.92	0.95	0.91	0.97	0.94
32	0.99	0.94	0.96	0.89	0.98	0.95
64	**0.99**	**0.95**	**0.97**	**0.91**	**0.98**	**0.96**
128	0.99	0.94	0.96	0.83	0.97	0.93

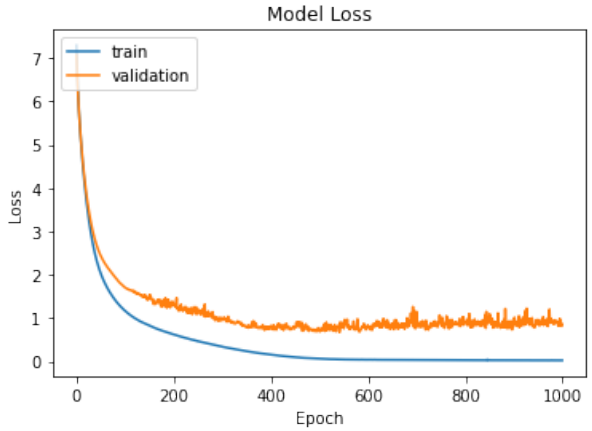
4.3.4 Number of Epochs Selection

Each training cycle of a neural network involves processing a batch of training samples, calculating the network loss, and finally updating the parameters of the network accordingly. These cycles are called epochs and govern the amount of training a model undergoes. High number of epochs result in an over-trained model that only performs well on the data that was used in training. Such a model is called an overfitted model and is said to have a high variance, in the literature. Conversely, an under-trained model, that arises due to a low number of training epochs, does not tune its parameters enough to fit the training data and performs poorly irrespective of the dataset. This is also referred to as a model with high bias (or an underfitted model). The balance between these two phenomena is referred to as bias-variance trade-off, and is necessary to produce a model that is competent at recognizing patterns relevant to the problem but does not fixate on the noise present in the training samples. In practice, a neural network's training and validation loss curves continue to decrease to a certain point after which the two diverge. Generally, this point is considered to be representative of the optimal number of epochs. This practice is also known as early stopping, which refers to stopping the training before the model starts overfitting.

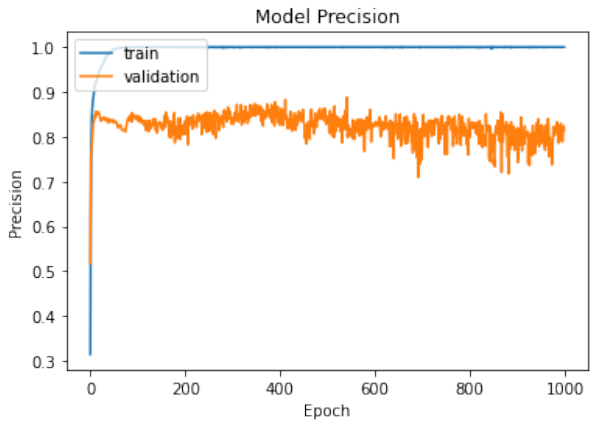
Thus, to determine the optimal number of epochs the network was trained for 1000 epochs and the training and validation losses were monitored. The model showed convergence shortly before the 400^{th} epoch as seen in Figure 4.6. Therefore, the number of epochs was set to 400.

4.4 Gait Emotion Recognition Experiments using BMDNN

Once the proposed BMDNN architecture's configuration was finalized through hyper-parameter tuning experiments, the network was used to perform an ablation study to validate the importance of its key elements. The network was also used to study the effects of the distinct LMA feature groups and their role in identifying the four emotions.



(a) The loss curves for the training and validation of BMDNN for 1000 epochs



(b) The precision curves for the training and validation of BMDNN for 1000 epochs

Figure 4.6: The proposed BMDNN's loss and precision graphs for the training and validation for 1000 epochs

4.4.1 Ablation Study

Each component of the neural network architecture is a crucial part of the proposed methodology. The input of raw gait sequences, JRAs, and JRDs to the LSTM subnetwork provides low-level features to the sequential network for producing higher-level features. The first subnetwork is responsible for condensing gait features sequentially to produce the rich feature set. The LMA features further enhance the performance by providing robust handcrafted features that make the model resilient to unbalanced data. Finally, the MLP subnetwork is responsible to process the combined features from the deep extracted features and the LMA-based handcrafted features, to produce high-level features for emotion recognition.

Effectiveness of Geometric Features

Classical machine learning approaches in the domain only explored a handful of features. Works based on deep learning incorporated affective features, however, the feature sets were still limited. Hence, the geometric features introduced for the proposed BMDNN architecture consider all possible angle and distance measures between two joints. To reduce the size of the set and ensure that only informative features are used for inference, the architecture exploits the ability of neural networks to automatically select and extract features. The geometric features contribute relevant information to the latent deep feature set that is formed using the information from body joint coordinates as well. This can be verified by the drop in the overall performance of the model observed on removing the geometric features from the input of the LSTM module (Table 4.4).

Table 4.4: Ablation Study for the various components of the proposed methodology

Architecture		**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
LSTM Subnetwork	✓	0.98	0.60	0.45	0.17	0.88	0.55
Geometric Features	×						
MLP Subnetwork	×						
LMA Features	×						
LSTM Subnetwork	✓	0.98	0.67	0.50	0.26	0.89	0.60
Geometric Features	×						
MLP Subnetwork	✓						
LMA Features	×						
LSTM Subnetwork	✓	0.99	0.87	0.82	0.61	0.95	0.82
Geometric Features	✓						
MLP Subnetwork	×						
LMA Features	×						
LSTM Subnetwork	×	0.99	0.93	0.97	0.86	0.98	0.94
Geometric Features	×						
MLP Subnetwork	✓						
LMA Features	✓						
LSTM Subnetwork	✓	0.99	0.88	0.86	0.52	0.95	0.81
Geometric Features	✓						
MLP Subnetwork	✓						
LMA Features	×						
LSTM Subnetwork	✓	0.99	0.93	0.97	0.85	0.97	0.94
Geometric Features	×						
MLP Subnetwork	✓						
LMA Features	✓						
LSTM Subnetwork	✓	0.99	0.95	0.97	0.91	0.98	0.96
Geometric Features	✓						
MLP Subnetwork	✓						
LMA Features	✓						

Effectiveness of LMA Features

The handcrafted features built using Laban Movement Analysis provide robustness and result in a well regularized model. These powerful features have never been used before for gait emotion recognition. Each of the feature groups contribute distinct information to the proposed architecture. From Table 4.4, excluding LMA feature groups results in a drastic decrease in the model precision for most emotions, reduction in the model's resilience towards data imbalance is observed as well. This experiment also serves as a validation for development of BMDNN architecture, and its superior performance over initial BMSNN model without LMA features. Performance of the model for emotion classes with lower data representation is affected strongly.

Effectiveness of LSTM and MLP subnetworks

The LSTM and MLP subnetworks are crucial building blocks of the proposed methodology. The main function of the LSTM module is to produce information-rich deep features from raw gait sequences and the geometric features (JRAs and JRDs). Since these low-level features are processed sequentially, information from both the temporal as well as the spatial domain is explored to produce the latent deep feature set. Further high-level features are extracted by the MLP module from this set. The MLP subnetwork also facilitates the fusion of the deep features and the LMA features. This is done by processing them together to extract powerful information which was never done before in the domain of GER.

The importance of these two subnetworks is validated by the results presented in Table 4.4. A decrease in the model precision value by 0.02 for the Happy class and by 0.05 for Neutral class, is recorded when the LSTM module is removed from the network. The effect is stronger with the removal of the MLP module from the network. On doing so, the precision values across all emotion classes except Angry decrease significantly and result in a macro mAP of 0.82.

4.4.2 Importance of LMA feature groups

The proposed method was trained and tested with different LMA feature groups to identify their impact on the overall model performance. Though all the features were necessary to achieve the highest performance, some feature groups were crucial for identifying certain emotions. Consequently, this experiment answered the sixth research question: "How do the Laban Movement Analysis feature groups affect the performance of the proposed network?".

As seen from Table 4.5, every LMA feature group was useful in the identification of Angry gaits. Furthermore, the most contributing feature groups for the identification of Happy and Neutral gaits were Body and Space features which contain angular, distance, and volumetric measures. This indicates that Happy and Neutral gaits contained in the dataset can be distinguished from other emotional gaits by geometric body positions and the space occupied by the subject during the gait. In contrast, Effort features proved to be important for recognizing Sad gaits because kinetic energies and velocities of various joints effectively capture the slower body movement of Sad emotions in comparison to other emotional gait. Another interesting observation is that the model precision for the neutral class is 0.46 with just shape features (density of the body joints); however, the model achieves a 0.91 precision with all the features combined, which is much higher than any individual feature group's precision. This indicates that the Shape feature group contributes unique information.

Table 4.5: Model performance with various LMA feature groups

Feature group used with BMDNN	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
No LMA Features	0.99	0.88	0.86	0.52	0.95	0.81
Body Features	0.99	0.94	0.88	0.87	0.97	0.92
Effort Features	0.99	0.88	0.90	0.80	0.96	0.90
Shape Features	0.99	0.89	0.80	0.46	0.95	0.78
Space Features	0.99	0.90	0.88	0.87	0.97	0.91
All LMA Features	**0.99**	**0.95**	**0.97**	**0.91**	**0.98**	**0.96**

4.4.3 Performance Comparison with the BMSNN Architecture

The proposed Bi-Modal Deep Neural Network (BMDNN) architecture was a result of improvements made to the Bi-Modular Sequential Neural Network (BMSNN). The BMSNN model utilized a dropout of 0.2 on the second layer in the MLP network. This enabled the model to be less biased against the under-represented classes, as seen in Table 4.6. In the BMDNN architecture this is achieved via L2 regularizers that keep the weight and bias parameters of the networks small. This eradicates the neural network's tendency to update parameters drastically when the loss is high. Rather, the updates are gradual and hence result in a stable learning. Another significant change made is the incorporation of the LMA features. In addition to decreasing the model's bias against under-represented classes, it improves the model's overall precision (refer to Table 4.5).

Table 4.6: Performance comparison of BMSNN and BMDNN

Architecture Configuration	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
BMSNN without Dropout	0.99	0.83	0.54	0.51	0.92	0.72
BMSNN without Batch Norm	0.99	0.90	0.84	0.46	0.96	0.80
BMSNN	0.99	0.91	0.90	0.65	0.97	0.86
BMDNN without L2 regularizers	0.99	0.92	0.94	0.87	0.98	0.93
BMDNN without Batch Norm	0.99	0.95	0.93	0.81	0.98	0.92
BMDNN	**0.99**	**0.95**	**0.97**	**0.91**	**0.98**	**0.96**

The batch normalization layer was used in BMSNN to ensure further resilience to the class imbalance, as shown in Table 4.6. This layer placed after the first MLP layer in BMSNN, has been moved to the beginning of the MLP subnetwork in BMDNN. This enables the MLP subnetwork of BMDNN to normalize the concatenated set of latent deep and LMA features before processing. Additionally, the first LSTM layer and the first MLP layer were removed from the LSTM subnetwork and the MLP subnetwork of BMSNN,

respectively. The removal of these layers did not result in any performance drop, but reduced the number of parameters of the entire architecture which reduces the model's susceptibility to high variance. The improvements made to the BMSNN architecture resulted in an increase in BMDNN's performance as seen in Table 4.6.

4.4.4 Performance Comparison with the State-Of-The-Art Methods

The best configuration of the proposed architectures were compared with the most recent state-of-the-art methods. The performances of Spatial Temporal Graph Convolutional Network (STGCN) [189], Affective and Deep Features (ADF) [144], STGCN for Emotion Perceeption (STEP) [20], Hierarchical Attention Pooling and Affective Mapping (HAPAM) [21], ProxEmo [129], and Bi-Modular Sequential Neural Network (BMSNN) [19] were compared with the proposed methods. Implementations of the comparators used in this work were done in Pytorch, and were provided by the respective authors via a publicly available GitHub repository. Additionally, all comparative experiments were performed on the standard version of Google Colaboratory Notebook which comprises of a Tesla K80 GPU with a 12 GB of GDDR5 VRAM, a dual-core Intel Xeon @ 2 GHz CPU and a memory size of 13.3 GB. The data split for all methods was also kept the same (80:10:10, for training, validation, and testing). Table 4.7 shows that the proposed architecture outperforms all other methods across all classes.

Both of the proposed models: BMSNN [19] and BMDNN [18] outperform the state-of-the-art methods in terms of micro mAP. The modifications made to the BMSNN architecture to produce BMDNN prove successful. In comparison to the best prior method, ProxEmo, the proposed Bi-Modal Deep Neural Network (BMDNN) architecture achieves an increase in the Average Precision by 10% for the Angry class, by 3.3% for the Happy class, and by 3.2% for the Sad class. Furthermore, the micro and mean Average Precision scores of 0.98 and 0.96 respectively, are also observed to be superior. Hence, the proposed BMDNN architecture outperforms the state-of-the-art methodologies in terms of overall precision, while maintaining high performance in all emotion classes.

The highest scores achieved by the proposed BMDNN architecture are attributed to the powerful LSTM and the MLP subnetworks. The high performance is also a result of the overall attenuated design of the neural network, the normalization techniques, and the regularization techniques employed. Another major contributing factor is the information-rich handcrafted features exploited by the BMDNN architecture.

Table 4.7: Comparison of the proposed method with state-of-the-art methods

Methods	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
STEP (2020)[20]	0.22	0.52	0.30	0.12	0.29	0.27
ADF (2019)[144]	0.22	0.59	0.30	0.12	0.31	0.27
STGCN (2018)[189]	0.06	0.97	0.20	0.01	0.34	0.41
HAPAM (2020)[21]	0.97	0.66	0.40	0.18	0.60	0.88
ProxEmo (2020)[129]	0.90	0.92	0.94	0.94	0.92	0.93
Proposed BMSNN (2022)	**0.99**	**0.91**	**0.90**	**0.65**	**0.97**	**0.86**
Proposed BMDNN (2022)	**0.99**	**0.95**	**0.97**	**0.91**	**0.98**	**0.96**

4.4.5 Number of Parameters and Inference Time

The last metric that was used to compare the performances of the various GER methods was their respective inference times for one gait sample. The implementation of the proposed Bi-Modular Neural Network (BMSNN) and Bi-Modal Deep Neural Network (BMDNN) was achieved through Keras, while all other methods were implemented using PyTorch. These experiments were run on the standard version of Google Colaboratory Notebook which uses a Tesla K80 GPU with a 12 GB of GDDR5 VRAM, a dual-core Intel Xeon @ 2 GHz CPU and a memory size of 13.3 GB. Each method was tested on the ELMD dataset.

The proposed Bi-Modular Sequential Neural Network (BMSNN) was observed to have the fastest inferences among all comparators of less than 17 milliseconds. This can be accredited to the fact that it has significantly fewer parameters in comparison to the other methods, as seen in Table 4.8.

Table 4.8: Comparison of number of parameters and inference time of the proposed methods with state-of-the-art methods

Method	**Number of parameters**	**Inference Time (in seconds)**
HAPAM [21]	40,444,854	4.66×10^{-2}
STGCN [189]	2,628,290	2.17×10^{-2}
STEP [20]	717,987	4.82×10^{-2}
ProxEmo [129]	334,849	1.08×10^{-1}
ADF [144]	310,978	3.91×10^{-2}
Proposed BMSNN	**295,940**	$\mathbf{1.63 \times 10^{-2}}$
Proposed BMDNN	**184,276**	$\mathbf{3.14 \times 10^{-1}}$

4.4.6 Experiments on BML dataset

To further validate the findings of this research, a second dataset known as the Body Motion Library (BML) dataset, was used. It was collected from 30 participants (15 male and 15 female). It was recorded for two walks from each participant, for each emotion. Therefore, it contains 60 gaits for each emotion class. It was

recorded as 3D coordinates for 15 body joints of participants of different age and gender groups, walking along linear paths. The proposed system was pre-trained on the ELMD dataset and was validated on both ELMD and BML datasets. These experiments served as an additional validation of the proposed architecture to showcase its potency in gait emotion recognition.

The conclusions derived from the experiment on the importance of LMA feature groups were validated on the BML dataset, which is a balanced dataset of a smaller size than ELMD (see Table 4.9). What is impressive, is that from this dataset, exactly the same conclusions on the impact of the LMA features were observed. All LMA features were found to be useful in recognizing Angry gaits for BML. Similarly, Body and Space features were useful in recognizing Happy gaits. Finally, combining all LMA features resulted in average precision scores higher than individual values. The only difference was the effect of Effort features was more significant in recognizing Neutral gaits. Therefore, the findings of the experiment that explored the importance of the various LMA feature groups for each emotion class was supported by two distinct datasets.

Table 4.9: Model performance with various LMA feature groups on the BML dataset

Feature group used with BMDNN	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
No LMA Features	0.72	0.94	0.76	0.95	0.83	0.84
Body Features	0.91	0.89	0.83	1	0.91	0.91
Effort Features	0.98	0.73	0.76	0.98	0.85	0.86
Shape Features	0.98	0.84	0.56	0.78	0.75	0.79
Space Features	0.98	0.91	0.75	0.81	0.88	0.86
All LMA Features	**1**	**1**	**0.79**	**1**	**0.93**	**0.95**

The proposed BMDNN's overall performance was also validated on the second dataset. The same data split of 80:10:10 with a stratified shuffling was used with the hyperparameter values optimized using ELMD. As seen in Table 4.10, the proposed BMDNN achieved a high average precision score in each emotion class, as well as high micro and macro mean average precision scores of 0.93 and 0.95, respectively. The formidable performance on two distinct datasets further supports the precise emotion recognition capabilities of the proposed BMDNN model.

Table 4.10: Model performance on the ELMD and BML datasets

Dataset Used	**Class AP Angry**	**Class AP Happy**	**Class AP Sad**	**Class AP Neutral**	**micro mAP**	**macro mAP**
ELMD	0.99	0.95	0.97	0.91	0.98	0.96
BML	1	1	0.79	1	0.93	0.95

4.4.7 Summary

This chapter convincingly demonstrated that the performance of the proposed architecture is superior to other emotion recognition methods. The hyper parameter tuning resulted in the final configuration of the Bi-Modal Deep Neural Network (BMDNN). Following the hyperparameter tuning experiments, the section presented the results of the ablation study that demonstrated the significance of the geometric handcrafted features, Laban Movement Analysis features, and the deep neural network of the BMDNN architecture. The final configuration of the proposed BMDNN architecture was used to compare its emotion recognition performance with the previously designed BMSNN architecture and the state-of-the-art methods. While the BMSNN proved to be the fastest method to infer a gait sample, the BMDNN architecture excelled at emotion recognition by outperforming all previous methodologies in terms of micro and macro mean average precision.

Chapter 5

Conclusion

5.1 Contribution Summary

Identifying emotions from gait is a new avenue in biometrics and has not witnessed much research. However, gait analysis and emotion recognition systems play a significant role industries such as smart homes, security, search and rescue, robotics, entertainment, and medicine. Some potential applications include behavioral surveillance in public spaces, fall detection and prevention, adaptive rehabilitation/education/entertainment, and identifying people in distress. The preliminary research consisted of works based on classical machine learning algorithms that provided a proof of concept, but were unable to achieve high accuracies. Lately, deep learning methodologies have been prevalent in the domain; however the methods published so far were sub-optimal. Consequently, this research was conducted to contribute to the domain by developing a powerful sequential neural network that harnesses the power of sequential deep neural networks and benefits from robust domain-specific handcrafted features for identifying emotions from gaits.

Recent works attempted to approach Gait Emotion Recognition (GER) using Graph Neural Networks and Convolutional Neural Networks. However, these works modelled gaits as rigid structures, which inhibited the ability of deep neural networks to extract low-level features from distant graph nodes or image pixels. A few methods employed Recurrent Neural Networks that explored all dependencies, but the implemented deep learning models had ineffective network designs. Furthermore, most of the prior works learned to reconstruct artificial gaits rather than learning features specific to emotion recognition. Hence, this thesis proposes unique approaches of fusing latent deep features with the robust handcrafted features for recognizing four classes of emotions from human gait sequences.

The proposed Bi-Modal Deep Neural Network (BMDNN) architecture uses delineating handcrafted fea-

tures based on the four components of human motion: Body, Effort, Shape, and Space, based on the Laban Movement Analysis. The novel fusion of these LMA features along with the deep features extracted from the optimized neural network mitigates the performance drop for under-represented classes. As a result, the proposed BMDNN architecture achieves a micro mean Average Precision of 0.98, and a macro mAP of 0.96 that outperforms all recent state-of-the-art methods, including BMSNN, on the ELMD dataset.

The results of this research were published in ICCI*CC'20 [17], Sensors [19] and ISVC'22 [18], and can be summarized as follows:

1. Previous works used sub-optimal approaches and could not harvest the full potential of such neural networks. Hence, this research proposed a novel hybrid deep learning architecture that utilized powerful Long Short Term Memory units and a Multi Layered Perceptron to effectively train and predict emotions from human gaits. The LSTM subnetwork sequentially extracted features that were mapped to one of four emotions using a MLP.

2. The proposed Bi-Modular Sequential Neural Network (BMSNN) model was designed with a tapered structure whose benefits are two-fold. Firstly, the attenuation forces the information to condense towards the end of the network by reducing the feature representation size. Secondly, it results in fewer network parameters unlike the networks used in other works. Hence, enabling the network to achieve an inference time of less than 17 milliseconds, which is faster than any prior method.

3. The BMSNN architecture also incorporates geometric features that consist of all possible Joint Relative Angles (JRAs) and Joint Relative Distances (JRDs). The network facilitates processing of raw gait sequences and these geometric features together. This eliminates any issues stemming from using an incomplete set of features in prior methods. Additionally, it enables extraction of relevant information from various combinations of geometric features and raw gait data. This extension enables the model to outperform prior methods with a micro mean Average Precision of 0.97.

4. The previous architecture was extended to include domain-specific robust Laban Movement Analysis (LMA) features. These body, effort, shape and space features provided characteristic descriptions for the gait samples that led to further increment in the model's overall performance. Additionally, the model's sensitivity towards the class distribution of the dataset was drastically reduced. As a result, the proposed Bi-Modal Deep Neural Network exhibited the highest macro mean Average Precision.

5. To accommodate the robust Laban Movement Analysis (LMA) features, another novel deep learning network was developed using BMSNN's architectural characteristics that were responsible for its high performance. The proposed Bi-Modal Deep Neural Network featured an overall attenuated design.

Additionally, BMDNN's design allowed a combined processing of LMA features and latent deep features to form higher-level features. The newly added batch-normalization layer normalized this combined feature set for optimal processing. Finally, L2 bias and kernel regularizers were introduced to promote regularization of the network parameters. These improvements lead to a significant increase in the network performance, and the BMDNN architecture outperformed all prior state-of-the-art methods with a micro mean Average Precision score of 0.98.

6. This thesis performed an ablation study to validate the importance of each component of the proposed BMDNN architecture. A decrease in the model performance was observed when any of its components were removed. Additional experiments were conducted to study the effects of various LMA features on the network performance. It was found that all feature groups contribute distinctive information to the comprehensive LMA feature set.

5.2 Future Research Directions

The problem of Gait Emotion Recognition has only recently started gaining popularity in biometric research. Subsequently, only a handful of datasets have been created for the problem. The novel deep learning methodologies proposed in this thesis were trained on the largest available dataset containing real-world gait sequences.

As a result of being trained on the emotionally labelled Edinburgh Locomotive MoCap Dataset, the proposed networks accept an input of 16 three-dimensional body joint coordinates. Furthermore, the dataset is not an exhaustive set that contains all variations (sickness, carrying an item, etc.) of gait sequences for a particular emotion. Hence, if the proposed methodologies are required to infer samples from a new dataset, the networks will have to be fine-tuned or trained from scratch.

The ELMD was originally recorded to produce artificial human motions and hence did not ensure an equal number of gaits for each emotion. The labels for perceived emotions were collected from a crowd-sourced website afterwards. This, in turn causes the previous models to identify samples from highly-represented classes with great precision but perform poorly for under-represented classes. This behavior was observed clearly in most of the prior works. To remedy this problem, the proposed network utilized robust handcrafted Laban Movement Analysis features that provide consistent emotion-relevant information for each gait sample. In the future, data augmentation techniques can be employed to generate samples for the under-represented classes to reduce the sensitivity of deep learning models towards the class distribution even further. Moreover, with appropriate data, researchers can investigate methods that identify and prevent spoofing. In addition,

having more datasets that contain diverse data on participants' age, ethnicity, height and body type, would be advantageous to develop further applications that can be deployed in the real-world.

Recent advances in deep learning have shown transformer networks to be proficient for processing long sequences. Self-attention based processing has the potential to extract even more information from the gait sequences. Hence, architectures based on such networks can be explored in the future. Another interesting approach for future investigation involves fusing networks that extract spatial features with networks that extract temporal features. Particularly, hierarchical deep learning models, that can process joint trajectories independently and subsequently combine the learned features, can be used in tandem with graph neural networks that process gaits structurally. Lastly, three-dimensional convolution on gaits embedded as images can further facilitate the extraction of spatial and temporal features for gait emotion recognition.

5.3 Potential Applications

The findings of this research can benefit industries like robotics, medicine, disaster management, security, entertainment, and education.

Emotion Recognition can be performed on-the-fly to develop emotionally-aware robots that adapt their interactions with humans accordingly. Smart home designs can include such robots, or other devices to help prevent and/or detect fall. This application will be particularly beneficial to nursing homes, rehabilitation centers, and hospitals. Additionally, identifying stress, pain, and difficulty in walking can be used to curate rehabilitation exercises to suit the comfort of the patient for a faster recovery. Similarly, physiological diseases such as Parkinson's disease can be diagnosed using gait analysis.

This work can also be adopted to detect panic in evacuation situations and used to notify disaster management authorities. Moreover, the gait analysis performed in this work can be extended to detect suspicious behavior in persons at international borders, public venues, and financial places of interest. Other potential applications include providing an adaptive virtual reality entertainment experience by monitoring the user's frustration. This can also be used to modify the difficulty of an educational course to ensure the student's stress levels do not exceed the healthy amount. The solution presented in this thesis for improving Gait Emotion Recognition performance may not only benefit the industries it can be applied to, but also help to uncover new information about the intricate relations of body joint movements and the expressed emotions.

Bibliography

[1] Foteini Agrafioti, Dimitris Hatzinakos, and Adam K Anderson. ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1):102–115, 2011.

[2] Faisal Ahmed, Padma Polash Paul, and Marina L Gavrilova. DTW-based kernel and rank-level fusion for 3d gait recognition using kinect. *The Visual Computer*, 31(6):915–924, 2015.

[3] Ferdous Ahmed, ASM Hossain Bari, and Marina L Gavrilova. Emotion recognition from body movement. *IEEE Access*, 8:11761–11781, 2019.

[4] Ferdous Ahmed, Brandon Sieu, and Marina L Gavrilova. Score and rank-level fusion for emotion recognition using genetic algorithm. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 46–53. IEEE, 2018.

[5] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE, 2017.

[6] Ashwaq Alhargan, Neil Cooke, and Tareq Binjammaz. Affect recognition in an interactive gaming environment using eye tracking. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 285–291. IEEE, 2017.

[7] Noor Almaadeed, Amar Aggoun, and Abbes Amira. Audio-visual feature fusion for speaker identification. In *International Conference on Neural Information Processing*, pages 56–67. Springer, 2012.

[8] Israa M Alsaadi. Physiological biometric authentication systems, advantages, disadvantages and future development: A review. *International Journal of Scientific & Technology Research*, 4(12):285–289, 2015.

[9] Saima Aman and Stan Szpakowicz. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[10] Suzan A Anwar. *Real Time Facial Expression Recognition and Eye Gaze Estimation System*. PhD thesis, University of Arkansas at Little Rock, 2019.

[11] Claudio Aracena, Sebastián Basterrech, Václav Snáel, and Juan Velásquez. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2632–2637. IEEE, 2015.

[12] Bishnu S Atal and Suzanne L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

[13] Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, 93:133–142, 2017.

[14] G Bapineedu, B Avinash, Suryakanth V Gangashetty, and B Yegnanarayana. Analysis of lombard speech using excitation source information. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] ASM Hossain Bari and Marina L Gavrilova. Artificial neural network based gait recognition using kinect sensor. *IEEE Access*, 7:162708–162722, 2019.

[16] Chiraz BenAbdelkader, Ross Cutler, and Larry Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 372–377. IEEE, 2002.

[17] Yajurv Bhatia, ASM Hossain Bari, and Marina Gavrilova. A LSTM-based approach for gait emotion recognition. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 214–221. IEEE, 2021.

[18] Yajurv Bhatia, ASM Hossain Bari, and Marina Gavrilova. Gait emotion recognition using a bi-modal deep neural network. In *International Symposium on Visual Computing (ISVC) 2022, Part I, Lecture Notes in Computer Science 13598, Chapter 4*. Springer, 2022.

[19] Yajurv Bhatia, ASM Hossain Bari, Gee-Sern Jison Hsu, and Marina Gavrilova. Motion capture sensor-based emotion recognition using a bi-modular sequential neural network. *Sensors*, 22(1):403, 2022.

[20] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1342–1350, 2020.

[21] Uttaran Bhattacharya, Christian Roncal, Trisha Mittal, Rohan Chandra, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping. In *European Conference on Computer Vision*, pages 145–163. Springer, 2020.

[22] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52(7-8):613–625, 2010.

[23] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[24] Elif Bozkurt, Engin Erzin, Cigdem Eroglu Erdem, and A Tanju Erdem. Use of line spectral frequencies for emotion recognition from speech. In *2010 20th International Conference on Pattern Recognition*, pages 3708–3711. IEEE, 2010.

[25] Ran Breuer and Ron Kimmel. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*, 2017.

[26] Annie Britton, Martin Shipley, Marek Malik, Katerina Hnatkova, Harry Hemingway, and Mlichael Marmot. Changes in heart rate and heart rate variability over time in middle-aged men and women in the general population. *The American Journal of Cardiology*, 100(3):524–527, 2007.

[27] Gerhard Budin, Marina L Gavrilova, Duane F Shell, Yingxu Wang, Rodolfo A Fiorini, Bernard Widrow, Lotfi A Zadeh, Newton Howard, Sally Wood, Virendrakumar C Bhavsar, and Christen Chan. Cognitive intelligence: Deep learning, thinking, and reasoning by brain-inspired systems. *International Journal of Cognitive Informatics and Natural Intelligence*, 10(4):1–20, 2016.

[28] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive Behavioural Systems*, pages 144–157. Springer, 2012.

[29] Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2):213–225, 2003.

[30] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. An overview of machine learning. *Machine Learning*, pages 3–23, 1983.

[31] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 71–82. Springer, 2007.

[32] Guillaume Chanel, Sunny Avry, Gaëlle Molinari, Mireille Bétrancourt, and Thierry Pun. Multiple users' emotion recognition: Improving performance by joint modeling of affective reactions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 92–97. IEEE, 2017.

[33] Ting-Yi Chang, Cheng-Jung Tsai, and Jyun-Hao Lin. A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices. *Journal of Systems and Software*, 85(5):1157–1165, 2012.

[34] Bo Cheng and Guangyuan Liu. Emotion recognition from surface EMG signal using wavelet transform and neural network. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pages 1363–1366. IEEE, 2008.

[35] Colin Cherry, Saif M. Mohammad, and Berry De Bruijn. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5s1:BII.S8933, 2012.

[36] Wenzheng Chi, Jiaole Wang, and Max Q-H Meng. A gait recognition method for human following in service robots. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9):1429–1440, 2017.

[37] Andrea Chiorrini, Claudia Diamantini, Alex Mircoli, and Domenico Potena. Emotion and sentiment analysis of tweets using BERT. In *EDBT/ICDT Workshops*, 2021.

[38] Mangtik Chiu, Jiayu Shu, and Pan Hui. Emotion recognition through gait on mobile devices. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 800–805. IEEE, 2018.

[39] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017.

[40] Peter M Corcoran, Florin Nanu, Stefan Petrescu, and Petronel Bigioi. Real-time eye gaze tracking for gaming design and consumer electronics systems. *IEEE Transactions on Consumer Electronics*, 58(2):347–355, 2012.

[41] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004.

[42] Pawel Czerwinski. A photo of a camera on a lookout tower in poland., Jun 2018. [Online; accessed October 19, 2022; Available at https://unsplash.com/photos/zBTYRFCeaS0].

[43] Charles Darwin and Phillip Prodger. *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA, 1872.

[44] Karishma Dasgaonkar and Swati Chopade. Analysis of multi-layered perceptron, radial basis function and convolutional neural networks in recognizing handwritten digits. *International Journal of Advance Research, Ideas and Innovations in Technology*, 4(3):2429–2431, 2018.

[45] Matthieu Destephe, Takayuki Maruyama, Massimiliano Zecca, Kenji Hashimoto, and Atsuo Takanishi. The influences of emotional intensity for happiness and sadness on walking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7452–7455. IEEE, 2013.

[46] Xuedan Du, Yinghao Cai, Shuo Wang, and Leijie Zhang. Overview of deep learning. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 159–164. IEEE, 2016.

[47] Andrew T Duchowski. *Eye Tracking Methodology: Theory and Practice*. Taylor & Francis, 1981.

[48] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 467–474. Association for Computing Machinery, 2015.

[49] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[50] Issam El Naqa and Martin J Murphy. What is machine learning? In *Machine Learning in Radiation Oncology*, pages 3–11. Springer, 2015.

[51] Clayton Epp, Michael Lippold, and Regan L Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 715–724, 2011.

[52] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[53] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Towards real-time speech emotion recognition using deep neural networks. In *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–5. IEEE, 2015.

[54] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057, 2007.

[55] Clinton Fookes, Anthony Maeder, Sridha Sridharan, and George Mamic. Gaze based personal identification. In *Behavioral Biometrics for Human Identification: Intelligent Applications*, pages 237–263. IGI Global, 2010.

[56] Afra Foroud and Ian Q Whishaw. Changes in the kinematic structure and non-kinematic features of movements during skilled reaching after stroke: A laban movement analysis in two case studies. *Journal of Neuroscience Methods*, 158(1):137–149, 2006.

[57] James R Gage. Gait analysis. an essential tool in the treatment of cerebral palsy. *Clinical Orthopaedics and Related Research*, 288:126–134, 1993.

[58] Hugo Gamboa and Vasco Ferreira. Widam-web interaction display and monitoring. In *Proceedings in the 5th International Conference on Enterprise Information Systems*, pages 21–27, 2003.

[59] Marina L Gavrilova, Ferdous Ahmed, ASM Hossain Bari, Ruixuan Liu, Tiantian Liu, Yann Maret, Brandon Kawah Sieu, and Tanuja Sudhakar. Multi-modal motion-capture-based biometric systems for emergency response and patient rehabilitation. In *Research Anthology on Rehabilitation Practices and Therapy*, pages 653–678. IGI global, 2021.

[60] Marina L Gavrilova, Fahim Anzum, ASM Hossain Bari, Yajurv Bhatia, Fariha Iffath, Quwsar Ohi, Md Shopon, and Zaman Wahid. A multifaceted role of biometrics in online security, privacy, and trustworthy decision making. In *Breakthroughs in Digital Biometrics and Forensics*, pages 303–324. Springer, 2022.

[61] Deepak Ghimire, Sunghwan Jeong, Joonwhoan Lee, and San Hyun Park. Facial expression recognition based on local region specific features and support vector machines. *Multimedia Tools and Applications*, 76(6):7803–7821, 2017.

[62] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013.

[63] Kiel M Gilleade and Alan Dix. Using frustration in the design of adaptive videogames. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, pages 228–232, 2004.

[64] Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. Keystroke dynamics with low constraints SVM based passphrase enrollment. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6. IEEE, 2009.

[65] Donald Glowinski, Antonio Camurri, Gualtiero Volpe, Nele Dael, and Klaus Scherer. Technique for automatic emotion recognition by body gesture analysis. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE, 2008.

[66] Khadidja Gouizi, F Bereksi Reguig, and Choubeila Maaoui. Emotion recognition from physiological signals. *Journal of Medical Engineering & Technology*, 35(6-7):300–307, 2011.

[67] Alex Graves, Christoph Mayer, Matthias Wimmer, Jürgen Schmidhuber, and Bernd Radig. Facial expression recognition with recurrent neural networks. In *Proceedings of the International Workshop on Cognition for Technical Systems*, 2008.

[68] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and Research Workshop on Affective Dialogue Systems*, pages 36–48. Springer, 2004.

[69] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017.

[70] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2005.

[71] SL Happy, Anjith George, and Aurobinda Routray. A real time facial expression classification system using local binary patterns. In *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, pages 1–5. IEEE, 2012.

[72] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 30–40, 2017.

[73] Shivani Hashia, Chris Pollett, and Mark Stamp. On using mouse movements as a biometric. In *Proceeding in the International Conference on Computer Science and its Applications*, volume 1, page 5. ICCSA, 2005.

[74] Ikuo Homma and Yuri Masaoka. Breathing rhythms and emotions. *Experimental Physiology*, 93(9):1011–1021, 2008.

[75] Mohammed E Hoque, Mohammed Yeasin, and Max M Louwerse. Robust recognition of emotion from speech. In *International Workshop on Intelligent Virtual Agents*, pages 42–53. Springer, 2006.

[76] Maryam Imani and Gholam Ali Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019.

[77] Yuichi Iwadate, Masayuki Inoue, Ryotaro Suzuki, Naoto Hikawa, Mao Makino, and Y Kanemoto. Mic interactive dance system-an emotional interaction system. In *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)*, volume 1, pages 95–98. IEEE, 2000.

[78] Carroll E Izard. *Human Emotions*. Springer Science & Business Media, 2013.

[79] Takamune Izui, Isabelle Milleville, Sophie Sakka, and Gentiane Venture. Expressing emotions using gait of humanoid robot. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 241–245. IEEE, 2015.

[80] Deepak Kumar Jain, Zhang Zhang, and Kaiqi Huang. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognition Letters*, 139:157–165, 2020.

[81] Joseph Jankovic. Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008.

[82] Daniel Janssen, Wolfgang I Schöllhorn, Jessica Lubienetzki, Karina Fölling, Henrike Kokenge, and Keith Davids. Recognition of emotions in gait patterns by means of artificial neural nets. *Journal of Nonverbal Behavior*, 32(2):79–92, 2008.

[83] K Jellinger, D Armstrong, HY Zoghbi, and AK Percy. Neuropathology of rett syndrome. *Acta Neuropathologica*, 76(2):142–158, 1988.

[84] Cai Jing, Guangyuan Liu, and Min Hao. The research on emotion recognition from ECG signal. In *2009 International Conference on Information Technology and Computer Science*, volume 1, pages 497–500. IEEE, 2009.

[85] Joshua Juen, Qian Cheng, Valentin Prieto-Centurion, Jerry A Krishnan, and Bruce Schatz. Health monitors for chronic disease by gait analysis with mobile phones. *Telemedicine and E-Health*, 20(11):1035–1041, 2014.

[86] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[87] Michelle Karg, Kolja Kühnlenz, and Martin Buss. Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):1050–1061, 2010.

[88] Pawel Kasprowski and Józef Ober. Eye movements in biometrics. In *International Workshop on Biometric Authentication*, pages 248–258. Springer, 2004.

[89] Christina Katsini, Marios Belk, Christos Fidas, Nikolaos Avouris, and George Samaras. Security and usability in knowledge-based user authentication: A review. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics*, pages 1–6, 2016.

[90] Christos D Katsis, Nikolaos Katertsidis, George Ganiatsas, and Dimitrios I Fotiadis. Toward emotion recognition in car-racing drivers: A biosignal processing approach. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(3):502–512, 2008.

[91] Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Björn Schuller. End-to-end learning for dimensional emotion recognition from physiological signals. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 985–990. IEEE, 2017.

[92] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236, 2017.

[93] Jonghwa Kim. Bimodal emotion recognition using speech and physiological changes. *Robust Speech Recognition and Understanding*, 265:280, 2007.

[94] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.

[95] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004.

[96] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2):401, 2018.

[97] Dana Kulic and Elizabeth A Croft. Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, 23(5):991–1000, 2007.

[98] Rudolf Laban and Lisa Ullmann. *The Mastery of Movement.* ERIC, 1971.

[99] Hosub Lee, Young Sang Choi, Sunjae Lee, and IP Park. Towards unobtrusive emotion recognition for affective social communication. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pages 260–264. IEEE, 2012.

[100] Matthias R Lemke, Thomas Wendorff, Brigitt Mieth, Katharina Buhl, and Martin Linnemann. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *Journal of Psychiatric Research*, 34(4-5):277–283, 2000.

[101] Jacqyln A Levy and Marshall P Duke. The use of laban movement analysis in the study of personality, emotional state and movement style: An exploratory investigation of the veridicality of" body language". *Individual Differences Research*, 1(1), 2003.

[102] Baobin Li, Changye Zhu, Shun Li, and Tingshao Zhu. Identifying emotions from non-contact gaits information based on microsoft kinects. *IEEE Transactions on Affective Computing*, 9(4):585–591, 2016.

[103] Lan Li and Ji-hua Chen. Emotion recognition using physiological signals from multiple subjects. In *2006 International Conference on Intelligent Information Hiding and Multimedia*, pages 355–358. IEEE, 2006.

[104] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, pages 1195–1215, 2020.

[105] Shun Li, Liqing Cui, Changye Zhu, Baobin Li, Nan Zhao, and Tingshao Zhu. Emotion recognition using kinect motion capture data of human gaits. *PeerJ*, 4:e2364, 2016.

[106] Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, and Bin Hu. Exploring eeg features in cross-subject emotion recognition. *Frontiers in Neuroscience*, 12:162, 2018.

[107] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

[108] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

[109] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv Preprint*, 2015.

[110] Alexander Lischke, Christoph Berger, Kristin Prehn, Markus Heinrichs, Sabine C Herpertz, and Gregor Domes. Intranasal oxytocin enhances emotion recognition from dynamic facial expressions and leaves eye-gaze unaffected. *Psychoneuroendocrinology*, 37(4):475–481, 2012.

[111] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based emotion recognition and its applications. In *Transactions on Computational Science XII*, pages 256–277. Springer, 2011.

[112] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and EEG to enhance emotion recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[113] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[114] Marko Lugger and Bin Yang. The relevance of voice quality features in speaker independent emotion recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–17. IEEE, 2007.

[115] Hai-Rong Lv, Zhong-Lin Lin, Wen-Jun Yin, and Jin Dong. Emotion recognition based on pressure sensor keyboards. In *2008 IEEE International Conference on Multimedia and Expo*, pages 1089–1092. IEEE, 2008.

[116] Choubeila Maaoui and Alain Pruski. Emotion recognition through physiological signals for human-machine communication. *Cutting Edge Robotics*, 2010(317-332):11, 2010.

[117] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70, 2010.

[118] Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. Affective computing in virtual reality: Emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1):1–15, 2018.

[119] Maximalfocus. The humanoid robot asimo of honda live in action at miraikan museum of emerging science and innovation., Jun 2020. [Online; accessed October 19, 2022; Available at https://unsplash.com/photos/eZWGK5sIiBM].

[120] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[121] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

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

[123] Albert Mehrabian. *Nonverbal Communication*. Routledge, 2017.

[124] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre GR Day, Clint Richardson, Charles K Fisher, and David J Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, 2019.

[125] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. Biometrics recognition using deep learning: A survey. *arXiv preprint*, 2019.

[126] Saif Mohammad. # Emotional Tweets. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, 2012.

[127] Joann M Montepare, Sabra B Goldstein, and Annmarie Clausen. The identification of emotions from gait information. *Journal of Nonverbal Behavior*, 11(1):33–42, 1987.

[128] Murugappn Murugappan, Mohamed Rizon, Ramachandran Nagarajan, S Yaacob, I Zunaidi, and D Hazry. EEG feature extraction for classifying emotions using FCM and FKM. *International Journal of Computers and Communications*, 1(2):21–25, 2007.

[129] Venkatraman Narayanan, Bala Murali Manoghar, Vishnu Sashank Dorbala, Dinesh Manocha, and Aniket Bera. Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8200–8207. IEEE, 2020.

[130] Fatma Nasoz, Kaye Alvarez, Christine L Lisetti, and Neal Finkelstein. Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1):4–14, 2004.

[131] Daniel Neiberg, Kjell Elenius, and Laskowski. Emotion recognition in spontaneous speech using gmms. In *Ninth International Conference on Spoken Language Processing (ICSLP)*, pages 101–104, 09 2006.

[132] Michael A Nielsen. *Neural Networks and Deep Learning*, volume 25. Determination press San Francisco, CA, USA, 2015.

[133] Lars Omlor and Martin A Giese. Extraction of spatio-temporal primitives of emotional body expressions. *Neurocomputing*, 70(10-12):1938–1942, 2007.

[134] Timo Partala and Veikko Surakka. Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, 59(1-2):185–198, 2003.

[135] Sananda Paul, Anwesha Banerjee, and DN Tibarewala. Emotional eye movement analysis using electrooculography signal. *International Journal of Biomedical Engineering and Technology*, 23(1):59–70, 2017.

[136] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):186–197, 2009.

[137] ST Pheasant. A Review of: "Human Walking". By V. T. INMAN, H.J. RALSTON and F. TODD. (Baltimore, London: Williams & Wilkins, 1981.) [Pp.154.]. *Ergonomics*, 24(12), 1981.

[138] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.

[139] Oudeyer Pierre-Yves. The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183, 2003.

[140] Robert Ed Plutchik and Hope R Conte. *Circumplex Models of Personality and Emotions.* American Psychological Association, 1997.

[141] Alexander P Pons and Peter Polak. Understanding user perspectives on biometric technology. *Communications of the ACM*, 51(9):115–118, 2008.

[142] Xiaojun Quan, Qifan Wang, Ying Zhang, Luo Si, and Liu Wenyin. Latent discriminative models for social emotion detection with emotional dependency. *ACM Transactions on Information Systems*, 34(1):1–19, 2015.

[143] Juan Carlos Quiroz, Elena Geangu, and Min Hooi Yong. Emotion recognition using smart watch sensor data: Mixed-design study. *JMIR Mental Health*, 5(3):e10153, 2018.

[144] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint*, 2019.

[145] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications*, 9(1):58–69, 2006.

[146] Vidas Raudonis, Gintaras Dervinis, Andrius Vilkauskas, Agnė Paulauskaitė-Taraševičienė, and Gintarė Keršulytė-Raudonė. Evaluation of human emotion from eye motions. *International Journal of Advanced Computer Science and Applications*, 4(8):79–84, 2013.

[147] Kenneth Revett. *Behavioral Biometrics: a Remote Access Approach*. John Wiley & Sons, 2008.

[148] Georgios Rigas, Christos D Katsis, George Ganiatsas, and Dimitrios I Fotiadis. A user independent, biosignal based, emotion recognition method. In *International Conference on User Modeling*, pages 314–318. Springer, 2007.

[149] James A Russell. Affective space is bipolar. *Journal of Personality and Social Psychology*, 37(3):345, 1979.

[150] Ali-Akbar Samadani, Sarahjane Burton, Rob Gorbet, and Dana Kulic. Laban effort and shape analysis of affective hand and arm movements. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 343–348. IEEE, 2013.

[151] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[152] Klaus R Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[153] Sebastian Scholz. Nuki smart lock (smarthome), May 2019. [Online; accessed October 19, 2022; Available at https://unsplash.com/photos/IJkSskfEqrM].

[154] Ann-Kathrin Seifert, Abdelhak M Zoubir, and Moeness G Amin. Radar-based human gait recognition in cane-assisted walks. In *2017 IEEE Radar Conference (RadarConf)*, pages 1428–1433. IEEE, 2017.

[155] Jerritta Selvaraj, Murugappan Murugappan, Khairunizam Wan, and Sazali Yaacob. Classification of emotional states from electrocardiogram signals: a non-linear approach based on hurst. *Biomedical Engineering Online*, 12(1):1–18, 2013.

[156] Mridula Sharma and Haytham Elmiligi. Behavioral biometrics: Past, present and future. In *Recent Advances in Biometrics*, chapter 4. IntechOpen, 2022.

[157] Md Shopon, ASM Hossain Bari, Yajurv Bhatia, Pavan Karkekoppa Narayanaswamy, Sanjida Nasreen Tumpa, Brandon Sieu, and Marina Gavrilova. Biometric system de-identification: Concepts, applications, and open problems. In *Handbook of Artificial Intelligence in Healthcare*, pages 393–422. Springer, 2022.

[158] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.

[159] Terry Shultz P.T. A person wearing a knee brace, December 2020. [Online; accessed October 19, 2022; Available at https://unsplash.com/photos/53WaTNC4].

[160] Brandon Sieu and Marina L Gavrilova. Person identification from audio aesthetic. *IEEE Access*, 9:102225–102235, 2021.

[161] Milan Sigmund. Spectral analysis of speech under stress. *IJCSNS International Journal of Computer Science and Network Security*, 7:170–172, 2007.

[162] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel. Joint speaker verification and antispoofing in the i-vector space. *IEEE Transactions on Information Forensics and Security*, 10(4):821–832, 2015.

[163] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[164] Benjamin Stephens-Fripp, Fazel Naghdy, David Stirling, and Golshah Naghdy. Automatic affect perception based on body gait and posture: A survey. *International Journal of Social Robotics*, 9(5):617–641, 2017.

[165] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1556–1560, 2008.

[166] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5688–5691. IEEE, 2011.

[167] Ioannis C Stylios, Olga Thanou, Iosif Androulidakis, and Elena Zaitseva. A review of continuous authentication using behavioral biometrics. In *Proceedings of the SouthEast European Design Automation, Computer Engineering, Computer Networks and Social Media Conference*, pages 72–79, 2016.

[168] Madeena Sultana, Padma Polash Paul, and Marina Gavrilova. A concept of social behavioral biometrics: Motivation, current developments, and future trends. In *2014 International Conference on Cyberworlds*, pages 271–278. IEEE, 2014.

[169] Kazuhiko Takahashi. Comparison of emotion recognition methods from bio-potential signals. *The Japanese Journal of Ergonomics*, 40(2):90–98, 2004.

[170] Rawesak Tanawongsuwan and Aaron Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[171] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint*, 2018.

[172] Vicky Tsang. Eye-tracking study on facial emotion recognition tasks in individuals with high-functioning autism spectrum disorders. *Autism*, 22(2):161–170, 2018.

[173] Sanjida Nasreen Tumpa and Marina L Gavrilova. Score and rank level fusion algorithms for social behavioral biometrics. *IEEE Access*, 8:157663–157675, 2020.

[174] Sanjida Nasreen Tumpa, KN Pavan Kumar, Madeena Sultana, Gee-Sern Jison Hsu, Orly Yadid-Pecht, Svetlana Yanushkevich, and Marina L Gavrilova. Social behavioral biometrics in smart societies. In *Advancements in Computer Vision Applications in Intelligent Systems and Multimedia Technologies*, pages 1–24. IGI Global, 2020.

[175] Gentiane Venture. Human characterization and emotion characterization from gait. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 1292–1295. IEEE, 2010.

[176] Gentiane Venture, Hideki Kadone, Tianxiang Zhang, Julie Grèzes, Alain Berthoz, and Halim Hicheur. Recognizing emotions conveyed by human gait. *International Journal of Social Robotics*, 6(4):621–632, 2014.

[177] Lisa M Vizer, Lina Zhou, and Andrew Sears. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies*, 67(10):870–886, 2009.

[178] Garima Vyas, Malay Kishore Dutta, Kamil Riha, Jiri Prinosil, and Chandni. An automatic emotion recognizer using mfccs and hidden markov models. In *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 320–324. IEEE, 2015.

[179] Wen Wan-Hui, Qiu Yu-Hui, and Liu Guang-Yuan. Electrocardiography recording, feature extraction and classification for emotion recognition. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 4, pages 168–172. IEEE, 2009.

[180] Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu. Fusion of static and dynamic body biometrics for gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(2):149–158, 2004.

[181] Yang Wang, Zhao Lv, and Yongjun Zheng. Automatic emotion perception using eye movement information for e-healthcare systems. *Sensors*, 18(9):2826, 2018.

[182] Ying Wang, Shoufu Du, and Yongzhao Zhan. Adaptive and optimal classification of speech emotion recognition. In *2008 Fourth International Conference on Natural Computation*, volume 5, pages 407–411. IEEE, 2008.

[183] Cynthia M Whissell. The dictionary of affect in language. In *The Measurement of Emotions*, pages 113–131. Elsevier, 1989.

[184] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):165–183, 2006.

[185] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

[186] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Yi Guo, Victor Leung, Jun Cheng, and Bin Hu. Emotion recognition from gait analyses: Current research and future directions. *arXiv preprint arXiv:2003.11461*, 2020.

[187] Ya Xu and Guang-Yuan Liu. A method of emotion recognition based on ECG signal. In *2009 International Conference on Computational Intelligence and Natural Computing*, volume 1, pages 202–205. IEEE, 2009.

[188] Ya Xu, Guangyuan Liu, Min Hao, Wanhui Wen, and Xiting Huang. Analysis of affective ECG signals toward emotion recognition. *Journal of Electronics (China)*, 27(1):8–14, 2010.

[189] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[190] Shanxiao Yang and Guangying Yang. Emotion recognition of EMG based on improved LM BP neural network and SVM. *Journal of Software*, 6(8):1529–1536, 2011.

[191] B Yegnanarayana, R Kumara Swamy, and K Sri Rama Murty. Determining mixing parameters from multispeaker data using speech-specific information. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1196–1207, 2009.

[192] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*, volume 2018, pages 3688–3692, 2018.

[193] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2016.

[194] Erhu Zhang, Yongwei Zhao, and Wei Xiong. Active energy image plus 2DLPP for gait recognition. *Signal Processing*, 90(7):2295–2302, 2010.

[195] Shiqing Zhang. Emotion recognition in chinese natural speech by combining prosody and voice quality features. In *International Symposium on Neural Networks*, pages 457–464. Springer, 2008.

[196] Zhan Zhang, Yufei Song, Liqing Cui, Xiaoqian Liu, and Tingshao Zhu. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ*, 4:e2258, 2016.

[197] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d CNN LSTM networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[198] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[199] Mu Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6, 2004.

[200] Xizhi Zhu. Emotion recognition of EMG based on BP neural network. In *Proceedings of International Symposium on Network Network Security*, pages 227–229. Citeseer, 2010.

[201] Jyun Rong Zhuang, Guan Yu Wu, Hee Hyol Lee, and Eiichiro Tanaka. Applying the interaction of walking-emotion to an assistive device for rehabilitation and exercise. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6489–6494. IEEE, 2019.

[202] Philippe Zimmermann, Patrick Gomez, Brigitta Danuser, and S Schär. Extending usability: Putting affect into the user-experience. *Proceedings of NordiCHI'06*, pages 27–32, 2006.