Multimodal Emotion Recognition Software to Facilitate Human-Robot Interaction

Joshua Bamforth

Sheffield Hallam University

A thesis submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of *Master of Philosophy*

October 2024

Abstract

This thesis explores the development and evaluation of a multimodal emotion recognition system, integrating both Facial Emotion Recognition (FER) and Audio Emotion Recognition (AER). The FER system employs various face detection algorithms, including Tiny YOLO, YOLO, dlib, and Haar Cascade, followed by emotion recognition models such as MobileNetV2, ResNet50, and VGG16, trained using transfer learning. These models are assessed for both accuracy and efficiency, and their real-time performance is tested on a robotic platform. For AER, off-the-shelf tools like OpenSMILE are considered alongside sentiment analysis systems such as IBM Watson and large language models, to evaluate emotion detection through vocal input. Both modalities are evaluated on metrics such as response time and emotion recognition accuracy, with a particular focus on their real-world applicability when deployed on a physical robot. The study also discusses the trade-off between accuracy and inference speed, highlighting the most effective combinations of models. Future work includes testing the system in live, uncontrolled environments with human participants, implementing on-robot AER systems, and exploring the fusion of facial and audio data for enhanced emotional understanding. These efforts aim to further optimise the system for real-time interaction and improved emotion recognition in human-robot interaction scenarios.

Keywords: Facial Emotion Recognition, Audio Emotion Recognition, Social Robot, Multimodal Emotion Recognition.

Author

Joshua Bamforth

Supervisory team: Prof. Alessandro Di Nuovo, Dr. Jing Wang

I hereby declare that:

1. I have not been enrolled for another award of the University, or other academic or professional organisation, whilst undertaking my research degree.

2. None of the material contained in the thesis has been used in any other submission for an academic award.

3. I am aware of and understand the University's policy on plagiarism and certify that this thesis is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.

4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy.

5. The word count of the thesis is 18900.

Name: Joshua Bamforth

Date: Tuesday 28th January, 2025

Award: Computing and Informatics

Director(s) of Studies:

Prof. Alessandro Di Nuovo

Dr. Jing Wang

Acknowledgement

I would like to extend my sincere gratitude to those who have supported and guided me throughout the completion of this thesis. First and foremost, I would like to thank Alessandro Di Nuovo and Jing Wang for their invaluable mentorship, guidance, and unwavering support. I am also deeply grateful to SITS Lab, whose expertise and feedback have been instrumental in shaping this work.

This research was made possible through the generous funding provided by the IBM Shared University Research Award. I extend my sincere gratitude for this support, which has been instrumental in enabling the successful completion of my MPhil research.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Emotion recognition plays a critical role in Human-Robot Interaction (HRI) by enabling robots to better understand and respond to the emotional states of humans. This is crucial for fostering natural and intuitive communication between robots and humans, as emotional cues guide social interactions [18]. When robots can recognise emotions, they can adapt their behaviour, tone [15], or responses that make the interaction more engaging, empathetic, and contextually appropriate. This capability is particularly important in healthcare and education where the robot's ability to sense and respond to emotions can significantly enhance the user experience.

In healthcare, robots that can detect patient emotions are better equipped to provide support and improve outcomes, especially for people with mental health conditions, autism [25], or older people. Emotional awareness helps robots engage with patients more effectively, offering personalised interactions that take into account mood fluctuations or stress [28].

Current approaches to emotion recognition face significant challenges, particularly when they rely on a single modality. Systems that analyse only facial expressions may struggle to accurately interpret emotions in cases where facial cues are not visible, such as when individuals are not facing the camera. Similarly, audio-based emotion recognition can be limited by variations in speech patterns, background noise, or linguistic differences, leading to potential inaccuracies.

Multimodal emotion recognition refers to identifying and interpreting human emotions by combining data from multiple sources, such as facial expressions, vocal intonations, and gestures. Unlike traditional methods that rely on a single modality, multimodal approaches offer a more comprehensive and nuanced understanding of emotional states. By integrating different types of data, these systems can achieve greater accuracy and reliability in emotion detection, overcoming the limitations inherent in relying on just one mode of input.

A more comprehensive and robust system integrating visual and auditory inputs would benefit the field. By combining face detection with audio emotion recognition, it becomes possible to cross-validate and reinforce the emotional insights derived from each modality. This multimodal approach makes the system more adaptable to real-world conditions, where emotions are often expressed through a complex interplay of visual and auditory signals [27].

This thesis explores the integration of facial and textual data to create a more robust emotion recognition system tailored for the more widely available robots which tend to be resource-limited, beginning with a comprehensive review of existing literature on emotion recognition, focusing on the challenges and advancements in gesture recognition, facial recognition, audio/text recognition, and multimodal approaches. In addition, a critical assessment of the role of invasive technologies such as EEG in emotion detection is performed, considering their ethical implications and practical limitations.

The Materials and Methods chapter details the robotic platform used in this study. There is then a discussion on implementing emotion detection algorithms, utilising established software libraries and frameworks, and describing the methodologies used for data preparation and system evaluation on the resource-limited robots.

In Chapter 4: Facial Emotion Recognition, the study explores a dual-model approach where separate models are used for face detection and emotion classification. For face detection, various algorithms such as Haar Cascade, Dlib, and more advanced models like Tiny YOLO and YOLO are evaluated. This division of tasks allows for more specialised processing, ensuring that each stage—detecting the face and classifying the emotions—can be optimised independently. The system's performance

is carefully analysed in resource-constrained environments, where computational efficiency is critical. By comparing models, the study highlights the trade-offs between accuracy and processing speed, particularly important for systems deployed in low-power devices or real-time applications.

Chapter 5: Audio Emotion Recognition delves into the use of IBM Watson's capabilities for analysing emotional content in speech, with a focus on sentiment analysis. Using IBM Watson's powerful natural language understanding, this system can perform sentiment analysis detecting emotions such as happiness, sadness, anger, and fear, based on textual input. The chapter assesses Watson's performance in terms of accuracy and speed, determining whether it is suitable for real-time speech emotion detection. Additionally, the role of large language models (LLMs), like ChatGPT, is explored in complementing emotion recognition by enhancing conversational capabilities, allowing for more natural and empathetic interactions. While IBM Watson remains central to the system's speech analysis, the study also investigates other tools like OpenSMILE, known for its advanced feature extraction capabilities in audio signal processing.

The Conclusion chapter provides a comprehensive summary of the research, highlighting the key findings and contributions of the multimodal emotion recognition system. Reflecting on the effectiveness of the combined facial and audio emotion detection approaches, highlighting the strengths and limitations of the models used in both modalities. In addition, the chapter offers insights into areas for future improvement. These include enhancing emotion detection accuracy, optimising computational efficiency, and exploring deeper integration of large language models for more natural and responsive interactions. The conclusion also outlines possible future directions, such as user studies to validate the system's performance and further development of multimodal fusion techniques to achieve even more accurate emotion recognition.

Chapter 2

Literature

2.1 Facial Emotion Recognition

Facial emotion recognition is a crucial aspect of affective computing [75] that involves analysing facial expressions to identify human emotions. This skill is essential for successful interactions between people and is particularly important in the realm of human-robot interaction. For robots to respond to human emotions promptly, facial emotion recognition is key. When aware of human emotions, robots can interact more naturally with humans by quickly and accurately recognising emotions. More natural interaction capability will favour acceptance and use of robots in peoples lives.

2.1.1 Neural Networks

Convolutional Neural Networks (CNNs) have emerged as the dominant approach in the realm of vision-based emotion recognition for robotic systems. Researchers typically adopt a two-phase methodology, first using CNNs for the extraction of features, followed by the implementation of classification techniques. One study introduces a multistep technique that aims to improve facial recognition accuracy. It begins with the application of a histogram equation to enhance image contrast, which is then succeeded by a bilateral filter to reduce noise while maintaining edge integrity. Then, the Viola-Jones (Haar Cascade) face detection algorithm in OpenCV

is utilised to pinpoint the facial area within the input image. The proposed technique further refines the extraction of features through an innovative variant of local binary pattern (LBP), which takes advantage of a convolution filter and a Kirsch operator to capture features that withstand variations in illumination, scaling, and rotation [62].

Innovative approaches, such as using Haar-Cascade for initial face detection and cropping the image to isolate the face before implementing more advanced CNNs [7] or SVMs [36] [81], show promise for working in resource-limited environments. One study used a Haar cascade to quickly locate faces, followed by a CNN to extract features, and finally a long-short-term memory network [52] to perform classification. Other studies have used a convolutional autoencoder and support vector regressors [4] or recurring neural networks [14] to incorporate temporal features into emotion recognition and establish correlations between facial expression transformations and the six basic emotions.

In a study by Kusuma et al. [49], VGG16 was effectively used for emotion recognition on the FER dataset. Their model achieved an overall accuracy of 69.4% after careful optimisations through specific configurations, such as using an imbalanced dataset (they did not use data augmentation to rectify the heavy imbalances within the dataset), Global Average Pooling (GAP), non-frozen layers, and the Stochastic Gradient Descent (SGD) optimiser.

ResNet50 was used to a high degree of accuracy on FER 2013 by Pramerdorfer et. al. [76] achieving a 72.4% with 5.3 million trainable parameters. No special modifications were made to the network except removing the initial CP (convolution and pooling layer) from the architecture and making the network slightly narrower, having 256 feature maps in the final residual group to reduce the number of parameters.

[86] presents the design of an adaptive learning system for real-time emotion recognition in humanoid robots. The system continuously updates individualised models based on user interactions, improving performance over time. It employs an ensemble of ResNet50 and Inception v3 networks, leveraging transfer learning to enhance emotion recognition from facial expressions. They performed a two-stage

user study featuring 75 participants using the results for stage one to personalise the experience in stage two. The robot's adaptive actions used the recognised emotions to engage users in social interactions and to elicit emotional responses, such as trust, empathy, and engagement. Results showed a 12% improvement in emotion recognition accuracy and an 8.28% increase in the success rate of emotion elicitation between stages, showcasing the system's ability to adapt and foster meaningful social interactions.

I-MobileNetV2, an enhanced version of MobileNetV2 proposed by Zhu et al. [103], aimed to improve facial emotion recognition tasks by addressing issues such as large parameter quantities, loss of feature information and low accuracy rates. Key modifications include the retention of depthwise separated convolution for computational efficiency, a reverse fusion mechanism to preserve negative features, the use of the SELU activation function to avoid gradient vanishing, and the integration of the SE-Net channel attention mechanism to improve feature recognition. These enhancements resulted in recognition accuracies of 68.62% on FER2013 and 95.96% on CK+, with an 83.8% reduction in parameter count.

Despite these improvements, the accuracy gains over the base MobileNetV2 are modest, with only a 0.72% increase on FER2013 and 6.14% on CK+. However, the reduction in parameters should significantly improve inference speeds over the MobileNetV2 base model.

Data augmentations have been employed in various approaches to enhance facial emotion recognition, often in conjunction with conventional CNNs. Several studies have demonstrated that augmenting training data can improve model performance by addressing challenges such as class imbalances and overfitting [84] [83]. A particularly successful technique involved Generative Adversarial Networks (GANs) for data augmentation, as demonstrated by Song and Kwon (2019). Their study also emphasised the importance of including the lower half of the face during training to improve the accuracy in detecting emotions through facial recognition.

2.1.2 Applications

The exploration of applications is also apparent, ranging from studies investigating and improving the effectiveness of emotion recognition in older adults [58] to those focusing on unconstrained environments [95] and those with the goal of creating a robot capable of helping speech therapy through the ability to articulate words similar to that of human speech [31], through the development of facial expression recognition and lip syncing capabilities, the RASA robot aims to engage children and enhance their learning outcomes.

The Nao robot is a popular choice in research exploring facial emotion recognition. Many studies have focused solely on robot cameras for recognition, with classification being handled by a separate laptop due to Nao's processing limitations [82]. Notably, [61] revealed Nao's constrained processing capacity, with live inference on the robot's cameras only achieving 0.25 frames per second (FPS). However, this limitation was significantly addressed by integrating the Neural Compute Stick 2 (NCS 2), a neural network preprocessor developed by Intel. Another solution involved reprogramming the NaoQI software of the Nao robot to be lighter to allocate more processing power to emotion recognition [54].

One study focused on developing a system capable of operating efficiently on limited computational power, specifically for use on the Ohmni robot. The Lightweight EMotion recognitiON (LEMON) model [26] used a residual learning-based technique that combined Dilated Convolutional layers with Standard 2D Convolutional layers. While the model did not achieve the highest accuracy, its strong performance in resource-constrained environments highlights its potential applicability in robotics.

Chih-Lyang [41] presented a study featuring an Omni-Directional Service Robot (ODSR) that uses a Faster-CNN to detect humans within its field of view. Once a person is identified, the robot assesses whether the individual is oriented toward the camera before applying a Haar Cascade to detect and crop the person's face. The cropped face is then analysed to deduce the individual's emotion using a Sinogram Super-Resolution and Denoising Convolutional Neural Network (SRCN). The identified emotion is used to select and play music that corresponds to the

detected emotion. Additionally, a second SRCN is employed for speech recognition, enabling the robot to respond and act upon verbal commands, such as 'follow.'

Facial muscle movements known as Action Units (AUs) are an integral part of the Facial Action Coding System (FACS). AUs serve as building blocks for describing facial expressions and play a critical role in the recognition of facial emotions. By breaking down expressions into discrete components, AUs are used to analyse and categorise them. Each AU corresponds to specific facial muscle movements and their combinations represent a diverse array of facial expressions. The primary goal is to deconstruct facial expressions into fundamental units, which enhances our comprehension and recognition of emotions [63]. Chinonso Paschal Udeh [92] aimed to create a system that provides more access to incorporating AUs into research using a multitask approach along with multiview co-regularisation frameworks as the baseline, the study achieves an average CNN recognition accuracy of 80% in seven emotion categories for reclassifying datasets based on seven main AU categorisations and expressions.

2.1.3 Datasets

Pierre Luc Carrier and Aaron Courville [34] introduced the Facial Expression Recognition 2013 (FER-2013) dataset as part of a larger project aimed at advancing emotion recognition research. Created using the Google image search API, it collected images matching 184 emotion-related keywords like 'blissful' and 'enraged.' The dataset includes nearly 36,000 images, processed using OpenCV for face detection and manually curated for accuracy. These images were resized to 48x48 pixels, converted to grayscale, and categorised into seven broad emotion classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral. It serves as a crucial resource for training and evaluating emotion recognition models, being the most used data set for review of articles.

AffectNet [66] is a large-scale facial expression dataset introduced to address the need for more diverse and comprehensive data in facial emotion recognition. It contains over one million images of faces collected using emotion-related keywords

translated into multiple languages on three different search engines (Google, Bing and Yahoo). These images were manually annotated into eight different emotion categories: neutral, happy, sad, surprise, fear, disgust, anger, and contempt, along with additional labels for valence and arousal. AffectNet stands out due to its extensive size, diversity in ethnicity, age, and conditions such as pose and lighting variations, making it a valuable resource for training and evaluating emotion recognition systems.

JAFFE [5] and CK+ are both smaller, highly curated datasets. JAFFE contains 213 images of posed facial expressions from 10 Japanese female models, labelled with six basic emotions plus neutral. It is often used for cross-cultural studies of emotion recognition. CK+, meanwhile, includes 593 video sequences from 123 subjects, with each sequence showing a transition from a neutral face to a peak expression. CK+ is notable for including both emotion and action unit labels, providing fine-grained information about facial muscle movements, making it a strong choice for both emotion and FACS-based studies. KDEF [56], a separate dataset of 4,900 images from 70 individuals, focuses on a broader demographic range and is often used for validation in emotion recognition systems.

2.2 Audio-based Emotion Recognition

The field of affective computing includes emotion recognition through audio, which involves analysing vocal cues and patterns to discern human emotions. In situations where visual cues are not available, such as when individuals are out of a robot's line of sight, audio-based emotion recognition becomes crucial. In human-robot interaction, accurately detecting emotions through audio signals is extremely important. Audio-based emotion recognition technologies complement facial emotion recognition and allow robots to understand the subtle emotional states of individuals through speech, intonation, and other auditory features.

2.2.1 Common Methods

Many different methods are employed when attempting to do feature extraction in the context of audio emotion recognition:

Mel-frequency Cepstral Coefficients (MFCC) [3] is a method that involves converting the frequency spectrum into the Mel-frequency scale and then computing cepstral coefficients to represent the spectral envelope of the signal.

Gammatone Frequency Cepstral Coefficients (GFCC) [87] is a variant of MFCC that uses a gammatone filterbank instead of the traditional Mel filterbank. This modification is believed to better align with auditory perception, capturing more important features related to human hearing. However, despite its theoretical advantages, GTCCs are not as commonly used compared to MFCCs in the context of emotion recognition research.

Linear Predictive Coding (LPC) [72], a technique used to represent the spectral envelope of a signal by modelling it as the output of a linear filter applied to a time-domain representation of the signal.

Linear Predictive Cepstral Coefficients (LPCC) [51] are a combination of the principles of LPC and cepstral analysis. Similar to MFCCs, LPCCs involve extracting cepstral coefficients from the linear predictive coding (LPC) coefficients. LPCCs aim to capture both the temporal dynamics and the spectral properties of the signal, making them suitable for tasks that require detailed representation of acoustic characteristics.

2.2.2 Algorithms

Methods such as MFCC are often used to extract features that are then classified into emotions using algorithms such as SVM [45], CNN, or DNN [85] [77]. When using GTCC feature extraction, KNN is a common choice as a classifier, while LSTM is preferred when the dataset is large enough for better results [102].

An emerging trend in audio-based emotion recognition involves the transformation of audio signals into visual representations [10], spectrograms, followed by the application of machine learning techniques such as CNNs [42] or Deep Belief Networks

(DBNs) [64]. This approach has gained considerable traction within the research community, with various studies adopting unique methodologies. For example, researchers have investigated the use of CNNs in conjunction with K-means clustering to identify spectrogram frames containing crucial information [37]. Furthermore, a study has expanded on this approach by integrating a bidirectional long-short-term memory (BiLSTM) network to analyse discriminative features extracted from spectrograms, allowing the inference of speaker emotional states [67].

Notably, the use of tools such as openSMILE toolkit by audEERING is observed in several studies. OpenSMILE is a feature extractor that can be configured to extract specific features from audio and music signals for signal processing and machine learning, emphasising features enabling emotion recognition from speech. The paper by [6] decided to test several of the extractable features for their application in emotion recognition, they tested: Intensity, Loudness, 12 MFCC, Pitch (f0), Probability of Voicing, F0 Envelope, 8 LSF and Zero-Crossing Rate. Once collected, they select the best features with a 'BestFirst' approach. These features were then classified, testing 3 different classification methods: multilayer perceptron neural networks, Rules Classifier oneR, and Tree Classifier J48. Their results showed that the multilayer Perceptron neural network had the best performance. A couple of other methods utilising this feature extraction method employed different networks for their classification, one choosing a Two-Layer Fuzzy Random Forest ensemble classifier [22], and another a SVM [9].

Attention-based speech emotion recognition models have garnered significant attention in recent years due to their ability to capture relevant features from audio data effectively, thus improving the precision of emotion classification tasks [74]. By dynamically weighting different segments of the input speech signal [101] based on their importance in expressing emotional content, these models offer a promising approach to discern subtle nuances in speech patterns associated with various emotional states [69]. Using mechanisms inspired by human attentional processes, such as self-attention and multi-head attention [79], these models excel in identifying salient acoustic cues indicative of specific emotions, thus paving the way for more nuanced and contextually rich emotion recognition systems.

2.2.3 Applications

In a study on speech emotion recognition in speaker-independent systems, specifically for the Mung robot [47], two key strategies were proposed to improve accuracy and reliability. The first strategy involves separating emotion recognition from consonants and obstruents to reduce text dependency and enhance adaptability. The second strategy introduces a rejection algorithm based on a confidence measure to ensure more reliable outcomes. Comparative analysis with conventional methods showed significant improvements, with recognition rates increasing from 6.9% to 27.6% across various emotional features through the separation algorithm and 73% to 92% with the confidence-based rejection mechanism for MFCCs [46].

In a paper by Carolis, B.D. [17], they use the Nao robot combined with a module that they developed called the VOCE 2.0. A module designed to classify speech used to send requests by features extracted according to dimensional models, valence, and arousal. They encountered low performance; however, they attribute this to the training data, the ∈motion dataset, and their target audience of elderly individuals, which have a different range of characteristics.

Research has highlighted the indispensable role of data augmentation techniques in strengthening speech-emotion recognition systems. Lakomkin et al. [50] carried out a study to evaluate the impact of data augmentation by testing two models, one with data augmentation and the other without. Their analysis, especially when using the iCub robot as a test bed, demonstrated a significant drop in overall resilience and effectiveness for the model that did not incorporate data augmentation. These results serve as a compelling reminder of the critical importance of utilising data augmentation strategies to improve the performance of speech-emotion recognition systems.

2.2.4 Datasets

The IEMOCAP database [16] is the leading resource for audio-based emotion recognition research. With approximately 12 hours of meticulously annotated audiovisual data, it covers a wide range of modalities including video recordings, speech samples,

and motion capture of facial expressions. In addition, the database includes detailed text transcriptions, making it a versatile platform for various emotion recognition projects. Research can use IEMOCAP for nuanced investigations of emotional expression, including facial emotion analysis and text-based sentiment classification.

The SAVEE (Surrey Audio-Visual Expressed Emotion) database [38] stands out as another crucial resource. It offers a varied selection of acted speech samples that depict emotional states such as happiness, sadness, anger, fear, disgust, and neutrality, delivered by four male post-graduate students. The text material consists of 15 TIMIT (Texas Instruments/Massachusetts Institute of Technology) sentences per emotion: 3 common, 2 emotion-specific, and 10 generic sentences that were different for each emotion and phonetically balanced.

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset [53] is a multimodal dataset that includes both speech and song recordings, designed to support research in emotion recognition. It contains recordings from 24 professional actors (12 male, 12 female) who vocalise emotions such as calm, happy, sad, angry, fearful, surprised, and disgusted. Each emotion is expressed at two levels of intensity, and the dataset provides both audio-only and audio-visual recordings, making it suitable for studies that involve both auditory and visual cues for emotion recognition.

2.3 Gesture-based Emotion Recognition

The analysis of human gestures to understand emotional states is a crucial aspect of affective human-human interaction. Body movements, hand gestures, and facial expressions are among the non-verbal cues that provide valuable emotional information. In situations where verbal or facial cues are limited, such as when someone is wearing a face mask, gesture-based emotion recognition is essential, leading to more intuitive and empathetic interactions between humans and machines.

Marinoiu et al. [59] explore the complexities of recognising emotions through gestures, focusing on adapting state-of-the-art RGB 3D human pose reconstruction methods that blend feedforward and feedback components. Their study compares

several baselines for recognising actions and emotions using 2D and 3D representations of both children and therapists. The results suggest that with proper adaptation, current RGB-based 2D and 3D reconstruction methods can rival industrial-grade RGB-D Kinect systems. They employed methods like DMHS (Deep Multitask Architecture for Integrated 2D and 3D Human Sensing) and a custom variant, DMHSPV, to extract features and introduced a new dataset, H80kPartial. While CNNs outperformed RNNs in emotion recognition, their findings highlight the ongoing efforts to enhance gesture-based emotion recognition, especially in contexts like robot-assisted therapy for children with autism.

Wang et al. [94] have introduced an innovative method for touch gesture and emotion recognition called Multi-Task Touch Gesture and Emotion Recognition (MUSCAT). This approach involves using a fabric embedded with touch sensors that mimic human skin, allowing accurate touch gesture recognition. The results of the TouchGET and CoST datasets demonstrate that the MUSCAT method significantly reduces computation costs while improving classification accuracy. Furthermore, the incorporation of Multi-Task Learning (MTL) further enhances classification performance, validating the effectiveness of the proposed MUSCAT method and MTL framework in touch gesture and emotion recognition.

Lyu and Sun [57] faced a unique challenge in the field of dance emotion recognition for robots, which presents numerous difficulties because video-based emotion detection is vulnerable to various external factors. To overcome these obstacles, the authors created a strong multi-feature fusion framework that combines global and local features using an LSTM mechanism. The study used three distinct data sets: RML, SAVEE, and a self-constructed dance video database. The experimental process involved training and testing on these datasets, which produced promising results that demonstrated the effectiveness of their proposed feature extraction algorithm. Notably, their approach surpassed single-feature methods, showing the viability of emotion recognition from dance.

The integration of multimodal sensory information presents a critical challenge in the development of advanced human-machine affective systems. A research article titled 'Deep Emotion Recognition through Upper Body Movements and Facial

Expression' by Aqdus et al. [8] delves into this challenge, focusing on spatial-temporal techniques for emotion analysis across visual modalities.

The study explores the fusion of two primary modalities: facial expressions and upper-body movements. The researchers aimed to develop a robust architecture capable of identifying emotions in real-time human-machine interaction systems. Their findings highlight the superiority of the bimodal approach over those of monomodal ones, regardless of the fusion method applied. In particular, the study achieved the best recognition rates for anger, happiness, and neutral emotions, while the worst recognition rate was observed for sadness, often misclassified as surprise. Evaluation metrics consistently demonstrated significant improvements in accuracy, moving from 77.7% and 76.8% for the recognition of emotion from facial and upper body movements, respectively, to 85.7% and 86.6% after the fusion of both modalities.

The study, aimed at improving Human-Robot Interaction (HRI) through emotion recognition, uses the self-created Body Expressions of Emotion (BEE) data set with a split training testing 60% - 40%. Using a Nao Robot with a depth camera for data collection, they employ a multi-layer neural network architecture that classifies emotional states from 3D skeleton data. The first layer uses Grow When Required (GWR) networks to learn pose and motion prototypes, while the second layer, a recurrent GWR variant, learns prototype sequences and associates them with unsupervised visual representations of emotions for classification.

2.3.1 Datasets

The Body Emotion Expression (BEE) dataset [30] captures six emotions—anger, fear, happiness, neutral, sadness, and surprise—from 19 participants with diverse cultural backgrounds. Using two Nao robots, one equipped with a depth sensor, body motions were recorded from frontal and side views. Participants performed both neutral and emotionally driven actions inspired by brief scenarios. The dataset includes 570 sequences, with 3D skeleton data extracted from 11 key body joints.

The H80kPartial dataset [59] is a subset of the larger H80k (Humans 80k) dataset, specifically designed for emotion recognition through human pose analysis. This

dataset focuses on capturing human body postures and gestures that are indicative of emotional states. It contains annotated data that emphasise partial body poses, such as the upper body, arms, and facial orientation, to help identify emotional expressions without requiring full-body information.

2.4 Table Of Robots

The following tables present a comprehensive compilation of robots featured in the literature related to emotion recognition and human-robot interaction. Each entry includes the robot's name, an accompanying image, and references to the studies that discuss its application or implementation. This collection highlights the diversity of robotic platforms employed in research, showcasing their varying capabilities and roles in facilitating emotional understanding in interactions with humans.

Table 2.1: Table of all Robots in Literature

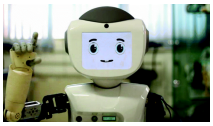
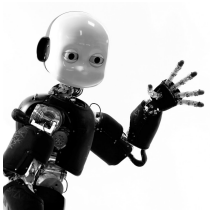
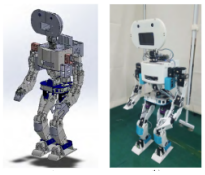
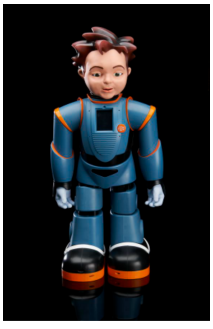
	Robot Name	**Ref**
	Pepper	[99] [100]
	Nao	[17] [33] [43] [54] [61] [81] [82] [83] [86] [95] [96]
	Mung	[46]
	Omnidirectional Service Robot	[41]
	Pioneer P3-DX robot, LARa robot	[60]
	RASA	[31]
	Ohmni	[26]

Table 2.2: Table of all Robots in Literature Cont.

	Robot Name	Ref
	iCub	[50]
	ROBIN	[9]
	ESRS (Emotional Service Robot System)	[19] [20] [23]
	Harley	[52]
	Robot eye	[7]
	Zeno	[59]

2.5 Multi-modal Emotion Recognition

Multimodal emotion recognition systems aim to improve the accuracy and reliability of emotion detection by incorporating multiple types of input data, such as visual, auditory, physiological, or textual information. By drawing from diverse sources, these systems can capture a more comprehensive understanding of human emotions. Some approaches integrate these modalities into a unified output, combining the strengths of each to enhance overall performance. Others treat each modality independently, using one to validate or back up the other in cases of ambiguity or failure. The fusion of modalities allows for more robust emotion recognition, particularly in complex, real-world settings where single modalities may fall short.

Studies such as [90] have investigated classification techniques for combining various modalities, artificial neural networks (ANN) with k-nearest neighbours (k-NN). Decision trees were also employed by [1] having a simple CNN for facial detection and a log-Mel spectrum for feature extraction from speech.

Kansizoglou et al. [44] used two CNNs, one for audio recognition and one for facial expression recognition, together with a DNN to fuse them; a long- and short-term memory (LSTM) layer and a Reinforcement Learning (RL) agent are trained in cascade, stopping feature extraction for final prediction. Additionally, using a Haar cascade for initial face cropping, they employ MobileNetV2 for image classification and a VGG architecture for audio, both retrained on emotion recognition datasets. The results indicate improved accuracy through fusion, albeit with some emotion confusion, tested on RML and BAUM-1 datasets.

In the pursuit of advancing multimodal emotion recognition within the realm of human-robot interaction, Yu and Tapus [99] present a study titled 'Interactive Robot Learning for Multimodal Emotion Recognition.' Their research employs a Pepper robot and a sophisticated experimental setup featuring a Kinect and an Optris thermal camera to capture human gait information and thermal facial images for emotion recognition. This study developed a multimodal emotion recognition model grounded in gait and thermal facial data, using a random forest (RF) model and modified confusion matrices of two individual models. A comparative analysis

between individual RF models and the hybrid decision-level model demonstrates the effectiveness of their integration method in classifying emotions during human-robot interaction. Moreover, the extensive experimentation involving online testing before and after Interactive Robot Learning (IRL) substantiates that interactive robot learning is a valuable technique, yielding a significant increase of more than 10% in the accuracy of multimodal emotion recognition with gait and thermal data. Yu and Tapus [100] even attempted to further improve upon their innovative thermal imaging plus human gait information for emotion recognition by implementing WaveNet to get more benefits from spatial and temporal information.

Chen et al. [21] have introduced a groundbreaking approach to multimodal emotion recognition in human-robot interaction called Coupled Multimodal Emotional Feature Analysis (CMEFA). This method utilises a Broad-Deep Fusion Network (BDFN) to extract emotional features from facial expressions and gestures. By applying Canonical Correlation Analysis (CCA) to capture the correlation between these features, CMEFA offers a more comprehensive understanding of emotional cues. A coupling network recognises emotions based on the extracted bi-modal features. Remarkably, simulation experiments conducted on the FABO database demonstrate the superiority of CMEFA over existing methods, outperforming the SVM Recursive Feature Elimination (SVMRFE) method by achieving a recognition rate of 1.15% higher and exceeding other approaches by significant margins.

In addition, the researchers conducted preliminary application experiments on an emotional social robot system, where the robot successfully recognised emotions based on the facial expressions and body gestures of the volunteers. This showcases the practical applicability of CMEFA in real-world scenarios.

A study by Aqdus et al. [8] emphasised the importance of integrating gesture recognition into emotion recognition systems. Their work focused on deep emotion recognition through upper body movements and facial expressions, exploring three distinct methods, feature-level fusion, compact bilinear pooling fusion, and decision-level fusion, to combine these modalities to improve accuracy and robustness.

Several studies have emerged focusing on sentiment analysis, addressing this gap in research. For example, Augello et al. [11] presented 'Multimodal Mood Recognition

for Assistive Scenarios', showcasing the effectiveness of their approach in detecting emotions from textual data. Additionally, Heredia et al. [39] proposed the 'Adaptive Multimodal Emotion Detection Architecture for Social Robots,' incorporating natural language processing (NLP) transformers and an emotion ontology to enhance emotion detection capabilities in social robots.

Temporal features have emerged as a significant benefit for multimodal emotion recognition systems. Research efforts such as those by Hung et al. [40] have focused on leveraging temporal feature learning to improve emotion recognition accuracy, highlighting the effectiveness of using multiple models to capture temporal dynamics in emotional expressions.

2.5.1 Datasets

The FABO (FAcial and BOdily Expression) dataset [35] plays a significant role in pose-based emotion recognition, as it captures both facial expressions and body postures across a range of emotions. With recordings from 23 subjects displaying ten different emotional states, the dataset emphasises the importance of bodily cues in recognising emotions. This makes FABO particularly valuable for research focused on pose-based emotion recognition, where body language and movement are key to identifying emotions. The combination of facial and bodily data allows for a more comprehensive analysis of how emotions are expressed physically, supporting the development of robust, multimodal recognition systems.

2.6 Critical Review of Invasive Technology for Emotion Recognition

Emotion recognition technologies have made significant advances in recent years, with various invasive and less invasive techniques being developed to better capture emotional states. These technologies can be broadly categorised into electro-based systems, physiological sensors, and non-contact methods. Although each approach has its strengths, they also present notable limitations, particularly in terms of user

comfort and practicality for real-world application. This review critically evaluates key technologies for emotion recognition, outlining their mechanisms, benefits, and drawbacks.

2.6.1 Electro-Based Technologies

Electro-based methods such as EEG (Electroencephalography), EMG (Electromyography), EOG (Electrooculography) and ECG (Electrocardiography) are commonly used in emotion recognition due to their ability to capture direct biosignals from the brain and body. EEG, for example, measures brain electrical activity and is used to classify emotional states based on asymmetries in the prefrontal cortex. High arousal emotions (e.g., joy, anger) are associated with greater activity of the left frontal cortex, whereas low arousal emotions (e.g., fear, sadness) show increased activity of the right frontal [65]. However, EEG signals vary significantly between individuals, making the development of universal models challenging [73]. Furthermore, EEG setups require sensors attached to the scalp, which limits practicality in nonlaboratory settings due to discomfort and susceptibility to movement artefacts, particularly head movements [91].

EMG, on the other hand, measures muscle activity and is used to detect emotional expressions through facial or bodily muscle movements. For example, negative emotions correlate with high activity in the corrugator supercilii (frowning muscles), while positive emotions exhibit reduced activity in this region [65]. Although it is an effective tool for detecting nuanced emotional expressions, EMG requires that sensors be placed directly on the skin, which can be intrusive and uncomfortable, especially in dynamic everyday situations.

The electrocardiogram (ECG) is another key method, used to measure the electrical activity of the heart, enabling the detection of emotional reactions by analysing the variability of heart rate (HRV). Similar problems arise with EOG, which tracks eye movement and pupil dilation, but again requires physical contact with sensors around the eyes [29].

Overall, these electro-based systems excel in providing high-quality, detailed emotion-related data, yet their invasive nature and reliance on stationary or controlled environments impede their adoption in daily life. In terms of precision, multimodal recognition systems, such as those that combine EEG and facial expression recognition through convolutional neural networks (CNNs), have been shown to perform better. For example, integrating EEG signals and facial data through plurality voting classifiers and the Monte Carlo method achieved an impressive accuracy of 83.33% [91].

2.6.2 Physiological Sensors

Recent developments in physiological sensors aim to reduce invasiveness while still collecting valuable emotion-related data. Devices such as the Microsoft Band 2 and smartphones use less invasive sensors to monitor heart rate, skin temperature, and galvanic skin response (GSR) through optical heart rate monitors, accelerometers, and UV sensors. These sensors are more practical for daily use, as they are worn externally and do not require direct contact with the skin in multiple locations [98].

For example, GSR measures skin conductivity and can indicate emotional arousal, with increased conductance correlated with emotional intensity. Skin temperature sensors also provide information on emotional states, as temperature tends to increase during negative emotions such as anger and decrease during positive emotions such as calm [68]. Furthermore, heart rate and breathing rate can be monitored non-invasively, with quicker, deeper breaths often associated with negative emotions and slower breaths with positive emotions [98].

Although these less invasive devices offer improved user comfort, they are limited in their ability to capture precise emotional nuances, often requiring advanced signal processing and analysis algorithms to compensate for lower signal-to-noise ratios. For example, remote photoplethysmography (rPPG) enables heart rate monitoring without direct skin contact by analysing light reflected from the skin. Although it increases user comfort, the reduced accuracy due to environmental factors such as lighting conditions and motion noise presents a challenge [29].

Similarly, a Doppler radar-based system, used to track chest movements to extract heart rate and breathing patterns, offers non-contact emotion recognition. However, this technology is more suitable for controlled environments, as everyday movements can distort signals [13].

2.6.3 Challenges of Invasive Emotion Recognition

The most significant challenge with invasive emotion recognition technologies is the discomfort and practicality issues associated with attaching sensors to the skin or body. The use of contact-based systems, such as EEG, ECG, EOG, and EMG, creates limitations for their use in real-world applications where people need to move around, introducing noise. These technologies excel in laboratory environments where user movements can be controlled and minimised.

In addition to comfort, generalisability remains a concern. EEG signals, for example, are subject-dependent, meaning that models developed for one person may not be easily transferable to others. This limits the scalability of EEG-based systems for broader applications, particularly in healthcare or consumer devices.

Invasive emotion recognition technologies, while highly effective in controlled environments, face significant barriers to widespread adoption in real-world scenarios. Electro-based methods like EEG and ECG provide detailed biosignals for emotion classification, but their need for skin contact and susceptibility to movement interference limits their practicality. Less invasive approaches, such as physiological sensors and non-contact methods such as rPPG and Doppler radar, offer greater user comfort but often sacrifice signal quality, requiring more advanced processing techniques to maintain accuracy.

2.7 Discussion

While current multimodal emotion recognition systems exhibit promising capabilities, several gaps remain that could enhance their performance and applicability in real-world scenarios. One notable gap is the reliance on Haar Cascade for facial detection

across various papers. Although Haar Cascade has been a staple in face detection due to its simplicity and efficiency, it may not offer the highest recognition rates compared to more advanced techniques. Models such as YOLO (You Only Look Once) [80], or HOG+Linear SVM [89] could provide superior accuracy and robustness in detecting faces in diverse conditions. By exploring these alternative models, it is plausible that the subsequent emotion recognition processes will also improve, ultimately leading to more effective human-robot interactions.

Furthermore, the integration of cloud-based tools presents a significant opportunity to enhance multimodal emotion recognition systems. Current approaches utilising sentiment analysis do not fully leverage the potential of cloud computing, which can significantly offload processing tasks from local systems. By moving computationally intensive operations to the cloud, valuable resources can be freed up for other essential tasks, such as real-time interaction and response generation. Such an architecture could enhance the scalability and flexibility of emotion recognition systems, enabling them to adapt more readily to varied user interactions and environmental contexts.

By addressing these gaps, this research intends to develop a more sophisticated emotion recognition system capable of accurately interpreting human emotions in a wider range of scenarios. The continued exploration of advanced face detection techniques and the adoption of cloud-based solutions could lead to significant advancements in the reliability and effectiveness of multimodal emotion recognition, ultimately enhancing the capabilities of robotic systems in understanding and responding to human emotional states.

Chapter 3

Materials & Methods

3.1 Overview

This chapter investigates the development of a multi-modal emotion recognition system for a robot that integrates both facial expression and speech analysis to determine human emotions. The system comprises two main components: image-based emotion recognition and speech-based emotion recognition. Three face detection methods (Haar Cascades, Dlib HOG + SVM, and YOLOv4) were compared for the image-based component to identify the most effective technique for isolating faces from images. Subsequently, three convolutional neural network (CNN) architectures (VGG16, ResNet50, and MobileNetV2) were evaluated for their performance in classifying emotions from the segmented facial images. The speech-based component utilises IBM Watson's speech-to-text API to transcribe and analyse user speech for emotion detection. The integration of these components allows the robot to accurately assess emotions through visual and auditory input, providing a robust and comprehensive understanding of human emotional states.

3.2 Materials

3.2.1 Robot Platform

The TurtleBot 4 is a sophisticated and versatile robotic platform designed for research, education, and experimentation in the fields of robotics and artificial intelligence (AI). It is an evolution of the TurtleBot series, integrating advanced hardware and software components to provide enhanced functionality and performance.

Hardware Components

The Turtlebot4 is equipped with an iRobot® Create3 mobile base, based on the Roomba®, a robot vacuum cleaner. At the front of the robot is a multizone bumper equipped with seven sets of IR proximity sensors, allowing for seamless obstacle detection. The OAK-D spatial AI stereo camera enables the robot to perceive the world in a human-like manner by combining a stereo depth camera and a high-resolution colour camera with on-device Neural Network inferencing and Computer Vision capabilities.

Figure 3.1: Turtlebot 4

The robot features a Raspberry PI 4 equipped with Broadcom BCM2711, Quad-core Cortex-A72 (ARM v8) 64-bit SoC running at 1.8GHz and XGB of LPDDR4-3200 SDRAM.

The robot uses a standard Lithium Ion Battery designed for Roomba® e & i series robots. The battery onboard with the robot is a 26 Wh, 4S Lithium-Ion smart battery pack, with a nominal voltage of 14.4 V (12 V min, 16.8 V max)

The TurtleBot 4 is built on ROS, a flexible framework for developing robotic applications. ROS provides a set of tools and libraries for various tasks such as

sensor data processing, navigation, and control. The TurtleBot 4, specifically, comes equipped with ROS2 Humble with the Raspberry PI 4 running on Ubuntu 22.04.

3.2.2 Training Computer

Since training efficiency is not the focus of this project, the HP Z8 G4 Workstation is its training PC, a high-performance computing solution tailored for intensive professional tasks. The system features dual Intel Xeon Gold 6244 CPUs with 8 cores and 16 threads operating at 3.60GHz, 512 GB of Samsung ECC RAM running at 2666MT/s, and two NVIDIA Quadro RTX 8000 GPUs with 48GB of GDDR6 memory each.

The development environment was set up using Python 3.10.12 on the training PC with Ubuntu 22.04. Image processing and face detection were handled using OpenCV 4.5, the Dlib library and darknet by Alexeyab [2]. TensorFlow 2.15.1 and Keras were used to develop and train CNN models. Speech analysis was performed using the IBM Watson Speech-to-Text API.

3.3 Methods

To implement emotion detection algorithms, we will utilise various software libraries and frameworks tailored to different aspects of the task. This section outlines the methodologies and tools that will be employed for both facial detection and emotion recognition.

3.3.1 Facial Detection Algorithms

Haar Cascade

The emotion recognition system considers the Haar Cascade model as a choice for face detection. This pretrained model is easily accessible and adept at identifying faces in images. The Haar Cascade algorithm, created by Viola and Jones [93], is a well-regarded technique for object detection, with a particular emphasis on detecting faces within images.

The Haar Cascade algorithm identifies a collection of rectangular features referred to as Haar-like features. These features are basic designs that exhibit variations in pixel intensities across adjacent sections of the image. To efficiently compute these Haar-like features, the algorithm leverages an integral image representation of the input image. The integral image enables swift computation of the total sum of pixel intensities within any given rectangular area of the image.

The following step entails instructing a series of weak classifiers with the Adaboost learning algorithm. Each of these classifiers is taught to recognise a particular Haar-like feature that is indicative of the intended object, such as a face. Throughout the training process, Adaboost allocates greater importance to incorrectly classified examples, directing subsequent iterations towards rectifying these mistakes.

The Haar Cascade classifier utilises a cascade structure to organise trained weak classifiers. Sequentially arranged, each stage of the cascade consists of multiple weak classifiers. The cascade design enables early stages to swiftly reject negative examples, while positive examples proceed to subsequent stages for further evaluation. During the detection phase, the Haar Cascade algorithm utilises a sliding window approach to scan the input image. At each position of the sliding window, the algorithm applies each stage of the cascade sequentially, rapidly discarding regions of the image that are unlikely to contain the target object based on the results of earlier stages.

After going through all the stages of the cascade, the regions of the image that meet the criteria for the target object are identified as positive detections. The algorithm provides the location and size of the detected objects in the image by producing bounding boxes around these regions.

One of the key advantages of the Haar Cascade model is its efficiency and ease of implementation. It is pre-trained on an extensive dataset of labeled face images, allowing for immediate use without the need for additional training. The training set comprises 4,916 hand-labeled faces, all scaled and aligned to a base resolution of 24x24 pixels, ensuring consistency and accuracy in detection. These faces were extracted from a diverse set of images collected through a random crawl of the World Wide Web, offering robustness across various scenarios. Additionally, the model is evaluated on the MIT+CMU test set, which includes 130 images and 507 faces,

demonstrating its capability to perform well even in complex, real-world conditions. Notably, Haar Cascade is renowned for its computational efficiency, making it ideal for real-time applications, especially in resource-constrained environments where rapid face detection is crucial.

YOLO

The You Only Look Once (YOLO) model for object detection is a highly efficient and accurate approach for real-time object detection in images and videos. This model is known for its unique architecture and approach, which enable it to detect objects with remarkable accuracy. The YOLO model has gained popularity in the academic and research communities due to its exceptional performance, and it has become an important tool for various applications in computer vision and machine learning.

YOLO's heart lies its single neural network architecture that operates directly on the full image, rather than using traditional sliding window or region proposal methods. This enables YOLO to simultaneously predict bounding boxes and class probabilities for multiple objects in a single forward pass through the network. This approach eliminates the need for multiple passes and significantly reduces computational overhead, making YOLO well-suited for a robotics application where available computational resources are small.

The YOLO algorithm employs a technique where the input image is partitioned into a grid of cells. Within each cell, YOLO predicts the bounding boxes and class probabilities for objects in that cell. In particular, every grid cell is responsible for predicting several bounding boxes, whether or not objects exist within that cell. This approach ensures that YOLO preserves spatial information and can accurately identify objects of diverse sizes and aspect ratios.

The YOLO model applies a regression approach to anticipate bounding boxes, which are denoted by a series of coordinates for the corresponding grid cell. In addition, the model estimates the confidence score for each bounding box, which signifies the probability of an object being present in the box and the predicted box's

accuracy. This score considers both the objectness probability (the probability of an object being present within the bounding box) and the precision of the box's coordinates. YOLO then forecasts class probabilities for each bounding box to recognise the occurrence of specific objects within the image. [80]

YOLO has a smaller variant called Tiny-YOLO. Although they share the same underlying principles and architecture, there are notable differences between the two in terms of model size, speed, and accuracy.

Tiny YOLO is a condensed version of YOLO that prioritises speed and efficiency. Its streamlined network architecture reduces the number of layers and parameters, resulting in a smaller model size. This makes Tiny YOLO an excellent choice for real-time applications on devices with limited computational resources. While it may sacrifice some accuracy compared to its larger counterpart, Tiny YOLO still delivers competitive performance in object detection tasks. Its balance between speed and accuracy makes it well suited for a robotics application.

HOG + Linear SVM

This consists of two components combined to make a method known for its robustness and efficiency in object detection tasks, including face detection.

The Histogram of Oriented Gradients (HOG) is a feature descriptor, it focuses on the structure or shape of an object by capturing the distribution of intensity gradients or edge directions. The first step is gradient computation, typically done using a filter, such as the Sobel operator.

The image is first divided into small spatial regions called cells, such as 8x8 pixels. A histogram of gradient directions is created for each cell, with the magnitude of each gradient used to vote into the histogram bins based on the orientation. Typically, 9 bins are used, covering 0 to 180 degrees.

Histograms are usually normalised to address differences in illumination and contrast. This normalisation involves grouping the cells into larger spatial regions called blocks, for example, 2x2 cells to a block. The histograms within a block are

then concatenated to form the block descriptor. The normalisation factor is then applied and typically includes options such as L2-norm, L2-Hys, L1-norm, or L1-sqrt.

A detection window is then moved across the image at multiple scales. For each window, a HOG descriptor is calculated and used in the linear SVM.

The linear Support Vector Machine (SVM) is a type of supervised learning algorithm specifically designed for binary classification tasks. In the context of face detection, the SVM is used to distinguish between face and non-face HOG descriptors. This involves training the SVM on a dataset containing labelled examples of both faces and non-faces, each represented by its corresponding HOG descriptor. During the detection process, the HOG descriptor of each detection window is computed and then entered into the trained SVM classifier. Based on the input, the SVM generates a score that indicates the likelihood of the window being a face or a non-face. Typically, windows with scores that surpass a certain threshold are classified as faces. [24]

3.3.2 Emotion Recognition Model

After successfully detecting faces, an emotion recognition system is created using a Convolutional Neural Network (CNN) implemented in TensorFlow. The CNN model will be taught to categorise facial expressions into specific emotion categories, including happiness, sadness, anger, surprise, fear, and disgust. Using the TensorFlow framework gives a flexible and effective platform for developing, training, and implementing deep learning models.

In summary, the methodology for emotion detection will involve testing multiple facial detection algorithms, including YOLOv4, dlib, and Haar Cascade. This will be followed by the implementation of a CNN model in TensorFlow for emotion recognition. This comprehensive approach aims to develop a robust and accurate system for real-time emotion detection from visual inputs.

Three models, VGG16, ResNet50 and MobileNetV2, will be tested.

MobileNetV2, ResNet50, and VGG16

Convolutional Neural Networks (CNNs) are by far the most widely used architectures in emotion recognition research. Their ability to automatically learn hierarchical feature representations from raw image data makes them highly effective for complex tasks like facial emotion detection. However, despite their popularity, many studies in the literature do not specify the exact CNN architecture used, leaving out important details about model choice and design. This lack of transparency can make it difficult to assess and compare the performance of different approaches across datasets and applications.

The three CNN architectures considered in this work—MobileNetV2, ResNet50, and VGG16—were all used to high degrees of success in the literature. Each offers different trade-offs in terms of accuracy, efficiency, and computational requirements.

MobileNetV2 is a lightweight CNN that uses depthwise separable convolutions to reduce the number of parameters and computations, making it highly efficient for real-time applications on devices with limited resources. Its simplicity makes it ideal for mobile and embedded systems, but this efficiency comes at the cost of potentially missing more complex emotional cues in images.

ResNet50, on the other hand, uses a much deeper network with residual learning to solve the problem of vanishing gradients, allowing it to learn more detailed and hierarchical features. This makes ResNet50 highly effective for recognising subtle facial expressions, although its deep architecture increases computational demand, making it less suitable for real-time systems without powerful hardware.

VGG16, known for its simplicity and effectiveness, uses small convolutional filters (3x3) across 16 layers. It is particularly good at capturing fine-grained visual details, but its large number of parameters makes it resource-intensive, resulting in slower processing times compared to more optimised models like MobileNetV2.

In this work, all models were trained using transfer learning, a technique where a pre-trained CNN is fine-tuned on a new dataset. Transfer learning leverages the knowledge these networks have already gained from training on large-scale image datasets, such as ImageNet, to accelerate learning on smaller, domain-specific

datasets. This approach significantly reduces computational resources and training time required, while still achieving high accuracy. By reusing learned features from earlier layers and adapting them to emotion recognition, transfer learning allows these models to generalise well, even when trained on limited data specific to facial emotions.

Chapter 4

Facial Emotion Detection

This section explores the facial recognition system utilised in the multimodal emotion recognition framework. It encompasses the detailed training methodologies, performance evaluations, and datasets used in the development of facial detection and emotion classification models. The section provides an in-depth analysis of the integration of various detection algorithms, including Haar cascades, dlib, and YOLO (You Only Look Once), alongside the implementation of CNN architectures MobileNetV2, VGG16, and ResNet50 for emotion detection. The comprehensive overview aims to elucidate the effectiveness and efficiency of the system in recognising and interpreting human emotions from facial expressions.

Considering the constrained computational resources inherent in robotic systems, the approach prioritises efficiency without compromising accuracy in emotion recognition. Robots often operate in resource-constrained environments, where computational overhead must be carefully managed to ensure smooth and efficient functioning. In this robot emotion recognition system, the approach is to balance accuracy and computational efficiency. Initially, the intent is to employ a Haar cascade, dlib's HOG + linear SVM, or the YOLO algorithm to locate the face within the robot's camera feed accurately. The use of these algorithms ensures that the subsequent emotion recognition model receives the expected input of only the facial region.

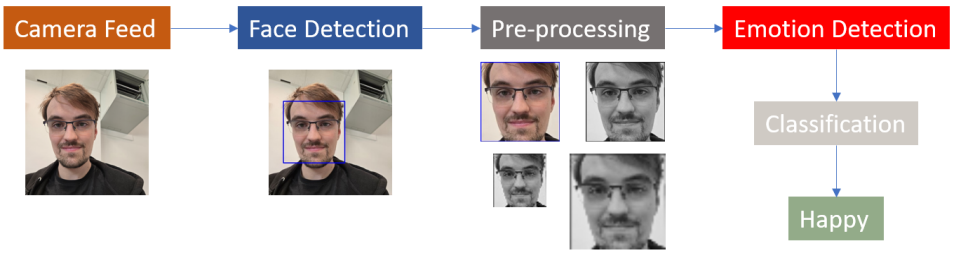


Figure 4.1: System Pipeline

4.1 Face Detection

4.1.1 Training

To effectively utilise YOLO, it is necessary to undergo training from scratch or fine-tuning on a specific dataset to suit the intended purpose. This entails collecting a vast dataset of labelled images where each object of interest is annotated with its bounding-box coordinates and class labels.

The models were trained using the recommended YOLOv4 settings from the Darknet GitHub page, by changing the config file which details the training settings: batch size set to 64, subdivisions set to 16, network size width and height both set to 416, and max_batches set to 6000. Although max_batches is typically calculated as the number of classes multiplied by 2000, which would result in 2000 for a single class, the minimum allowable value is 6000. Therefore, this value was adjusted accordingly to meet the training requirements.

Then the layers are slightly altered to match the number of classes. Each [yolo] layer has a classes section, this results in changing 3 total layers to have 1 instead of the default 80. Finally, just before each [yolo] layer is a [convolutional] layer that needs the filter settings lowering from the default 255 to (classes + 5) * 3, thus these are set to 18. Changing this setting allows the model to predict only one class.

Changing the Tiny-YOLO config follows the same process as full YOLO; however, there are only 2 [yolo] layers instead of 3. Lastly, both models have their own pre-trained weights file that was included to assist with training. The final command used in each is shown below.

See the following commands:

```
$ ./darknet detector train data/obj.data \
      yolov4_face.cfg data/yolov4.conv.137 -map -gpus 0,1
$ ./darknet detector train data/obj.data \
      yolov4_tiny_face.cfg data/yolov4-tiny.conv.29 \
          -map -gpus 0,1
```

4.1.2 Dataset

Both models are trained on the WIDER FACE [97] dataset, a well-known dataset for face recognition and detection tasks. This data set consists of a large number of images annotated with faces at various scales, positions, and levels of occlusion, making it a challenging and diverse data set that is ideal for testing object detection models.

The WIDER FACE dataset has been meticulously crafted to support face detection and recognition algorithms research. It has 12,878 images in the training set and 3,224 images in the validation set sourced from different locations, featuring an extensive assortment of scenes, lighting conditions, poses, and occlusions. This data set is uniquely annotated with precise boundary boxes around each face in every image, providing a reliable basis for training and evaluation purposes.

Each image in this dataset has been annotated with one or more bounding boxes, which capture information about each face's position, orientation, and scale. These annotations effectively encompass the diversity of face sizes and poses in different images, including occluded faces. This data set is exceptional in its capacity to evaluate the robustness of face detection algorithms under challenging conditions.

The WIDER FACE dataset is distinguished by its rich diversity in subjects, scenes, and environmental conditions. It contains images captured in different settings, such as indoor and outdoor scenes, crowded environments, street scenes, and surveillance footage. It features faces of numerous genders, ages, ethnicities, and facial expressions, ensuring a comprehensive representation of real-world scenarios.

An evaluation of YOLO and Tiny YOLO's performance will be done on both the complete dataset and a couple chosen subsets. One subset comprises images

containing only one face per image, providing a more focused and simplified training scenario. This approach allows for targeted experimentation and analysis, enabling insights into the model's performance and robustness in scenarios with minimal occlusions or distractions from multiple faces. The resulting size of this dataset is 1,342 images in the training set and 334 images in the validation set. The other subset contains all the images that featured more than one face; this should reveal if the models have any issues with higher complexity images. This set has 8245 in the training set and 2104 images in the validation set.

4.1.3 Performance

To ensure the effectiveness of the YOLO model in detecting faces for subsequent emotion recognition tasks, its performance is evaluated using a variety of metrics. These metrics offer a comprehensive view of the accuracy, speed, and robustness of the model.

In this section, the performance of the YOLO and Tiny YOLO object detection models was evaluated on a face detection dataset: the Wider Face dataset. This dataset was then split into images containing multiple faces, and images containing only one face. The metrics used for the evaluation include precision, recall, F1 score, average Intersection over Union (IoU), and Average Precision (AP).

Precision measures the accuracy of positive predictions. It is calculated as the ratio of true positives (correctly detected faces) to the sum of true positives and false positives (incorrect detections). High precision means that most of the faces detected by the model are actual faces.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall, or sensitivity, measures the ability of the model to find all relevant instances. It is the ratio of true positives to the sum of true positives and false negatives (missed detections). High recall means that the model can detect most if not all of the faces present in a given image.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 score is the harmonic mean of precision and recall, providing a single metric to evaluate the model's overall performance. It balances the trade-off between precision and recall and is especially useful as an evaluation metric in binary classification.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Intersection over Union measures the overlap between the predicted bounding box and the ground truth bounding box. It is calculated by dividing the overlap area by the union area between the two boxes. A higher IoU means the predicted bounding box closely matches the actual bounding box.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Average Precision is a weighted mean of precision scores, where the weight is the increase in recall from the previous threshold. A high AP means that a model has a low false negative and false positive rate.

Table 4.1: Performance of YOLO on Full Wider Face and Single Face Datasets

Metric	Full Wider Face	Multi Face	Single Face
Precision	0.61	0.61	0.95
Recall	0.64	0.63	0.92
F1 Score	0.63	0.62	0.94
Average IoU	45.77%	45.12%	80.90%
AP	63.17%	62.11%	97.39%

Table 4.2: Performance of Tiny YOLO on Full Wider Face and Single Face Datasets

Metric	Full Wider Face	Multi Face	Single Face
Precision	0.48	0.47	0.95
Recall	0.44	0.43	0.91
F1 Score	0.46	0.45	0.93
Average IoU	35.32%	34.46%	78.08%
AP	36.63%	35.04%	92.33%

Tables 4.1 and 4.2 summarise the performance of the YOLO and Tiny YOLO models, respectively, on the 3 datasets.

The results in the tables show an interesting development. Whilst the full YOLO model performs markedly better when there are multiple faces in the image compared to its Tiny counterpart, when it comes to detections where only a single face is present in the image, there is no discernible difference between the two except a slight reduction in the average IoU and average precision. The results on the version of the dataset with only multiple faces backs up the hypothesis that Tiny YOLO struggles with the higher complexity images. Thus, in this system's intended use case (a human-robot 1-on-1), Tiny YOLO is more than sufficient to perform face detection.

In addition to YOLO and Tiny YOLO, the performance of Haar Cascade and dlib's HOG+Linear SVM was evaluated on the same datasets. However since these two models do not provide confidence scores along with their predictions, the metric AP cannot be calculated. AP is derived from the precision-recall curve, where predictions are ranked by their confidence scores. The curve is created by varying a confidence threshold, recalculating precision and recall at different confidence levels. Without confidence scores, every prediction is treated equally, leading to an uninformative or flat curve, which won't reflect the model's ability to prioritise better predictions.

Table 4.3: Performance of Haar Cascade on Full Wider Face and Single Face Datasets

Metric	**Full Wider Face**	**Multi Face**	**Single Face**
Precision	0.69	0.74	0.45
Recall	0.15	0.14	0.69
F1 Score	0.25	0.24	0.55
Average IoU	69.64%	69.63%	69.36%

Tables 4.3 and 4.4 present the performance of the Haar Cascade and HOG + Linear SVM models, respectively, on the full Wider Face dataset and a single face dataset.

These results indicate that both the Haar Cascade and HOG+Linear SVM models, unlike YOLO and Tiny YOLO, perform better on the more complex detection task

Table 4.4: Performance of HOG+Linear SVM on Full Wider Face and Single Face Datasets

Metric	Full Wider Face	Multi Face	Single Face
Precision	0.96	0.97	0.94
Recall	0.14	0.12	0.78
F1 Score	0.24	0.22	0.86
Average IoU	69.61%	69.91%	66.35%

of the Full Wider face dataset. The high precision but low recall of both models on the full Wider Face dataset and the multiple face subset suggests that they are more conservative in detections, leading to fewer false positives but more false negatives. These initial tests were conducted on the high-powered training machine, clearly highlighting the difference in model complexity by the time taken to complete the tests on the validation sets, the Haar Cascade took a total of 295.36 seconds. In comparison, the HOG+Linear SVM took 1402.19 seconds.

4.2 Emotion Detection

4.2.1 Datasets

Training

The datasets selected for emotion detection training were FERPlus [12] and an altered version of CK+ [55]. Both datasets are downloaded from publicly available databases online. FERPlus and CK+ both contain images of unconstrained facial expressions. A sample of images from FERPlus and CK+ are shown in figure 4.2.

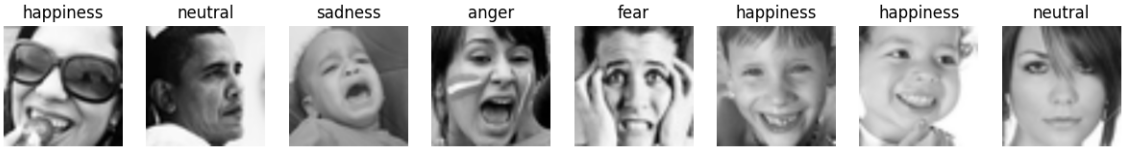


Figure 4.2: Sample of images from the combined dataset of FER+ and CK+

The FERPlus database is a derivative of the original FER2013 dataset. FER2013 suffered from several issues that made the dataset difficult to perform recognition tasks on, these issues include non-face data and false labels resulting in even state-

of-the-art models reaching around 60%-70% accuracy. FERPlus results from 10 crowd-sourced taggers relabelling the original dataset to provide better quality ground truth for still image emotion than the original FER labels. The taggers categorised each image into one of 10 different labels: 8 emotions (happiness, neutral, sadness, surprise, fear, disgust, contempt, and anger) and 2 additional categories representing cases where the emotion is indeterminate or the image does not contain a human face (unknown and non-face). To remove uncertain images and filter each image into a single labelled folder, a maximum voting method was implemented. This results in 28,386 samples in the training set, 3546 in the private test set and 3,553 in the public test set. 4.5 indicates the number of images in each emotion category.

Table 4.5: Emotion distribution of the training dataset

Emotion	PrivateTest	PublicTest	Training	Total
Anger	325	319	2,466	3,110
Contempt	27	24	165	216
Disgust	23	34	191	248
Fear	93	74	652	819
Happiness	928	899	7,528	9,355
Neutral	1,262	1,335	10,308	12,905
Sadness	444	412	3,514	4,370
Surprise	444	456	3,562	4,462
Total	3,546	3,553	28,386	35,485

The dataset is heavily skewed towards the 'Happiness' (9,355 images) and 'Neutral' (12,905 images) categories. These categories collectively represent over 62% of the entire dataset. Such a high representation can result in a bias or overfitting [78] towards these emotions, reducing its sensitivity and accuracy in detecting less represented emotions.

However, emotions such as 'Contempt' (216 images), 'Disgust' (248 images), and 'Fear' (819 images) are significantly under-represented. The model might not learn to recognise these emotions effectively due to the limited examples available, leading to poor performance in real-world scenarios where these emotions are present. Emotions like 'Anger' (3,110 images), 'Sadness' (4,370 images), and 'Surprise' (4,462 images) have a moderate number of samples but still far less compared to 'Happiness' and

'Neutral'. Although better than the least represented categories, the imbalance still poses a risk of the model underperforming in these emotions.

CK+ or Extended Cohn-Kanade dataset consists of 593 video sequences from 123 subjects ranging from 18 to 50 years old. Gender and heritage of participants varied with 69% female, 81%, Euro-American, 13% Afro-American, and 6% other groups. Each video shows a facial shift from a neutral expression to a targeted peak expression, recorded at 30 frames per second (FPS) at a resolution of 640x480 pixels. These videos are labelled into one of seven emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. This data needs to be in a format similar to the FERPlus dataset, 48x48 in greyscale. On a database-sharing website, Kaggle, there is a version of CK+ in this format, below is the information provided alongside the database:

- Contains adapted data up to 920 images from 920 original CK+ dataset
- Data is already reshaped to 48x48 pixels, in grayscale format and facecropped using haarcascade_frontalface_default.
- Noisy (based on room light/hair format/skin colour) images were adapted to be clearly identified using Haar classifier.
- Columns from file are defined as emotion/pixels/Usage

The resulting number of images in each emotion category is detailed in 4.6.

Table 4.6: Image counts for each emotion in CK+

Emotion	**Count**
Anger	135
Contempt	54
Disgust	177
Fear	75
Happiness	207
Sadness	84
Surprise	249

Preprocessing

To ensure compatibility and consistency across VGG16, ResNet50, and MobileNetV2 models, the FER and CK+ datasets undergo specific preprocessing steps.

Firstly, all images are resized from 48x48 to 224x224 pixels. This resizing is necessary because, while VGG16 and MobileNetV2 can be adjusted to accept 48x48 images, ResNet50 could not process the images at 48x48 and required a minimum size of 224x224. Resizing all images to 224x224 ensures uniformity across all models.

Secondly, the images, originally in greyscale, need to be converted to have three channels as required by the models. This is achieved using OpenCV to convert single-channel greyscale images to three-channel images using cv2.cvtColor() with the constant cv2.COLOR_GRAY2BGR.

Lastly, normalisation is specifically required for VGG16. The images are reshaped for normalisation using the StandardScaler from the sklearn.preprocessing Python library and then reshaped back to its original dimensions. This normalisation step ensures that the input data is standardised, which is crucial for the performance of VGG16.

In addition to these pre-processing steps, data augmentation techniques are applied to enhance the robustness of the models, prevent overfitting, and alleviate the issue of the unbalanced data set. Data augmentation involves artificially increasing the size of the training dataset by generating new training samples from the original data. This can be done through geometric transformations such as width and height shifts, horizontal flips, and zooming as well as many other techniques such as GAN (General Adversarial Networks) and Photometric Transformations [88]. These transformations help the models generalise better by exposing them to various image conditions and distortions they might encounter in real-world scenarios.

Specifically, the following augmentations are applied:

- Width and Height Shifts: Images are randomly shifted horizontally and vertically by up to 10% of the image width and height (width_shift_range = 0.1 and height_shift_range = 0.1).

- Horizontal Flip: Images are randomly flipped horizontally to simulate different viewing angles (horizontal_flip = True).
- Zoom: Random zooms in and out within a range of 0.8 to 1.2 times the original size are applied (zoom_range = 0.2).

These augmentations are performed using the ImageDataGenerator class from the Keras library, which allows for real-time data augmentation during the training process. By applying these augmentations, the diversity of the training data is significantly increased.

4.2.2 Training

After processing the base convolutional layers of each pretrained model (VGG16, ResNet50, and MobileNetV2), the feature maps were flattened using the flatten() function.

The output layer of each model was configured to match the number of emotion classes in the dataset. This layer used a dense layer with the number of neurones equal to the number of classes and a softmax activation function to provide a probability distribution as the resulting prediction.

The models were then trained using the Adam optimiser with an initial learning rate of 0.001 and a batch size of 64. The models were trained for 50 epochs. The categorical cross-entropy loss function was used to optimise the model.

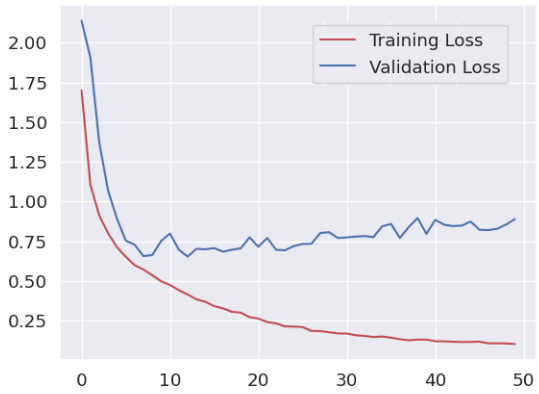


Figure 4.3: The loss graph for the first successful training of MobileNetV2

The graph 4.3 illustrates the training and validation loss for the MobileNetV2 model. As training progresses, both losses decrease sharply, demonstrating that the model is learning from the data. Around the 20-epoch mark, the training loss continues to decline steadily, indicating that the model is fitting well to the training data. The validation loss begins to see a slight upward trend after around 11 epochs suggesting that the model is overfitting as the training loss continues to decrease.

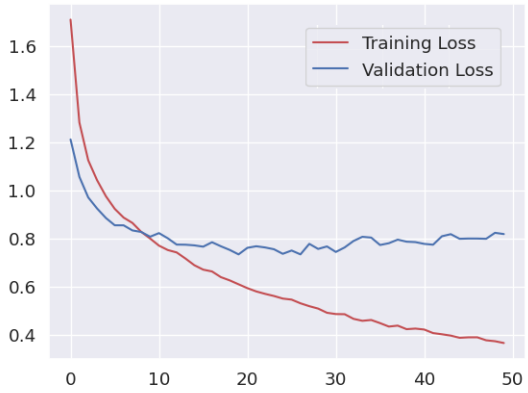


Figure 4.4: The loss graph for the first successful training of ResNet50

Graph 4.4 shows the training and validation loss for the ResNet50 model. Similar to MobileNetV2, both losses start high and decrease significantly in the early epochs. The training loss for ResNet50 drops more quickly and smoothly compared to the validation loss, reaching a much lower value as epochs progress. The validation loss shows a decreasing trend but with more pronounced fluctuations, indicating some instability in performance on the validation set. After around 30 epochs the validation loss starts a slight upward trend. By the end of the 50 epochs, the training loss is significantly lower than the validation loss, which might suggest slight overfitting.

Figure 4.5: The loss graph for the first successful training of VGG16

Finally, graph 4.5 represents the training and validation loss for the VGG16 model. Both losses start high and decrease rapidly in the initial epochs, similarly to the other models. However, the training loss for VGG16 continues to decrease more steeply and steadily, reaching very low values, indicating a strong fitting to the training data. The validation loss decreases initially but starts to exhibit more fluctuation and even an upward trend after around 11 epochs. This divergence between training and validation loss suggests that VGG16 might be overfitting to the training data, capturing noise and details that do not generalise well to the validation set.

To further mitigate the impact of overfitting in the three emotion recognition models a few more techniques were added. Firstly an early stopper was added, this, with a patience set at 10, stops the training of the model if no improvements are made after 10 epochs of training and a checkpointer that will restore the model to the best weights. Alongside this, a reduced learning rate was implemented that lowers the learning rate if the training starts to hit a plateau in accuracy.

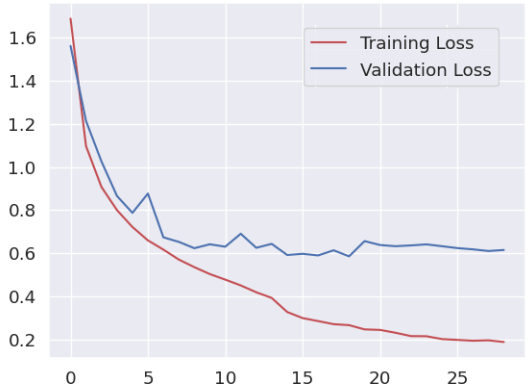


Figure 4.6: The loss graph for the second successful training of MobileNetV2

MobileNetV2s second training loss graph is shown in figure 4.6. The graph shows that the training only got to 29 epochs before the early stopper function stopped it. The application of techniques to prevent overfitting seems effective. The gap between training and validation loss is relatively small. The model continues to improve on both training and validation data, indicating that it is learning useful patterns rather than just memorising the training data. The stabilisation of the validation loss suggests that the model has reached a point where further training may yield diminishing returns.

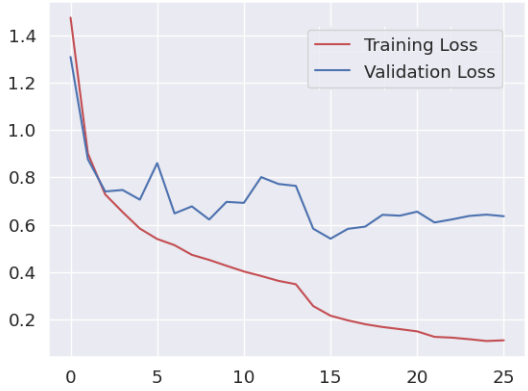


Figure 4.7: The loss graph for the second successful training of ResNet50

ResNet50's second training results in a similar graph to its first run. After only 5 epochs validation loss begins to fluctuate increasing and decreasing until around epochs 15 to 25, where the validation loss shows a slight upward trend.

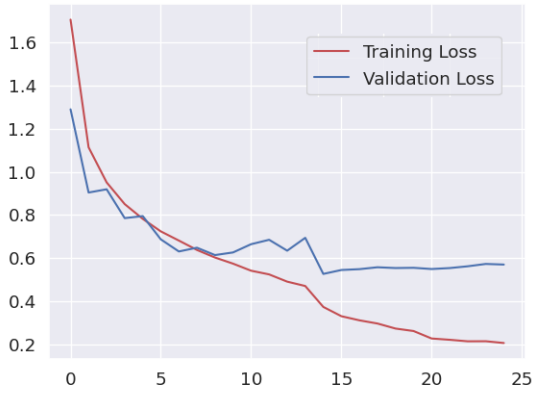


Figure 4.8: The loss graph for the second successful training of VGG16

In the second training run of VGG16 after about 5 epochs, the validation loss begins to fluctuate, while the training loss continues to decrease steadily. From Epochs 10 to 25, the validation loss shows a slight downward trend with occasional fluctuations, indicating potential minor overfitting. However, the training loss continues to decrease smoothly, suggesting that the model is still learning effectively. Overall, the model demonstrates good generalisation, and the techniques applied seem to mitigate the severe overfitting it saw in the first run.

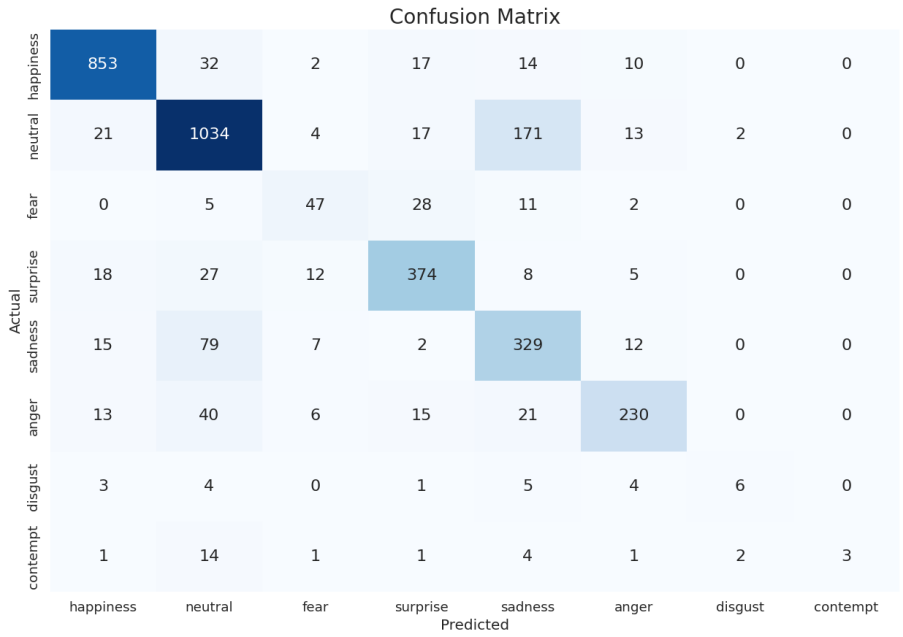


Figure 4.9: The confusion matrix detailing the performance of MobileNetV2 on the PrivateTest set

The confusion matrix 4.9 illustrates the performance of MobileNetV2 across the 8 emotions. The model accurately classifies the 'happiness' and 'neutral' expressions, with 853 and 1034 correct predictions, respectively. However, it struggles with 'fear' and 'contempt', frequently misclassifying them as other emotions. There is notable confusion between 'sadness' and 'neutral' with it incorrectly classifying them as each other.

Figure 4.10 shows an example of images from sadness and neutral. The confusion could be attributed to the subtle facial differences between 'sadness' and 'neutral' expressions. Both emotions tend to exhibit minimal facial muscle movement, and the lack of exaggerated features such as smiles or frowns can make it challenging for models to distinguish between them.

Figure 4.10: Example images showing the very slight variation between sadness and neutral

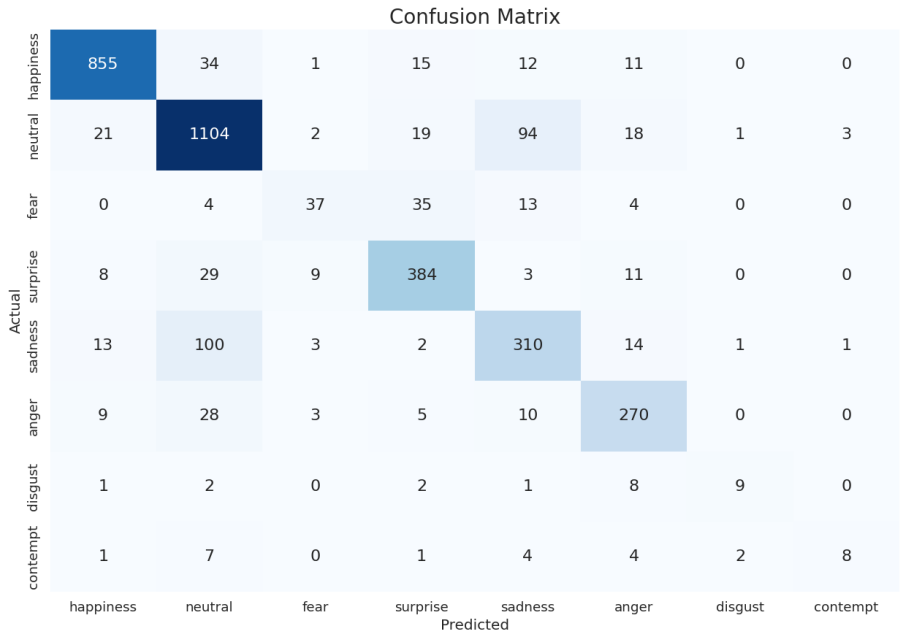


Figure 4.11: The confusion matrix detailing the performance of ResNet50 on the PrivateTest set

ResNet50 performed very similarly to MobileNetV2 but had a slightly higher recognition rate for each emotion except 'fear' and 'sadness'. It suffers from the same misclassification of 'sadness' and 'neutral' as MobileNetV2.

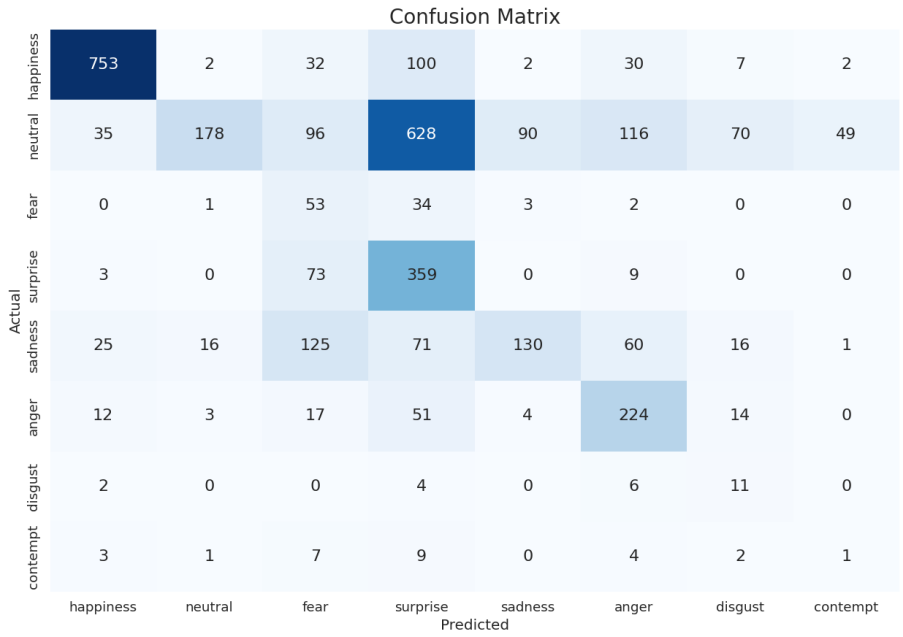


Figure 4.12: The confusion matrix detailing the performance of VGG16 on the PrivateTest set

VGG16 misclassified most of the 'neutral' pictures, only getting 178 correct, mainly classifying them as 'surprise' and 'anger'. However, VGG16 achieved the highest results of the three models in the fear category. The model also struggles to differentiate between 'sadness' and 'fear'. Overall, the model does perform well; however, in comparison to ResNet50 and MobileNetV2, it misclassifies too many emotions to be considered reliable.

4.2.3 Testing

This section evaluates the performance of three emotion recognition models in conjunction with face detection algorithms: Haar cascades, dlib, Tiny YOLO and YOLO. The evaluation is conducted using the Expression in-the-Wild (ExpW) dataset, which contains facial images captured in diverse and unconstrained environments. This comprehensive testing aims to assess the robustness and accuracy of the models and algorithms in recognising emotions under varied and challenging scenarios.

A selection of random images from ExpW can be seen in 4.13.

Figure 4.13: A random selection of images from the ExpW dataset

The ExpW dataset consists of 106,962 images, almost all (91,793) are annotated with the coordinates of the face present (some images contain multiple faces, however, only one face is annotated), and every annotated face shows one emotion out of happiness, neutral, sadness, surprise, fear, disgust, and anger. Images that are not annotated are not included in any of the following tests.

To determine the optimal combination of face detection and emotion detection algorithms for use in a resource-constrained robotic system, a comprehensive test was performed. This test involved pairing each face detection algorithm with each emotion detection algorithm to evaluate their performance. The primary goal was to find the best performing combination in terms of speed and accuracy for real-time applications on the robot. Each face detection and emotion detection instance was measured to calculate the average time taken for detection and prediction. Successful detections of faces and correct emotion predictions were meticulously recorded and compared to the actual emotions presented in the images of the data set.

Table 4.7: Average Detection Times in Milliseconds for Face and Emotion Detection Algorithms

Algorithm	Algorithm Name	Avg. Inf Time (ms)	Model Size (MB)
Face	Tiny YOLO	25.0	22.4
	Haar Cascade	40.7	1.19
	dlib	45.6	0.696
	YOLO	161.0	244
Emotion	MobileNetV2	13.6	69.8
	ResNet50	95.3	187
	VGG16	312.2	80.8

Table 4.8: Accuracy and Number of Face Detection for Model Combinations, out of a possible 91,793 faces

Model Combination	Accuracy	No. of Face Detections
dlib + MobileNetV2	37.06%	56098
dlib + ResNet50	37.77%	56098
dlib + VGG16	34.52%	56098
Haar + MobileNetV2	46.63%	76416
Haar + ResNet50	47.81%	76416
Haar + VGG16	44.52%	76416
Tiny YOLO + MobileNetV2	55.93%	88332
Tiny YOLO + ResNet50	57.36%	88332
Tiny YOLO + VGG16	51.94%	88332
YOLO + MobileNetV2	57.54%	90773
YOLO + ResNet50	59.02%	90773
YOLO + VGG16	53.43%	90773

Since only one face in each image is annotated, all faces detectable in the image are compared to the one in the labels file, and the detected face that is closest (using Euclidian distance) to the listed face is considered the valid face for further analysis.

Tiny YOLO and YOLO, as face detection methods, provide the highest number of face detections, leading to improved overall accuracy when combined with emotion recognition models. Specifically, YOLO paired with ResNet50 achieves the highest accuracy in emotion recognition (59.02%). Haar Cascade and dlib both achieved a lower number of face detections (resulting in lower overall accuracy) and a slower detection speed than Tiny YOLO which achieved a detection time of 0.0250 seconds on average. YOLO does provide a higher accuracy when it comes to the number of face detections and as a result, a higher overall accuracy; however, this increase in accuracy comes at the downside of 0.1610 seconds per face detection, more than 6 times slower than its Tiny counterpart for an overall accuracy increase of only 1.66%. Thus, Tiny YOLO is the clear choice for facial recognition models. Among the emotion recognition models, ResNet50 generally outperforms MobileNetV2 and VGG16 across all face detection methods; however, considering the difference in detection times for ResNet50 (95.3 milliseconds) and MobileNetV2 (13.6 milliseconds) the choice is not as obvious and both could be left as suitable options depending on the need for higher accuracy or higher detection speed. Another consideration in a

resource-constrained environment is the size of the models. ResNet50 has the largest model size at 187MB leaving less memory for the robots other processes to work with. This leaves the best combination for overall accuracy, memory efficiency, and speed to be Tiny YOLO + MobileNetV2 with it only requiring 92.2 MB of memory for both models.

Finally, the performance of the model is evaluated directly on a robot. The system is designed to be standalone and operate without relying on a connected PC, so testing the models in this context is essential.

Table 4.9: Average Detection Times in Milliseconds for Face and Emotion Detection Algorithms performed on the TurtleBot4

Algorithm Type	**Algorithm Name**	**Average Detection Time (ms)**
Face Detection	Tiny YOLO	697.4
	Haar Cascade	491.7
	dlib	216.2
	YOLO	6846.5
Emotion Detection	MobileNetV2	124.7
	ResNet50	1334.3
	VGG16	5384.2

The results of the Turtlebot4 tests are summarised in the table above. It should be noted that the performance of face detection methods exhibits a significant reversal when implemented on a robot with limited resources. Surprisingly, Tiny YOLO, which was initially the fastest, was surpassed by both dlib and Haar Cascade. Even Haar Cascade, was outperformed by dlib by a considerable margin. This behaviour could be due to a couple things, the Turtlebot 4 does not possess any GPUs unlike the training PC and it is possible that YOLO is using GPU acceleration to improve detection times. The other possibility is that YOLO leverages the vastly increased number of CPU cores on the training PC to perform parallel computations more efficiently, distributing the workload across multiple cores. However, when deployed on the Turtlebot4, which has significantly fewer cores, YOLO's speed advantage diminishes, leading to slower inference times compared to simpler models like dlib and Haar Cascade.

4.2 Emotion Detection

The results for the emotion models remained consistent with the original tests conducted on the training PC. Among these models, MobileNetV2 stands out as the fastest and most accurate choice for direct implementation on a robot. However, the selection of a face detection model is more nuanced. For maximum accuracy, Tiny YOLO is the preferred option, closely trailing full YOLO in accuracy, but with each detection taking less than a second as opposed to 6.8 seconds. If rapid detections are a priority, dlib is a clear choice, while Haar Cascade provides a balanced compromise between accuracy and inference speed.

Chapter 5

Audio Emotion Detection

This chapter examines the audio emotion recognition component of the multimodal framework, focusing on the use of IBM Watson's capabilities. The analysis includes performance tests to assess the accuracy and speed of IBM Watson in detecting emotions from audio input. Additionally, this chapter discusses the role of large language models (LLMs) in the context of emotion recognition and the limitations that prevent the use of OpenSMILE for this project. Through a detailed evaluation, this chapter aims to provide insight into the effectiveness of IBM Watson as an audio emotion recognition tool and the considerations involved in choosing appropriate technologies for audio analysis.

5.1 IBM Watson

Given the current computational demands of core functionalities and concurrent facial emotion recognition processes, it would be beneficial to consider offloading speech emotion recognition to an external entity. Using cloud-based solutions, such as IBM Watson's API, presents an appealing option. By interfacing with Watson, recorded human speech can be remotely processed, allowing emotional predictions based on textual analysis. This approach not only reduces the computational burden on the robot, but also harnesses the advanced emotion analysis capabilities offered by cloud services.

5.1 IBM Watson

IBM Watson is a cognitive computing platform developed by IBM that uses artificial intelligence (AI) techniques to analyse and interpret large amounts of data. It includes a range of AI-powered services and tools designed to help businesses gain insights, make informed decisions, and improve user experiences in various industries. Watson's capabilities include natural language processing, machine learning, and data analytics, making it a versatile solution for addressing complex challenges.

One useful feature of IBM Watson is its conversational abilities, which allow for structured dialogue between the robot and the user. By integrating Watson's Conversation service, the robot can engage in structured conversations with users, responding to prompts and queries based on predefined conversation trees. This approach allows the robot to guide the conversation along predetermined paths, collecting specific information, or addressing user inquiries within predefined topics.

Ultimately, the choice to use cloud-based emotion recognition is a strategic decision that weighs computational efficiency against the aim of developing a flexible and emotionally intelligent robotic system. By tapping into external resources, we not only enhance the robot's performance but also pave the way for integrating state-of-the-art emotion analysis capabilities into the HRI framework, thereby enhancing the user experience and pushing the boundaries of human-robot interaction. [48]

Initially, a system was created to incorporate IBM Watson as a chatbot to inform people about ongoing public health issues. The goal was to provide accurate and up-to-date responses to common questions and to give people peace of mind. The system was designed to use IBM's sentiment analysis to detect fear and point people to resources that could help them. In addition, the chatbot provided the ability to connect an Aldebaran robot, which provided a physical presence that people could interact with. The robot would use its microphones to pick up user speech, which could then be sent to IBM for analysis, after which the robot would respond with what IBM Watson sent back.

This system also leveraged IBM Watson's text-to-speech capabilities, allowing users to fully customise the generated voice based on a variety of parameters. In addition to selecting the gender of the voice, users could choose accents from different regions, making the interaction more personalised and culturally relevant. This

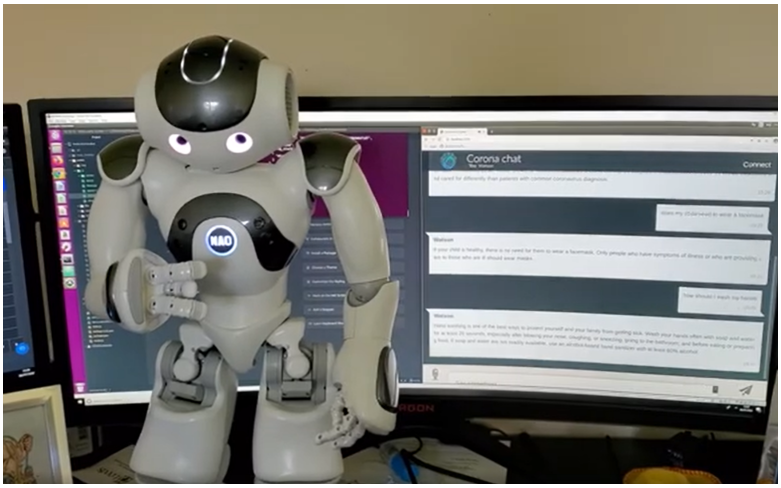


Figure 5.1: Aldebaran robot Nao with IBM Watson ChatBot

also allows adjustments to pitch, enabling a higher or lower tone depending on user preference or specific application needs.

5.2 LLM

It is important to note that while IBM Watson's Conversation service provides a structured approach to dialogue management, it may lack the spontaneity and flexibility of natural human conversation. Relying on conversation trees imposes constraints on the flow of interaction, limiting the opportunity for open-ended dialogue and real-time adaptation to user input. As a result, interactions with the robot may feel scripted or constrained, potentially detracting from the overall user experience in certain situations.

To achieve a natural and engaging human-robot interaction, it is imperative to develop a comprehensive system that integrates a Large Language Model (LLM). ChatGPT, a state-of-the-art language model, plays a crucial role in enabling seamless human-like conversations between the robot and the user [71]. Its ability to generate contextually relevant responses allows for a more natural dialogue exchange that closely resembles human conversation patterns.

By incorporating ChatGPT into this system, we can create a more interactive and emotionally responsive dialogue experience. In this setup, the speech-emotion

recognition system handles the analysis of user speech to detect emotions such as happiness, sadness, anger, or neutrality. These detected emotional states can then be communicated to ChatGPT alongside the text of the user's speech. This allows ChatGPT to factor in both the content of the conversation and the emotional context provided by the speech-emotion recognition system, helping it generate more empathetic and contextually appropriate responses.

For example, if the speech-emotion recognition system detects frustration in the user's voice, this emotional information can be fed to ChatGPT, allowing it to adapt its responses in real time to address the user's emotional state more sensitively. This synergy allows for a deeper and more emotionally aware interaction, where ChatGPT can tailor the flow of the conversation based on both the user's words and their emotional tone.

Additionally, ChatGPT can use the emotional feedback from the speech-emotion system to adjust the direction of the conversation, perhaps steering toward topics that might alleviate negative emotions or enhance positive ones. This makes it possible to create a more engaging and emotionally intelligent interaction, where the robot can respond in a way that feels more human and responsive to the user's mood. Using ChatGPT for dialogue and the emotion recognition system for emotional analysis, we enable a hybrid approach where each component focuses on its strengths, resulting in a more robust and user-centric interaction.

Thus, ChatGPT was integrated into the system. This allows users to interact with ChatGPT seamlessly through a browser, making it accessible from virtually any device, whether a desktop, laptop, or mobile device. This integration ensures that users can engage with the system without the need for specialised software or hardware, broadening its utility and accessibility. The Web Messenger supports both text- and speech-based interactions allowing the user to converse with the robot in a natural way.

One of the core strengths of this system lies in how it expands ChatGPT's capabilities, making it a far more versatile assistant. Using the function-calling mechanism, ChatGPT can now fetch real-time data such as weather reports, time, and date, or even retrieve specific data from databases. This transforms it from being

a static question-answering system to an interactive, real-time assistant. Moreover, the system has been designed to allow future scalability, enabling developers to integrate additional functions based on evolving user needs, such as connecting to more advanced AI models, adding new APIs, or enhancing its conversational context-awareness.

The system could also leverage locally run language models, such as GPT4All, to enhance its natural language understanding and response capabilities. Running these models locally ensures full control over data privacy and security. This approach allows for greater flexibility, as the models can be fine-tuned to better suit the system's specific needs without reliance on external cloud services. In addition, the system can operate independently of an internet connection, making it more reliable in environments with limited or unstable connectivity. This setup offers both customisation options and scalability, ensuring robust performance for complex, language-driven tasks.

5.3 OpenSMILE

OpenSMILE [32], which stands for "Open-Source Speech and Music Interpretation by Large-Space Extraction", is a powerful open-source toolkit widely used in audio signal processing. Its primary function is to extract an extensive range of acoustic features from audio signals, providing a versatile platform for various applications that include speech recognition, emotion recognition, speaker identification, and music analysis. One of the key strengths of OpenSMILE lies in its modular architecture, which allows the customisation of the feature extraction process to suit specific requirements. This modularity is achieved through a collection of feature extraction components known as "functionals," each responsible for computing a particular set of features. It is possible to choose from a rich library of functionals and combine them as needed to create tailored feature sets.

Moreover, OpenSMILE is designed for real-time processing of audio streams, making it suitable for applications that demand low-latency feature extraction, such as real-time speech recognition systems or interactive multimedia applications. Its

cross-platform compatibility ensures that it can seamlessly integrate into various environments running on major operating systems, including Windows, macOS, and Linux. Additionally, the toolkit offers extensive configuration options that allow one to specify parameters such as frame size, overlap, and feature selection, thus providing flexibility to adapt to various audio processing tasks.

OpenSMILE facilitates the integration of extracted features with machine learning algorithms, serving as a crucial preprocessing step for tasks such as classification. The features computed by OpenSMILE capture essential characteristics of audio signals, enabling accurate modelling and interpretation of audio data. However, given the limited resources available on a robotic platform, it could run into performance issues that severely limit its capabilities. The memory requirements of OpenSMILE can also be significant, particularly when extracting a large number of features from lengthy audio streams, with each feature set taking up to 100MB for a short 18-second audio clip. Robotics platforms typically have limited memory capacity, and allocating resources to OpenSMILE may strain the system, potentially impacting overall system stability and reliability. This limitation is purely hardware based and future robots that can afford more powerful systems would be able to utilise OpenSMILE feature extraction plus a classification model to determine emotions.

5.4 IBM Waston Performance

In this section, the performance of IBM Watson's Natural Language Understanding (NLU) service is evaluated in analysing the emotional content of various phrases. To ensure a robust assessment, each phrase was tested five times, and response times were recorded. The table 5.2 presents the response times (in seconds) for each test run with ten different phrases shown in 5.1.

Table 5.1: Phrases and their expected emotions

Text	Expected Emotion
I am so happy today! Everything is going great.	Joy
I am very sad and disappointed by the news.	Sadness
I am so angry at the situation!	Anger
This is so scary and frightening.	Fear
I am just so disgusted by what happened.	Disgust
The sun is shining and the birds are singing. It's a beautiful day to be alive. I feel so grateful for all the wonderful things in my life. I have a loving family, great friends, and a job that I am passionate about. Days like today make me feel like all the hard work has paid off and I can truly appreciate the beauty of life.	Joy
Today I received some heartbreaking news. A dear friend of mine passed away unexpectedly. The shock and sorrow I feel are overwhelming. We had so many plans together, so many dreams left unfulfilled. It's hard to imagine life without them. This loss leaves a void that can never be filled.	Sadness
I am furious about the latest policy changes at work. They were implemented without any consultation with the staff, and they make our jobs much harder. It feels like management doesn't care about our well-being or input. This kind of disregard is unacceptable, and I won't stand for it.	Anger
Walking through the dark alley, I could feel my heart racing. Every sound seemed amplified, and the shadows looked like they were moving. I couldn't shake the feeling that someone was following me. It was one of the most terrifying experiences I've ever had. I just wanted to get out of there as quickly as possible.	Fear
The food at that restaurant was absolutely disgusting. The meat was undercooked, the vegetables were soggy, and there was a strange smell coming from the kitchen. I felt nauseous just being there. It's unacceptable to serve such poor quality food to customers.	Disgust

Table 5.2: Test results for IBM Watson's response times on 10 phrases across 5 runs in seconds, alongside the average time and standard deviation. The phrases in this table match the phrases in table 5.1 in order.

Test number	Run 1	Run 2	Run 3	Run 4	Run 5	Average	SD
Phrase 1	1.81	2.07	1.36	1.58	1.45	1.654	0.257
Phrase 2	0.69	0.51	0.38	0.40	0.36	0.468	0.123
Phrase 3	0.43	0.47	0.39	0.36	0.40	0.41	0.037
Phrase 4	0.69	0.63	0.38	0.39	0.43	0.504	0.130
Phrase 5	0.38	0.38	0.40	1.23	0.48	0.574	0.330
Phrase 6	0.47	0.40	0.60	0.52	0.43	0.484	0.071
Phrase 7	0.56	0.43	0.40	0.61	0.41	0.482	0.086
Phrase 8	0.43	0.42	0.39	0.44	0.39	0.414	0.021
Phrase 9	0.47	0.41	0.40	0.40	0.45	0.426	0.029
Phrase 10	0.42	0.40	0.36	0.45	0.39	0.404	0.030

IBM Watson NLU demonstrates efficient and consistent performance in emotion analysis, with response times typically under one second for any given phrase. However, there is a noticeable delay for the first phrase of each session, likely due to the system establishing an initial connection to IBM Watson. To investigate this, five additional tests were conducted with the phrases in reverse order. The first, more complex, phrase took an average of 1.3601 seconds to process, while subsequent phrases averaged just 0.3713 seconds. This indicates that the first phrase consistently experiences a longer response time. To optimise performance, it would be beneficial for the program to send a throwaway phrase to minimise delays for subsequent inputs.

Table 5.3: The resulting output probability for each emotion for each phrase. The phrases in this table match the phrases in table 5.1 in order.

Test number	Sadness	Joy	Fear	Disgust	Anger
Phrase 1	0.025	0.983	0.008	0.002	0.006
Phrase 2	0.960	0.014	0.025	0.017	0.008
Phrase 3	0.070	0.031	0.036	0.005	0.861
Phrase 4	0.017	0.003	0.999	0.006	0.010
Phrase 5	0.225	0.001	0.019	0.894	0.067
Phrase 6	0.093	0.921	0.009	0.004	0.011
Phrase 7	0.672	0.129	0.076	0.010	0.098
Phrase 8	0.309	0.150	0.089	0.038	0.212
Phrase 9	0.245	0.269	0.419	0.011	0.044
Phrase 10	0.390	0.026	0.131	0.452	0.067

Each row in table 5.3 shows the predicted probabilities for sadness, joy, fear, disgust, and anger for each phrase in table 5.1. Overall, IBM Watson's predictions match well with the expected emotions. However, the only phrase that did not meet the expected emotion is the one about changes in workplace policies (phrase 8), which was predicted mainly as sadness when the expected emotion was anger, anger was the next highest prediction.

5.5 ChatGPT Performance

In this section, we evaluate the performance of several models, including GPT-3.5-turbo, GPT-4o, and GPT-4o-mini [OpenAI], by measuring their response times to a set of predefined phrases. Each model was tested multiple times to ensure a thorough assessment of efficiency. The recorded response times (in seconds) for each test run are presented in three separate tables. This analysis aims to provide insights into the responsiveness of each model and compare their performance under consistent testing conditions.

This section also provides an overview of the key differences among three language models: GPT-3.5-turbo, GPT-4o, and GPT-4o Mini. Each model is built on advanced architectures, but they vary significantly in performance and intended use cases.

5.5.1 GPT-3.5-turbo

ChatGPT 3.5-turbo is based on the GPT-3.5 engine, which was trained on over 175 billion parameters. While it represents a significant advancement in natural language processing, it has notable downsides. One major issue is its accuracy and reliability; ChatGPT 3.5 is more prone to 'hallucinations', which means it can generate incorrect or non-sensical information, especially when faced with ambiguous queries. These limitations can lead to inappropriate outputs, which may affect user trust and satisfaction. Despite these challenges, ChatGPT 3.5-turbo remains effective for many applications, such as basic content generation and straightforward chatbot interactions.

Table 5.4: Test results for GPT-3.5-turbo response times over 10 phrases, alongside the average time and standard deviation. The phrases in this table match the phrases in table 5.1 in order.

Test number	Run 1	Run 2	Run 3	Run 4	Run 5	Average	SD
Phrase 1	0.91	0.91	0.86	1.05	0.99	0.944	0.067
Phrase 2	0.65	0.69	0.82	0.91	0.85	0.784	0.098
Phrase 3	1.26	0.95	1.95	1.70	1.20	1.412	0.362
Phrase 4	1.24	1.41	1.54	1.31	1.09	1.318	0.152
Phrase 5	0.96	1.07	1.11	1.02	1.04	1.040	0.050
Phrase 6	3.33	2.96	3.57	1.67	3.06	2.918	0.659
Phrase 7	2.93	3.43	2.73	2.70	1.84	2.726	0.514
Phrase 8	1.69	1.94	1.30	1.55	2.06	1.708	0.272
Phrase 9	2.55	3.44	2.66	4.42	2.76	3.166	0.700
Phrase 10	1.10	1.46	1.80	1.62	1.45	1.486	0.231

From the results, it is noticeable that shorter, simpler phrases (like Phrase 1 and Phrase 2) tend to have lower response times, averaging around 0.94 and 0.78 seconds, respectively. These phrases exhibit relatively low standard deviations, indicating stable performance between runs. As the complexity of the phrases increases, the response times also increase, as seen in phrases such as Phrase 6 and Phrase 9, which have significantly higher averages of 2.91 and 3.16 seconds, respectively. These more complex phrases also show greater variation between runs, reflected by higher standard deviations (e.g., 0.659 for Phrase 6 and 0.700 for Phrase 9), suggesting that complexity affects both the processing time and consistency.

Interestingly, certain phrases such as Phrase 3 and Phrase 7 also show a marked increase in response time and standard deviation, indicating that as the task becomes more complex, the model requires more time and produces more varied results. Overall, the data demonstrates that GPT-3.5-turbo's response times correlate with phrase complexity, with the model generally being faster on simpler phrases and taking longer on more intricate ones.

5.5.2 GPT-4o

GPT-4o is the full-fledged version of the GPT-4 architecture, representing a significant upgrade over GPT-3.5-turbo. This model features enhanced accuracy and reliability,

being trained on more than a trillion parameters, which allows it to generate more precise responses and significantly reduce the likelihood of hallucinations. GPT-4o excels at understanding nuanced contexts and producing coherent, contextually appropriate text. In addition, it is designed for complex tasks that require high computational power, making it suitable for applications in industries such as finance, healthcare, and research, where precision and depth are crucial.

Table 5.5: Test results for GPT-4o response times over 10 phrases, alongside the average time and standard deviation. The phrases in this table match the phrases in table 5.1 in order.

Test number	Run 1	Run 2	Run 3	Run 4	Run 5	Average	SD
Phrase 1	1.63	1.00	0.86	0.83	1.31	1.126	0.304
Phrase 2	0.82	0.87	1.24	0.87	0.70	0.900	0.181
Phrase 3	0.86	1.36	1.37	1.24	2.00	1.366	0.367
Phrase 4	1.74	0.81	1.36	1.07	1.10	1.216	0.315
Phrase 5	0.80	5.37	1.05	0.82	0.96	1.800	1.787
Phrase 6	1.49	1.97	1.22	2.31	3.63	2.124	0.842
Phrase 7	1.64	2.95	1.73	3.02	2.42	2.352	0.583
Phrase 8	4.34	4.62	3.42	4.14	5.76	4.456	0.763
Phrase 9	5.39	4.37	2.76	4.65	5.73	4.580	1.033
Phrase 10	2.13	1.58	1.57	2.98	2.58	2.168	0.554

For simpler phrases, such as Phrase 1 and Phrase 2, GPT-4o performs relatively quickly, with average response times of 1.13 and 0.90 seconds, respectively. The standard deviations are low, indicating consistent performance across runs. As complexity increases, response times gradually increase.

Phrase 5 sees a large jump in SD because one run takes 5.37 seconds to get a response. It is not clear as to why this happened, this could have been due to a momentary drop in internet quality or an issue with OpenAI's servers. Without this outlier, the average response time was 0.908 seconds and the standard deviation is only 0.103, which is the expected result.

5.5.3 GPT-4o Mini

GPT-4o Mini is a compact and efficient version of GPT-4o that balances performance with accessibility. It is smaller and more resource efficient than its larger counterpart,

sacrificing some performance for greater accessibility. Despite this, GPT-4o Mini remains effective for various applications where the full capabilities of GPT-4o are not required.

Table 5.6: Test results for GPT-4o-mini response times over 10 phrases, alongside the average time and standard deviation. The phrases in this table match the phrases in table 5.1 in order.

Test number	Run 1	Run 2	Run 3	Run 4	Run 5	Average	SD
Phrase 1	1.48	1.10	0.97	1.04	0.78	1.074	0.230
Phrase 2	1.22	1.07	1.13	1.16	0.84	1.084	0.131
Phrase 3	0.97	1.76	1.19	0.86	1.08	1.172	0.314
Phrase 4	0.95	1.53	0.65	1.34	0.80	1.054	0.331
Phrase 5	1.09	1.35	0.73	1.62	1.08	1.174	0.298
Phrase 6	1.89	2.02	1.63	1.69	1.25	1.696	0.263
Phrase 7	2.54	2.34	2.04	2.29	2.51	2.344	0.180
Phrase 8	3.25	5.00	3.66	2.25	4.51	3.734	0.964
Phrase 9	5.49	6.12	5.41	7.28	6.61	6.182	0.702
Phrase 10	1.68	1.57	2.10	1.82	1.35	1.704	0.251

In general, GPT-4o-mini tends to respond slightly faster than GPT-4o, with a couple of exceptions (phrase 2 and phrase 9). GPT-4o-mini also shows that it is more consistent with its response times having generally a lower standard deviation on all phrases except phrase 8.

5.6 Discussion

The audio emotion recognition system has largely met performance expectations, showing robust capabilities in real-time emotion detection. IBM Watson's consistent response times, typically under one second, alongside its reliable accuracy in detecting a range of emotional tones, make it a valuable tool, especially in contexts where facial emotion recognition may not be possible or practical. Its ability to quickly process audio input and return relevant emotional insights ensures that it can seamlessly complement or even substitute facial emotion recognition when required.

The responses generated by ChatGPT models to the input phrases, detailed in the Appendices, show a noticeable variance in quality. The GPT-4o and GPT-4o-mini models consistently produced more coherent and relevant responses compared to

the GPT-3.5-turbo model. Notably, while GPT-3.5-turbo tended to elaborate on Phrase 7 as if it were part of a narrative, discussing various actions of a friend, both GPT-4o and GPT-4o-mini focused on providing more relevant and contextually appropriate replies. With the response times of GPT-3.5-turbo and GPT-4o-mini being comparable the better responses garnered from GPT-4o-mini make it the clear choice for engaging the user in conversation.

Chapter 6

Conclusion

This thesis aimed to explore the development and evaluation of a multimodal emotion recognition system that uses facial and audio inputs, with a focus on achieving a balance between speed and accuracy in emotion detection, particularly in real-time robotic applications. The evaluation of different face detection models revealed significant performance differences in various metrics.

For overall accuracy, YOLO exhibited the best performance, attaining an F1 score of 0.63 in the wider face dataset and 0.94 in the single face dataset. In contrast, Tiny YOLO faced challenges with the more intricate Wider Face data set, which produced an F1 score of 0.46. However, it showed similar performance to YOLO on the simpler single-face data, achieving an F1 score of 0.93. Notably, Tiny YOLO outperformed YOLO in inference times, averaging 25 ms compared to YOLO's 161 ms. This observation was further validated during robotic platform tests, where Tiny YOLO required 697.4 ms and YOLO required 6846.5 ms. Although dlib and the Haar cascade did not achieve the same levels of accuracy as the two YOLO models, they did achieve significant results in terms of inference speeds on the robotic platform, outperforming Tiny YOLO with 216.2 ms and 491.7 ms, respectively. This presents the choice between prioritising speed or accuracy based on the selected face recognition model.

For emotion recognition, ResNet50 demonstrated the highest accuracy, delivering the best results in the confusion matrices, closely followed by MobileNetV2.

MobileNetV2 significantly outpaced other models in terms of speed, with 13.6 ms per detection, while ResNet50 required 95.3 ms. These performance differences were evident in the TurtleBot tests, with MobileNetV2 increasing to 124.7 ms and ResNet50 to 1334.3 ms for inference, highlighting the trade-off between speed and accuracy.

When combining models, YOLO+ResNet50 proved to be the best, achieving an overall precision of 59.02%, with TinyYOLO + ResNet50 closely following at 57.36%. For efficiency, the fastest combination of models when running on a PC, is TinyYOLO + MobileNetV2, which requires only 38.6ms per emotion detection. However, if running on a Turtlebot, the dlib+MobileNetV2 combination offers the best speed, albeit at the expense of accuracy, with only 37.06%.

The facial and audio emotion recognition findings show the promise of a multimodal system that can harness the strengths of each method. While facial recognition demonstrates greater speed in specific scenarios, audio recognition can offer emotion recognition when facial recognition is not feasible. These combined systems lay the foundation for more resilient and adaptive human-robot interaction.

Although this study provides valuable information on multimodal emotion recognition using both facial emotion recognition (FER) and audio emotion recognition (AER), there are several limitations that must be recognised.

First, the face detection models were tested on a dataset that varies significantly in complexity. Although the results for face detection and emotion recognition were promising, the generalisability of these findings to real-world environments, especially those with dynamic and unstructured scenes, remains a challenge. The models performed well under controlled conditions, but their accuracy and inference times can be degraded in settings where multiple faces, occlusions, or extreme lighting conditions are present. This limitation is particularly relevant when deploying these models on robotic platforms that operate in uncontrolled environments.

The facial emotion recognition system has only been tested in the Expressions-in-the-Wild database. Although this data set attempts to simulate real-world scenarios, the images do not present problems such as poor lighting conditions and obstructed

faces. They may not fully capture the complexities and variability of emotions expressed in truly dynamic and unstructured real-world conditions.

It was decided early on that the system would use an off-board-based analysis to perform audio emotion recognition. This resulted in, after careful research, that the audio emotion recognition capabilities of OpenSMILE were not tested. However, with the efficiency that was achieved by the facial emotion recognition system, it could be possible to run this system on the robot as well.

Future research will focus on testing the system in live environments with human participants to gain a deeper understanding of its real-world performance. Conducting trials in uncontrolled dynamic environments will allow evaluation under conditions where lighting, background noise, and other factors that affect detection accuracy are not as easily managed. This will provide more robust data on how well the system performs in daily human-robot interactions and will help identify any performance bottlenecks, usability issues, or edge cases that were not encountered in the initial tests.

In addition to live testing, integrating a full audio emotion recognition system, such as OpenSMILE, directly onto the robot will be a key step. Currently, reliance on cloud tools limits the flexibility of the system and its ability to process audio data in real time, particularly in environments with background noise or multiple speakers. By embedding a native AER system on the robot, this aim is to improve both processing speed and accuracy, enabling more seamless and interactive communication between the robot and its users.

Exploring techniques such as Generative Adversarial Networks (GANs), temporal feature analysis, and the Facial Action Coding System (FACS) could further enhance the performance of emotion recognition systems. GANs could be leveraged for data augmentation, addressing the challenge of limited and imbalanced datasets by generating synthetic examples, which would improve the model's robustness and generalisation to real-world scenarios. Temporal feature analysis, on the other hand, allows for the recognition of emotions over time, making it possible to capture subtle shifts in facial expressions that may not be evident in static images, providing a deeper understanding of emotional states. Finally, FACS offers a systematic way

to encode facial movements into Action Units (AUs), which are the fundamental components of facial expressions. By focusing on these underlying muscle movements, models could achieve more accurate emotion classification, particularly for complex or nuanced expressions. Each of these approaches could contribute uniquely to improving the overall performance of emotion recognition systems.

Moreover, exploring the fusion of facial and audio modalities represents a critical next step in advancing the system's emotional understanding. By combining data from facial expressions and vocal tones, the system could achieve more nuanced and context-sensitive emotion detection. Multimodal fusion has the potential to improve accuracy, especially in cases where one modality may be ambiguous or unavailable, such as situations where the face is partially obscured or the audio is noisy. Different fusion techniques, including early, late, or hybrid fusion, should be investigated to determine which method yields the most reliable and consistent results under various conditions.

Further optimisations for real-time performance on robotic platforms will also be pursued. As evidenced by the current study, inference times can increase substantially when models are deployed on hardware-constrained systems like the TurtleBot4. Future work will involve refining the system's computational efficiency through model optimisations, to ensure that real-time emotion recognition remains feasible without sacrificing accuracy.

The integration of multimodal emotion recognition systems represents a substantial advancement toward developing robots capable of understanding and responding to human emotions in real-time, thereby paving the way for more empathetic and engaging human-robot interactions.

References

[1] Adiga, S., Vaishnavi, D. V., Saxena, S., and Tripathi, S. (2020). Multimodal emotion recognition for human robot interaction. In *2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE.

[2] Alexey (2021). darknet: YOLOv4 / Scaled-YOLOv4 / YOLO - neural networks for object detection (windows and linux version of darknet).

[3] Ali, S., Tanweer, S., Khalid, S., and Rao, N. (2021). Mel frequency cepstral coefficient: A review. In *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India*. EAI.

[4] Allognon, S. O. C., Koerich, A. L., and Britto, Jr, A. d. S. (2020). Continuous emotion recognition via deep convolutional autoencoder and support vector regressor.

[5] Alshamsi, H., Këpuska, V., and Meng, H. (2017). Real time automated facial expression recognition app development on smart phones.

[6] Anjum, M. (2019). Emotion recognition from speech for an interactive robot agent. In *2019 IEEE/SICE International Symposium on System Integration (SII)*. IEEE.

[7] Appuhamy, E. J. G. S. and Madhusanka, B. G. D. A. (2018). Development of a GPU-based human emotion recognition robot eye for service robot by using convolutional neural network. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE.

[8] Aqdus, C., Nunes, R., Kamal, Rehm, M., and Moeslund, T. (2021). *Deep Emotion Recognition through Upper Body Movements and Facial Expression*.

[9] Ashok, A., Pawlak, J., Paplu, S., Zafar, Z., and Berns, K. (2022). Paralinguistic cues in speech to adapt robot behavior in human-robot interaction. In *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE.

[10] Augello, A., Bella, G. D., Infantino, I., Pilato, G., and Vitale, G. (2022a). Multimodal mood recognition for assistive scenarios. *Procedia Comput. Sci.*, 213:510–517.

[11] Augello, A., Bella, G. D., Infantino, I., Pilato, G., and Vitale, G. (2022b). Multimodal mood recognition for assistive scenarios. *Procedia Comput. Sci.*, 213:510–517.

[12] Barsoum, E., Zhang, C., Canton Ferrer, C., and Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*.

[13] Boric-Lubecke, O., Massagram, W., Lubecke, V. M., Host-Madsen, A., and Jokanovic, B. (2008). Heart rate variability assessment using doppler radar with linear demodulation. In *2008 38th European Microwave Conference*, pages 420–423.

[14] Brandizzi, N., Bianco, V., Castro, G., Russo, S., and Wajda, A. (2021). Automatic rgb inference based on facial emotion recognition. In *System (Linköping)*.

[15] Breazeal, C. (2003). Emotion and sociable humanoid robots. *Int. J. Hum. Comput. Stud.*, 59(1-2):119–155.

[16] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, 42(4):335–359.

[17] Carolis, B. D., Ferilli, S., Palestra, G., and Redavid, D. (2016). *Emotion-Recognition from Speech-based Interaction in AAL Environment*.

[18] Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*, New York, NY, USA. ACM.

[19] Chen, L., Li, M., Lai, X., Hirota, K., and Pedrycz, W. (2020a). Cnn-based broad learning with efficient incremental reconstruction model for facial emotion recognition. *IFAC-PapersOnLine*, 53(2):10236–10241. 21st IFAC World Congress.

[20] Chen, L., Li, M., Su, W., Wu, M., Hirota, K., and Pedrycz, W. (2021). Adaptive feature selection-based AdaBoost-KNN with direct optimization for dynamic emotion recognition in human–robot interaction. *IEEE Trans. Emerg. Top. Comput. Intell.*, 5(2):205–213.

[21] Chen, L., Li, M., Wu, M., Pedrycz, W., and Hirota, K. (2023). Coupled multimodal emotional feature analysis based on broad-deep fusion networks in human-robot interaction. *IEEE Trans. Neural Netw. Learn. Syst.*, PP:1–11.

[22] Chen, L., Su, W., Feng, Y., Wu, M., She, J., and Hirota, K. (2020b). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Inf. Sci. (Ny)*, 509:150–163.

[23] Chen, L., Zhou, M., Su, W., Wu, M., She, J., and Hirota, K. (2018). Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Information Sciences*, 428:49–61.

[24] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

[25] Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 362(1480):679–704.

[26] Devaram, R. R., Beraldo, G., De Benedictis, R., Mongiovì, M., and Cesta, A. (2022). LEMON: A lightweight facial emotion recognition system for assistive robotics based on dilated residual convolutional neural networks. *Sensors (Basel)*, 22(9).

[27] Dhanith, J., Venkatraman, S., Narendra, M., Sharma, V., Malarvannan, S., and Gandomi, A. H. (2024). Multimodal emotion recognition using audio-video transformer fusion with cross attention.

[28] Dhuheir, M., Albaseer, A., Baccour, E., Erbad, A., Abdallah, M., and Hamdi, M. (2021). Emotion recognition for healthcare surveillance systems using neural networks: A survey.

[29] Dzedzickis, A., Kaklauskas, A., and Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors (Basel)*, 20(3):592.

[30] Elfaramawy, N., Barros, P., Parisi, G. I., and Wermter, S. (2017). Emotion recognition from body expressions with a neural network architecture. In *Proceedings of the 5th International Conference on Human Agent Interaction*, New York, NY, USA. ACM.

[31] Esfandbod, A., Rokhi, Z., Meghdari, A. F., Taheri, A., Alemi, M., and Karimi, M. (2023). Utilizing an emotional robot capable of lip-syncing in robot-assisted speech therapy sessions for children with language disorders. *Int. J. Soc. Robot.*, 15(2):165–183.

[32] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

[33] Faria, D. R., Vieira, M., Faria, F. C. C., and Premebida, C. (2017). Affective facial expressions recognition for human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

[34] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests.

[35] Gunes, H. and Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1148–1153.

[36] Gupta, S. (2018). Facial emotion recognition in real-time and static images. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE.

[37] Hajarolasvadi, N. and Demirel, H. (2019). 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy (Basel)*, 21(5):479.

[38] Haq, S. and Jackson, P. (2009). Speaker-dependent audio-visual emotion recognition. In *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Norwich, UK*.

[39] Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W., and Aguilera, A. (2022). Adaptive multimodal emotion detection architecture for social robots. *IEEE Access*, 10:20727–20744.

[40] Hung, H. M., Kim, S.-H., Yang, H.-J., and Lee, G.-S. (2020). Multiple models using temporal feature learning for emotion recognition. In *The 9th International Conference on Smart Media and Applications*, New York, NY, USA. ACM.

[41] Hwang, C.-L., Deng, Y.-C., and Pu, S.-E. (2023). Human–robot collaboration using sequential-recurrent-convolution-network-based dynamic face emotion and wireless speech command recognitions. *IEEE Access*, 11:37269–37282.

[42] Jaiswal, S., Jain, A., and Nandi, G. C. (2020). Image based emotional state prediction from multiparty audio conversation. In *2020 IEEE Pune Section International Conference (PuneCon)*. IEEE.

[43] Jaiswal, S. and Nandi, G. C. (2022). Optimized, robust, real-time emotion prediction for human-robot interactions using deep learning. *Multimedia Tools Appl.*, 82(4):5495–5519.

[44] Kansizoglou, I., Bampis, L., and Gasteratos, A. (2022). An active learning paradigm for online audio-visual emotion recognition. *IEEE Trans. Affect. Comput.*, 13(2):756–768.

[45] Khan, A. (2023). Improved multi-lingual sentiment analysis and recognition using deep learning. *J. Inf. Sci.*, page 016555152211372.

[46] Kim, B. S., Korea Institute of Industrial Technology, Ansan-si, Gyeongi-do, South Korea, and Kim, E. H. (2018). Speaker-independent emotion recognition for interstate measuring of user based on separation and rejection. *Int. J. Mach. Learn. Comput.*, 8(2):152–157.

[47] Kim, E. H., Kwak, S. S., Hyun, K. H., Kim, S. H., and Kwak, Y. K. (2009). Design and development of an emotional interaction robot, mung. *Adv. Robot.*, 23(6):767–784.

[48] Kumar, A., Tejaswini, P., Nayak, O., Kujur, A. D., Gupta, R., Rajanand, A., and Sahu, M. (2022). A survey on IBM watson and its services. *J. Phys. Conf. Ser.*, 2273(1):012022.

[49] Kusuma, G. P., Jonathan, J., and Lim, A. P. (2020). Emotion recognition on FER-2013 face images using fine-tuned VGG-16. *Adv. Sci. Technol. Eng. Syst. J.*, 5(6):315–322.

[50] Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., and Wermter, S. (2018). On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

[51] Li, P., Hu, F., Li, Y., and Xu, Y. (2014). Speaker identification using linear predictive cepstral coefficients and general regression neural network. In *Proceedings of the 33rd Chinese Control Conference*, pages 4952–4956.

[52] Li, T.-H. S., Kuo, P.-H., Tsai, T.-N., and Luan, P.-C. (2019). Cnn and lstm based facial expression analysis model for a humanoid robot. *IEEE Access*, 7:93998–94011.

[53] Livingstone, S. R. and Russo, F. A. (2018). The ryerson Audio-Visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS One*, 13(5):e0196391.

[54] Lopez-Rincon, A. (2019). Emotion recognition using facial expressions in children using the NAO robot. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*. IEEE.

[55] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). 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.

[56] Lundqvist, D., Flykt, A., and Öhman, A. (2015). Karolinska directed emotional faces. Title of the publication associated with this dataset: PsycTESTS Dataset.

[57] Lyu, Y. and Sun, Y. (2022). Global and local feature fusion via long and short-term memory mechanism for dance emotion recognition in robot. *Front. Neurorobot.*, 16:998568.

[58] Ma, K., Wang, X., Yang, X., Zhang, M., Girard, J. M., and Morency, L.-P. (2019). ElderReact: A multimodal dataset for recognizing emotional response in aging adults. In *2019 International Conference on Multimodal Interaction*, New York, NY, USA. ACM.

[59] Marinoiu, E., Zanfir, M., Olaru, V., and Sminchisescu, C. (2018). 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2158–2167.

[60] Mazzoni Ranieri, C., Vicentim Nardari, G., Moreira Pinto, A. H., Carnieto Tozadore, D., and Francelin Romero, R. A. (2018). Lara: A robotic framework for human-robot interaction on indoor environments. In *2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE)*, pages 376–382.

[61] Melinte, D. O. and Vladareanu, L. (2020). Facial expressions recognition for human-robot interaction using deep convolutional neural networks with rectified adam optimizer. *Sensors (Basel)*, 20(8):2393.

[62] Mistry, K., Rizvi, B., Rook, C., Iqbal, S., Zhang, L., and Joy, C. P. (2020). A Multi-Population FA for automatic facial emotion recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

[63] Mohammadpour, M., Khaliliardali, H., Hashemi, S. M. R., and AlyanNezhadi, M. M. (2017). Facial emotion recognition using deep convolutional networks. In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*. IEEE.

[64] Mohammed, S. and Alia, H. (2020). Speech emotion recognition using MELBP variants of spectrogram image. *Int. J. Intell. Eng. Syst.*, 13(5):257–266.

[65] Mohammed, S. N. and Karmin, A. (2021). A survey on emotion recognition for human robot interaction. *J. Comput. Inf. Technol.*, 28(2):125–146.

[66] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild.

[67] Mustaqeem, Sajjad, M., and Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, 8:79861–79875.

[68] Nawasalkar, R. K. and Butey, P. K. (2017). Study of comparison of human bio-signals for emotion detection using HCI.

[69] Nie, W., Chang, R., Ren, M., Su, Y., and Liu, A. (2022). I-GCN: Incremental graph convolution network for conversation emotion detection. *IEEE Trans. Multimedia*, 24:4471–4481.

[OpenAI] OpenAI. ChatGPT Models. https://platform.openai.com/docs/models/o1. [Accessed 22-10-2024].

[71] OpenAI (2022). ChatGPT. https://chat.openai.com/chat. Accessed: 2023-9-10.

[72] O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE Potentials*, 7(1):29–32.

[73] Pal, S., Mukhopadhyay, S., and Suryadevara, N. (2021). Development and progress in sensors and technologies for human emotion recognition. *Sensors (Basel)*, 21(16):5554.

[74] Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., and Akagi, M. (2020). Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, 8:16560–16572.

[75] Picard, R. W. (2000). *Affective Computing*. The MIT Press. MIT Press, London, England.

[76] Pramerdorfer, C. and Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art.

[77] Qayyum, A., Arefeen, A. B., Shahnaz, A., and Ieee Xplore, C. (2019). *Convolutional Neural Network (CNN) Based Speech-Emotion Recognition*.

[78] Rangulov, D. and Fahim, M. (2020). Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network.

[79] Rasendrasoa, S., Pauchet, A., Saunier, J., and Adam, S. (2022). Real-time multimodal emotion recognition in conversation for multi-party interactions. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, New York, NY, USA. ACM.

[80] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2015). You only look once: Unified, real-time object detection.

[81] Rosula Reyes, S. J., Depano, K. M., Velasco, A. M. A., Kwong, J. C. T., and Oppus, C. M. (2020). Face detection and recognition of the seven emotions via facial expression: Integration of machine learning algorithm into the NAO robot. In *2020 5th International Conference on Control and Robotics Engineering (ICCRE)*. IEEE.

[82] Ruiz-Garcia, A., Elshaw, M., Altahhan, A., and Palade, V. (2018a). A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. *Neural Comput. Appl.*, 29(7):359–373.

[83] Ruiz-Garcia, A., Webb, N., Palade, V., Eastwood, M., and Elshaw, M. (2018b). Deep learning for real time facial expression recognition in social robots. In *Neural Information Processing*, Lecture notes in computer science, pages 392–402. Springer International Publishing, Cham.

[84] Saxena, S., Tripathi, S., and Sudarshan, T. S. B. (2022). An intelligent facial expression recognition system with emotion intensity classification. *Cogn. Syst. Res.*, 74:39–52.

[85] Shanta, S. S., Sham-E-Ansari, M., Chowdhury, A. I., Shahriar, M. M., and Hasan, M. K. (2021). A comparative analysis of different approach for basic emotions recognition from speech. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*. IEEE.

[86] Shenoy, S., Jiang, Y., Lynch, T., Manuel, L. I., and Doryab, A. (2022). A self learning system for emotion awareness and adaptation in humanoid robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 912–919.

[87] Shi, X., Yang, H., and Zhou, P. (2016). Robust speaker recognition based on improved GFCC. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE.

[88] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data*, 6(1).

[89] Singh, S., Singh, D., and Yadav, V. (2020). Face recognition using HOG feature extraction and SVM classifier. *Int. J. Emerg. Trends Eng. Res.*, 8(9):6437–6440.

[90] Song, K.-S., Nho, Y.-H., Seo, J.-H., and Kwon, D.-S. (2018). Decision-level fusion method for emotion recognition using multimodal emotion recognition information. In *2018 15th International Conference on Ubiquitous Robots (UR)*. IEEE.

[91] Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., and Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomed. Signal Process. Control*, 70(103029):103029.

[92] Udeh, C. P., Chen, L., Du, S., Li, M., and Wu, M. (2022). A co-regularization facial emotion recognition based on multi-task facial action unit recognition. In *2022 41st Chinese Control Conference (CCC)*. IEEE.

[93] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I.

[94] Wang, Y.-X., Li, Y.-K., Yang, T.-H., and Meng, Q.-H. (2022). Multitask touch gesture and emotion recognition using multiscale spatiotemporal convolutions with attention mechanism. *IEEE Sens. J.*, 22(16):16190–16201.

[95] Webb, N., Ruiz-Garcia, A., Elshaw, M., and Palade, V. (2020). Emotion recognition from face images in an unconstrained environment for usage on social robots. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

[96] Yang, P., Cao, L. M., Zhu, L. L., and Luo, S. N. (2022). Design of attendance system based on nao face, speech and emotion recognition. In *2022 10th International Conference on Orange Technology (ICOT)*, pages 1–3.

[97] Yang, S., Luo, P., Loy, C. C., and Tang, X. (2016). Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[98] Younis, E. M. G., Zaki, S. M., Kanjo, E., and Houssein, E. H. (2022). Evaluating ensemble learning methods for multi-modal emotion recognition using sensor data fusion. *Sensors (Basel)*, 22(15):5611.

[99] Yu, C. and Tapus, A. (2019). Interactive robot learning for multimodal emotion recognition. In *Social Robotics*, Lecture notes in computer science, pages 633–642. Springer International Publishing, Cham.

[100] Yu, C. and Tapus, A. (2020). Multimodal emotion recognition with thermal and RGB-D cameras for human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA. ACM.

[101] Zhichao, P., Wenhua, H., Hongji, T., Minlei, X., and Ruwei, L. (2020). Attention-based sequence modeling for categorical emotion recognition with modulation spectral feature. In *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE.

[102] Zhu, C. and Ahmad, W. (2019). Emotion recognition from speech to improve human-robot interaction. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE.

[103] Zhu, Q., Zhuang, H., Zhao, M., Xu, S., and Meng, R. (2024). A study on expression recognition based on improved mobilenetv2 network. *Sci. Rep.*, 14(1):8121.

Appendix A

For ChatGPT Webpage that allows connecting to a robot:
https://github.com/Terramet/ChatNao
The same thing as above with ChatGPT having the ability to access realtime apis and runs in Python instead:
https://github.com/Terramet/ChatNaoPython
For the standalone Watson Sentiment analysis:
https://github.com/Terramet/WatsonSentimentAnalysis
Standalone facial emotion detection:
https://github.com/Terramet/MPhilFacialEmotionDetection
For the ChatGPT results, including reponse times and the actual response:
https://github.com/Terramet/MPhilDataStorage/tree/main/ChatGPTTests
For all the models used for testing the facial emotion recognition, yolo, vgg16, resnet50 and mobilenetv2:
https://github.com/Terramet/MPhilDataStorage/tree/main/Models
For all the results from every detected face in the Exp_W dataset:
https://github.com/Terramet/MPhilDataStorage/tree/main/results_closest_face_pc
For all the results from every detected face in the Exp_W dataset on the robot:
https://github.com/Terramet/MPhilDataStorage/tree/main/results_closest_face_robot
For a video showing the different voices available through IBM Watson
https://youtube.com/shorts/Qqz03HE4MFg