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

Emotion recognition is a key enabler of effective human-robot interaction (HRI), allowing robots to respond appropriately to users' emotional states. However, many current approaches rely on a single modality or multimodal fusion techniques which are computationally intensive and unsuitable for widely available, resource-constrained robotic platforms. This presents a significant barrier to deploying emotionally aware robots in real-world settings such as healthcare, education, and assistive technology.

This thesis addresses this challenge by evaluating two independent, low-resource emotion recognition approaches: facial emotion recognition and text-based sentiment analysis. The goal is to assess their individual effectiveness, feasibility, and potential to support emotionally intelligent behaviour without relying on full multimodal integration.

A literature review contextualises the work within existing research on visual, auditory, and gesture-based emotion recognition. Experimental evaluations explore the accuracy and efficiency of both modalities in constrained environments using a robotic platform.

Results demonstrate that both facial and text-based emotion recognition methods can operate effectively in isolation, offering practical solutions for real-time deployment on low-power systems. These findings suggest that strategic use of unimodal methods can enhance robot emotional responsiveness while avoiding the complexity of multimodal systems. The thesis concludes by identifying future research directions, including real-world testing, improved on-device processing, and lightweight integration strategies.

Keywords: Facial Emotion Recognition, Sentiment Analysis, 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 21211.

Name: Joshua Bamforth

Date: Monday 28th July, 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 human emotional states. This capacity is essential for creating more natural, empathetic, and socially appropriate interactions, as emotional cues are fundamental to human communication [20]. When robots can recognise emotions, they can adjust their tone, behaviour, or responses to match the user's affective state [17], thereby improving user experience and engagement particularly in sensitive domains such as healthcare and education.

In healthcare, robots capable of recognising patient emotions can offer better support and improve outcomes for groups such as individuals with autism, mental health challenges, or elderly populations [28]. Emotional awareness allows for personalised, adaptive interaction, and can help robots respond more appropriately to shifts in mood or stress levels [31].

However, existing emotion recognition systems often rely on a single input modality, such as facial expressions or speech. These approaches are inherently limited: facial-based systems fail when faces are obscured or out of frame, while audio-based systems can be disrupted by noise or speech variation. These limitations present challenges to real-world deployment, especially in unstructured environments.

To address these limitations, researchers have explored multimodal emotion recognition, which combines data from multiple sources, typically facial expressions, vocal intonation, and gestures, to achieve greater accuracy and robustness. While

true multimodal fusion, where data streams are tightly integrated at the algorithmic level, has shown promise in lab settings, it also introduces high computational complexity and integration challenges that limit its use on resource-constrained robotic platforms.

This thesis is motivated by the need to enhance emotion-aware interactions on widely available, resource-limited robots, which often lack the computational power or hardware needed for complex multimodal fusion.

The aim of this work is to evaluate and implement individual unimodal emotion recognition channels, specifically facial emotion recognition and sentiment-based emotion recognition, and to assess their effectiveness and feasibility in real-world HRI contexts. Rather than developing a fused multimodal system, this thesis focuses on understanding the strengths, weaknesses, and trade-offs of each modality when applied independently on low-power, real-time robotic systems. To achieve this, the research objectives are as follows: review current literature on facial, audio, gesture-based, and multimodal emotion recognition methods, implement and evaluate facial emotion recognition methods suitable for low-resource environments, explore the use of text-based sentiment analysis using existing cloud-based tools, and assess the performance and suitability of each method on a resource-limited robot platform, considering constraints such as computational load and response time.

The literature review chapter presents a comprehensive survey of emotion recognition methods across various modalities, including facial expressions, gestures, speech, and multimodal systems. This chapter highlights the strengths, limitations, and applicability of each approach within the context of human-robot interaction, with special attention given to systems designed for resource-constrained environments. It also considers the role of more invasive techniques, such as EEG, addressing their practical implications.

Materials and Methods outlines the technical setup used throughout this study. It describes the robotic platform, datasets, and the implementation of emotion recognition systems across both facial and audio modalities. The chapter also details the selection and configuration of software libraries and tools, emphasising their suitability for low-power robotic platforms and real-time processing constraints.

Next the results chapter combines the evaluation of both facial and sentiment analysis systems. It presents a comparative analysis of various face detection and classification algorithms, as well as response times for the sentiment analysis. The performance of each system is assessed in terms of accuracy, computational efficiency, and suitability for deployment in resource-limited environments. This chapter also includes reflections on the challenges encountered during implementation and testing.

Chapter 5, Discussion, interprets the results presented in the previous chapter, analysing their significance in relation to the research objectives. It explores the trade-offs between accuracy and efficiency, the practical challenges of integrating emotion recognition into real-world robotic systems, and the implications of using unimodal rather than multimodal approaches.

The conclusion summarises the key findings of the thesis and reflects on how the research meets its stated aims and objectives. It evaluates the effectiveness of the unimodal systems developed, and identifies areas for improvement, including model performance, hardware integration, and user validation. The chapter concludes by proposing directions for future research, such as lightweight multimodal fusion strategies and extended testing in real-world HRI scenarios.

Chapter 2

Literature

2.1 Psychological Foundations of Emotion in HRI

2.1.1 Essential Emotions for Human-Robot Interaction

Emotions are fundamental to natural social communication, and social robots are increasingly designed with emotional intelligence so that they can "infer and interpret human emotions" during interaction. A widely adopted framework is Ekman's model of six basic emotions - happiness, surprise, fear, disgust, sadness, and anger - which are thought to have universal facial signals [91]. Later on Ekman's model of emotions was expanded at include a seventh basic emotion, contempt [66], though most omit this. Empirical HRI systems typically choose a subset of basic emotions that are most relevant and reliably detectable. For example, Alonso-Martín et al. [5] deliberately limited their NAO-based system to neutral, happiness, sadness, and surprise, noting that these four cover the key dialog cases and are easier to recognise by camera and microphone.

It is essential for HRI systems to detect and respond appropriately to happiness. Smiling or laughter from a user typically means things are going well, and robots that recognise happiness can respond by building rapport. In practice, happy expressions are readily recognised by current systems: for instance, a NAO robot using a CNN achieved approximately 91% accuracy on happy faces [37].

Sadness usually signals that something is wrong. Robots that recognise sadness can respond empathetically (e.g. speaking softly or offering help). Many HRI experiments explicitly use sadness as a target emotion [106]. Detection systems also handle sadness well, for example, the same NAO running a CNN as above, recognised sad faces with approximately 90% accuracy [37]. Sadness is a core negative state for social robots to sense; it is included in nearly every emotion set, and responding to it (e.g. consolation) supports natural interaction.

Anger detection is critical for managing conflict or danger. An angry expression can indicate frustration, disagreement, or risk. Anger is a high-arousal signal of dissatisfaction [106]. Practically, if a robot detects that a person is angry, it can take steps to defuse the situation (apologise or give space). Recognition of anger tends to be good (85% accuracy in one study [87]). Since anger often requires adjustment of robot behavior, it's widely treated as an essential emotion in HRI [106].

HRI systems also consider fear and surprise important, but usually to avoid. Inducing fear or startling people is known to harm trust. In HRI safety research, for instance, Sisbot et al. [103] emphasise that a socially acceptable robot must never trigger human fear, surprise or discomfort. Thus, some HRI systems include fear/surprise detection mainly to monitor safety (e.g. pausing if a user appears startled), though it is sometimes harder to detect (about 65% accuracy in the NAO study [37]).

Disgust is the least-studied of the six in social robotics, but it is included for completeness. In human social signaling, disgust usually means "this is aversive" (e.g. a bad smell or morally repugnant content). Few HRI applications explicitly focus on disgust, but it is part of standard emotion sets [106]. When detected, a robot might interpret disgust similarly to anger/avoidance (e.g. stop a disagreeable action). Recognition of disgust tends to be lower (it was grouped with sadness and fear at approximately 65% accuracy [37]). Thus, while disgust is acknowledged as a basic human emotion, most HRI systems prioritise the others; it is included mainly to round out the universal facial expression categories.

A "neutral" state is normally treated as another category. Most human interaction is emotionally neutral, thus systems include it as a catch-all [5].

Finally, contempt is rarely a focus in HRI systems. Most HRI emotion-recognition work uses Ekman's orignial basic categories (anger, disgust, fear, happiness, sadness, surprise, and a neutral catagory) and often omits contempt. Some commercial APIs (e.g. Microsoft's Face API) do output contempt, but empirical HRI studies find almost no contemptful expressions in practice [26]. In short, contempt is technically included in some classifiers, but it's not commonly reported or reliably detected in real HRI data.

Overall, recent HRI research converges on Ekman's six basic emotions as the "core" set that robots should recognise, most commonly ommiting the seventh emotion, contempt. Happiness and surprise are seen as especially beneficial for positive user experience, whereas anger, sadness, and fear are included so robots can handle conflict or distress appropriately [26]. Disgust is rarely targeted. Overall, the "essential emotions" for effective HRI turn out to be those that (a) humans naturally express strongly in social settings and (b) robots can reliably sense the core Ekman emotions plus a neutral baseline.

2.1.2 Appraisal Theory

Appraisal theory accounts for the elicitation of emotion by linking it to an individual's cognitive evaluation of events, particularly in terms of goal relevance, perceived control, certainty, and agency. Thus it explains why identical events may evoke divergent emotional responses across individuals [73]. This theoretical perspective has informed several computational approaches to emotion modeling. For instance, knowledge-based systems such as EmotiNet represent prototypical event-action sequences and their associated affective outcomes based on predefined appraisal rules [13]. In the domain of emotion analysis, appraisals have been used as intermediate representations to support more interpretable and robust classification. Troiano et al. [109] introduced a corpus comprising event descriptions annotated with both emotion labels and appraisal dimensions, demonstrating that appraisal features can be automatically inferred from text with near-human reliability. Critically, they showed that incorporating appraisal representations alongside emotion categories

enhanced classification accuracy, suggesting that appraisal-based reasoning provides a complementary pathway for modeling affective meaning, particularly in cases involving implicit cues, such as goal obstruction leading to anger or fear.

Within affective HRI, appraisal-informed models enable artificial agents to infer users' emotional states by assessing interaction context, such as whether the agent's actions support or hinder user goals. Some implementations explicitly encode appraisal variables (e.g., task success, goal congruence, user feedback) alongside perceptual cues. Demutti et al. [29], for example, proposed a cloud-based HRI framework wherein user-reported appraisals (e.g., satisfaction with conversational topics and task outcomes) were integrated with facial and gaze data. A Random Forest classifier trained on both appraisal and sensory features outperformed models relying solely on perceptual input when classifying affective valence.

In broader HRI contexts, appraisal theory has been leveraged to imbue robots with more human-like emotion reasoning capabilities. For example, Tang et al. [108] describe an architecture wherein robots continuously evaluate the valence, relevance, and goal impact of human behaviors in real time. Drawing on models such as the Ortony, Clore, and Collins (OCC) framework, the robot applies appraisal rules to determine whether events are congruent with user goals and generates contextually appropriate affective responses. In experimental scenarios, appraisal-enabled robots demonstrated improved social responsiveness; for instance, one robot correctly interpreted a student's disappointment, despite incongruent verbal cues, and provided empathetic feedback, whereas a non-appraisal baseline failed to recognise the underlying emotion and responded inappropriately.

2.1.3 Valence-Arousal model

Alongside cognitive appraisal, many emotion-aware systems use dimensional models of affect, which map emotions onto continuous scales. The most common framework is the valence-arousal (VA) space (often called Russells circumplex model [95]). In this model, valence corresponds to pleasantness (positive vs. negative affect), while arousal corresponds to activation or intensity (calm vs. excited). Every emotional

state can be placed as a point in this 2D space: for example,"joy" is high-valence, high-arousal, while "sadness" is low-valence, low-arousal. These two dimensions capture the core affective quality of most emotions [65].

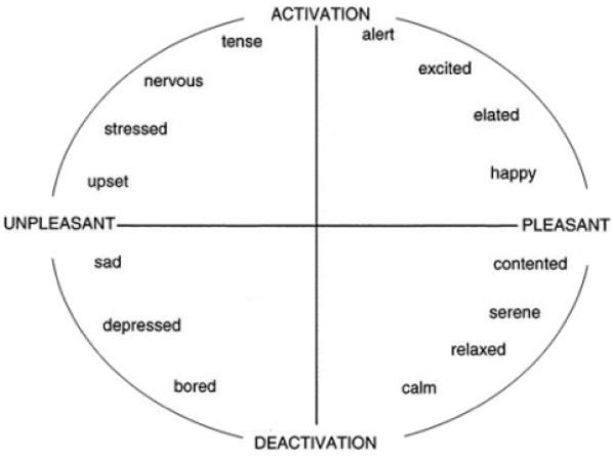


Figure 2.1: A Graphical representation of Russell's Circumplex Model of Affect [95].

Dimensional labeling is popular in HRI because it provides a common language across modalities and datasets. For instance, researchers often annotate facial expression or physiological data with valence-arousal values so that different sensors can be merged or compared. In fact, Spezialetti et al. [105] note that "several dimensional annotated datasets share a common valence-arousal (VA) representation, which allows comparing and merging data from different datasets". As a result, VA labels are often provided as a good practice in emotion datasets. By covering a broad range of VA values, such datasets enable training models that predict continuous affect.

In affective computing systems (vision, speech, multimodal recognition), VA is also apparent. Modern facial emotion recognition benchmarks (like the ABAW challenge) include frame-by-frame valence/arousal estimation as official tasks. Multi-task deep models extract features from faces (e.g. with EfficientNet) and output both discrete expression labels and continuous VA scores [96]. These continuous predictions allow systems to gauge nuanced changes in affect (e.g. gradual smiles) rather than hard categories. Similarly, speech-based emotion models often regress valence and arousal from audio features.

2.2 Facial Emotion Recognition

Facial emotion recognition is a crucial aspect of affective computing [84] 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 HRI. 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.

2.2.1 Datasets

Pierre Luc Carrier and Aaron Courville [38] 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 48×48 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 [75] 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 [6] 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 [61], 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.

2.2.2 Algorithms

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 [69].

Figure 2.2 summarises the underlying network types used in all reviewed studies, showing the prevalence of CNN-based methods and the relative frequency of other algorithmic choices.

Facial expression analysis remains one of the most practical non-invasive modalities for emotion recognition in HRI. However, most systems rely on outdated or

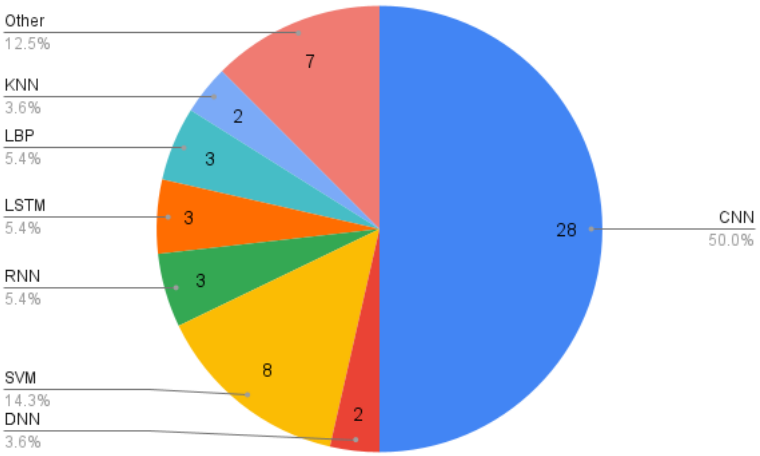


Figure 2.2: A summary of the algorithmic choices used in the reviewed studies.

limited face detection methods like using Haar-Cascade for initial face detection and cropping the image to isolate the face before implementing more advanced CNNs [8] or SVMs [40] [92], 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 [57] to perform classification. Other studies have used a convolutional autoencoder and support vector regressors [4] or recurring neural networks [16] 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. [54], 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. This model has potential for real-time applications in HRI.

ResNet50 was used to a high degree of accuracy on FER-2013 by Pramerdorfer et al. [85], achieving 72.4% accuracy with 5.3 million trainable parameters. No special modifications were made to the network, except removing the initial convolution and pooling layer and narrowing the architecture by using 256 feature maps in the final

residual group, thereby reducing the parameter count. While the model remains relatively large by embedded deployment standards, its high reported accuracy makes it a strong candidate for emotion recognition in HRI. However, its real-world feasibility depends on whether its inference speed and computational demands can be tolerated on resource-constrained robotic platforms.

Shenoy et al. [99] 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. [122], 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. However, MobileNetV2 alone shows good performance on

FER2013, achieving 67.9% accuracy with 2.2 million parameters, making it a strong candidate for real-time applications in HRI.

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 [97] [94]. 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.

2.2.3 Applications

The exploration of applications is also apparent, ranging from studies investigating and improving the effectiveness of emotion recognition in older adults [63] to those focusing on unconstrained environments [113] 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 [34], 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 [93]. Notably, [68] 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 [59].

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 [30] 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 [45] 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 [70]. Chinonso Paschal Udeh [110] 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.

2.3 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 HRI, 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.

2.3.1 Common Methods

A variety of signal processing techniques have been developed to extract meaningful features from speech for emotion recognition. These methods aim to capture spectral, prosodic, and temporal characteristics that correlate with emotional expression. This section outlines some of the most commonly used approaches.

The Mel-Frequency Cepstral Coefficients (MFCCs) [3] are among the most widely used features in audio emotion recognition. They represent the short-term power spectrum of a signal, computed by applying a discrete cosine transform to the logarithm of the Mel-scaled power spectrum. The Mel scale spaces frequency bands non-linearly to approximate the human auditory system's sensitivity to pitch and tone.

By capturing perceptually relevant spectral information, MFCCs emphasise frequency regions where emotional prosody, such as pitch, timbre, and intensity, varies most. This makes them highly effective for distinguishing vocal expressions of emotion.

Gammatone Frequency Cepstral Coefficients (GFCCs) [100] are an alternative to MFCCs that employ a gammatone filterbank, which more closely models the frequency selectivity of the human cochlea. This biologically inspired design enhances sensitivity to perceptual cues relevant to speech, potentially improving robustness to noise and capturing finer-grained auditory information. Despite these theoretical advantages, GFCCs remain less widely adopted than MFCCs in emotion recognition tasks, with their use more common in robust speech recognition and noisy environments.

Linear Predictive Coding (LPC) [81] models each speech sample as a weighted sum of previous samples, assuming that recent history can predict the present. The resulting coefficients effectively capture the spectral envelope of the signal, offering a compact representation of vocal tract resonances over time.

Linear Predictive Cepstral Coefficients (LPCCs) [56] are obtained by applying cepstral analysis to the LPC model, resulting in a feature representation that, like MFCCs, captures the spectral envelope in a decorrelated form. LPCCs encode both spectral shape and temporal structure, making them useful in speech and emotion recognition tasks, though they are generally more sensitive to noise compared to MFCCs.

2.3.2 Datasets

The IEMOCAP database [18] 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 [42] 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 [58] 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.

2.3.3 Algorithms

Methods such as MFCC are often used to extract features that are then classified into emotions using algorithms such as SVM [50], CNN, or DNN [98] [86]. 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 [121].

An emerging trend in audio-based emotion recognition involves the transformation of audio signals into visual representations [11], spectrograms, followed by the application of machine learning techniques such as CNNs [46] or Deep Belief Networks (DBNs) [72]. 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 [41]. 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 [76].

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 [7] 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

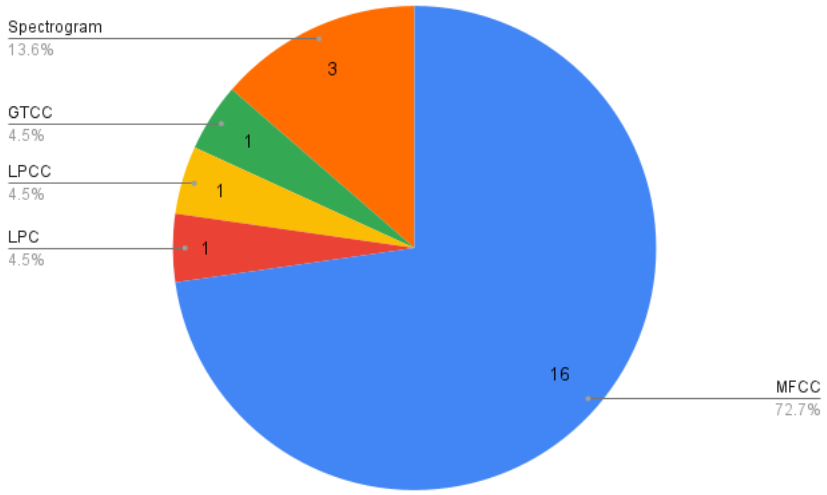


Figure 2.3: A chart showing the frequency of methods used to detect/extract emotional features from audio.

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 [24], and another a SVM [10].

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 [83]. By dynamically weighting different segments of the input speech signal [120] 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 [78]. Using mechanisms inspired by human attentional processes, such as self-attention and multi-head attention [89], 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.

2.3.4 Applications

In a study on speech emotion recognition in speaker-independent systems, specifically for the Mung robot [52], 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 [51].

In a paper by Carolis, B.D. [19], 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 $\in$ 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. [55] 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.

2.4 Gesture-based Emotion Recognition

Gesture-based emotion recognition focuses on interpreting emotional states from non-verbal body cues such as posture, movement, and hand gestures. These cues play a critical role in affective communication, particularly in scenarios where facial expressions or speech are unavailable or unreliable, such as when a person is wearing a mask, facing away, or in silence.

This section reviews both the algorithmic approaches used to extract and classify emotional information from gestures, as well as practical applications in interactive robotic settings.

2.4.1 Datasets

The Body Emotion Expression (BEE) dataset [33] 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 [64] 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.

2.4.2 Algorithms

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. [64] 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.

Wang et al. [112] 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 [62] 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. [9] 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.

While effective in experimental settings, these methods often require a fixed camera and precise calibration, making them difficult to scale or adapt to a robot intended to move around.

2.4.3 Applications

Marinoiu et al. [64] present a gesture-based emotion recognition approach with strong implications for robot-assisted therapy, particularly in recognizing emotions expressed by children and therapists. Their system enables non-invasive emotional monitoring in developmental contexts such as autism therapy. The researchers employ the Zeno robot, a humanoid platform specifically designed to support children with autism. By recognizing emotions through gestures and body movements, Zeno enhances its interactive capabilities, making it a promising tool for empathetic and adaptive therapeutic interventions.

Elfaramawy et al. [33] apply their gesture-based emotion recognition framework in a HRI setting involving two Nao humanoid robots. One Nao is equipped with an Asus Xtion depth sensor, positioned to capture depth-map video sequences from multiple angles, while the second robot facilitates interaction from a side perspective. This setup enables the collection of frontal, side, and rear views of participants as they express emotions through full-body movement. To elicit spontaneous emotional behavior, participants were prompted with realistic roleplay scenarios (e.g., for fear,

imagining a break-in and requesting the robot to call for help). This experimental design yielded a rich, multi-view dataset of emotional body expressions.

2.5 Multi-modal Emotion Recognition

Multi-modal emotion recognition systems combine various data sources, such as visual, auditory, and gesture signals, to enhance the accuracy and robustness of emotion detection. By integrating multiple modalities, these systems can capture a more comprehensive understanding of human emotions, aiming to improve performance in real-world applications. This section reviews the datasets and algorithms used in multi-modal emotion recognition, highlighting their significance in advancing the field.

2.5.1 Datasets

The FABO (FAcial and BOdily Expression) dataset [39] plays a significant role in multimodal 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.

2.5.2 Algorithms

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 [104] 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. [49] 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 HRI, Yu and Tapus [117] present a study titled 'Interactive Robot Learning for Multimodal Emotion Recognition.' Their research employs 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 HRI. 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 [118] 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. [23] have introduced a groundbreaking approach to multimodal emotion recognition in HRI 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.

Several studies have emerged focusing on sentiment analysis, addressing this gap in research. For example, Augello et al. [12] presented 'Multimodal Mood Recognition for Assistive Scenarios' showcasing the effectiveness of their approach in detecting emotions from textual data. Additionally, Heredia et al. [43] 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. Sentiment analysis shows promise in working in a resource constrained environment, this can also be offloaded to a cloud service, allowing for more complex models to be used without the need for high computational power on the robot itself.

Temporal features have emerged as a significant benefit for multimodal emotion recognition systems. Research efforts such as those by Hung et al. [44] 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.

2.5.3 Applications

While multimodal emotion recognition has advanced rapidly in theory, its deployment on robotic platforms remains sparse. Practical constraints, such as real-time processing, hardware limitations, and environmental variability, have limited its adoption. However, emerging applications show promise in enriching robot perception and interaction.

In one of the few real-world implementations of multimodal emotion recognition on a robotic platform, Yu and Tapus [117] employed the Pepper robot in a controlled laboratory setting. Their system combined thermal facial imaging, captured via an Optris thermal camera, with gait analysis using a Kinect sensor, allowing the robot to infer emotional states from both body movement and facial temperature cues. The experimental procedure involved 8 participants, each completing 24 sessions across three phases: initial online testing, an intermediate interactive robot learning (IRL) phase, and post-IRL testing. A total of 192 sessions were conducted over the course of one month. Environmental variables such as lighting and temperature were tightly controlled to ensure data consistency.

Another real-world deployment of multimodal emotion recognition in robotics was demonstrated in a preliminary application experiment involving an emotional social robot system. The setup utilized a Kinect sensor to capture full-body and facial data under natural indoor lighting conditions, enabling emotion inference based on both posture and facial expressions. Eight postgraduate participants (balanced by gender) were recruited to express the seven basic emotions, with peak emotional states used for training. During testing, an emotion was considered correctly recognized if the system identified the apex state accurately. The experiment achieved an average recognition accuracy of 75.85%. While misclassifications were largely attributed to environmental factors and limitations in the face and body detection algorithms,

the results highlight the feasibility and promise of deploying multimodal emotion recognition in affective HRI scenarios [23].

2.6 Table Of Robots

The following table provides a consolidated overview of robotic platforms used in emotion recognition and HRI research. For each robot, the table includes its name, a representative image, and references to studies detailing its application or implementation. References are annotated according to the modalities employed in each study. This compilation illustrates the diversity of robots explored in the literature, highlighting their distinct capabilities and roles in enabling emotionally aware and socially responsive interactions with humans.

2.6 Table Of Robots

Table 2.1: Table of all Robots in Literature

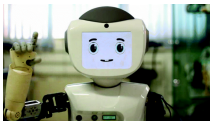
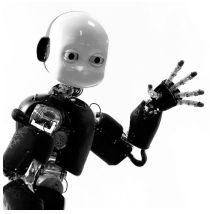
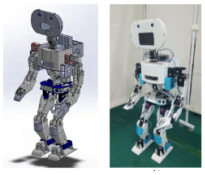
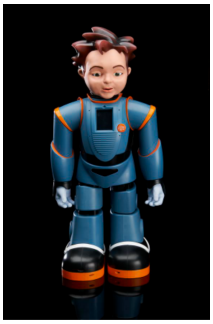
Robot Name	Facial	Audio	Gesture
 Pepper	[117] [118]		[117] [118]
 Nao	[36] [47] [59] [68] [92] [93] [94] [99] [113] [114]	[19]	[114]
 Mung		[51]	
 Omnidirectional Service Robot	[45]	[45]	
 Pioneer P3-DX robot, LARa robot	[67]	[67]	
 RASA		[34]	
 Ohmni	[30]		

Table 2.2: Table of all Robots in Literature Cont.

	Robot Name	Facial	Audio	Gesture
	iCub		[55]	
	ROBIN		[10]	
	ESRS (Emotional Service Robot System)	[21] [22] [25]		
	Harley	[57]		
	Robot eye	[8]		
	Zeno			[64]

2.7 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.

2.7.1 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 [74]. However, EEG signals vary significantly between individuals, making the development of universal models challenging [82]. 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 [107].

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 [74]. 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 [32].

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% [107].

2.7.2 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 [116].

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 [77]. 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 [116].

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 [32].

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 [15].

2.7.3 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.

2.8 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) [90], or HOG+Linear SVM [102] 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 HRIs.

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

In this section, an overview of the system architecture for the emotion recognition system developed for a resource-limited robotic platform is presented. The architectural diagram (see 3.1) illustrates how data flows from sensor acquisition through processing to decision-making, enabling the robot to determine human emotional states in real time. The system is composed of two primary modules: the image-based emotion recognition module and the speech-based emotion recognition module, each designed to operate independently while allowing the system to select the most appropriate modality based on contextual demands.

The image-based module begins with the acquisition of visual input via a camera. This input is processed through a face detection component where three methods—Haar Cascades, Dlib HOG+SVM, and YOLOv4—are implemented and compared to determine the most effective approach for isolating faces under varying conditions. Once faces are detected, the resulting regions of interest are forwarded to an emotion classification sub-module. Here, several convolutional neural network (CNN) architectures, including VGG16, ResNet50, and MobileNetV2, are evaluated for their ability to classify the emotional expressions accurately. The CNN models are trained using transfer learning on a dataset of labelled facial images representing seven basic emotions: happiness, sadness, anger, surprise, fear, disgust and contempt.

The trained models are then deployed on the robot to process real-time video streams and classify the emotions of human subjects.

Concurrently, the speech-based module captures auditory input through an onboard microphone. This audio stream is first transcribed into text by the `speech_recognition` Python library. The transcribed text is then analysed for emotional content using sentiment analysis provided by IBM Watson Cloud Services. By processing these two streams separately, the system can dynamically choose between visual and auditory modalities, ensuring that the most reliable source of emotion data is used based on the situational context.

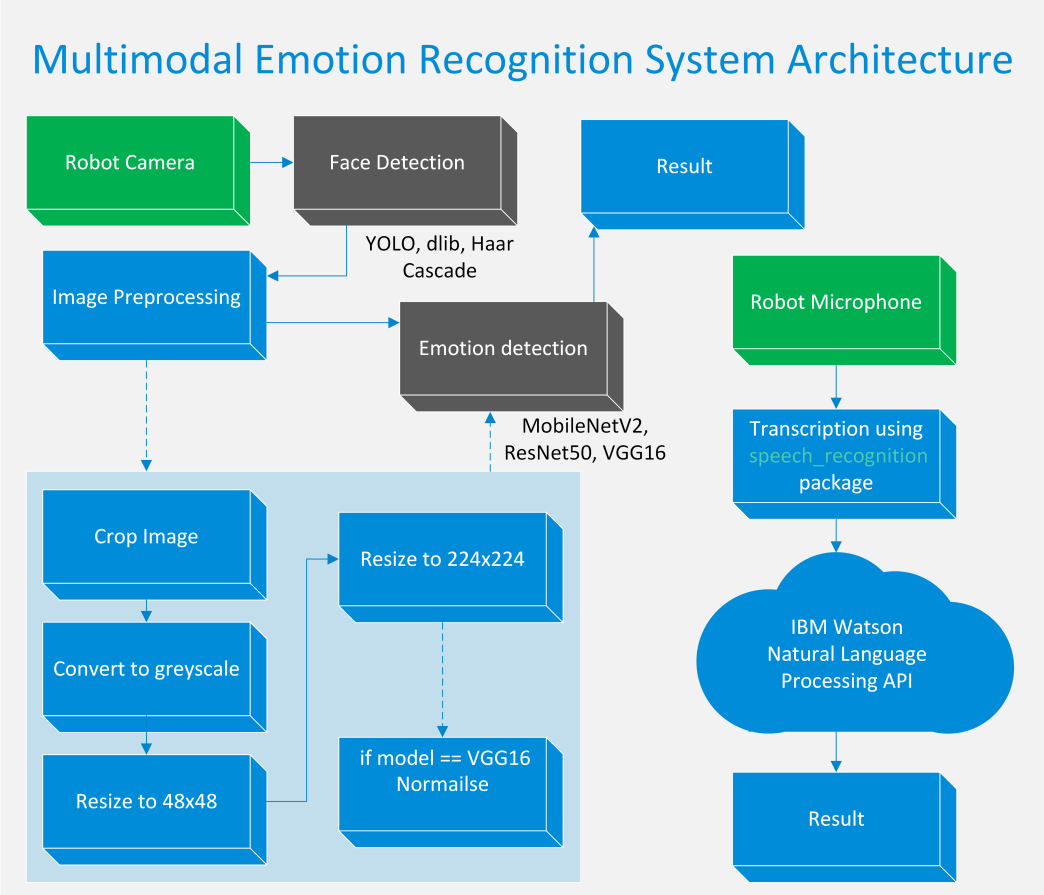


Figure 3.1: Architecture diagram showing the integration of facial and sentiment analysis components

The development process was iterative, beginning with the independent design and validation of each module using established datasets for face detection and visual emotion classification. Following individual validation, the modules were

integrated into a unified framework capable of processing real-time data from the robot's sensors. Rigorous testing was conducted using annotated datasets designed to simulate real-world conditions to assess the system's accuracy, processing speed, and overall resource utilisation.

Overall, the architectural design emphasises modularity and flexibility, allowing for individual components to be updated or replaced without impacting the entire system. This design not only streamlines development and testing but also positions the system for future enhancements and adaptations as the requirements of human-robot interaction evolve.

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.

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.

Figure 3.2: Turtlebot 4

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 [111], 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

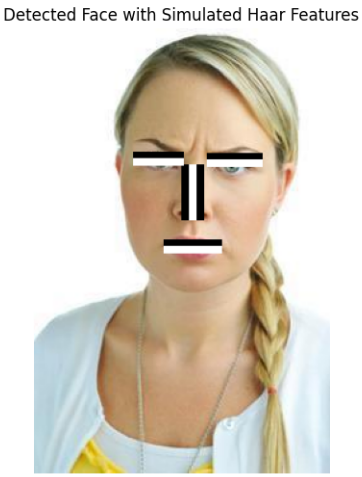


Figure 3.3: Image showing Haar-like features used in the Haar Cascade algorithm

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.

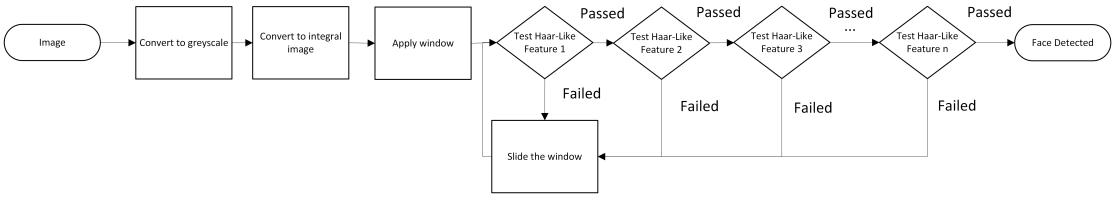


Figure 3.4: Image showing how the Haar Cascade algorithm rapidly discards regions of the image that are unlikely to contain the target object

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 24×24 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.

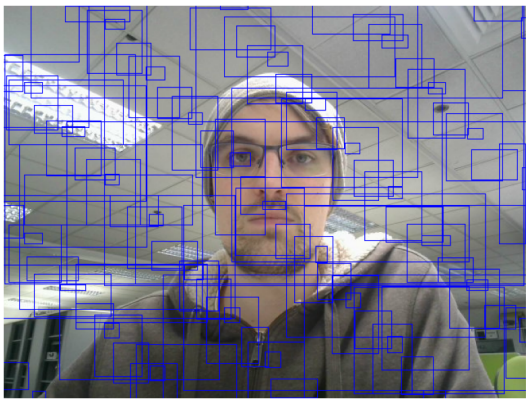


Figure 3.5: The image after being processed by the YOLO model, showing a significant amount of the bounding boxes predicted by the model, even ones with zero confidence

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. [90]

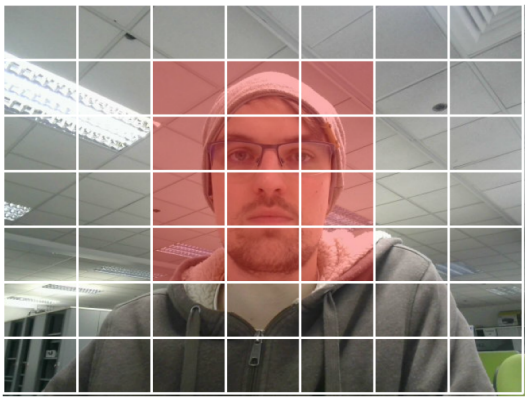


Figure 3.6: Image showing the grid cells used by the YOLO model to predict confidence scores and bounding boxes, red boxes signifies the grid cells with the highest probability of containing the object

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.

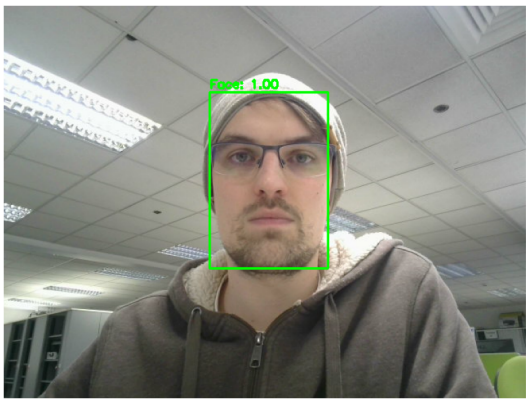


Figure 3.7: The image with the final predicted bounding boxes after applying non-maximum suppression

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 8×8 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

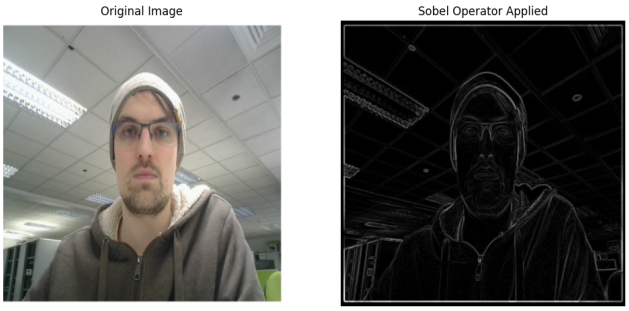


Figure 3.8: The resulting image after the Sobel operator

called blocks, for example, 2×2 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.

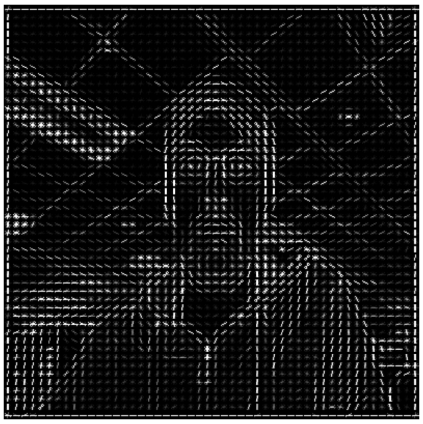


Figure 3.9: Visualisation of the HOG descriptor

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. [27]

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 involved 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, were 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.

3.3 Methods

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 (3×3) 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.

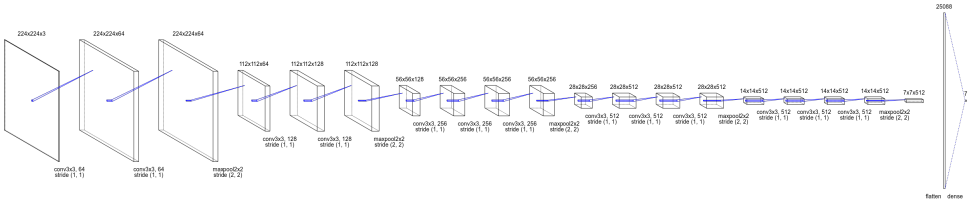


Figure 3.10: Visualisation of VGG16

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.

3.3.3 Datasets

Face Detection Dataset

The WIDER FACE dataset [115] has been meticulously curated to support research in face detection and recognition tasks. It comprises 12,878 images in the training set and 3,224 images in the validation set, sourced from a wide variety of environments. Each image is annotated with one or more bounding boxes that precisely capture the position, orientation and scale of each face, thereby accommodating the extensive variability found in real-world scenarios.

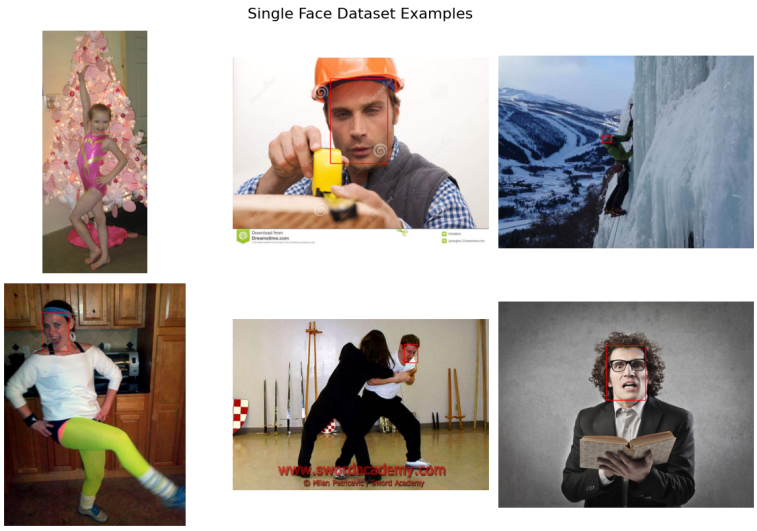


Figure 3.11: A sample of images from the single-face subset of the WIDER FACE dataset

This dataset is distinguished by its rich diversity in subjects, scenes, and environmental conditions. It features images captured in both indoor and outdoor settings, including crowded environments, street scenes, and surveillance footage. The annotations reflect a broad range of lighting conditions, facial poses, and occlusions, making the dataset highly challenging and suitable for evaluating the robustness of face detection algorithms.

For the experiments, the dataset was further subdivided into two distinct evaluation subsets. The first subset consists of images containing only one face per image (1,342 training and 334 validation images), providing a more controlled scenario with minimal distractions. The second subset comprises images with multiple faces (8,245 training and 2,104 validation images), designed to assess the model's performance

Figure 3.12: A sample of images from the multi-face subset of the WIDER FACE dataset

under more complex conditions with higher face density. Figures 3.11 and 3.12 illustrate examples from these subsets, highlighting the diversity within the dataset.

Emotion Recognition Dataset

The datasets used for training the emotion recognition models were FERPlus [14] and an adapted version of CK+ [60]. Both datasets are publicly available and contain images depicting a range of facial expressions in unconstrained environments. Sample images from these datasets are presented in Figure 3.13.

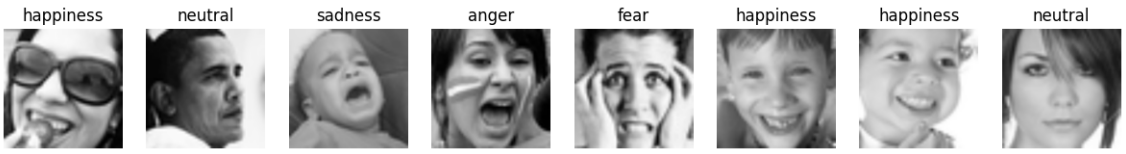


Figure 3.13: Sample of images from the combined FERPlus and CK+ dataset

FERPlus is an improved version of the FER2013 dataset, addressing issues such as mislabeled samples and non-face images that previously led to limited recognition accuracy. The dataset was re-annotated using 10 crowd-sourced labelers, categorising each image into one of ten classes: eight emotion categories (happiness, neutral, sadness, surprise, fear, disgust, contempt, and anger) and two additional categories (unknown' for indeterminate emotions and non-face' for images that do not contain

a human face). A maximum voting method was used to assign a single label to each image. The dataset consists of 28,386 training images, 3,546 private test images, and 3,553 public test images, with the distribution of emotions detailed in Table 3.1.

Table 3.1: 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 Extended Cohn-Kanade (CK+) dataset consists of 593 video sequences from 123 subjects aged 18 to 50 years. The dataset includes individuals of diverse gender and ethnicity backgrounds, with 69% female, 81% Euro-American, 13% Afro-American, and 6% from other groups. Each sequence progresses from a neutral facial expression to a peak expression, recorded at 30 frames per second at a resolution of 640×480 pixels. The dataset includes seven labeled emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise.

Table 3.2: Image counts for each emotion in CK+

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

For consistency with FERPlus, an altered version of CK+ was used, where frames were preprocessed to be grayscale and resized to 48×48 pixels. The adapted dataset was retrieved from the database sharing website Kaggle and included the following modifications:

- Contains adapted data up to 920 images from 920 original CK+ dataset
- Data is already reshaped to 48×48 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 final dataset contains the emotion distribution summarised in Table 3.2.

The Expressions in the Wild (ExpW) [119] dataset was used to evaluate the performance of the emotion recognition system. ExpW consists of 106,962 images, almost all (91,793) are annotated facial images collected from the web, covering a diverse range of real-world conditions such as varying lighting, occlusions, and head poses. Each image is labeled with one of seven emotion categories: neutral, happiness, sadness, surprise, fear, disgust, and anger. ExpW contains highly unconstrained facial expressions, making it a valuable benchmark for testing the robustness of the trained model in real-world scenarios.

Figure 3.14: A random selection of images from the ExpW dataset

A selection of random images from ExpW can be seen in 3.14.

3.3.4 Small Sentiment Dataset

To evaluate the performance of IBM Watson's text-based emotion recognition, a small ad-hoc dataset was constructed. This dataset consists of ten manually written phrases, each designed to strongly reflect a specific emotional state: joy, sadness,

anger, fear, or disgust. The phrases range from short, direct expressions to longer, more context-rich scenarios. This approach was chosen due to its flexibility and control, allowing for tailored testing of Watson's ability to identify clearly defined emotional content across varying text lengths and complexity.

The main goal of this dataset was not to replicate large-scale sentiment corpora but to conduct a focused, qualitative probe into Watson's classification behaviour. Given the service's limits on free-tier usage, a small dataset was both practical and sufficient for demonstrating trends in how Watson interprets emotional tone.

This method also enabled direct comparison between the system's predictions and a known, human-assigned ground truth. While the dataset is limited in scale, its controlled design supports clear interpretability of results and highlights how well Watson performs under idealised textual conditions.

Table 3.3: 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

Chapter 4

Results

4.1 Overview

The developed system integrates Facial Emotion Recognition (FER) and sentiment analysis to interpret user emotions, with both modalities working independently but optimised for their respective domains. The FER system combines face detection and emotion classification, both optimised to enhance accuracy and efficiency. For face detection, Tiny YOLO, YOLO, HOG+Linear SVM, and Haar Cascade were evaluated based on performance metrics including precision, speed, and robustness. While Tiny YOLO and YOLO were specifically trained to achieve the capability to detect faces, HOG+Linear SVM and Haar Cascade were utilised directly from their respective libraries, Dlib and OpenCV. A comparative analysis of these models identified the most effective solution for real-time emotion recognition tasks.

The emotion classification component used advanced models, including MobileNetV2, ResNet50, and VGG16, leveraging transfer learning techniques. These pre-trained architectures were fine-tuned using labelled emotion datasets to improve their performance for the specific task. During the development process, data augmentation techniques were applied to the dataset to enhance classification accuracy. A comparative analysis of these models highlighted the balance between computational efficiency and classification accuracy.

The sentiment analysis system, using text-based emotion recognition, utilised the cloud-based IBM Watson platform. Rather than being developed from the ground up, this component was selected for its advanced capabilities and tested to ensure seamless integration with the broader system. ChatGPT, a Large Language Model (LLM), was employed to generate speech via a synthesiser, facilitating conversational interactions between the system and users. This user speech was subsequently analysed by IBM Watson to extract sentiment information. Both components were evaluated for their response times, with IBM Watson additionally assessed for its accuracy in identifying sentiments.

4.2 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.

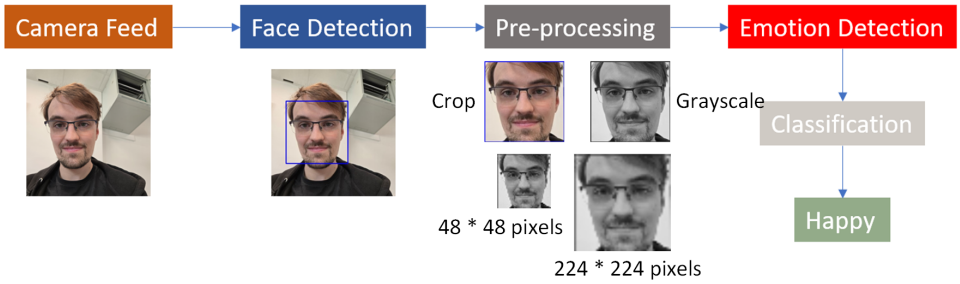


Figure 4.1: System Pipeline

4.2.1 Face Detection

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. Each [yolo] layer has a `classes` parameter, which is set to 80 by default in the cloned repository, as the dataset was originally trained on the COCO dataset. This parameter needs to be adjusted to 1 to match the single class in our dataset. Consequently, the filter settings in the [convolutional] layer preceding each [yolo] layer must also be updated. The number of filters is calculated as:

$$\text{filters} = (\text{classes} + 5) \times 3$$

Substituting classes $= 1$, the filters are set to:

$$\text{filters} = (1 + 5) \times 3 = 18$$

These changes ensure that the model predicts 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
```

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 using the WIDER Face dataset, a widely used benchmark for face detection tasks. This dataset contains a diverse range of face images, including variations in scale, pose, occlusion, and illumination, making it an effective testbed for assessing model robustness. To further analyse performance under different conditions, the dataset was split into images containing multiple faces and those containing only one face.

The evaluation employed several key performance metrics: precision, recall, F1 score, average Intersection over Union (IoU), and Average Precision (AP). Precision measures the accuracy of positive predictions, calculated as the ratio of true positives (correctly detected faces) to the sum of true positives and false positives (incorrect

detections). High precision indicates that most of the detected faces are actual faces. The inclusion of the WIDER Face dataset ensured a rigorous assessment of the face detection models, highlighting their strengths and weaknesses in varying real-world scenarios.

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

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. An example of the area of overlap can be seen in figure 4.2 and the area of union can be seen in 4.3.

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

Classification models often output a probability score indicating the likelihood that a given input belongs to a particular class. To make a definitive class prediction, this probability is compared against a predetermined threshold. For instance, in a binary classification scenario, if the threshold is set at 0.5, inputs with a probability above

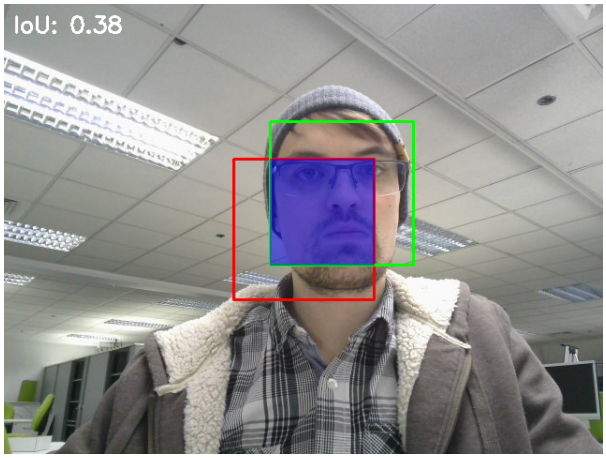


Figure 4.2: An example of IoU overlap, the red box is the predicted bounding box and the green box is the ground truth bounding box. The blue area is the overlap.

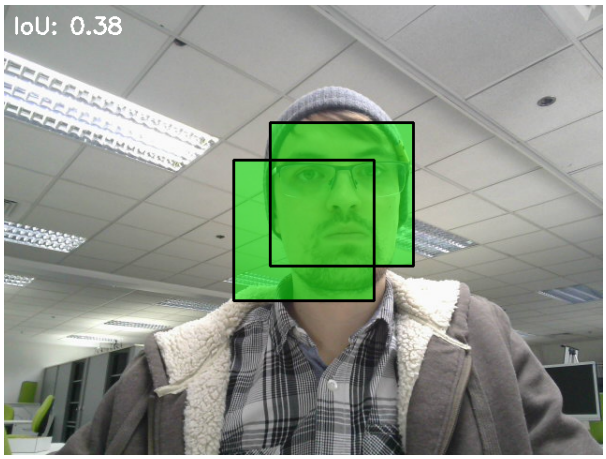


Figure 4.3: The union area of the predicted bounding box and the ground truth bounding box, represented by the total green highlighted area, used in the IoU calculation.

0.5 are classified as positive, while those below are classified as negative. Adjusting this threshold affects the model's sensitivity (true positive rate) and specificity (true negative rate), allowing practitioners to balance between false positives and false negatives based on the application's requirements.

By varying the confidence threshold and observing the resulting performance, a Receiver Operating Characteristic (ROC) curve can be constructed. This curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at different threshold levels. Specificity, or the true negative rate, measures the proportion of actual negatives correctly identified by the model and is calculated as:

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

The area under the ROC curve (AUC) quantifies the model's performance across all thresholds, with a higher AUC indicating better discrimination ability. An AUC of 0.5 suggests no discrimination (random guessing), while an AUC of 1.0 indicates perfect discrimination.

The precision-recall curve is another evaluation metric, particularly useful in scenarios with imbalanced datasets. It plots precision against recall across different threshold values. This curve helps in understanding the trade-off between precision and recall for different threshold settings, providing insights into how well the model balances between identifying positive instances and avoiding false positives.

In the context of face detection models like YOLO and Tiny YOLO, confidence scores are provided with each prediction, enabling the construction of ROC and precision-recall curves by varying the threshold and observing changes in performance metrics. However, models such as Haar Cascade and dlib's HOG+Linear SVM do not output confidence scores with their detections. This absence makes it challenging to adjust thresholds and generate the corresponding curves. Thus neither a precision-recall curve nor an ROC curve can be generated for these models.

Average Precision (AP) is a metric used to evaluate classification models, especially in imbalanced datasets. It summarises the precision-recall curve into a single value, reflecting the model's ability to balance precision and recall across thresholds. AP is calculated by first sorting the predicted scores in descending order, then computing precision and recall at each threshold. Precision measures the proportion of true positives among positive predictions, while recall shows the proportion of true positives among all actual positives.

The AUC represents the AP, which can be calculated as:

$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n$$

Where P_n and R_n are the precision and recall at the nth threshold. AP values range from 0 to 1, with higher values indicating better performance. An AP of

1.0 signifies perfect performance, while closer to 0 suggests poor performance. AP provides a comprehensive measure of a model's ability to identify positive instances while minimising false positives.

Table 4.1: Performance of YOLO on on the Wider Face dataset and the single face and multi face subsets

Metric	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 the Wider Face dataset and the single face and multi face subsets

Metric	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.

Table 4.3: Performance of Haar Cascade on the Wider Face dataset and the single face and multi face subsets

Metric	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 Wider Face dataset and the single face and multi face subsets.

Table 4.4: Performance of HOG+Linear SVM on the Wider Face dataset and the single face and multi face subsets

Metric	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%

4.2.2 Emotion Detection

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 48×48 to 224×224 pixels. This resizing is necessary because, while VGG16 and MobileNetV2 can be adjusted to accept 48×48 images, ResNet50 could not process the images at 48×48 and required a minimum size of 224×224. Resizing all images to 224×224 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 and prevent overfitting. 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 [101]. These transformations help the models generalise

better by exposing them to various image conditions and distortions they might encounter in real-world scenarios.

```
datagen = ImageDataGenerator(  width_shift_range = 0.1,
                               height_shift_range = 0.1,
                               horizontal_flip = True,
                               zoom_range = 0.2)
testgen = ImageDataGenerator(  width_shift_range = 0.1,
                               height_shift_range = 0.1,
                               horizontal_flip = True,
                               zoom_range = 0.2)
```

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.

4.2.3 Training

All of the models were obtained pre-trained on the ImageNet dataset. The base models were loaded without their fully connected layers `include_top=False`, allowing it to act as a feature extractor for the emotion recognition task.

```
base_model = tf.keras.applications.VGG16(
```

```
    input_shape=(width, height, 3),
    include_top=False,
    weights="imagenet"
)
```

Where `input_shape=(width, height, 3)` are the size of the images that the model should accept. In this case, the images were 224×224 pixels with 3 channels.

The output of all the models were flattened before being passed through a fully connected `Dense` output layer with a softmax activation function, which classified the facial expression into one of the predefined emotion categories.

```
model = base_model.output
model = Flatten()(model)
output_layer = Dense(num_of_classes,
    activation='softmax')(model)
model = Model(inputs=base_model.input, outputs=output_layer)
```

The model was compiled using categorical cross-entropy as the loss function, given the multi-class classification nature of the task, and the Adam optimizer with a learning rate of 0.0001. Accuracy was selected as the primary evaluation metric.

```
model.compile(loss='categorical_crossentropy',
              optimizer=Adam(learning_rate=0.0001),
              metrics=['accuracy'])
```

The model was trained for 50 epochs with a batch size of 64, using the augmented training dataset. Validation was performed using the augmented test dataset to assess generalisation performance. The number of steps per epoch was determined by the dataset size, ensuring that all samples were processed within each epoch.

Then the following code is used to train the model:

```
history = model.fit(train_generator,
                    epochs=50,
                    batch_size=64,
                    verbose=1,
```

```
validation_data=test_generator,
steps_per_epoch=len(X_train_scaled) // 64,
validation_steps=len(X_test_scaled) // 64)
```

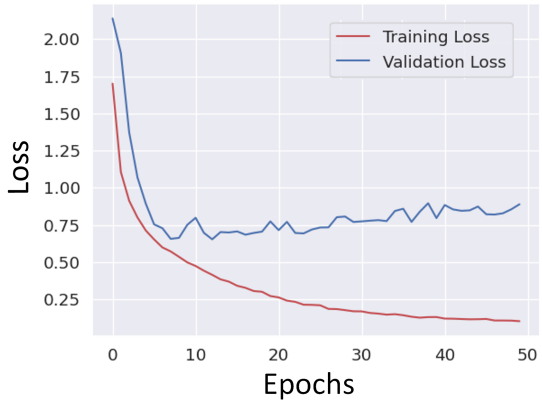


Figure 4.4: The loss graph for the first successful training of MobileNetV2

The graph 4.4 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.

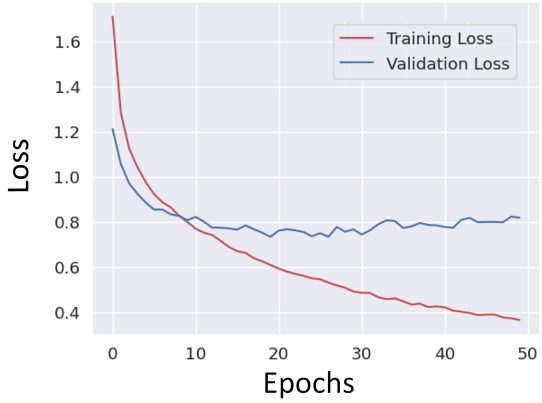


Figure 4.5: The loss graph for the first successful training of ResNet50

Graph 4.5 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.

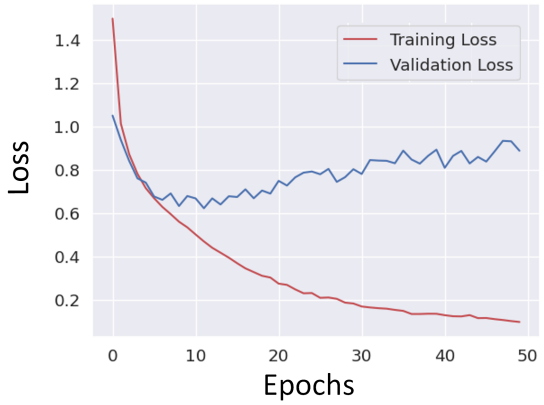


Figure 4.6: The loss graph for the first successful training of VGG16

Finally, graph 4.6 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.

```
checkpointer = ModelCheckpoint(fle_s, monitor='val_loss',
    verbose=1, save_best_only=True,
    save_weights_only=False, mode='auto')
early_stopping = EarlyStopping(monitor='val_loss',
    patience=10, restore_best_weights=True)
reduce_lr = ReduceLROnPlateau(monitor='val_loss',
    factor=0.2, patience=5, min_lr=1e-6)
```

With the updated model fit code looking like this:

```
history = model.fit(train_generator,
            epochs=50,
            batch_size=64,
            verbose=1,
            callbacks=[checkpointer, early_stopping, reduce_lr],
            validation_data=test_generator,
            steps_per_epoch=len(X_train_scaled) // 64,
            validation_steps=len(X_test_scaled) // 64)
```

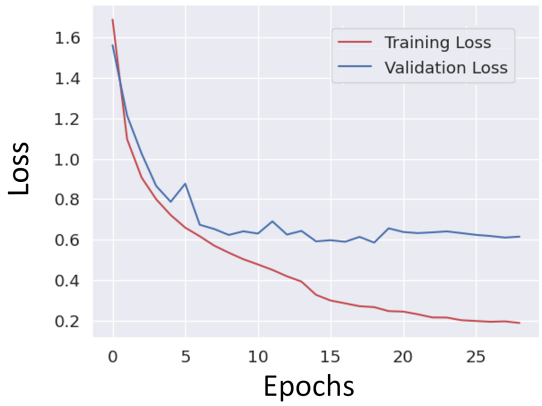


Figure 4.7: The loss graph for the second successful training of MobileNetV2

MobileNetV2s second training loss graph is shown in figure 4.7. 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.

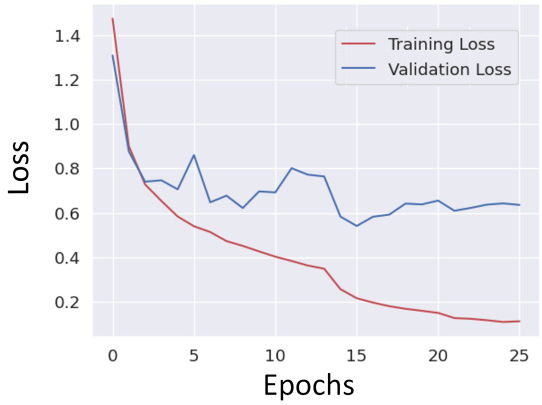


Figure 4.8: The loss graph for the second successful training of ResNet50

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.

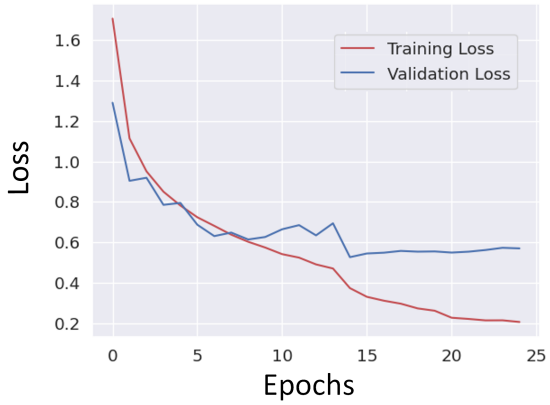


Figure 4.9: The loss graph for the second successful training of VGG16

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.

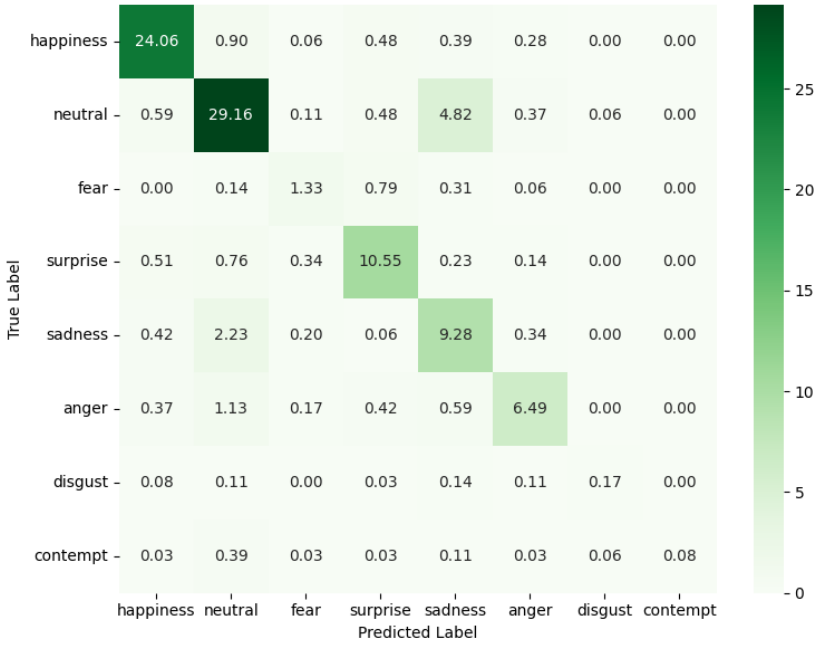


Figure 4.10: The confusion matrix detailing the performance of MobileNetV2 on the PrivateTest set

The confusion matrix 4.10 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.

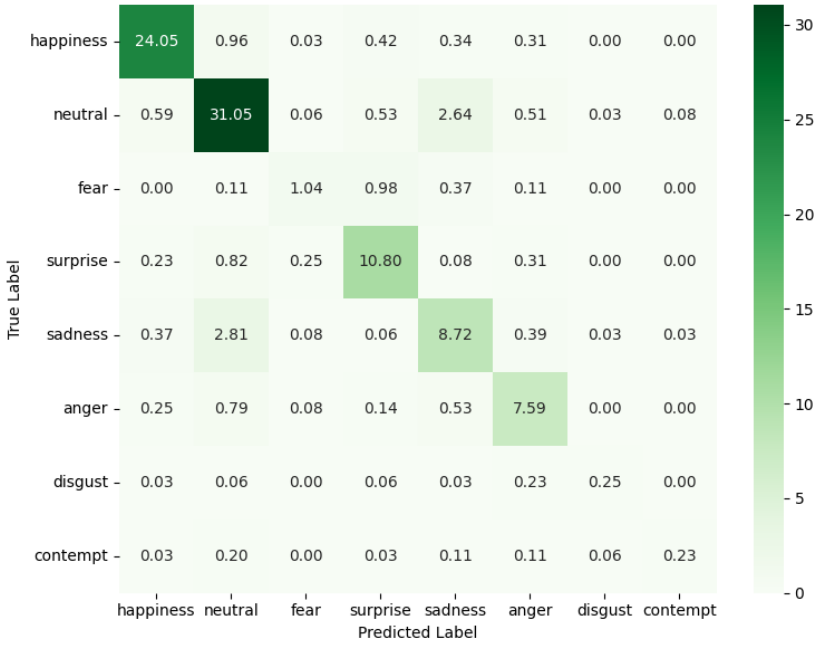


Figure 4.11: The confusion matrix detailing the performance of ResNet50 on the PrivateTest set

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.

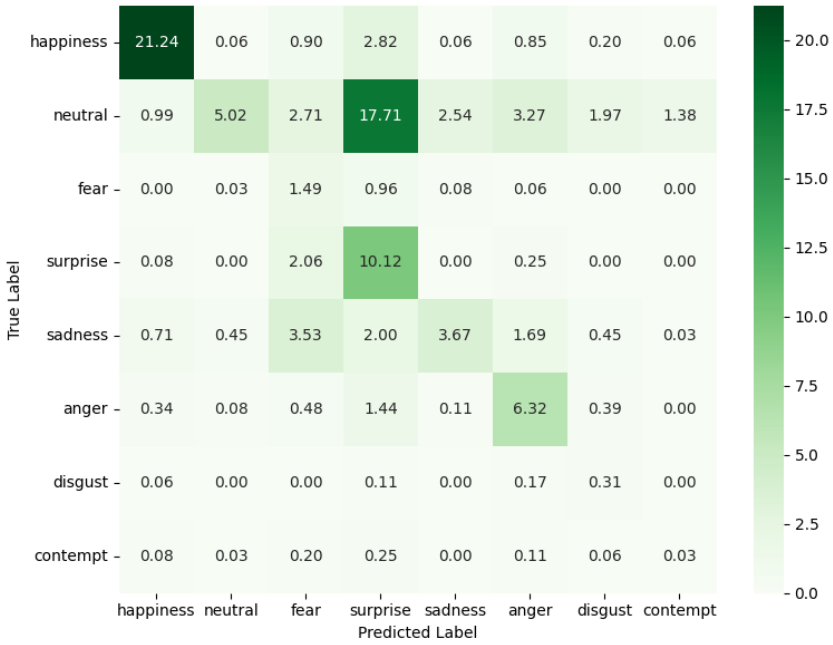


Figure 4.12: The confusion matrix detailing the performance of VGG16 on the PrivateTest set

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.

4.2.4 Testing Combined Face and Emotion Recognition

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.

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.

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.

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.

The results of the Turtlebot4 tests are summarised in the table 4.7.

Table 4.5: 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 4.6: 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

Table 4.7: 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

4.3 Sentiment Analysis

This chapter examines the sentiment analysis 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 text input. Additionally, this chapter discusses 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 text emotion analysis tool and the considerations involved in choosing appropriate technologies for text/audio analysis.

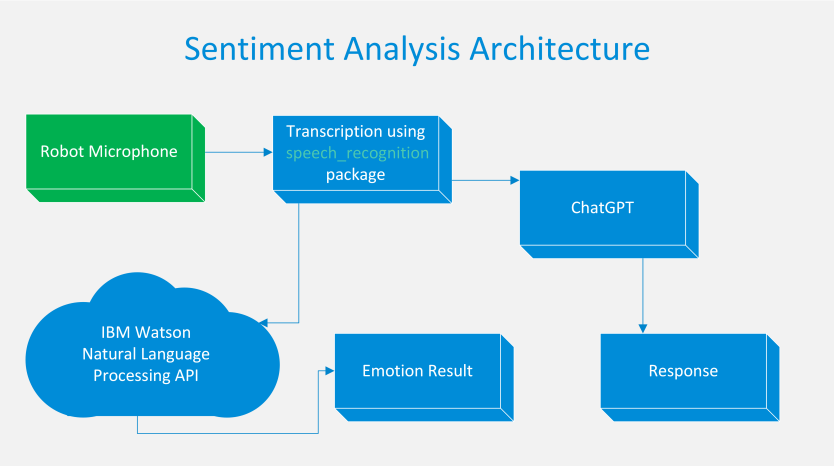


Figure 4.13: The architecture of the sentiment analysis system

4.3.1 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.

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. [53]

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

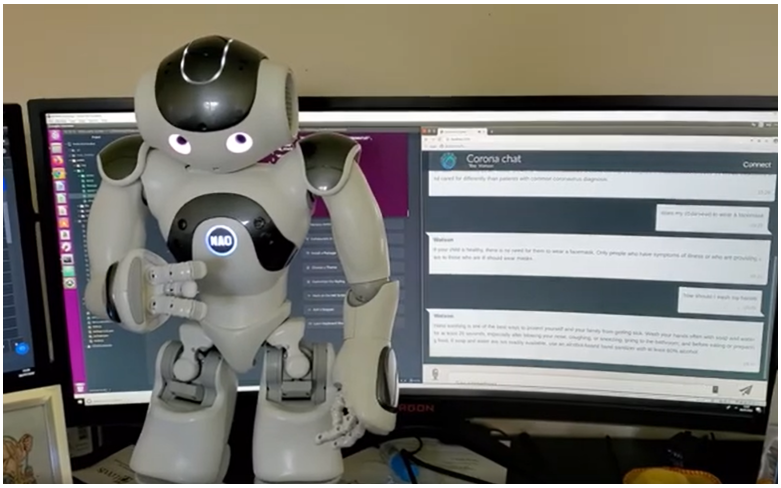


Figure 4.14: Aldebaran robot Nao with IBM Watson ChatBot

4.3.2 OpenSMILE

OpenSMILE [35], 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.

4.3.3 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 4.8 presents the response times (in seconds) for each test run with ten different phrases shown in 3.3.

Table 4.8: 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 3.3 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 4.9: The resulting output probability for each emotion for each phrase. The phrases in this table match the phrases in table 3.3 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 4.9 shows the predicted probabilities for sadness, joy, fear, disgust, and anger for each phrase in table 3.3. 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.

4.4 LLM

This section presents an evaluation of a Large Language Model (LLM) that was incorporated into the system as a conversational partner. Its purpose was to generate

natural language responses during interactions, providing the user with meaningful dialogue. These responses then served as input for IBM Watson's text-based emotion recognition system. It is important to note that the LLM itself was not responsible for interpreting or classifying emotional content, it was solely used to produce conversational material that could be analysed by other components of the system.

It is important to note that while IBM Watson does provide a conversation service, it is a structured approach to dialogue management, meaning it lacks 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 [80]. 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.

4.4.1 LLM 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.

In this system ChatGPT is only providing a way of interacting with the user to prompt speech and conversation and is not being used to analyse the speech for emotion. The role of ChatGPT in this system is strictly limited to facilitating natural and engaging dialogue with the user. It does not perform any emotion recognition or analysis on the user's speech. Instead, the emotion recognition task is handled entirely by the IBM Watson system, which processes the text input to detect emotional states.

This distinction is crucial because the tests conducted for ChatGPT focus solely on response times, rather than its ability to interpret or analyse emotional content. By separating these functionalities, the system leverages the strengths of each component: ChatGPT for conversational interaction and IBM Watson for emotion analysis.

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 4.10: 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 3.3 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

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 4.11: 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 3.3 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.

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.

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

Table 4.12: 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 3.3 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

more consistent with its response times having generally a lower standard deviation on all phrases except phrase 8.

Chapter 5

Discussion

5.1 Overview

This section critically examines the key findings, limitations, and challenges encountered throughout the project. The performance of the implemented emotion recognition models is analysed in light of the results obtained, with particular attention given to factors that may have influenced accuracy, generalisability, and reliability. In addition, practical constraints such as dataset quality, class imbalance, and system implementation limitations are explored to provide a balanced assessment of the work. Finally, potential improvements and future directions are suggested to address these issues and guide further development.

5.2 Face Detection

The comparative results from the face detection experiments provide valuable insights into the strengths and weaknesses of each model within the context of this project. Notably, the full YOLO model demonstrated superior performance in scenarios involving multiple faces, clearly outperforming its Tiny YOLO counterpart in terms of precision and robustness. However, in simpler cases where only a single face was present, the performance gap between the two models narrowed considerably, with Tiny YOLO showing only a minor drop in average Intersection over Union (IoU) and

average precision. This finding supports the hypothesis that Tiny YOLO struggles more with higher-complexity images but remains highly effective for simpler detection tasks.

Given the intended application of this system, a 1-on-1 human-robot interaction scenario, these results suggest that Tiny YOLO offers an efficient and sufficiently accurate solution for real-world deployment. Its lower computational demands make it a practical choice without a significant sacrifice in detection quality for single-face contexts.

In contrast, both the Haar Cascade and HOG+Linear SVM models exhibited a different pattern of strengths. These classical methods outperformed YOLO-based models on the more complex Full Wider Face dataset, particularly in precision. Their high precision but relatively low recall on the full dataset and multiple-face subset indicate a more conservative detection strategy, successfully avoiding false positives but at the cost of missing some true positives. This trade-off highlights their potential utility in applications where false positives must be minimised, although their lower recall limits their suitability for comprehensive face detection tasks.

5.3 Facial Emotion Recognition

A consistent challenge observed across all tested models was the frequent misclassification of sadness as neutral, as highlighted in the confusion matrices. This trend points to a significant limitation in the models' ability to differentiate between these two emotions. Figure 5.1 illustrates typical examples of sadness and neutral expressions, emphasizing the subtle visual distinctions between them. This confusion likely stems from the fact that both emotions involve minimal facial muscle movement, lacking the exaggerated features such as broad smiles or deep frowns that make other emotions more visually distinct. As a result, even high-performing models struggled to make reliable distinctions in these cases.

Figure 5.1: Example images showing the very slight variation between sadness and neutral

This finding suggests an inherent challenge in relying solely on facial cues for emotion recognition, especially when working with subtle expressions. One potential avenue to mitigate this limitation could involve augmenting training datasets with more diverse and nuanced examples of sadness and neutral expressions, ideally with high-quality labeling to capture fine-grained differences. Alternatively, leveraging the sentiment analysis capabilities of IBM Watson could provide a complementary approach to recognising these specific emotions, particularly in cases where facial cues are ambiguous or difficult to interpret.

A significant limitation encountered during this project relates to the class imbalance present in the facial emotion dataset used for training. The dataset is heavily skewed towards the Happiness (9,355 images) and Neutral (12,905 images) categories, which together account for more than 62% of the total images. This overrepresentation likely introduced bias into the training process, potentially causing the model to overfit to these dominant classes and reducing its sensitivity to less frequently represented emotions. As noted in previous research [88], such imbalance can adversely affect the model's generalisability and lead to poor performance when detecting underrepresented emotions.

In contrast, categories such as Contempt (216 images), Disgust (248 images), and Fear (819 images) were severely underrepresented. The limited number of examples in these categories means the model may not have learned the relevant features needed to recognise these emotions reliably. Even emotions like Anger (3,110 images), Sadness (4,370 images), and Surprise (4,462 images), while better represented than the minority classes, still appear in far lower quantities than Happiness and Neutral, which may have resulted in the model's reduced predictive power for these classes

To mitigate this issue in future, several strategies could be employed. One potential solution is the use of data augmentation targeted specifically at the minority classes, rather than on the entire dataset, artificially increasing their representation

by generating variations of the existing images through techniques such as rotation, flipping, cropping, or brightness adjustment. Another approach would involve resampling methods, undersampling the overrepresented ones, to achieve a more balanced training set [71]. Alternatively, class weighting [48] could have be implemented during training to penalise misclassifications in minority classes more heavily, thereby forcing the model to pay more attention to those examples. In future work, more balanced or curated datasets should be prioritised to help improve overall model performance and robustness across all emotion categories.

5.4 Combined Face and Emotion Recognition models

The results highlight key considerations when selecting face detection and emotion recognition models for real-world robotic systems. Tiny YOLO and full YOLO both provided the highest number of face detections, which directly improved the overall accuracy of emotion recognition when paired with models such as ResNet50. While full YOLO achieved the highest overall accuracy (59.02%), this came at a significant computational cost, with each face detection taking 0.1610 seconds, over six times slower than Tiny YOLO's 0.0250 seconds. The relatively small gain in accuracy (just 1.66% higher than Tiny YOLO) raises questions about whether this trade-off is worthwhile in practical applications that require real-time processing.

In terms of emotion recognition, ResNet50 consistently outperformed MobileNetV2 and VGG16 across all face detection methods. However, despite its superior accuracy, ResNet50's high inference time (95.3 milliseconds) and large model size (187MB) introduce challenges, particularly in resource-constrained environments such as mobile robots. By comparison, MobileNetV2 delivered much faster detection times (13.6 milliseconds) and a smaller model size (69.8MB), making it a compelling alternative in situations where speed and memory use are critical concerns. Based on these factors, the combination of Tiny YOLO and MobileNetV2 appeared to offer the best

overall balance of accuracy, speed, and memory efficiency, requiring just 92.2MB of memory for both models combined.

A particularly noteworthy finding emerged when these models were deployed on the Turtlebot 4 robot. Here, the performance trends observed during initial tests on the high-powered training PC did not hold: both dlib and Haar Cascade outperformed Tiny YOLO in face detection speed, despite Tiny YOLO being the fastest model during prior evaluations. This shift in performance likely stems from hardware limitations of the Turtlebot 4, which lacks a GPU and has significantly fewer CPU cores than the training machine. These constraints likely prevent Tiny YOLO from leveraging hardware acceleration and parallel processing, key advantages it relies on for fast performance. As a result, its speed advantage diminished in the robot environment.

The emotion recognition models, in contrast, remained consistent across both platforms. MobileNetV2, in particular, maintained its status as the fastest and most efficient option, reinforcing its suitability for deployment in resource-limited robotic systems.

In reflecting on these findings, it is clear that hardware context plays a decisive role in determining which models are optimal. While Tiny YOLO remains the top performer in terms of detection accuracy, its slowed performance on the Turtlebot 4 highlights a key limitation for real-time robotic applications. In scenarios where speed is a higher priority, dlib stands out as the better option despite its lower detection accuracy, while Haar Cascade offers a balanced compromise between speed and performance.

Although the models were ultimately tested in their standard forms, one way to address Tiny YOLO's diminished speed on the robot would have been to explore more hardware-optimised versions of the model, or alternative deployment strategies such as using external compute units, like the Neural Compter Stick 2 or offloading processing to a server. These options could help maintain high accuracy while alleviating on-board resource constraints, an avenue worth considering for future work.

5.4.1 Sentiment Discussion

While the use of a custom, hand-crafted dataset enabled targeted evaluation of IBM Watson's emotion recognition, this method comes with several important limitations. Firstly, the dataset lacks inter-rater reliability; all emotional labels were assigned by a single individual, which introduces subjectivity into the ground truth. In contrast, standardised emotion datasets typically rely on annotations from multiple human raters, reducing personal bias and improving validity.

Secondly, the constructed phrases may overrepresent clear or exaggerated emotional expressions that do not reflect the nuance and ambiguity found in real-world language. As such, the system's strong performance on these examples may not generalise well to less overt emotional content.

Additionally, the small scale of the dataset limits statistical significance and does not support comprehensive evaluation across diverse linguistic contexts. The use of only two examples per emotion class restricts the ability to assess variation within categories or identify borderline or mixed emotional states.

Finally, IBM Watson's emotion analysis API imposes a cap on the number of free requests, meaning large-scale experimentation is not feasible without incurring additional costs. This constraint reinforces the need for a compact, efficient dataset but also limits the scope of evaluation.

The sentiment 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 input and return relevant emotional insights ensures that it can seamlessly complement or even substitute facial emotion recognition when required.

5.5 LLM Discussion

Although the large language model (LLM) component was not responsible for any aspect of emotion recognition, it was tested as part of the system to assess how well it could support natural language interaction. The LLM served as a conversational front-end, providing responses to user inputs that were later analyzed by a separate sentiment analysis tool.

Three models were evaluated for this role, GPT-3.5-turbo, GPT-4o, and GPT-4o-mini, with a focus on response time across a fixed set of phrases. Among the three, GPT-3.5-turbo demonstrated the most consistent and responsive performance, with the lowest average response times and smallest standard deviations across all phrases. This makes it particularly suitable for systems where quick turn-taking is essential. Note, however, that every model exhibited some variability in response times, notably the variation increases as the complexity of the phrases increased. For example, Phrase 1 had an average response time of 0.944 seconds with a standard deviation of 0.067 for GPT-3.5-turbo, while Phrase 9 had an average of 3.166 seconds and a standard deviation of 0.700.

GPT-4o generally offered higher processing times, with occasional latency spikes (e.g., over 5 seconds for Phrase 5), which could interrupt the fluidity of interaction. GPT-4o-mini exhibited similar variability, particularly in later phrases, with average times exceeding six seconds in some cases. This variability may be problematic in real-time applications, particularly on resource-constrained systems.

Given the limited role of the LLM in this architecture, acting purely as a user-facing conversational agent, the results suggest that smaller, faster models like GPT-3.5-turbo are preferable, particularly where consistent, low-latency performance is more important than advanced linguistic ability. These findings help clarify the trade-offs between model complexity and practical responsiveness in a modular system.

Chapter 6

Conclusion

This thesis set out to evaluate emotion recognition methods suitable for resource-constrained robots, focusing specifically on facial emotion recognition and text-based sentiment analysis. It aimed to assess the accuracy and efficiency of different face detection and facial emotion classification models and to explore how off-the-shelf sentiment analysis tools could supplement emotion detection when visual input is limited or unavailable. Through comparative evaluation across multiple platforms, including robotic hardware, the study met its objectives by identifying optimal model combinations for real-time deployment, clarifying the trade-offs between speed and accuracy, and demonstrating how each modality can operate independently to support more flexible emotion-aware interactions.

The project contributes to the field in several key ways. First, it offers a comparative evaluation of widely used face detection and emotion recognition models in the context of robotic deployment, providing insight into the trade-offs between speed and accuracy. Second, it shows the potential of deploying facial and sentiment analysis systems in parallel to improve system robustness. The work also demonstrates that cloud-based processing can be effective in the short term, while identifying the limitations this approach introduces, especially regarding latency and real-time performance. Together, these contributions lay the groundwork for more responsive, emotionally aware human-robot interactions.

Future work will involve extending this system beyond controlled test environments. In particular, the next phase will focus on evaluating the system's performance in live interactions with human participants and in dynamic, real-world settings. This will help assess its practical effectiveness and identify usability issues or technical constraints not evident during initial testing. Further developments will also include integrating audio emotion recognition directly onto the robotic platform to reduce reliance on cloud services and improve real-time responsiveness. In addition, exploring multimodal fusion techniques, advanced facial analysis methods, and data augmentation strategies will be critical for increasing recognition accuracy and robustness across varying conditions.

In summary, this thesis has delivered on its aim of developing a functional multimodal emotion recognition system, provided meaningful contributions to the field of affective robotics, and identified several promising directions for future advancement.

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] Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., and Salichs, M. A. (2013). A multimodal emotion detection system during human-robot interaction. *Sensors (Basel)*, 13(11):15549–15581.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] Balahur, A., Hermida, J. M., Montoyo, A., and Muñoz, R. (2011). EmotiNet: A knowledge base for emotion detection in text built on the appraisal theories. In *Natural Language Processing and Information Systems*, Lecture notes in computer science, pages 27–39. Springer Berlin Heidelberg, Berlin, Heidelberg.

[14] 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)*.

[15] 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.

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

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Chuah, S. H.-W. and Yu, J. (2021). The future of service: The power of emotion in human-robot interaction. *J. Retail. Consum. Serv.*, 61(102551):102551.

[27] 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.

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

[29] Demutti, M., D'Amato, V., Oneto, L., Sgorbissa, A., and Recchiuto, C. (2022). A cloud architecture for emotion recognition in human-robot interaction based on the appraisal theory.

[30] 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).

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Filippini, C., Perpetuini, D., Cardone, D., and Merla, A. (2021). Improving human-robot interaction by enhancing NAO robot awareness of human facial expression. *Sensors (Basel)*, 21(19):6438.

[38] 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.

[39] 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.

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

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

[42] 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*.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *J. Big Data*, 6(1).

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] Marmpena, M., Lim, A., and Dahl, T. S. (2018). How does the robot feel? perception of valence and arousal in emotional body language. *Paladyn*, 9(1):168–182.

[66] Matsumoto, D. (1992). More evidence for the universality of a contempt expression. *Motiv. Emot.*, 16(4):363–368.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE.

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

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

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

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

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

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

[78] 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].

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

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

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

[83] 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.

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

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

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

[87] Ramis, S., Buades, J. M., and Perales, F. J. (2020). Using a social robot to evaluate facial expressions in the wild. *Sensors (Basel)*, 20(23):6716.

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

[89] 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.

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

[91] Reyes, M. E., Meza, I. V., and Pineda, L. A. (2019). Robotics facial expression of anger in collaborative human–robot interaction. *Int. J. Adv. Robot. Syst.*, 16(1):172988141881797.

[92] 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.

[93] 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.

[94] 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.

[95] Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.*, 39(6):1161–1178.

[96] Savchenko, A. V. (2024). HSEmotion team at the 6th ABAW competition: Facial expressions, valence-arousal and emotion intensity prediction.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

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

[102] 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.

[103] Sisbot, E. A., Marin-Urias, L. F., Broquère, X., Sidobre, D., and Alami, R. (2010). Synthesizing robot motions adapted to human presence. *Int. J. Soc. Robot.*, 2(3):329–343.

[104] 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.

[105] Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Front. Robot. AI*, 7:532279.

[106] Stock-Homburg, R. (2022). Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. *Int. J. Soc. Robot.*, 14(2):389–411.

[107] 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.

[108] Tang, C., Tang, C., Gong, S., Kwok, T. M., and Hu, Y. (2025). Robot character generation and adaptive human-robot interaction with personality shaping.

[109] Troiano, E., Oberländer, L., and Klinger, R. (2023). Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Comput. Linguist. Assoc. Comput. Linguist.*, 49(1):1–72.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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)*.

[116] 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.

[117] 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.

[118] 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.

[119] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2016). From facial expression recognition to interpersonal relation prediction.

[120] 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.

[121] 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.

[122] 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