University of Nevada, Reno

Emotion Recognition: An Integration of Different Perspectives

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Computer Science and Engineering

by

Derek Stratton

Dr. Emily Hand - Thesis Advisor
May 2022

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

DEREK STRATTON

entitled

Emotion Recognition: An Integration of Different Perspectives

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Emily M. Hand, Ph.D
Advisor

Frederick C. Harris Jr., Ph.D
Committee Member

Bethany Contreras, Ph.D
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

May, 2022

Abstract

Automatic emotion recognition describes the computational task of predicting emotion from various inputs including visual information, speech, and language. This task is rooted in principles from psychology such as the model used to categorize emotions and the definition of what constitutes an emotional expression. In both psychology and computer science, there is a plethora of different perspectives on emotion. The goal of this work is to investigate some of these perspectives about emotion recognition and discuss how these perspectives can be integrated to create better emotion recognition systems. To accomplish this, we first discuss psychological concepts including emotion theories, emotion models, and emotion perception, and how this can be used when creating automatic emotion recognition systems. We also perform emotion recognition on text, visual, and speech data from different datasets to show that emotional information can be expressed in different modalities.

Acknowledgments

I would first like to thank my advisor, Dr. Emily Hand, for her mentorship. It was with her support and guidance that I was able to create this thesis. I would also like to thank Dr. Fred Harris for his support during my education and for serving on my committee. I would like to thank Dr. Bethany Contreras for her advice on my research and for serving on my committee. Finally, I would like to thank my family, friends, and all of my instructors for all that they've done that has helped me get to where I am today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Emotions are an important part of the human experience. They are an integral part of the way we perceive the world and communicate with one another. Despite their omnipresence, there is still a lot that researchers in multiple fields aim to learn about emotions. In psychology, gaining a deeper understanding of the nature of emotions, how they are expressed, and what this means for human interaction has inspired both research and discussion. In computer science, the development of automatic emotion recognition systems has been a compelling topic with many potential applications. For example, emotion recognition can be used in assistive technologies and for creating better human computer interfaces. Deep learning has only amplified interest in this topic, with systems quickly improving at predicting emotions.

In psychology, the study of emotion is an ongoing research area with a variety of developments and ideas. A large amount of emotion recognition systems utilize the same model of emotion and often fail to consider various factors that are crucial in human emotion recognition. Understanding a diverse set of theories, models, and concepts from psychology about emotion can only help computer scientists studying the field of emotion recognition. The goal of the first contribution is to describe

psychological concepts about emotions and facial expressions and explain how these perspectives can be applied to automatic emotion recognition systems.

Emotion recognition systems have been developed in the context of different domains in the machine learning community. This problem has been addressed by different perspectives, but the three dominant modalities are computer vision, speech processing, and natural language processing. More recently, some efforts have been made to use information from multiple modalities to predict emotions, and this is known as multimodal emotion recognition. Multimodal emotion recognition is a complex problem because it sits at the cross-section of multiple fields of machine learning, and integrating them together is difficult. Emotion data can also vary widely in what modality the most useful emotion information is encoded. To handle this complexity, many approaches to multimodal emotion recognition are created from the perspective of a single domain, with the other domains being added on for increased accuracy. The goal of the second contribution of this work is to highlight the limitation of that approach by showing two multimodal emotion datasets that hold a majority of their emotion information in different domains.

In Chapter 2, we discuss background on emotion recognition in computer science. In Chapter 3, we analyze emotion perception from a psychology perspective, emphasizing the importance of context. In Chapter 4, we compare two multimodal emotion datasets and perform experiments on them to show the contextual importance of different modalities on emotion recognition accuracy. Finally, in Chapter 5, we summarize our contributions and talk about future work.

Chapter 2

Background

In computer science, emotion recognition broadly refers to the task of predicting a person's emotion based on some input data (e.g. face, body, speech, etc.) [63]. This is typically accomplished with machine learning methods, and work on this problem has increased with the advent of deep learning for a myriad of classification problems.

Emotion recognition has typically been addressed in the specific context of the modality of the input data. Three of the most common modalities included in datasets for emotion recognition are vision, speech, and natural language. While some work has been done on other inputs such as biometric indicators [61], the former three modalities are the most researched because they can be approached as specialized tasks in the context of their larger fields. They are also all easily available when recording a video of someone, while other signals often require specialized equipment to obtain. Some approaches to emotion recognition also leverage multiple input modalities to achieve better emotion recognition accuracy, and this is known as multimodal emotion recognition [95].

In this chapter, we discuss speech, text, and vision approaches for emotion recognition as distinct problems. Then, we discuss multimodal emotion recognition and

how it combines different modalities to make better emotion predictions.

2.1 Speech Emotion Recognition

The goal of speech emotion recognition is to classify speech data with emotion labels [5]. To accomplish this, relevant features are first extracted from the raw audio. Then, a classification model is used to predict emotions based on these extracted features. The quality of these features is critical to successfully predicting emotions from speech. Traditionally, prosodic and spectral features have been used in speech processing [5], but recently self-supervised learning has become the dominant approach for extracting features from speech data [105].

Prosody refers to linguistic structural properties of speech that are important in conveying both meaning and emotion [57]. Some examples of prosodic features of speech include pitch, energy, and duration. Prosodic features have the advantage of being more interpretable than spectral features, but there are challenges in reliably computing prosodic features [57].

Spectral features are based on transforming a raw audio signal to the frequency domain using a Fourier transform [5]. The most popular feature representations of speech using a spectral approach are Mel Frequency Cepstral Coefficients (MFCC). MFCCs are computed based on a Mel-Filter Bank and inverse Fourier transform [5]. MFCCs have been considered more reliable and practical than prosodic features [98]. However, both prosodic and spectral features have been recently outperformed by self-supervised learning for speech emotion recognition.

Self-supervised learning is the machine learning paradigm of learning representations of unlabeled data by leveraging the structure of the data to create pseudo-labels for training. This approach has shown to be useful in speech processing, with self-

supervised models achieving state-of-the-art results on many common speech processing tasks [105].

Self-supervised learning approaches start with pretraining on a large, unlabeled speech corpus to learn the structure of the data. Then, these models can be fine-tuned to various downstream tasks. The SUPERB benchmark was created as a leaderboard and challenge for speech processing tasks, including emotion recognition [105]. The SUPERB benchmark of the emotion recognition task uses the IEMOCAP [24] dataset, and only includes the four balanced emotion classes (neutral, happy, sad, and angry). The wav2vec model showed the power of a self-supervised approach by using a convolutional neural network (CNN) and contrastive loss function to predict future segments during speech pretraining [94]. The wav2vec 2.0 model improved on the core idea of wav2vec by using a transformer to mask parts of the input [7]. HuBERT is a similar pretraining approach that includes an offline clustering step to generate better representations [51].

2.2 Text Emotion Recognition

Emotion recognition in natural language processing (NLP) is often viewed as a branch of the sentiment analysis task [2]. Sentiment analysis is the task of determining if an opinion is positive or negative about a subject, and the data evaluated usually consists of online opinions from websites like Twitter, IMDB, or Amazon [2]. Sentiment analysis and text emotion recognition are both viewed as classification problems where text is labeled with sentiment and emotion labels, respectively.

Traditionally, emotion recognition in text used various approaches that did not rely on machine learning methods [2]. One of these is keyword recognition, which involves the construction of an emotion dictionary that is used to directly predict an emotion.

WordNet-Affect was a dictionary for text emotion recognition [99]. Another approach is lexical affinity, which assigns probabilities to keywords in an emotion dictionary. This approach generally rates words on a probabilistic scale ranging from negative to positive, while keyword recognition generally categorizes words into a set of emotion classes [2].

Feature extraction is important in NLP because it is necessary to convert text to numeric representations before applying machine learning methods [73]. The Bag of Words approach is a simple approach that creates a word vector based on the number of word occurrences. N-grams are a method that helps encode the order of words by representing text based on groups of n adjacent words. Term Frequency Inverse Document Frequency (TF-IDF), is another approach that embeds a word based on its relative frequency in a document [73]. There are also deep-learning based approaches to generating word embeddings such as Word2Vec [71] and GloVe [75] that often outperform traditional methods.

Before deep learning, classical machine learning approaches had comparable results to dictionary-based approaches for text emotion recognition. The approaches generally involved applying a feature extraction approach and applying an algorithm like naive bayes or support vector machines [73]. With the adoption of deep learning, machine learning approaches have shown to outperform traditional methods in text emotion recognition [2]. Long-Short Term Memory Networks (LSTMs) and Convolutional Neural Networks (CNNs) achieved better results than traditional machine learning approaches on this task.

The state-of-the-art approach for text emotion recognition uses transformer-based models because of their ability to encode context [3]. Specifically, Bidirectional Encoder Representations from Transformers (BERT) has become a popular base model for this task [31]. BERT learns context in text with its self-supervised learning ap-

proach, and then the BERT model can be adapted to learn a specific NLP task, which in this case is emotion recognition. BERT is pretrained by reconstructing masked words from surrounding text, which has shown to be useful in learning contextual representations.

2.3 Visual Emotion Recognition

In computer vision, most emotion recognition research has focused on predicting emotion from facial expressions, and this field is known as facial expression recognition [63]. Facial expression recognition has grown alongside the task of facial recognition, where faces are labeled with their identity. Both of these fields have greatly benefited from deep learning, and they share similar pre-training pipelines that include face alignment, data augmentation, and pose normalization [63]. Facial expression recognition tasks can broadly be categorized based on whether the inputs are static images or videos.

Static images are easier to process and have been researched more than videos [63]. The most widely used and most successful approach for predicting emotion classes from images uses Convolutional Neural Networks (CNNs) [63]. CNNs have achieved high levels of performance at extracting features for many image processing problems, including facial expression recognition. Because many facial expression datasets are relatively small for deep learning, it is common to train by starting with a model trained on a larger face recognition dataset, and fine-tuning to emotion recognition [63].

Videos are composed of a sequence of images, and this generally means that videos provide more information about the context of an emotion than a static image does. However, the structure of the data can make it more complex to fully utilize this

additional data [63]. A simple technique is to predict emotion on each frame in the video and aggregate these predictions together [58]. 3D-CNNs are CNNs with 3D convolutional kernels, and they have been applied to emotion prediction in videos [81]. While they have shown some positive results, they have also been found to struggle on fine-grained tasks like expression detection [79].

While facial expressions are a major visual indicator of emotion, body language is another visual signal that has been analyzed for emotion prediction [107]. Pose estimation is the task of determining a person's joint positions in an image at a given time [107]. [107] and [96] use pose estimations as inputs to emotion prediction models.

2.4 Multimodal Emotion Recognition

Multimodal emotion recognition is the task where multiple modalities are combined to predict emotion. Because of the nature of this task, it gives the most possible information for emotion prediction and is the closest to human emotion perception. This is challenging because it requires both good features to be extracted from different input sources and an efficient mechanism to combine these separate features. The simplest way to do this is to process the modalities independently and combine these outputs at the end to predict an emotion [95]. However, an ideal system would account for the dependencies that different modalities have on each other [95].

Research has been done on different techniques for fusing the information of different modalities to make better predictions. [66] fuses text, visual, and audio features hierarchically by combining each pair of modalities first and then fusing these 3 pairs for a final prediction. [110] fuses each pair of modalities, but also all 3 modalities so that the emotion prediction model can account for all interaction combinations between the modalities.

In many multimodal emotion recognition approaches, automatically extracted features are used in place of the raw data. This has the advantage of being simpler and providing the same set of starting features for comparing different multimodal fusion methods. The CMU-Multimodal SDK provides extracted features for various multimodal emotion datasets and tools to temporally align modalities [111].

Chapter 3

Psychology of Emotion Recognition

This chapter appears as a conference publication at the CVPR 2022 ABAW Workshop [100].

D. Stratton and E. Hand. "Bridging the gap between automated and human facial emotion perception." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 2022.

3.1 Introduction

Emotions are a core part of the human experience. From the pleasant joy of engaging in a favorite hobby to the sorrow of struggling through a difficult period in life, emotions add color to the many experiences we all have. All kinds of emotions have been invoked in art, literature, music, speech, dance, and countless other mediums throughout history [97]. Emotions can change our perception of the world and influence the actions we take every day [21]. A large focus of emotion research is the relationship between emotions and facial expressions. Facial expressions are viewed as a key to nonverbal communication and for expressing emotion [4].

Despite the omnipresence of emotion and expressions throughout life, there is a lack of scientific consensus regarding many key questions. A large number of theories and body of research on emotion emerged in the 20th and 21st centuries [45, 89]. The study of facial emotion perception in psychology is continuously developing and dynamic. This complexity can make it challenging for researchers outside the field to understand emotion science with its many nuanced and contradictory ideas.

As psychologists have increased their interest in studying emotion, so too have computer scientists, focusing on developing automated emotion perception systems [63]. Automatic emotion perception has applications in assistive technology, human-computer interaction, as well as many others [62]. Unfortunately, the nuances of human emotion perception are often lost when translated automated emotion perception. Some of this stems from the practical requirements of these automated systems (e.g. limited data) while some stems from a fundamental lack of knowledge of the psychological perspective.

The goal of this work is to align both perspectives (psychology and computer science, i.e. human and automated) of emotion perception, facilitating future discussions and research in this interdisciplinary area. It is essential that facial emotion perception be researched and discussed through an interdisciplinary lens as findings can help both fields of psychology and computer science.

The remainder of our paper is organized as follows. Section 2 details prominent emotion theories, while Section 3 details the models used to classify emotions. Section 4 focuses on facial expressions and their relationship to emotion. Section 5 provides a discussion of the psychology and computer science perspectives with suggestions for future research and Section 6 summarizes and concludes our work.

3.2 Emotion Theories

Emotion theory has intrigued philosophers, academics, and psychologists for hundreds of years [45]. Psychologists have contemplated many questions about the fundamental properties of emotion, aiming to determine what functions emotions serve, how emotions are related to memory and what separates emotions from mood and temperament [1, 47]. In this paper, we focus on one question: "Are there basic emotions?" Basic emotion theory states that there is a small set of emotions that are innate to humans. The alternative explanation – emotion construction theory – is that humans have created categories of emotion to help better understand the subject. The underlying emotion theory dictates the choice of emotion model and therefore has significant impacts on downstream research tasks.

3.2.1 Basic Emotion Theory

Basic Emotion Theory states that there exists a small set of distinct, fundamental emotions that are biologically innate to humans. The concept of a facial expression is important when describing a basic emotion because many supporters of Basic Emotion Theory directly link emotions and facial expressions. Basic emotions are described as having "distinct physiology" and "brief duration" [36]. Basic emotions are usually grouped into a small, discrete set that are thought to be hardwired into the brain [27].

Charles Darwin is often considered the inspiration for the Basic Emotion Theory, having stated that basic emotions are innate and based on specific facial and bodily expressions [45, 30]. He asserted that these emotions and emotional expressions came about due to their adaptive functions. For example, fear can facilitate the survival of the organism against a hostile attack. Darwin's perspective later influenced the work

of other psychologists [32, 101, 54].

A cross-cultural study of the Fore Tribesmen in Papua New Guinea was one of the first major studies used to provide large-scale evidence for basic emotions [39]. They found that the study subjects were able to match pictures of facial expressions to a set of 6 discrete emotion labels, supporting the idea that basic emotions exist across cultures. Analyses of the many cross-cultural studies on emotion have found both consistency in the expression and recognition of emotion across cultures, and variance that can be explained by cultural and other factors [93].

Some have shown that there is neurological evidence for basic emotions, but there is not a one-to-one mapping between a brain responses and basic emotions [26]. While there is undisputed evidence that there are some biological underpinnings for emotion, there is still much debate over what this means. Those who support Basic Emotion Theory believe that there exists some mapping between neurological responses and basic emotions. Others believe that this neurobiological evidence simply points to similar core patterns that are observed from emotion words [90].

3.2.2 Emotions as Social Constructs

Social Constructionists argue that emotions do not exist in a discrete set of biologically-innate categories, but instead that societies create emotion categories as a way to better understand affect and facilitate communication about feelings [13]. Emotion and affect are often used interchangeably. Affect is "any experience of feeling or emotion," and "both mood and emotion are considered affective states" [103]. These conceptions of emotion need not be universal, which is an important distinction from Basic Emotion Theory. Constructionists often emphasize a mix of both neurological mechanisms and social factors like context and culture to explain the variability of emotion[12].

William James introduced ideas that would later serve as the foundation for Constructed Emotion Theory, describing emotion as a product of elementary processes [67, 56, 45]. Constructed Emotion Theory gained popularity in the 21st century [34, 45].

[88] distinguishes two components of an emotion: core affect and prototypical emotion episodes. Core affect is an internal state characterized by "elemental processes of pleasure and activation," and prototypical emotion episodes are rare, complex sequences of events that match a typical emotion category like anger or fear. [90] explains that emotion categories are not clearly defined, and only some emotional episodes would fit well with a prototypical category. [11] posits that concepts like fear, anger, and sadness are social constructs used to classify these emotional episodes since people experience similar patterns of emotion [11]. [64] found the combination of core affect and conceptual knowledge to be important in the experience of fear. [10] used neuroimaging studies to show that different neural structures are activated by different emotion-related stimuli, showing evidence for complex processes for constructing emotion.

Many believe that a combination of innate and constructed emotions exist. [53] states that Basic Emotion Theory is still valid if basic emotions are restricted to emotions that are characterized by "evolutionary adaptions that are involuntarily and automatically triggered." Some take research supporting Constructed Emotion Theory as evidence for Basic Emotion Theory, acknowledging that there are more categories of emotion that have more complex relationships with the brain than previously thought [59].

The most important takeaway from emotion theory is that there is support for both innate and environmental explanations of emotion. Researchers in automated emotion perception should be aware of these theories and how they impact critical

research design choices including data, models and applications.

3.3 Emotion Models

A host of models have been created that attempt to provide a system for comparing and measuring emotions [22]. Emotion models can be divided into two broad classification approaches: categorical – representing the space of all emotions as a finite set – and dimensional representing emotions by continuous values on multiple axes [22]. These two classes of emotion models align with Basic Emotion Theory and Constructed Emotion Theory, respectively.

Choosing an emotion model is dependent on the emotion theory underlying the work and the problem the model is applied to. Categorical models consist of semantic categories, which are generally more intuitive. Dimensional models have a larger range of representations and are better at quantifying the relationship between different emotional states.

3.3.1 Categorical Emotion Models

Categorical emotion models classify emotions into distinct categories, which aligns well with Basic Emotion Theory. These models started with very few categories, but compound facial expressions have been studied to create more emotion categories [33]. Table 3.1 details various categorical emotion models.

Ekman's original emotion model contains 6 basic emotions [35]. This model was revised two more times to include a total of 15 basic emtions [41, 37]. Beyond Ekman's models, some categorical models have additional structure. Robert Plutchik describes emotions on a wheel, with opposite emotions being on opposite sides of the wheel [78]. Figure 3.1 shows Plutchik's emotion wheel, with 8 basic emotions that have

Table 3.1: Categorical Emotion Models. Emotions common to all models are shown in bold. It should also be noted that each model contains either "happiness" or "joy".

Author(s)	**Count**	**Emotions**
Ekman [35]	6	**Anger**, Disgust, **Fear**, Happiness, **Sadness**, **Surprise**
Ekman and Friesen [41]	7	**Anger**, Contempt, Disgust, **Fear**, Happiness, **Sadness**, **Surprise**
Ekman [37]	15	Amusement, **Anger**, Contempt, Contentment, Disgust, Embarrassment, Excitement, **Fear**, Guilt, Happiness, Pride in achievement, Relief, **Sadness**, Satisfaction, Sensory pleasure, Shame, **Surprise**
Plutchik [78]	8	**Anger**, Anticipation, Disgust, **Fear**, Joy, **Sadness**, **Surprise**, Trust
Parrott [74]	6	**Anger**, **Fear**, Joy, Love, **Sadness**, **Surprise**
Cowen and Keltner [29]	27	Admiration, Adoration, Aesthetic appreciation, Amusement, **Anger**, Anxiety, Awe, Awkwardness, Boredom, Calmness, Confusion, Craving, Disgust, Empathic pain, Entrancement, Excitement, **Fear**, Horror, Interest, Joy, Nostalgia, Relief, Romance, **Sadness**, Satisfaction, Sexual Desire, **Surprise**

mild forms, intense forms, and combinations [77].

Parrott created a tree structure of emotions based on finding the prototypicality of various emotion keywords rated by students [74]. The first 2 layers of the tree are shown in figure 3.2. The tree-like structure emphasizes how the differences between emotions can be distinct or subtle.

Cowen and Kelter utilized responses to emotion-eliciting videos to created a discrete list of 27 emotions, noting that emotion states have fuzzy boundaries [29]. Their emotion model is visualized in figure 3.3.

Automated emotion perception has relied heavily on categorical emotion models, and specifically Ekman's original model with 6 emotions. Using other categorical models is an obvious next step in the automated perception of more subtle emotions.

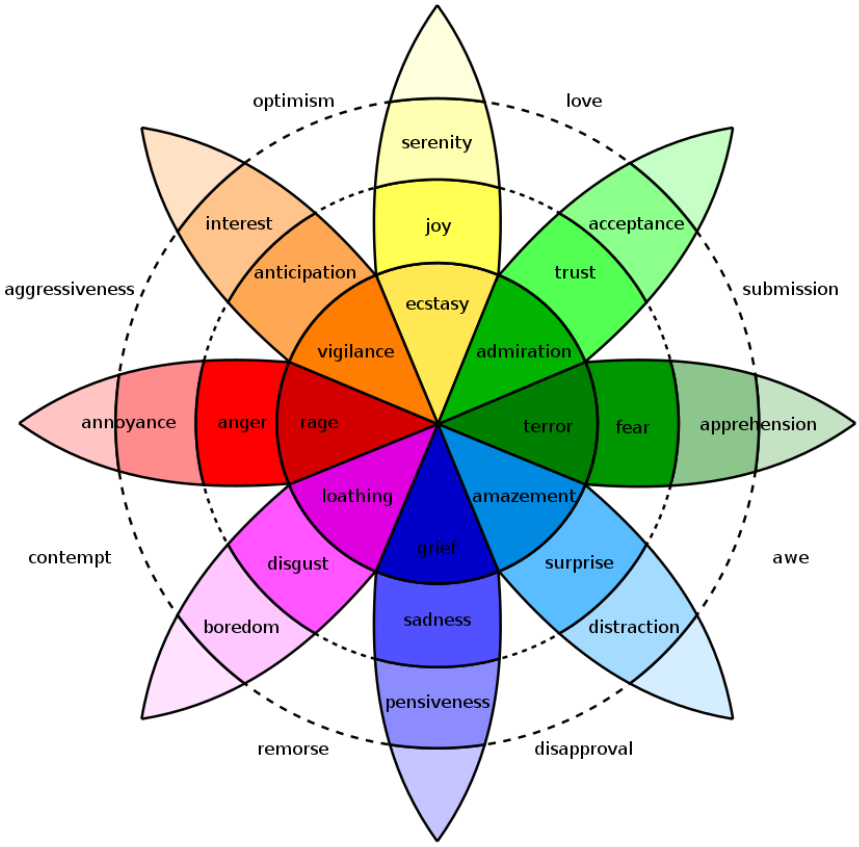

Figure 3.1: Plutchik's Wheel of Emotions shows 8 basic emotions represented by leaves on a wheel. Words closer to or farther from the center represent higher or lower intensities of the emotion, respectively. Adjacent petals represent similar emotions and opposing petals represent opposing emotions. The words between the petals describe emotions related to the adjacent petals [78]

3.3.2 Dimensional Emotion Models

Dimensional emotion models represent emotions with a set of real values scored on independent axes aligning well with Constructed Emotion Theory. Table 3.2 details common dimensional models.

James Russell introduced the first dimensional model – the Circumplex Model [87]. The Circumplex Model was constructed by plotting various emotion keywords on a circle that contained two real-valued axes: valence and arousal. Valence captures

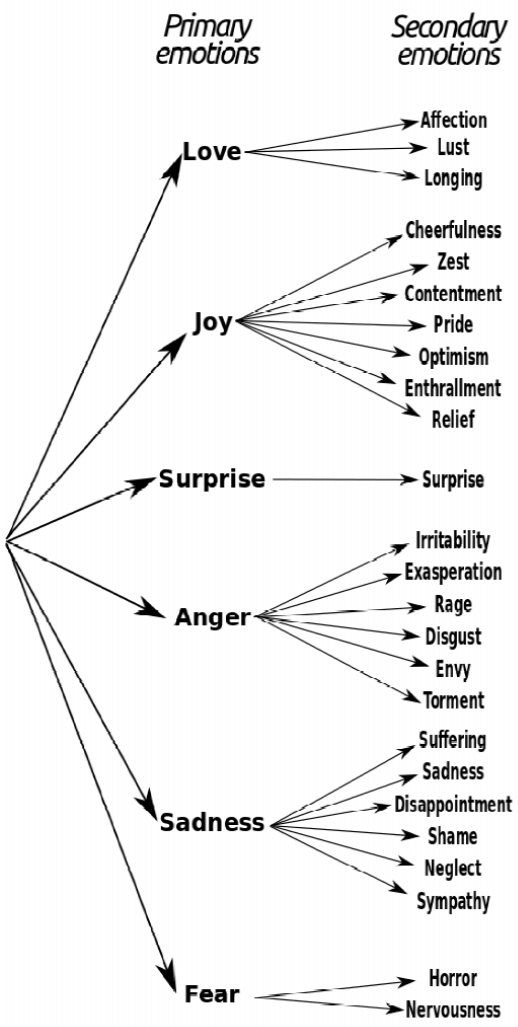

Figure 3.2: Parrott's Tree of Emotions define a list of 6 primary emotions, and various secondary emotions that stem from a primary emotion. Tertiary emotions are not shown [17].

the positivity or negativity of an affect while arousal captures the intensity of the affect [87]. Figure 3.4 shows this model, and how different emotion keywords were represented on it. The Circumplex Model has been successfully applied to other languages and cultures [91].

Watson and Tellegen introduced the Positive-Affect Negative-Affect (PANA) model for mood with two dimensions of positive and negative affect, shown in figure 3.5 [104]. The PANA model is similar to a rotated Circumplex Model.

Figure 3.3: Cowen and Kelter's Mapping of Emotional Videos plotted with t-SNE. The colors of the points represent the emotion of the video, which they also grouped into 27 categories. [29]

Table 3.2: Dimensional Emotion Models

Author(s)	**Count**	**Dimensions**
Russell [87]	2	Valence and Arousal
Watson and Tellegen [104]	2	Positive-Affect and Negative-Affect
Mehrabian [70]	3	Pleasure, Arousal, and Dominance

Mehrabian's Pleasure, Arousal and Dominance (PAD) model is a 3-dimensional emotion model [70]. The first two dimensions are similar to the Circumplex Model, but the added third dimension represents the dominance or submissiveness of an emotion. For example, fear and anger both represent emotions with a low valence and a high arousal, but fear is a more submissive emotion and anger is a more dominant

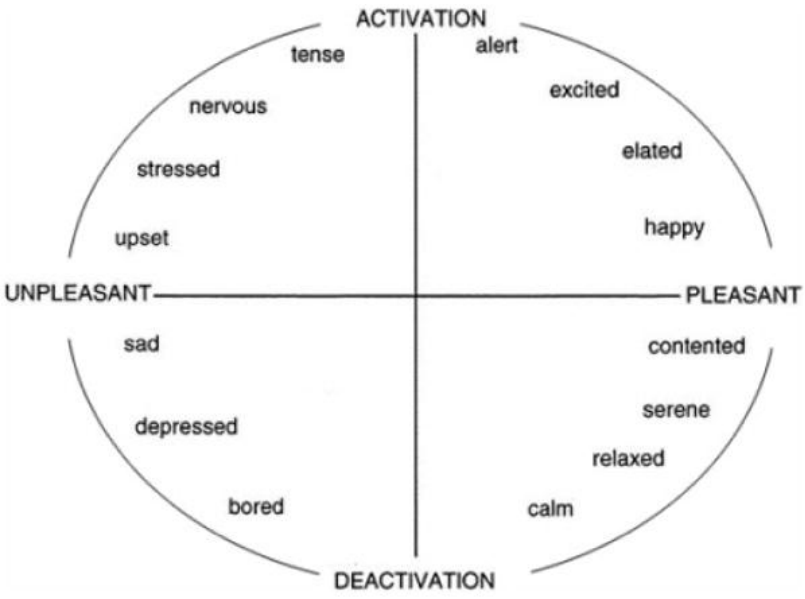

Figure 3.4: The Circumplex Model describes emotions in two dimensions: valence (x-axis) and arousal (y-axis). [83]

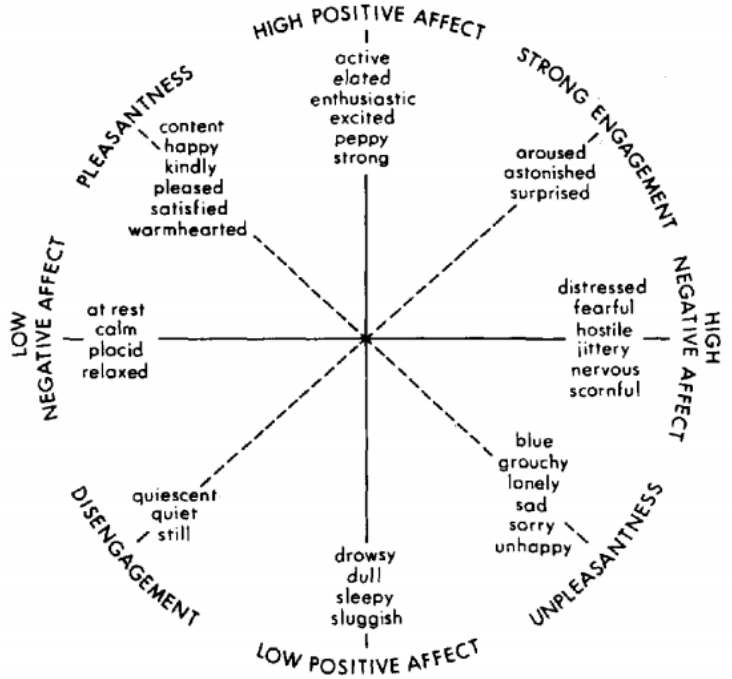

Figure 3.5: The PANA Model is a 2-dimensional model with the x-axis representing the level of negative affect and the y-axis representing the level of positive-affect.[104]

emotion [70].

All emotion models share a similar goal of providing a representation for a set of emotions. The choice of emotion model is largely a product of the specific goals of a research experiment. Categorical models can be easier to label than dimensional models because words are often more intuitive than numbers when conceptualizing emotion. Dimensional models have the advantage of being able to express subtle changes in emotion more effectively because they utilize real values. Understanding the limitations of an emotion model and experimenting with other models will lead to the development of more robust automated emotion perception systems.

3.4 Facial Expressions and Perception

Facial expressions are the configurations created by the movement of muscles in the face. They are known to play a vital role in nonverbal communication and widely believed to convey emotional information [4]. Out of 149 scientists, 80% believe that there are universal signals of emotion exhibited in the face or voice [38], but the extent to which facial expressions are linked to an underlying emotion is not fully understood. In this section we provide an overview of the relationship between facial expressions and emotions. For a more thorough survey we recommend [16].

3.4.1 Encoding the Face

Ekman and Friesen created the Facial Action Coding System (FACS) in 1978, updated in 2002, which has become one of the most ubiquitous systems for encoding facial features [102, 44]. FACS is based on mapping facial muscle movements to a set of action units that represent nearly all possible facial movements. Figure 3.6 shows examples of commonly used action units in FACS. Each of these action units are also

paired with an intensity value that represents how present any given action unit is in a face, measured on a scale of A-E, where E is the most intense. Various studies have verified the reliability of FACS measures [92, 28].

Figure 3.6: Some common Action Units in the Facial Action Coding System, along with a visual example and description. [108]

3.4.2 Posed and Spontaneous Facial Expressions

Posed facial expressions are expressions that are deliberately created, such as a smile for a photograph. Spontaneous facial expressions, on the other hand, are created in response to some stimulus, such as a smile that occurs after someone hears a funny joke. Smiles are the most well-studied expression when comparing posed versus genuine, with genuine smiles being commonly referred to as Duchenne smiles [48]. The presence of FACS action unit 6 is a distinguishing characteristic of Duchenne smiles,

as shown in figure 3.7 [18].

Figure 3.7: Comparison of Duchenne and Non-Duchenne Smiles from two people. Duchenne smiles exhibit AU6, cheek raiser, while Non-Duchenne smiles do not. The letters A-E represent the intensity of the AU. [18]

There are key differences between posed and spontaneous expressions [16]. [85] hypothesizes that posed expressions more closely resemble stereotypical ideas about facial expressions than spontaneous expressions. There is some evidence for voluntary and involuntary facial expressions being linked to different neural circuits, which may affect their expression [84, 20, 86].

Posed and spontaneous facial expressions are distinct in both their creation method and resulting muscular configuration. Spontaneous expressions have been shown to be more difficult for people to match with emotion keywords than posed expressions [72]. The use of posed expressions in emotion studies is criticized because of the

bias it can introduce towards emotions that more closely match our preconceptions of what emotion expressions should look like [16]. Despite the differences in posed and spontaneous expressions, [50] found that people do poorly at identifying an expression as posed or spontaneous, with people rating expressions as spontaneous more often than they truly are.

3.4.3 Microexpressions

Microexpressions are generally regarded as brief facial expressions that occur when people are trying to conceal true emotions [52, 40]. Duration is a key defining aspect of a microexpression. Microexpressions have a duration of 170-500ms, similar to that of a blink [106]. The other defining aspect of microexpressions is that they occur when an emotion's expression is inhibited, and there is evidence to support this behavior [43, 82]. However, microexpressions have not been found to consistently exist when someone is concealing their emotion [82].

Microexpressions have been touted for their ability to detect deception in popular culture with the 2009 TV series Lie to Me, and they have been incorporated into law enforcement training through the Wizards project [19]. However, there have been very few empirical studies of microexpressions. And results from studies in microexpression research do not provide much support for the ability of microexpressions to reveal deception.

[69] claims to find the first systematic evidence for the ability of microexpressions to differentiate liars and truthtellers [69], but [23] and [82] find that short, involuntary expressions occur in both liars and truthtellers, and that the findings are so inconsistent that conclusions should not be drawn from these involuntary expressions. Overall, the body of work on microexpressions is small and there is still more evidence needed for the creation, duration, and form of microexpressions.

3.4.4 Perception and Culture

Supporters of the universality hypothesis believe that some facial expressions are universal indicators of certain emotions, independent of other factors like culture [39]. Others argue that contextual factors play a large role in emotion perception [15].

In 1971, Ekman and Friesen performed a seminal study in New Guinea that provided evidence of a universal relationship between facial expressions and emotions [39]. The study identified a group of individuals with minimal exposure to Western culture. A translator told the group an emotional story, and the participants were asked to choose the picture which best represented the emotion in the story from a set with various facial expressions. The researchers found high percentages of respondents choosing the face that matched the intended emotion of the story. Ekman *et al.* performed a similar experiment comparing judgments of emotions and their perceived intensity across 10 cultures, and found consensus on the classification of emotion and classification of relative intensity from the facial expression picture [42].

Some argue that a forced-choice format, within-subjects test design, and use of posed expressions have biased results in favor of universality [89]. Gendron *et al.* compares the results of a remote culture and a Western culture tasked with sorting facial expression pictures into their own clusters of emotion categories, and found that the clusters did not follow a universal pattern [46]. Jack *et al.* showed animations to two groups of different cultures, and found that the mental representations associated with emotion categories differ according to culture [55].

3.4.5 Perception and Context

In addition to culture, there is also research investigating the importance of context in perceiving emotion from facial expressions. This context can include location,

background information, body, and any voice. Figure 3.8 illustrates the effect that context can have on emotion perception by showing Serena William's facial expression with and without context [15]. Without context, it is possible to perceive anger or pain from the face. But with context, it is much clearer that she is overjoyed and triumphant.

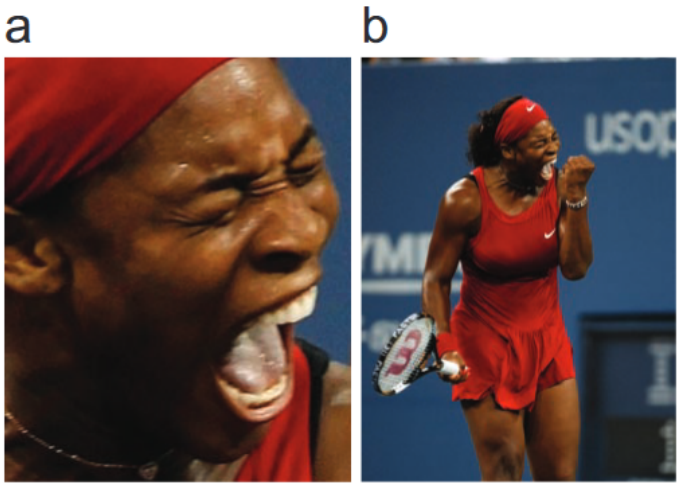

Figure 3.8: Comparison of Serena William's expression with and without context [15]: (a) without context can signal various emotions and (b) with context is very likely to signal joy.

People are extremely limited in recognizing emotion without context, highlighting the crucial role that context plays in emotion perception. [49] finds that configurations of facial muscles are ambiguous in determining emotion, and we recognize emotions in face-context combinations. [14] finds that when people are asked to recognize emotion, they better remember the context, which provides evidence that context is encoded in memory when an emotion is being perceived. The effect of context has been shown to be dependent on culture, as Japanese subjects were found to look at surrounding expressions when perceiving emotion, while Western subjects were more likely to look at an individual's expression [68]. Contextual factors are also considered to affect emotion perception on an individual level, as it has been found that each individual labeler has their own patterns when labeling emotions [6].

3.5 Discussion

It is essential that advances in automated emotion perception take into account the nuances of human emotion perception to foster a better understanding of emotion and to create more robust systems built on solid scientific foundations. This work aims to align the computational and psychology domains of emotion science to have a common ground moving forward in automated and human emotion perception. In this section, we distill our findings and provide recommendations for future research in automated emotion perception.

3.5.1 Standards for Facial Expression Recognition

There is some ambiguity in the term "facial expression" as it used in emotion recognition, often being conflated with "emotion." Instead, a standard should be used, such as the Facial Action Coding System (FACS) for facial expression recognition. This prevents confusion between the terms "expression" and "emotion," and gives the problem of "facial expression recognition" a clear evaluation standard in action units. For example, CK+ provides FACS codings for facial expressions with additional emotion labels [65].

3.5.2 Emotion Labels as Evidence not Truth

Evidence links facial expressions, body language, situational context, cultural information, and information about the observer as factors in emotion perception. The exact role and influence that each factor has on emotion perception is not currently known. This more holistic view of emotion perception brings about important research questions regarding the relationship between facial expressions and emotions such as 1) "what facial expressions result in more consensus in emotion perception?",

2) "which expressions are more ambiguous without additional information?" and 3) "do the demographic characteristics of the observer impact the perceived emotion?" We propose that future research in automated emotion perception measure uncertainty in emotion labels, allow for multiple emotion labels for a given sample and collect observer information to identify patterns in responding.

3.5.3 Incorporate Culture and Context

Research highlights the importance of culture and context in human emotion perception [49, 14]. Data used for automated emotion perception fall into one of two categories: constrained (lab created or controlled) or unconstrained (real-world). While unconstrained data is preferred for real-world applications, there are many unanswered research questions as to the role of context in both human and automated emotion perception, which leaves constrained data as the best path forward to answer these questions. A structured approach to adding context would allow for proper evaluation of context in emotion perception.

While some work has been done on multi-modal emotion perception (e.g. a combination of audio, text and visual), most work to date focuses on improving performance using a single modality or simply combining all modalities with deep neural networks without identifying the factors that contribute to a particular decision. A first step in incorporating context is to utilize these datasets to pinpoint the individual and combined effects of audio, text and visual information on human and automated emotion perception.

In order to study the effects of context it is important to systematically add context to the data. For audio and text, adding context might simply involve introducing the audio or text from time-points before the current utterance. For visual, transitioning from the most to least restrictive setting might involve 1) a tightly cropped face

making direct eye contact with the background removed, 2) adding changes in gaze, 3) adding movement, 4) adding background, 5) adding upper body, and finally 6) adding full body context. Combinations of the above contexts may also be used, resulting in a large number of settings.

Such a systematic and controlled study will require the collection of data in a lab using actors. While spontaneous expressions and unconstrained data are preferred, if we are to pinpoint the role of context in emotion perception, a certain level of precision is required. Otherwise, the field of automated emotion perception will continue on applying state-of-the-art deep learning models to the problem without any useful insights or true innovation.

With respect to the effect of culture on emotion perception, it is essential to tackle this problem with an interdisciplinary team with representatives from psychology, sociology, natural language processing and computer vision. A clear next step is to collect emotion perception data (videos and labels) from a wide variety of cultures and to maintain cultural information for both the data and the observers. With a large-scale dataset for this problem, significant analysis can be performed to begin to answer questions about the effect of culture on the presentation and perception of emotion. As with the context studies, it is essential to approach this problem systematically and in very controlled settings in order to isolate the variables of interest (i.e. the culture of the individuals presenting the emotion and of the individuals perceiving the emotion).

3.5.4 Posed vs. Spontaneous Emotion Expressions

Posed and spontaneous facial expressions are different in both appearance and in the context that they are created [16, 18]. However, humans are not very good at determining the difference between posed and spontaneous expressions [50]. The elicitation

of an expression is a key advantage of many lab-controlled datasets where that information is known. Unconstrained datasets (often collected from the internet using keyword search) do not discern between posed and spontaneous expressions, which can be problematic due to their inherent differences. We suggest that researchers avoid the use of web-collected data for the problem of emotion perception as the conditions under which the data was collected (e.g. posed vs. spontaneous, context and cultural information, etc.) are unknown.

When trying to perceive emotion, spontaneous expressions are likely to be better indicators of emotion because they are often elicited by an emotional response. Still, spontaneous expression datasets have constraints that are not present in the real world. One example of this is that many lab-controlled elicitations of emotion involve the subject starting from a neutral state. As emotion perception is essential in everyday interactions, data that represents everyday interactions (i.e. conversational dyads and groups) should be used for the study of human and automated emotion perception. We propose that future researchers collect data in controlled environments with more realistic scenarios. Specifically, using egocentric cameras to collect conversational data from the perspective of each participant will allow for the analysis of emotion elicitation and perception in real world situations without adding additional confounding factors from completely unconstrained data.

3.5.5 Subtle Expressions over Microexpressions

There is very little research investigating microexpression, despite their prevalence in popular culture [23]. Additionally, there is little evidence to support their nature, their implications, and even if they are a useful concept for human emotion perception. We recommend research in automated emotion perception explore paths that have stronger scientific foundations than microexpressions which have relatively little

empirical validation. Specifically, many automated approaches perform better with exaggerated, posed expressions than they do with subtle, spontaneous expressions [63]. We propose that future work focus on detecting the subtle expressions humans exhibit in real-world conditions rather than microexpressions. As we encourage the use of conversational scenarios for their ability to produce spontaneous expressions, we also encourage them for their ability to produce subtle expressions.

3.5.6 Interdisciplinary Approaches

It can be challenging for researchers in automated emotion perception to have a deep understanding of human emotion perception, and psychologists have been critical of work in automated emotion perception, highlighting the importance of understanding human emotion perception [16]. Automated emotion perception is a field which requires perspectives from psychology, sociology, natural language processing and computer vision at the very least. Previous automated emotion perception works that have been created with interdisciplinary perspectives tend to be more aligned with human emotion perception research than works lacking that perspective. For example, the CK+ dataset was created as a collaborative effort between researchers with both a psychology and computer science background, and they were careful to describe the limitations of emotion labels and only accept labels as valid if they met specific criteria [65]. We should pursue interdisciplinary study in this area as it will improve the quality of research being done and will advance the fields of automated and human emotion perception. As a crucial next step, we propose that interdisciplinary workshops be held to bridge the gap between research in automated and human emotion perception.

3.6 Conclusion

Emotion perception is a challenging field to study as it investigates complex human behaviors that are not fully understood. This work provides an interdisciplinary discussion of automated and human emotion perception. The goal of this work is to align both perspectives (psychology and computer science, i.e. human and automated) of emotion perception, facilitating future discussions and research in this interdisciplinary area.

Human emotion perception research is shifting away from the idea that emotions exist in a discrete, biologically-based set, and instead the concept might be constructed based on a human experience with emotion that is shaped by both natural and environmental factors. The extent to which facial expressions reflect emotion is still being studied, with the theory that they are universally linked being questioned by evidence of culture and contextual factors impacting emotion perception.

Current automated emotion perception systems can detect stereotypical, exaggerated expressions and assign an emotion from this, but tend to struggle to perceive emotion from more subtle and realistic examples of emotion. The underlying assumption that an emotion can always be perceived from a facial expression is not valid, and emotion perception systems must grow past this idea – ideally using multi-modal data – to achieve better performance in real-world scenarios.

We provide a series of recommendations for automated emotion perception based on some of the disparities we have identified between automated and human emotion perception research. In our discussion, we focus on identifying a standard for facial expressions, quantifying uncertainty in emotion labels, systematically incorporating culture and context, understanding limitations associated with posed and spontaneous emotion expressions, utilizing subtle expressions in place of microexpressions and finally approaching the problem of emotion perception as an interdisciplinary one

incorporating perspectives from psychology, sociology, natural language processing and computer vision.

It will be challenging to construct datasets and develop methods that begin to capture the complexities of human emotion perception. However, it is vital that a modern social science perspective be internalized by researchers in automated emotion perception. This perspective can greatly improve the capacity for machines to perceive emotions as we humans do naturally, and ultimately produce higher quality research than is possible without an understanding of human emotion perception.

Chapter 4

Emotion Recognition in Different Modalities

Humans perceive emotion by analyzing a combination of different indicators, including but not limited to visual cues, sound cues, and understanding the language being spoken. Multimodal emotion recognition uses multiple indicators to improve accuracy when predicting emotion. However, the relative amount of emotional information from each of these sources varies in different situations.

To illustrate this, we perform emotion recognition on the CMU MOSI and IEMOCAP datasets. For both of these datasets, we describe the different contexts of how the data was created, perform emotion recognition on the visual, audio, and text modalities independently, and discuss how the different contexts may have led to the different results.

4.1 CMU MOSI Dataset Analysis

The first multimodal dataset analyzed is CMU MOSI (Multimodal Corpus of Sentiment Intensity) [109]. This dataset was created for sentiment analysis, and it consists of people giving video reviews of movies. Within these videos, subjective sentences were manually extracted and labeled with a value from -3 to 3 based on if the segment had a positive or negative sentiment. For our experiments, we convert the labels to binary values to represent positive and negative sentiments as two classes. Figure 4.1 shows the distribution of labels used in the experiments.

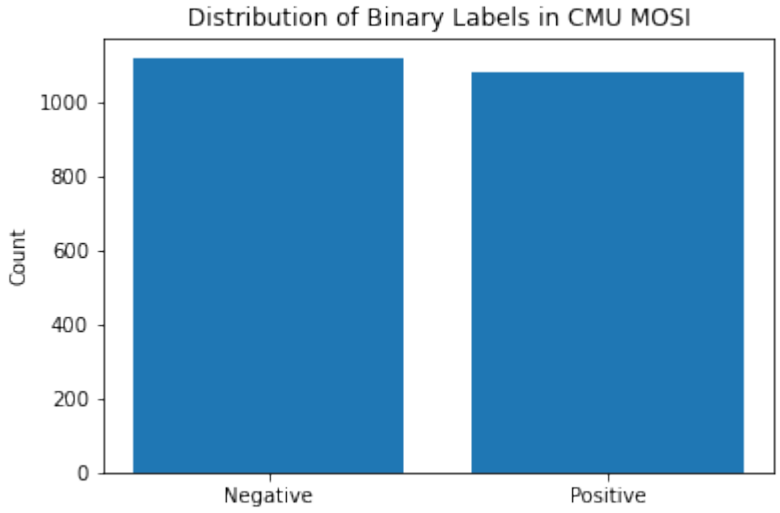

Figure 4.1: Accuracy on the evaluation dataset while training the text emotion recognition model on CMU MOSI.

It is important to note that this dataset is based on an NLP perspective, with additional modalities being used to help supplement it. This is because sentiment analysis originated as an NLP problem and the opinion segments were extracted based on the language. The content is also in the form of monologues, which contrasts with the dialogues in the IEMOCAP dataset.

CMU MOSI consists of 2,198 labeled segments, with standard splits of 1,283

training segments, 229 validation segments, and 686 test segments. In the following subsections, we discuss the experiments where I train text, speech, and vision models to predict emotion labels on CMU MOSI.

4.1.1 Text Emotion Recognition on CMU MOSI

To predict emotion from text on CMU MOSI, we start with Bidirectional Encoder Representations from Transformers (BERT) because of its rich text representations. Then, we fine tune the BERT transformer to the emotion recognition task on CMU MOSI. When training the model, we use a learning rate of 1e-5 and train for 20 epochs. Table 4.1 shows sample utterances from the dataset.

Table 4.1: Sample utterances from CMU MOSI. [109]

Utterance	**Label**
anyhow it was really good	Positive
they didnt really do a whole bunch of background info on why she has to fight and be prepared	Negative
i mean they did a little bit of it	Negative
a lot of sad parts	Negative
but it was really really awesome	Positive
he carried it	Positive

Figure 4.2 shows the learning curve of the experiment. On this experiment, we got 79.4% accuracy on the test set which is close to other works and shows that training was able to improve the prediction accuracy.

4.1.2 Speech Emotion Recognition on CMU MOSI

For speech emotion recognition on CMU MOSI, we utilize transfer learning by starting with a pretrained transformer and fine tuning it on the CMU MOSI data. The speech data is in the form of wav files, which encode an audio waveform. The pretrained

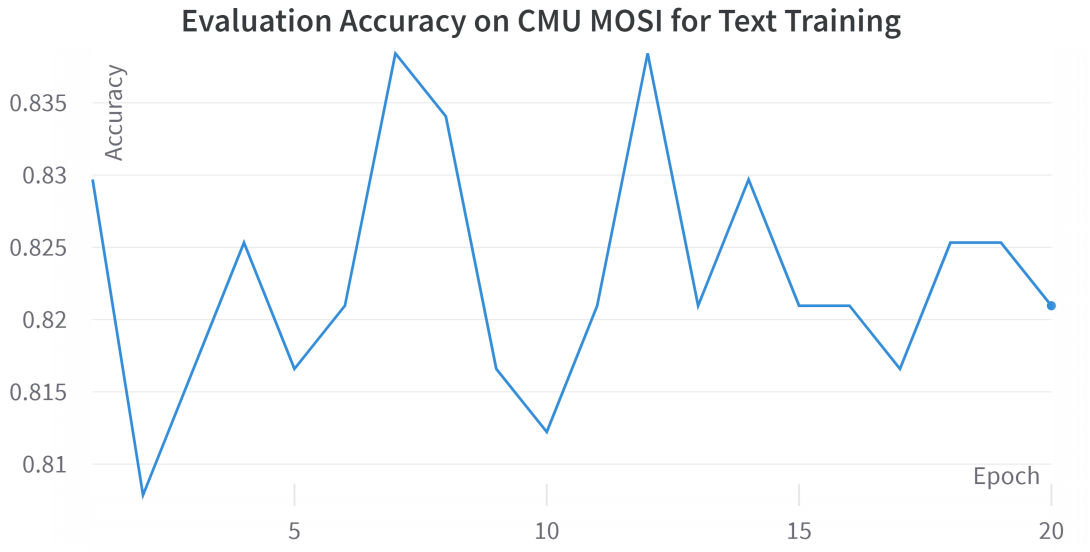

Figure 4.2: Accuracy on the evaluation dataset while training the text emotion recognition model on CMU MOSI.

transformer we start with is wav2vec 2.0, which was pretrained on a large speech corpus. We also remove segments longer than 10 seconds for training performance. We trained the model for 20 epochs with a learning rate of 1e-5.

When training the model, the performance on the evaluation dataset did not improve and the evaluation on the test dataset did not get above 50%. Figure 4.3 shows the evaluation accuracy during training. Considering this is a binary classification problem, it was unable to learn. Repeated experiments with different learning rates returned similar results. We discuss these results in Section 4.3, but the hypothesized reason for the low performance of this model is the lack of emotion expressed in speech of the speakers in the clip.

4.1.3 Visual Emotion Recognition on CMU MOSI

To predict emotion from visual data of CMU MOSI, we use visual data extracted from OpenFace [9]. OpenFace automatically extracts 427 features of facial information such

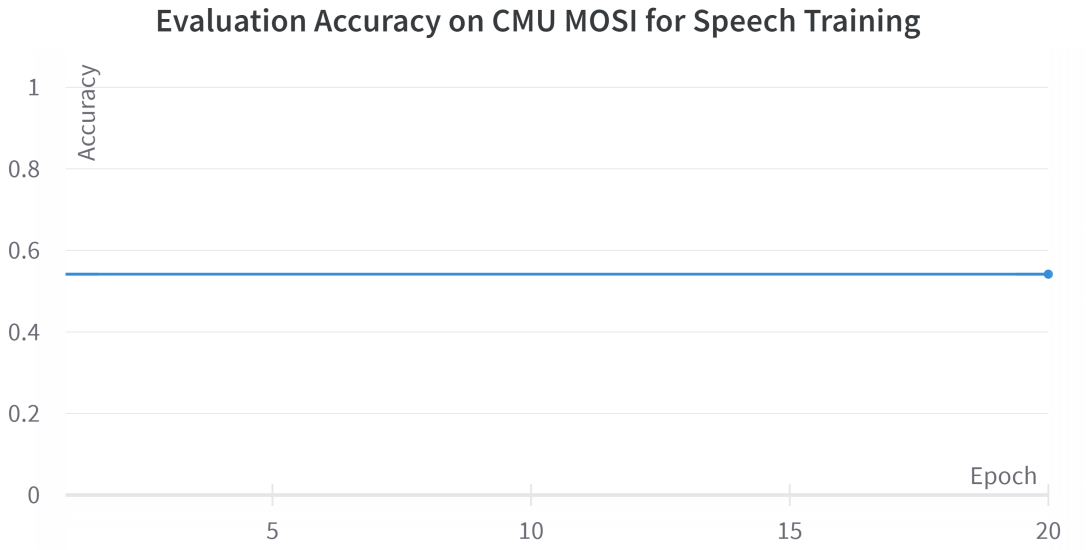

Figure 4.3: Accuracy on the evaluation dataset while training the speech emotion recognition model on CMU MOSI.

as gaze direction, action units, and facial landmarks. These features are standard and available in the CMU Multimodal SDK for both CMU MOSI and IEMOCAP [111]. Since the data consists of a single person, there is a single stream available for each utterance. Figure 4.4 shows example frames from CMU MOSI.

Figure 4.4: Sample frames from CMU MOSI [109].

For the experiment, we follow [110] and average the OpenFace features for each utterance over time and then standardize the resulting values. Then, we use a Support

Vector Machine (SVM) to predict the emotion based on these features. We fit the SVM with various parameters, and the best accuracy we obtained was 54.8% on the test set with a regularization parameter of 100. It should be noted that most experiments did not get above 50% accuracy for this binary prediction problem. The model did not perform well, which is in part due to the temporal averaging of features, but this will be analyzed in depth in Section 4.3.

4.2 IEMOCAP Dataset Analysis

The second multimodal dataset analyzed is IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) [24]. This dataset was created for emotion recognition, with all of the utterances having both categorical and dimensional emotion labels. For the following experiments, we only consider the categorical labels. And of the 7 categorical labels, we follow [105] and only utilize the balanced classes, which leave 4 labels including happy, sad, angry, and neutral. Figure 4.5 shows the distribution of the labels used for the experiments.

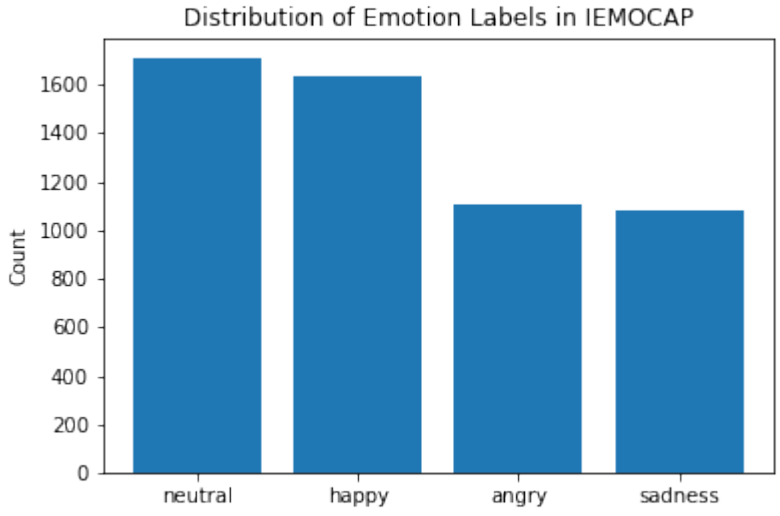

Figure 4.5: Distribution of labels used in IEMOCAP.

IEMOCAP was created to show acted dialogues that portray an emotional interaction between the actors. A variety of scenarios are performed, and they are all labeled with emotions for every utterance in the context of the dialogue. The emotion labels include 7 categorical emotions including happiness, sadness, anger, surprise, fear, disgust, and neutral. There are also dimensional labels, which are real-valued labels for valence, arousal, and dominance. IEMOCAP is largely used in speech emotion recognition, where it is the baseline emotion recognition dataset for the SUPERB speech benchmark [105].

The dataset does not come with standard splits, so I follow [60] and randomly assign the data to splits of 80% training data, 10% validation data, and 10% test data. In the following subsections, we discuss the experiments where we train text, speech, and vision models to predict emotion labels on IEMOCAP.

4.2.1 Text Emotion Recognition on IEMOCAP

To predict emotion from text on IEMOCAP, we follow the same approach as our experiment on CMU MOSI. We fine tune BERT to the IEMOCAP dataset, use a learning rate of 1e-5, and train for 20 epochs. Table 4.2 shows sample utterances from the dataset. The differences in the type of utterances from CMU MOSI is important and will be analyzed in Section 4.3.

Figure 4.6 shows the evaluation accuracy over time. The model was not able to improve on the evaluation set and did not get above 30% accuracy on the test set. These are poor results for a 4-class classification problem, and in Section 4.3 we discuss further reasons for why the model did so poorly.

Table 4.2: Sample utterances from IEMOCAP. [24]

Utterance	**Label**
What?	Angry
You seem kind of down.	Neutral
I feel like I'm always just standing here waiting. I feel like this night is going to be the night, but it never is.	Sadness
You whispered the sweetest, most intimate things to me right into my ear so I could feel them as much as hear them. And I remember thinking, this is it. You know, finally, I am as happy as I'm supposed to be.	Happy
What do you mean?	Neutral
I guess we don't need glasses.	Happy

4.2.2 Speech Emotion Recognition on IEMOCAP

For speech emotion recognition on IEMOCAP, we follow the same approach as for CMU MOSI by fine tuning the wav2vec 2.0 base model to the IEMOCAP data. We train the model for 20 epochs with a learning rate of 1e-5.

In the experiment, the trained model achieved 73.3% accuracy on the test set. Figure 4.7 shows the learning curve of the experiment. The model performed well to achieve the accuracy it did given that the problem is a 4-class classification problem and the upward trend of the evaluation accuracy during training.

4.2.3 Visual Emotion Recognition on IEMOCAP

We follow the same approach from CMU MOSI of using OpenFace extracted features for visual emotion recognition on IEMOCAP [9]. The important difference is that IEMOCAP consists of two people having a dialogue, and there are OpenFace features available for each person. For experiments on IEMOCAP, we predict the emotion based on only one face per stream to ensure that the model does not have more training data than the other experiments. Figure 4.8 shows a sample frame from IEMOCAP [24].

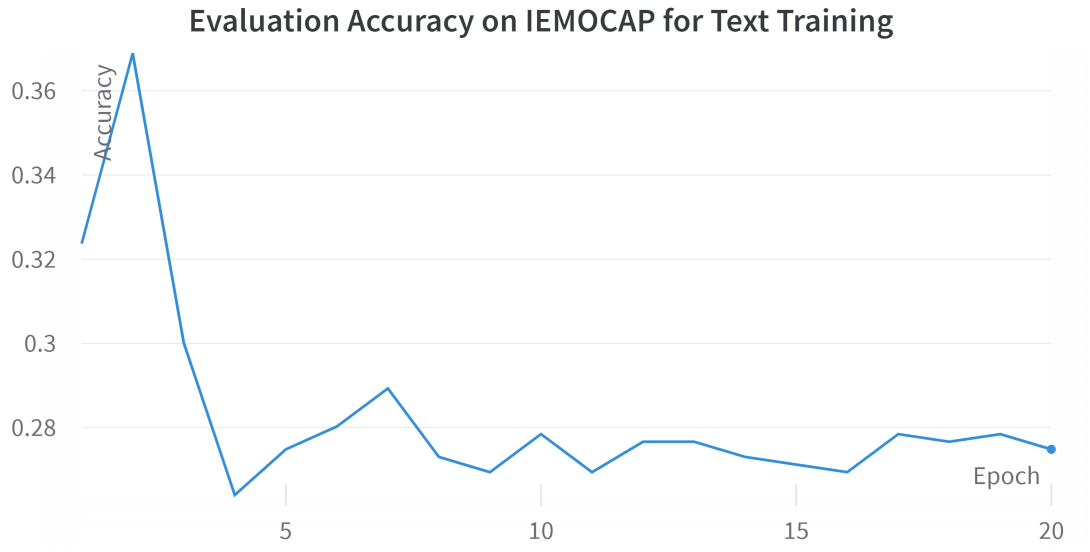

Figure 4.6: Accuracy on the evaluation dataset while training the text emotion recognition model on IEMOCAP.

After processing the data and finding the best fitting SVM as we did for CMU MOSI, the model gets an accuracy of 64.2% on the test set. This learning is significantly better than CMU MOSI, especially considering that there are 4 prediction classes rather than 2 and that these values come from only 1 of the 2 faces. The SVM would have even more data to train on if the left and right sets were combined.

4.3 Discussion

To summarize the experiments, we train an emotion prediction model for the speech, text, and vision modalities on both the CMU MOSI and IEMOCAP datasets. All of these models predict emotions for each independent utterance, without consideration of the nearby utterances for context. For speech prediction, a transformer architecture based on wav2vec 2.0 is used. For text prediction, a transformer architecture based on BERT is used. And finally for visual prediction, a SVM is trained on fea-

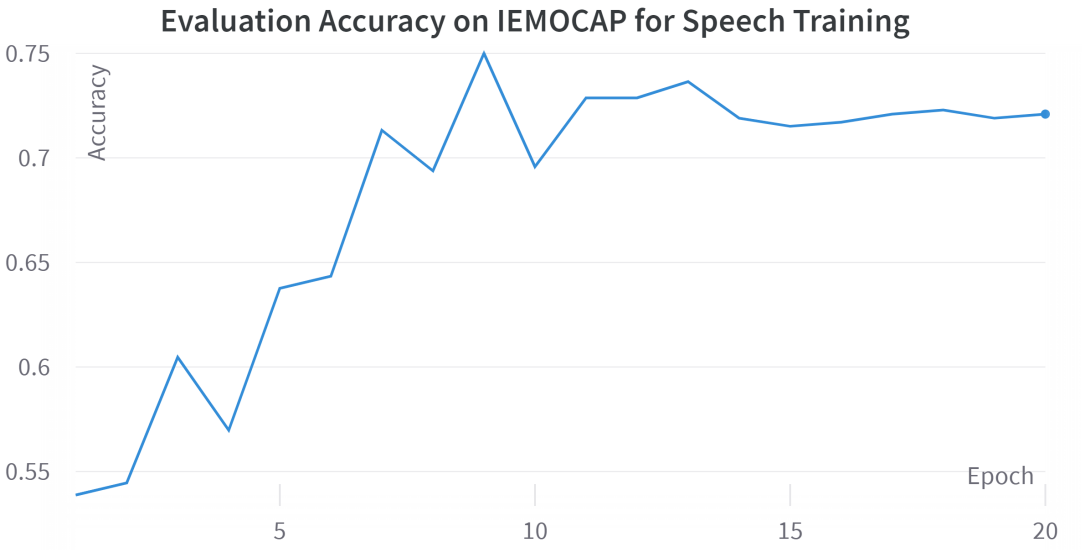

Figure 4.7: Accuracy on the evaluation dataset while training the speech emotion recognition model on IEMOCAP.

Figure 4.8: Sample frame from IEMOCAP [24]

tures extracted from OpenFace. And all of these models are trained with the same hyperparameters for each dataset.

The results show that the text model performed well on CMU MOSI but poorly on IEMOCAP. Conversely, the speech and vision models performed well on IEMOCAP but poorly on CMU MOSI. This surprising difference in relative performance may be explained by analyzing the differences in how the emotions are expressed in the two multimodal datasets. There are also potential explanations involving the nature of how each dataset is constructed, and we will discuss those as well.

The scenarios in the datasets are very different. CMU MOSI consists of online

review videos where people discuss their impressions of movies in a monologue. The context of when someone is giving a review is generally relaxed, and the subject is less personal and less likely to invoke animated emotional responses. In contrast, IEMOCAP consists of acted scenarios where emotions are expressed in a dialogue. The acted scenarios are designed to be situations that invoke large emotional responses, such as losing a driver's license [24]. These larger emotional responses are more pronounced in vision and speech, and this may factor into why the utterances in IEMOCAP were easier to classify with speech and vision models than the utterances in CMU MOSI.

It is necessary to analyze how context factors into predicting the emotion of an utterance. This is especially important when recognizing emotion in text. It can be beneficial in many cases to understand previous or future utterances when perceiving an emotion, but some cases may require context more than others. Specifically, understanding the utterances in a dialogue may require more context than a monologue because the utterances are often responses that have less independent meaning. For example, some independent utterances in IEMOCAP are fairly meaningless without context such as "No." and "What do you mean?" (see Figure 4.2). These types of contextual statements are less common in CMU MOSI.

When creating CMU MOSI, utterances were selected that were considered subjective statements, so they were more likely to be interpretable for independent predictions. On IEMOCAP, all utterances were labeled within the context of their dialogue, which also contributes to the importance in context for this text. In fact, [80] shows that emotion prediction can be done on the text on IEMOCAP if the contextual utterances are factored into the model.

For the visual models, using averaged OpenFace features will make emotion prediction more challenging than considering all features. However, this would also improve

recognition for IEMOCAP, and the point is that extracting emotion from visual data is more difficult in CMU MOSI than IEMOCAP because the visual features are more subtle.

The main takeaway of the experiments is to demonstrate that emotional indicators can appear in different places in multimodal emotion datasets. A facial expression may be a major indicator of emotion in one segment while specific word choices may be a more important factor in a different segment. While CMU MOSI and IEMOCAP consist of very different sets of scenarios, understanding how emotion expressions vary is crucial for creating robust emotion recognition systems that can accurately predict emotion in all scenarios.

The implication of these experiments for future work is to consider different emotional scenarios when training a multimodal emotion recognition model. There are various ways this can be accomplished. For example, a model may take inputs like the number of speakers or a descriptor of the type of content to help it generalize to different situations. Alternatively, the confidence of a prediction for each modality can be tracked and weighted to make a final prediction biased towards the modality with the highest confidence.

Chapter 5

Conclusion and Future Work

In this work, we explore different perspectives on emotion recognition and argue for the integration of multiple perspectives to create more robust emotion recognition systems. We discuss emotion recognition from vision, audio, text, and multimodal perspectives. Then, we analyze how psychology research on facial expressions and emotions can be applied to automatic emotion recognition. Finally, we perform emotion recognition experiments on multiple modalities to show that emotional expression can vary based on the situation.

In our first contribution, we discuss the perception of emotion from facial expressions from both psychological and computer science perspectives and discuss how these perspectives can be aligned. We describe basic emotion theory and the theory of emotions as social constructs to introduce different ideas about the nature of emotions. Then, we discuss a host of emotion models and highlight their similarities and differences so that emotion recognition researchers understand the variety of options that can be utilized.

Next, we describe the psychology of facial expressions and their perception. We discuss the Facial Action Coding System and how it is used to represent facial muscle

configurations, the differences between posed and spontaneous facial expressions, and the current state of research on microexpressions. We analyze the perception of emotion from facial expressions, talking both about the universality hypothesis and the impact of culture and context on facial emotion perception.

To conclude the first contribution, we provide a discussion of steps that can be taken to help align psychological and computer science perspectives for perceiving emotion from facial expressions. We argue for facial expressions being viewed as objective configurations rather than by their emotion, and we argue that emotion labels should be considered as evidence of an emotional state rather than a ground truth itself. We then recommend for the incorporation of context and culture into emotion prediction models to move machine perception further towards human perception. We encourage the use of spontaneous expressions because of their connection with emotional expressions, and that future work aims to analyze subtler expressions rather than strictly defined microexpressions. Finally, we advocate for interdisciplinary research in this domain to help align these perspectives in future research.

In our second contribution, we compare emotion recognition on the CMU MOSI and IEMOCAP datasets to show that different modalities have different performances across datasets. To accomplish this, we train a vision, speech, and text model for each dataset using the same set of parameters, and we compare their performances.

For the speech models, we start with a transformer model that has been pretrained on a large speech corpus and fine-tune it to perform speech emotion recognition. The model performs poorly on CMU MOSI but well on IEMOCAP. Our hypothesized reason for this discrepancy is that acted scenarios have more salient emotion information than content reviews.

For the text models, we fine-tune a transformer model that was pretrained using BERT. This model performs well on CMU MOSI but poorly on IEMOCAP, contrary

to the speech models. Some explanations for this include the language in reviews to have more emotional indicating words than in dialogues, and the lack of context when predicting emotion from utterances in a dialogue.

For the vision models, we train an SVM on facial feature extractions from OpenFace. Like the speech models, the vision models perform well on IEMOCAP but poorly on CMU MOSI. The speculated reason for this is similar, as emotional information is expressed more vividly in acted dialogues rather than in review monologues.

For future work, the first contribution can be expanded by investigating more specific impacts of factors like context and culture have on emotional expression. For example, analyses can be done on how Western and Eastern cultures express happiness, and that can be used based on the cultures of the people depicted in a dataset. The work can also be expanded by discussing methods for incorporating context and culture into emotion prediction algorithms.

To expand on the second contribution, different multimodal datasets can be explored to compare how emotion recognition is affected in more scenarios. Some potential options include CMU-MOSEI [8], MOUD [76], and MSP-IMPROV [25]. Another expansion can be done by investigating the context between utterances and how the inclusion of context affects prediction accuracy on each modality. The implications of this work can also be used to develop multimodal emotion recognition systems that account for different scenarios and variations of how emotional information is expressed.

Bibliography

[1] *The nature of emotion: Fundamental questions.* Series in affective science. Oxford University Press, New York, NY, US, 1994. ISBN 0-19-508943-X (Hardcover); 0-19-508944-8 (Paperback).

[2] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2 (7):e12189, 2020. doi: https://doi.org/10.1002/eng2.12189. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.12189.

[3] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen. Transformer models for text-based emotion detection: A review of bert-based approaches. *Artif. Intell. Rev.*, 54(8):5789–5829, dec 2021. ISSN 0269-2821. doi: 10.1007/s10462-021-09958-2. URL https://doi.org/10.1007/s10462-021-09958-2.

[4] R. Adolphs. Social cognition and the human brain. *Trends in Cognitive Sciences*, 3(12):469–479, Dec 1999. ISSN 1364-6613. doi: 10.1016/S1364-6613(99)01399-6. URL https://doi.org/10.1016/S1364-6613(99)01399-6.

[5] M. B. Akçay and K. Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2019.12.001. URL https://www.sciencedirect.com/science/article/pii/S0167639319302262.

[6] H. Aviezer, N. Ensenberg, and R. R. Hassin. The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, 17:47–54, 2017. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2017.06.006. URL https://www.sciencedirect.com/science/article/pii/S2352250X1730043X. Emotion.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[8] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208.

[9] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. doi: 10.1109/FG.2018.00019.

[10] L. Barrett and T. Wager. The structure of emotion: Evidence from neuroimaging studies. *Current Directions in Psychological Science*, 15:79–83, 04 2006. doi: 10.1111/j.0963-7214.2006.00411.x.

[11] L. F. Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46, 2006. doi: 10.1207/s15327957pspr1001_2. URL https://doi.org/10.1207/s15327957pspr1001_2. PMID: 16430327.

[12] L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 10 2016. ISSN 1749-5016. doi: 10.1093/scan/nsw154. URL https://doi.org/10.1093/scan/nsw154.

[13] L. F. Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[14] L. F. Barrett and E. A. Kensinger. Context is routinely encoded during emotion perception. *Psychological Science*, 21(4):595–599, 2010. doi: 10.1177/0956797610363547. URL https://doi.org/10.1177/0956797610363547. PMID: 20424107.

[15] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. doi: 10.1177/0963721411422522. URL https://doi.org/10.1177/0963721411422522.

[16] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. doi: 10.1177/1529100619832930. URL https://doi.org/10.1177/1529100619832930. PMID: 31313636.

[17] I. Bisio, A. Delfino, F. Lavagetto, and M. Marchese. *Opportunistic Detection Methods for Emotion-Aware Smartphone Applications*, pages 53–85. 11 2013.

ISBN 9781466646964. doi: 10.4018/978-1-4666-4695-7.ch003.

[18] Y. Bogodistov and F. Dost. Proximity begins with a smile, but which one? associating non-duchenne smiles with higher psychological distance. *Frontiers in Psychology*, 8:1374, 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.01374. URL https://www.frontiersin.org/article/10.3389/fpsyg.2017.01374.

[19] C. F. Bond Jr. and A. Uysal. On lie detection "wizards.". *Law and Human Behavior*, 31(1):109–115, 2007. doi: 10.1007/s10979-006-9016-1. URL https://doi.org/10.1007/s10979-006-9016-1.

[20] J. C. Borod, E. Koff, and B. White. Facial asymmetry in posed and spontaneous expressions of emotion. *Brain and Cognition*, 2(2):165–175, 1983. ISSN 0278-2626. doi: https://doi.org/10.1016/0278-2626(83)90006-4. URL https://www.sciencedirect.com/science/article/pii/0278262683900064.

[21] T. Brosch, K. Scherer, D. Grandjean, and D. Sander. The impact of emotion on perception, attention, memory, and decision-making. *Swiss medical weekly*, 143:0, 05 2013. doi: 10.4414/smw.2013.13786.

[22] A. F. Bulagang, N. G. Weng, J. Mountstephens, and J. Teo. A review of recent approaches for emotion classification using electrocardiography and electrodermography signals. *Informatics in Medicine Unlocked*, 20:100363, 2020. ISSN 2352-9148. doi: https://doi.org/10.1016/j.imu.2020.100363. URL https://www.sciencedirect.com/science/article/pii/S2352914820301040.

[23] J. Burgoon. Microexpressions are not the best way to catch a liar. *Frontiers in Psychology*, 9, 09 2018. doi: 10.3389/fpsyg.2018.01672.

[24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335, Nov 2008. ISSN 1574-0218. doi: 10.1007/s10579-008-9076-6. URL https://doi.org/10.1007/s10579-008-9076-6.

[25] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017. doi: 10.1109/TAFFC.2016.2515617.

[26] A. Celeghin, M. Diano, A. Bagnis, M. Viola, and M. Tamietto. Basic emotions in human neuroscience: Neuroimaging and beyond. *Frontiers in Psychology*, 8:1432, 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.01432. URL https://www.frontiersin.org/article/10.3389/fpsyg.2017.01432.

[27] A. Celeghin, M. Diano, A. Bagnis, M. Viola, and M. Tamietto. Basic emotions

in human neuroscience: Neuroimaging and beyond. *Frontiers in psychology*, 8: 1432–1432, Aug 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.01432. URL https://pubmed.ncbi.nlm.nih.gov/28883803. 28883803[pmid].

[28] J. F. Cohn, Z. Ambadar, and P. Ekman. *Observer-based measurement of facial expression with the Facial Action Coding System.*, pages 203–221. Series in affective science. Oxford University Press, New York, NY, US, 2007. ISBN 978-0-19-516915-7 (Hardcover).

[29] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1702247114. URL https://www.pnas.org/content/114/38/E7900.

[30] C. Darwin. *The Expression of the Emotions in Man and Animals*. Cambridge Library Collection - Darwin, Evolution and Genetics. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511694110.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[32] J. Dewey. The theory of emotion: I: Emotional attitudes. *Psychological Review*, 1(6):553–569, 1894. doi: 10.1037/h0069054.

[33] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1322355111. URL https://www.pnas.org/content/111/15/E1454.

[34] K. Dunlap. Are emotions teleological constructs? *The American journal of psychology*, 44(3):572–576, 1932. ISSN 0002-9556.

[35] P. Ekman. Universal facial expressions of emotion. *California Mental Health Research Digest*, 8(4):151–158, 1970.

[36] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4): 169–200, 1992. doi: 10.1080/02699939208411068. URL https://doi.org/10.1080/02699939208411068.

[37] P. Ekman. *Basic Emotions*, chapter 3, pages 45–60. John Wiley & Sons, Ltd, 1999. ISBN 9780470013496. doi: https://doi.org/10.1002/

0470013494.ch3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch3.

[38] P. Ekman. What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11(1):31–34, 2016. doi: 10.1177/1745691615596992. URL https://doi.org/10.1177/1745691615596992. PMID: 26817724.

[39] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. doi: 10.1037/h0030377. URL https://app.dimensions.ai/details/publication/pub.1011996504.

[40] P. Ekman and W. V. Friesen. Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29(3):288–298, 1974. doi: 10.1037/h0036006. URL https://doi.org/10.1037/h0036006.

[41] P. Ekman and W. V. Friesen. A new pan-cultural facial expression of emotion. *Motivation and Emotion*, 10(2):159–168, Jun 1986. ISSN 1573-6644. doi: 10.1007/BF00992253. URL https://doi.org/10.1007/BF00992253.

[42] P. Ekman, W. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Pitcairn, and P. Ricci Bitti. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53:712–7, 11 1987. doi: 10.1037/0022-3514.53.4.712.

[43] P. Ekman, E. T. Rolls, D. I. Perrett, and H. D. Ellis. Facial expressions of emotion: An old controversy and new findings [and discussion]. *Philosophical transactions. Biological sciences*, 335(1273):63–69, 1992.

[44] H. J. Ekman P, Friesen WV. *Facial Action Coding System: The Manual on CD ROM*. A Human Face, 2002.

[45] M. Gendron and L. F. Barrett. Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review : journal of the International Society for Research on Emotion*, 1(4):316–339, Oct 2009. ISSN 1754-0739. doi: 10.1177/1754073909338877. URL https://pubmed.ncbi.nlm.nih.gov/20221412. 20221412[pmid].

[46] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, 14(2):251–262, 2014. doi: 10.1037/a0036052. URL https://doi.org/10.1037/a0036052.

[47] L. Greenberg and J. Safran. Emotion in psychotherapy. *The American psychologist*, 44:19–29, 02 1989. doi: 10.1037/0003-066X.44.1.19.

[48] S. D. Gunnery and M. A. Ruben. Perceptions of duchenne and non-duchenne smiles: A meta-analysis. *Cognition and Emotion*, 30(3):501–515, 2016. doi: 10.1080/02699931.2015.1018817. URL https://doi.org/10.1080/02699931.2015.1018817. PMID: 25787714.

[49] R. R. Hassin, H. Aviezer, and S. Bentin. Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, 5(1):60–65, 2013. doi: 10.1177/1754073912451331. URL https://doi.org/10.1177/1754073912451331.

[50] U. Hess and R. E. Kleck. The cues decoders use in attempting to differentiate emotion-elicited and posed facial expressions. *European Journal of Social Psychology*, 24(3):367–381, 1994. doi: https://doi.org/10.1002/ejsp.2420240306. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420240306.

[51] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1, 10 2021. doi: 10.1109/TASLP.2021.3122291.

[52] C. M. Hurley, A. E. Anker, M. G. Frank, D. Matsumoto, and H. C. Hwang. Background factors predicting accuracy and improvement in micro expression recognition. *Motivation and Emotion*, 38(5):700–714, 2014. doi: 10.1007/s11031-014-9410-9. URL https://doi.org/10.1007/s11031-014-9410-9.

[53] D. D. Hutto, I. Robertson, and M. D. Kirchhoff. A new, better bet: Rescuing and revising basic emotion theory. *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.01217. URL https://www.frontiersin.org/article/10.3389/fpsyg.2018.01217.

[54] C. E. Izard. *Human emotions / [edited] by Carroll E. Izard.* Plenum Press New York, 1977. ISBN 0306309866.

[55] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012. doi: 10.1073/pnas.1200155109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1200155109.

[56] W. James. What is an emotion? *Mind*, 9(34):188–205, 1884. ISSN 00264423, 14602113. URL http://www.jstor.org/stable/2246769.

[57] D. Jouvet. Speech Processing and Prosody. In *TSD 2019 - 22nd International Conference of Text, Speech and Dialogue*, Ljubljana, Slovenia, Sept. 2019. URL https://hal.inria.fr/hal-02177210.

[58] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gulcehre, R. Memisevic,

P. Vincent, A. Courville, Y. Bengio, R. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, and Z. Wu. Combining modality specific deep neural networks for emotion recognition in video. pages 543–550, 12 2013. doi: 10.1145/2522848.2531745.

[59] D. Keltner, D. Sauter, J. Tracy, and A. Cowen. Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior*, 43, 06 2019. doi: 10.1007/s10919-019-00293-3.

[60] J. Kim, G. Englebienne, K. P. Truong, and V. Evers. Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning. In *Proc. Interspeech 2017*, pages 1113–1117, 2017. doi: 10.21437/Interspeech.2017-736. URL `http://dx.doi.org/10.21437/Interspeech.2017-736`.

[61] T. Kim, D. Shin, and D. Shin. Towards an emotion recognition system based on biometrics. In *2009 International Joint Conference on Computational Sciences and Optimization*, volume 1, pages 656–659, 2009. doi: 10.1109/CSO.2009.497.

[62] J. Kumari, R. Rajesh, and K. Pooja. Facial expression recognition: A survey. *Procedia Computer Science*, 58:486–491, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.08.011. URL `https://www.sciencedirect.com/science/article/pii/S1877050915021225`. Second International Symposium on Computer Vision and the Internet (VisionNet'15).

[63] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, page 1–1, 2020. ISSN 2371-9850. doi: 10.1109/taffc.2020.2981446. URL `http://dx.doi.org/10.1109/TAFFC.2020.2981446`.

[64] K. A. Lindquist and L. F. Barrett. Constructing emotion: The experience of fear as a conceptual act. *Psychological Science*, 19(9):898–903, 2008. doi: 10.1111/j.1467-9280.2008.02174.x. URL `https://doi.org/10.1111/j.1467-9280.2008.02174.x`. PMID: 18947355.

[65] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101. IEEE, 2010. ISBN 9781424470297.

[66] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161, 07 2018. doi: 10.1016/j.knosys.2018.07.041.

[67] G. Mandler. William james and the construction of emotion. *Psychological Science*, 1(3):179–180, 1990. doi: 10.1111/j.1467-9280.1990.tb00193.x. URL

https://doi.org/10.1111/j.1467-9280.1990.tb00193.x.

[68] T. Masuda, P. Ellsworth, B. Mesquita, J. Leu, S. Tanida, and E. Veerdonk. Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology*, 94:365–381, 04 2008. doi: 10.1037/0022-3514.94.3.365.

[69] D. Matsumoto and H. C. Hwang. Microexpressions differentiate truths from lies about future malicious intent. *Frontiers in Psychology*, 9:2545, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.02545. URL https://www.frontiersin.org/article/10.3389/fpsyg.2018.02545.

[70] A. Mehrabian. *Basic dimensions for a general psychological theory: implications for personality, social, environmental, and developmental studies*. Oelgeschlager, Gunn and Hain, Cambridge, 1980. ISBN 9780899460048;0899460046;.

[71] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

[72] M. T. Motley and C. T. Camden. Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Speech Communication*, 52(1):1–22, 1988. doi: 10.1080/10570318809389622. URL https://doi.org/10.1080/10570318809389622.

[73] P. Nandwani and R. Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81, Aug 2021. ISSN 1869-5469. doi: 10.1007/s13278-021-00776-6. URL https://doi.org/10.1007/s13278-021-00776-6.

[74] W. G. Parrott. *Emotions in social psychology: essential readings*. Psychology Press, Hove [England];Philadelphia;, 2001. ISBN 9780863776823;0863776833;9780863776830;0863776825;.

[75] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[76] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https:

//aclanthology.org/P13-1096.

[77] R. Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553, 1982. doi: 10.1177/053901882021004003. URL https://doi.org/10.1177/053901882021004003.

[78] R. Plutchik. The nature of emotions. *American scientist*, 89(4):344–344, 2001.

[79] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, 2016. doi: 10.1109/ICDM.2016.0055.

[80] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1081. URL https://aclanthology.org/P17-1081.

[81] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1081. URL https://aclanthology.org/P17-1081.

[82] S. Porter and L. ten Brinke. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19 (5):508–514, 2008. doi: 10.1111/j.1467-9280.2008.02116.x. URL https://doi.org/10.1111/j.1467-9280.2008.02116.x. PMID: 18466413.

[83] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. doi: 10.1017/S0954579405050340. URL https://doi.org/10.1017/S0954579405050340.

[84] W. E. Rinn. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, 95(1):52–77, 1984. doi: 10.1037/0033-2909.95.1.52. URL https://doi.org/10.1037/0033-2909.95.1.52.

[85] M. D. Robinson and G. L. Clore. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological bulletin*, 128(6):934–960, 2002. doi: 10.1037/0033-2909.128.6.934.

[86] E. D. Ross and V. K. Pulusu. Posed versus spontaneous facial expressions are modulated by opposite cerebral hemispheres. *Cortex*, 49(5): 1280–1291, 2013. ISSN 0010-9452. doi: https://doi.org/10.1016/j.cortex.2012.05.002. URL https://www.sciencedirect.com/science/article/pii/S0010945212001621.

[87] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

[88] J. Russell and L. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology*, 76:805–19, 06 1999. doi: 10.1037//0022-3514.76.5.805.

[89] J. A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, 1994. doi: 10.1037/0033-2909.115.1.102. URL https://doi.org/10.1037/0033-2909.115.1.102.

[90] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003. doi: 10.1037/0033-295X.110.1.145. URL https://doi.org/10.1037/0033-295X.110.1.145.

[91] J. A. Russell, M. Lewicka, and T. Niit. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57(5):848–856, 1989. doi: 10.1037/0022-3514.57.5.848. URL https://doi.org/10.1037/0022-3514.57.5.848.

[92] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior*, 25(3):167–185, 2001. doi: 10.1023/A:1010671109788. URL https://doi.org/10.1023/A:1010671109788.

[93] K. R. Scherer, E. Clark-Polner, and M. Mortillaro. In the eye of the beholder? universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology*, 46(6):401–435, 2011. doi: https://doi.org/10.1080/00207594.2011.626049. URL https://onlinelibrary.wiley.com/doi/abs/10.1080/00207594.2011.626049.

[94] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. pages 3465–3469, 09 2019. doi: 10.21437/Interspeech.2019-1873.

[95] N. Sebe, I. Cohen, and T. Gevers. Multimodal approaches for emotion recognition: A survey. *Proceedings of SPIE - The International Society for Optical Engineering*, 5670, 12 2004. doi: 10.1117/12.600746.

[96] Z. Shen, J. Cheng, X. Hu, and Q. Dong. Emotion recognition based on multi-view body gestures. pages 3317–3321, 09 2019. doi: 10.1109/ICIP.2019.8803460.

[97] P. Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology*, 9:342–357, 12 2005. doi: 10.1037/1089-2680.9.4.342.

[98] N. Singh, P. R. Khan, and R. S. Pandey. Mfcc and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54:9–13, 09 2012. doi: 10.5120/8529-2061.

[99] C. Strapparava and A. Valitutti. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf`.

[100] D. Stratton and E. Hand. Bridging the gap between automated and human facial emotion perception. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 2022.

[101] S. S. Tomkins. *Affect, imagery, consciousness: Vol. I. The positive affects.* Affect, imagery, consciousness: Vol. I. The positive affects. Springer, Oxford, England, 1962.

[102] E. P. . F. W. V. *The Facial Action Coding System.* Consulting Psychological Press, 1978.

[103] G. R. VandenBos. *APA Dictionary of Psychology.* American Psychological Association, 2007.

[104] D. Watson and A. Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235, 1985. doi: 10.1037/0033-2909.98.2.219. URL `https://doi.org/10.1037/0033-2909.98.2.219`.

[105] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

[106] W.-J. Yan, Q. Wu, Y.-H. Chen, J. Liang, and X. Fu. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37, 12 2013. doi: 10.1007/s10919-013-0159-8.

[107] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo. Pose-based body language recognition for emotion and psychiatric symptom interpretation. pages 294–

301, 01 2021. doi: 10.1109/ICPR48806.2021.9412591.

[108] D. B. M. Yin, S. Omar, B. A. Talip, A. Muklas, N. A. M. Norain, and A. T. Othman. Fusion of face recognition and facial expression detection for authentication: A proposed model. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, IMCOM '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348881. doi: 10.1145/3022227.3022247. URL https://doi.org/10.1145/3022227.3022247.

[109] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. 2016. doi: 10.48550/ARXIV.1606.06259. URL https://doi.org/10.48550/arXiv.1606.06259.

[110] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115.

[111] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.