UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

WILLAMS DE LIMA COSTA

Perceived emotion recognition from nonverbal communication cues in images and videos

Recife

2025

WILLAMS DE LIMA COSTA

Perceived emotion recognition from nonverbal communication cues in images and videos

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Mídia e Interação

Orientadora: Veronica Teichrieb

Coorientadores: Lucas Silva Figueiredo
Estefania Talavera Martinez

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Costa, Willams de Lima.
   Perceived emotion recognition from nonverbal communication cues in images and videos / Willams de Lima Costa. - Recife, 2024.
   124 f.: il.

   Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2024.
   Orientação: Veronica Teichrieb.
   Coorientação: Lucas Silva Figueiredo.
   Coorientação: Estefania Talavera Martinez.
   Inclui referências.

   1. Reconhecimento de emoções; 2. Reconhecimento de comportamento humano; 3. Visão computacional. I. Teichrieb, Veronica. II. Figueiredo, Lucas Silva. III. Martinez, Estefania Talavera. IV. Título.

UFPE-Biblioteca Central

To Pedro Oliveira de Lima.

You are the love of my life. Everything I have and everything I am is yours. Forever.

ACKNOWLEDGEMENTS

Ao SENHOR, Deus dos Exércitos, Deus forte e poderoso, que me sustentou e me guiou por todo este tempo. Louvado seja o nome do SENHOR para sempre.

À minha esposa Isabel pelo grande apoio, motivação e compreensão nesta jornada. Foram momentos difíceis, noites longas, mentes cheias, mas você nunca saiu do meu lado. Ao meu filho Pedro, que na data da defesa desta tese, terá apenas 6 meses de vida, mas que me foi por motivação durante todos estes 4 anos. À minha mãe, que me ensinou que a educação é a base e pelo apoio que sempre me deu. Aos meus amigos-irmãos que sempre me apoiaram, perseverando em oração, especialmente Gustavo e Karyta, Bruno e Candinha, e tantos outros que estiveram comigo neste período.

Agradeço profundamente ao Voxar Labs e seus membros que me ajudaram a pavimentar este caminho, especialmente Maria, pois sem Maria estaríamos todos loucos, e também aos membros do EmotionRAM: Malu, Renato e Isabela, por todo o apoio na pesquisa. Foi um grande prazer!

Agradeço também aos meus orientadores e mentores. Como já dizia Isaac Newton, "*se eu vi mais longe, foi por estar sobre ombros de gigantes*". Agradeço à Prof. Veronica Teichrieb por todo o apoio nestes anos e também pelas imensas palavras de motivação. Ao Prof. Lucas Figueiredo, que desde o final da graduação me apoiou fortemente e me ajudou a direcionar minha carreira acadêmica.

A la Prof. Estefanía Talavera - lamentablemente, no nos conocíamos antes, pero estoy muy agradecido de que eventualmente haya sucedido. Gracias por todas las mentorías, consejos y apoyo durante este período. Mi carrera académica fue y es considerablemente mejor gracias a su aporte y colaboración, y diría que nuestros momentos de trabajo fueron los puntos más altos de aprendizaje y desarrollo durante mi doctorado. También gracias a usted, hoy formo parte de una comunidad que me ayuda a desarrollarme cada vez más. Estaré agradecido por toda mi vida.

I would also like to say thanks to the whole LatinX in AI community. Being able to meet and interact with everyone was an awesome experience. This community truly has opened doors for me, and hopefully, it will open many more doors for others in the future.

No geral, agradeço a quem esteve ao meu lado durante esta caminhada. Por quem procurou me entender nas diversas situações que eu passei. Essa vitória é nossa.

"The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions." (MINSKY, 1988, p. 163)

RESUMO

Identificar emoções permite que sistemas inteligentes monitorem o comportamento de usuários, levando a uma compreensão mais profunda da pessoa. A percepção emocional é um processo que ocorre naturalmente em humanos por meio da comunicação de sinais não verbais, nos quais características emocionais são comunicadas implicitamente por meio de múltiplos canais. Nesta tese, propomos três técnicas que são respaldadas por evidências da literatura de psicologia de comportamento para o reconhecimento de emoções: (1) uma abordagem de reconhecimento de emoções baseada exclusivamente no contexto situacional, (2) um modelo de linguagem corporal que utiliza características de marcha para prever emoções a partir de estilos de caminhada em vídeos, e (3) um modelo que recebe múltiplos sinais extraídos de expressões faciais, contexto situacional e linguagem corporal para perceber emoções em imagens. Os resultados obtidos por nossos modelos se igualam ao estado da arte, mas com melhorias significativas relacionadas ao custo computacional.

Palavras-chaves: Reconhecimento de emoções. Reconhecimento de comportamento humano. Visão computacional. Comunicação não verbal.

ABSTRACT

Identifying emotions enables intelligent systems to monitor user behavior, leading to a deeper understanding of the person. Perceiving emotion occurs naturally in humans through the communication of nonverbal cues, in which emotional features are communicated implicitly through multiple channels. In this thesis, we propose three automatic frameworks supported by evidence from the behavioral psychology literature for emotion recognition: (1) an emotion recognition approach based solely on situational context, (2) a body-language model that uses gait features to predict emotion from walking styles from videos, and (3) a multi-cue model that combines facial expression, situational context, and body language to perceive emotions in images. The obtained results by our proposed models equal the state of the art but with severe improvements related to computational cost.

Keywords: Emotion recognition. Human behavior recognition. Deep learning. Computer vision. Nonverbal communication.

LIST OF ABBREVIATIONS AND ACRONYMS

ACC	Anterior Cingulate Cortex
AfeW	Acted Facial Expressions in the Wild Dataset
AI	Artificial Intelligence
AMT	Amazon Mechanical Turk
BoLD	Body Language Dataset
CAER	Context-Aware Emotion Recognition
CAER-S	Context-Aware Emotion Recognition (Static)
COCO	Common Objects in Context
DL	Deep Learning
EiLA	Emotions in LatAm dataset
EMOTIC	Emotions in Context
ERP	Event-Related Potential
FACS	Facial Action Coding System
FAU	Facial Action Unit
FER	Facial Expression Recognition
GCN	Graph Convolutional Neural Network
GIN	Graph Isomorphism Networks
HCI	Human-Computer Interaction
iMiGUE	Micro-Gesture Understanding and Emotion Analysis
mAP	Mean Average Precision
ML	Machine Learning
PCA	Principal Component Analysis
rPPG	Remote Photoplethysmography
STS	Superior Temporal Sulcus
VAD	Valence-Arousal-Dominance

YOLO	You Only Look Once
ZSL	Zero-shot Learning

CONTENTS

1 INTRODUCTION

Twenty-four years have passed since Rosalind Picard published her groundbreaking work "Affective Computing" (PICARD, 2000). It is compelling to see how many of her complaints are utterly unnatural for readers from this new generation, such as how the software would take an absurd amount of memory or how the weights of manuals were equivalent to the weight of hardware. These complaints have been *mostly* solved for today's typical user; however, this also points out the speed at which different disciplines in computer science have evolved. Software engineering has evolved to allow better usage of resources such as memory, and even if we do need more, hardware engineering has evolved to the point that it allows us to plug multiple RAM sticks into our computers, expanding their memory whenever we feel that we need them. Web pages such as LinkedIn use over 700 MB of RAM while in the background, and this occurs entirely invisible to the user. User experience has evolved towards natural interactions, in which kids can interact fluidly with computers or handheld devices. However, the same central point from Picard's work is still not so evolved today: recognizing affective behavior through computers and/or intelligent systems.

Recognizing affective behavior could be the key to improving Human-Computer Interaction (HCI), especially when looking through the eyes of paradigms such as *Natural Interaction*, which focus on improving how people interact with machines, allowing them to interact as they interact with each other, without the sense of interacting with a machine (VALLI, 2007; VALLI, 2008). The human factor is an essential part of this and many other HCI paradigms since it decides how to present information, impacting the individual experience for the user. Applications are changing focus from computer-centered approaches (in which the limitations of the machine imposed how the system would act) to user-centered approaches, which focus on interaction by design. The scope of these applications is also changing, not necessarily focusing on desktop software or mobile applications but especially on smart cities, smart environments, and autonomous/semi-autonomous vehicles. This task, however, is not easy, as it differs from other computer vision tasks such as object detection and scene recognition since it is not perception that plays a significant role but cognition.

Perception and cognition are two mental processes crucial in how we sense and perceive the world; although they work together in our brains, allowing us to understand the world we live in, they have fundamental differences. Perception refers to acquiring, interpreting, and organizing

sensory information from the environment, e.g., seeing colors and locations of nearby objects. In contrast, cognition refers to higher-level mental processes that involve thinking, reasoning, problem-solving, and decision-making (HALFORD; HINE, 2016; NES; SUNDBERG; WATZL, 2023).

Given how emotion recognition is fundamentally a cognitive task, how can we develop and design models that could have a take into this problem? One possible solution is to look at how humans approach this problem, designing biologically-inspired (and psychologically-inspired) models. The formal definition of this task varies from "recognizing specific emotional states in people" (KOSTI et al., 2017b) to "a process that uses low-level signal cues to predict high-level emotion labels" (RANGANATHAN; CHAKRABORTY; PANCHANATHAN, 2016). In this work, we propose a more applicable definition that can be more related to several specific applications: the task of inferring perceived emotion through nonverbal cues that can be linked to emotion, mood and thought.

First, it needs to be *perceived*. It is difficult to say that a model predicts or classifies human emotions because emotion is an internal, subjective, and cognitive representation of the person. Therefore, we place ourselves in the role of the perceiver, in which we, given a series of information, need to judge someone else's emotion. However, which information could be so helpful as to allow us to make these judgments?

Researchers from the behavioral sciences have been studying how humans make these judgments for a while and have found that humans communicate emotions even without intention through how we behave during interactions. In these cases, the communication of emotions that someone does not explicitly say is called *nonverbal communication*, in which we communicate our feelings through facial expressions, body language, speech tonality, and many other forms (ROUAST; ADAM; CHIONG, 2019; PATEL, 2014). For example, Darwin (1872) investigated how body movements and facial expressions communicate emotions and exemplify how emotions can act directly on bodily behavior. Joy, for instance, has a strong tendency to purposeless movements since this feeling causes our blood circulation to accelerate, stimulating the brain and leading to visible reactions in the body. Nonverbal communication is an efficient way of communicating affective information between people because humans can recognize signals and decode them to understand the mood and emotion of others (LHOMMET; MARSELLA, 2014).

These kinds of expressions, movements, and actions that are displayed and, therefore, communicated without intention are referred to as "nonverbal cues" because they give hints that are not communicated through speech and are the information that we, as humans, need

Figure 1 – An example of how context can contribute to the overall recognition of image. In (a), we see a facial crop that could lead to a negative emotion, such as pain or anger, while in (b) we can see from context that they are actually celebrating a victory in a competition.

(a) (b)

Source: BARRETT; MESQUITA; GENDRON (2011)

to judge the emotions of others. Therefore, if we can develop intelligent systems that can extract or identify these cues, this knowledge could be used to judge people's emotions.

Facial expressions contain prominent nonverbal cues and are the most natural for humans to perceive; researchers state that understanding emotions in a face could be to perceivers as if they are reading words on pages (BARRETT; LINDQUIST; GENDRON, 2007; BARRETT; MESQUITA; GENDRON, 2011). We display emotions as particular arrangements of facial actions, and most of us can easily recognize their affective meaning based on the facial structure. This behavior led to a task called Facial Expression Recognition (FER), in which researchers design models to extract affective information from facial expressions.

FER works well for some scenarios, especially controlled scenarios. However, in unrestricted in-the-wild scenes, we are also influenced by other multisensory cues, such as context and body language. We display an example in Figure 1, which shows Serena Williams after the 2008 U.S. Open Tennis finals, illustrating the importance of context when inferring meaning from an image. We can then argue that for some applications, relying only on FER might not be enough and that by expanding to other nonverbal cues that are present and visible, a robust

affective assessment is possible.

In this doctoral thesis, we propose a comprehensive study of emotion recognition in different scenarios, aspects, and nonverbal cues. We also propose a diverse set of models that can be used for a wide variability of scenarios, almost in a toolbox manner, based on research on behavioral psychology.

1.1 RESEARCH GOALS

- **RG1**: To propose approaches for emotion recognition that are supported by evidence from the behavioral psychology literature;
- **RG2**: To assess different possible approaches for emotion recognition based on cue availability;
- **RG3**: To develop new, fast frameworks for emotion recognition that are compatible with the state of the art.

In this work, we addressed the research goals stated above which led to the following contributions:

- A study on behavioral psychology to support emotion recognition methods that can serve as a path towards the development of new systems in the future that are based on evidence related to behavior (Chapter 2) (RG1, RG2);
- An overview of the theory that supports emotion recognition systems, as well as a discussion on the possible applications (Chapter 3) (RG1);
- From the best of our knowledge, the first definition of a taxonomy for emotion recognition (Section 3.1) (RG3);
- A comprehensive study on the datasets for emotion recognition (Chapter 4) and their current limitations (Subsection 4.1.6) (RG3);
- A new benchmark for emotion recognition with a focus on the Latin American culture, tackling the bias that is currently present on datasets recorded in the USA or Europe (Subsection 4.2.1) (RG2, RG3);
- A fast, novel framework for emotion recognition based on high-level features extracted from context that surpasses multiple techniques using more than one cue (Section 5.1) (RG1, RG3);

- A robust approach for emotion recognition based on gait (Section 5.2) that is the current state-of-the-art in the E-Gait dataset, surpassing multiple other methods from the literature (RG1, RG3);
- We propose EmotionRAM, a new approach that relies on facial expressions, context, and body language for emotion recognition in images, that is up to nine times faster than the current state-of-the-art method and only 0.12% worse in accuracy (Section 5.3) (RG1, RG3).

2 THE HUMAN PERSPECTIVE

In this chapter, we will overview emotion from the human perspective, based on the literature for psychology, behavioral psychology, and other disciplines that study and discuss emotion perception and communication. Please bear in mind that, given how this thesis is related to the computer science field, the objective of this discussion is to set a baseline for how humans perceive emotion and allow reasoning behind the design of the models and experiments proposed later in Chapter 5. Therefore, this is not a comprehensive literature review for this field, but it will contain findings that guided the development of this doctoral work during these years.

2.1 NONVERBAL COMMUNICATION

The ability to communicate our ideas and thoughts is a strong requirement for living in society since they play a significant role in social interactions, influencing behavioral responses and the construction of relationships with peers. In human interactions, there are two primary forms of communicating emotion: verbal communication, in which one verbally states feelings, and nonverbal communication, in which affective information is communicated through signals sent naturally by the sender.

While verbal communication is important, nonverbal cues are the decisive factor in judging the emotional states of others (JACOB et al., 2016). Looking back to your most recent interaction with a peer, it is very uncommon to continuously state one's emotion during conversations, for example, as it might happen more punctually. This behavior occurs mainly due to the challenges of verbalizing negative and positive nuances of emotion. Therefore, nonverbal signals are suitable for communicating affective information because humans can decode them effectively and naturally.

A vast majority of nonverbal communication happens without intention or interference from the sender or the receiver; this is why this type of communication is also known as *implicit* (BUCK, 1991; LHOMMET; MARSELLA, 2014). Kinesics, defined as *the study of body motion as related to the nonverbal aspects of inter-personal communication* (BIRDWHISTELL, 1952) is a core aspect of this nonverbal exchange, which reveals subconscious attitudes and reactions based on the result of an interaction with a 3rd party.

Table 1 – Body movements observed with specific emotions (citations from (DARWIN, 1872))

Joy	Various purposeless movements, jumping, dancing for joy, clapping of hands, stamping, while laughing head nods to and fro, during excessive laughter whole body is thrown backwards and shakes or almost convulsed, body held erect and head upright (pp. 76, 196, 197, 200, 206, 210, 214)
Sadness	Motionless, passive, head hangs on contracted chest (p. 176)
Pride	Head and body held erect (p. 263)
Shame	Turning away the whole body, more especially the face, avert, bend down, awkward, nervous movements (pp. 320, 328, 329)
Fear	Head sinks between shoulders, motionless or crouches down (pp. 280, 290) convulsive movements, hand alternately clenched and opened with twitching movement, arms thrown wildly over the head, whole body often turned away or shrinks, arms violently protruded as if to push away, raising both shoulders with the bent arms pressed closely against sides or chest (pp. 291, 305)
Anger/rage	Whole body trembles, intend to push or strike violently away, inanimate objects struck or dashed to the ground, gestures become purposeless or frantic, pacing up and down, shaking fist, head erect, chest well expanded, feet planted firmly on the ground, one or both elbows squared or arms rigidly suspended by the sides, fists are clenched, shoulders squared (pp. 74, 239, 243, 245, 271, 361)
Disgust	Gestures as if to push away or to guard oneself, spitting, arms pressed close to the sides, shoulders raised as when horror is experienced (pp. 257, 260)
Contempt	Turning away of the whole body, snapping one's fingers (pp. 254, 255, 256)

Source: Wallbott (1998)

Between nonverbal signals, Darwin (1872) investigates the role of body movements and facial expressions in communicating emotions and exemplifies how they can directly act on bodily behavior. In Table 1, we present citations from Darwin's work regarding body movements and emotions.

2.1.1 Perception of nonverbal cues

Given how humans can easily send and receive these nonverbal cues, we may ask a follow-up question: how do humans encode these cues to judge emotion? Researchers have studied various human communication and interaction aspects to understand how these cues convey emotional information between individuals. We will focus this discussion on the cues evaluated in this work: facial expressions, situational context, and body language.

Facial expressions

Among multiple researchers, the work by Ekman and Friesen in decoding the complex relationship between facial expressions and emotions stands as a foundation in various fields and affective computing. Their introduction of the Facial Action Coding System (FACS) (HJORTSJÖ, 1970; FRIESEN; EKMAN, 1978), which prominently features Facial Action Units (FAUs), marked a significant advancement in understanding emotional expression in faces. Each FAU is linked to muscle movements that respond to specific emotions. There is evidence that supports that humans perceive facial expressions by solving the "inverse problem of production," identifying which underlying FAUs are present (MARTINEZ, 2017).

Therefore, can there be emotions without facial expressions? As Ekman (1993) discussed, individuals may not show any visible evidence of emotion in the face. Tassinary and Cacioppo (1992) evaluated in their work that by employing surface electromyography, they could record slight electrical variations of the face. This means that, even when someone does not display emotions in their face, they do so in an *unobservably* manner.

People can also try to fabricate expressions when they do not feel any emotion to mislead the observer. Evidence suggests that emotions such as *enjoyment*, *anger*, *fear*, and *sadness* contain muscular actions that most people are unable to perform voluntarily. This evidence points out that, although this is possible, it is difficult to do so in an undistinguishable manner (EKMAN, 1993; EKMAN; ROPER; HAGER, 1980). The same evidence can be expanded into other aspects of behavior that one may want to disguise. For example, suppose one deliberately intends to deceive their visible emotion by forcing expressions on their face. In that case, the observer could still perceive the cues from the felt emotion from other sources, especially body language (EKMAN et al., 1991).

Situational context

Even though we can easily perceive emotions in others, primarily through the ease facial expressions provide for most of us, there is significant evidence that context influences emotion perception. In other words, although faces carry affective information, the emotional meaning of these facial actions is constructed from the context in which they are embedded (BARRETT; MESQUITA; GENDRON, 2011).

These contextual influences are perceived early and automatically and were validated in the past decades using fMRI[1] in an experiment by Mobbs et al. (2006) in which given identical faces across different contextual backgrounds, perceivers would judge these faces with different emotional features based on the contextual framing. fMRI results revealed greater activation in regions involved in social cognition, such as the Superior Temporal Sulcus (STS), an area located in the temporal lobe involved in social perception and understanding other's intentions and emotions; the temporal pole, that integrates sensory information to represent complex social concepts like attitudes and treats; the amygdala, involved in processing emotions, especially fear, by tagging sensory stimuli with emotional significance, and the Anterior Cingulate Cortex (ACC), which is activated when processing emotions, pain and conflict, and regulates emotional responses. These findings provide evidence that contextual framing can alter social perceptions and attributions by modulating activity in brain regions involved in emotion, social cognition, and contextual processing.

Righart and Gelder (2008) extends on these experiments to understand how emotional contexts influence early stages of face processing when explicitly categorizing facial expressions of fear and happiness using Event-Related Potentials (ERPs). ERPs is a neuroscience tool that studies the brain's electrical activity for specific stimuli. Two ERPs are especially relevant in this scenario: the N170, which is a component related to face encoding and happens $\approx$ 170 ms after the stimuli, and the P1, which is also associated with facial processing and happens at around 100 ms after the stimuli.

The authors propose an experiment using electroencephalography where they present facial expressions of fear and happiness within the context of emotional scenes, and participants categorize the facial expressions. They measured the P1 and N170 ERP components. They found that the P1 amplitude was slightly modulated by the emotional scenes, meaning that the emotional contribution of context influences early visual processing. However, the facial expression itself is not fully encoded yet. In other words, the context's encoding happened before the facial expression's encoding. As for the N170, facial expressions are now being encoded and influencing brain activity, but the emotional context is still influential. In fearful contexts, the modulation amplitude was higher, suggesting the need for fast action from the person.

These studies corroborate that context influences emotion perception from a neuroscience point of view. Barrett and Kensinger (2010) has approached this discussion from a more

[1] An imaging technique that is used to map brain activity by detecting changes in blood flow and oxygenation

practical point of view in a study to verify if context is more likely to be encoded when a person's task is to perceive emotion in the face of another person rather than to judge the face's affective value. They show that although facial expressions are a beacon to indicate the affective state of others, perceivers will routinely encode the context when asked to make a more specific inference about someone else's emotions. In another study, Aviezer et al. (2008a) used eye-tracking and detected changes in the patterns depending on how faces and the context are congruent, providing evidence that facial expression perception is malleable by context. This study is a strong driver towards designing a context encoding pipeline for our deep learning approach.

Body language

Past research indicates that emotions often result in purposeless movements and physical manifestations. These nonverbal cues, as demonstrated in Table 1, are correlated with emotions. In a study by Wallbott and Scherer (1986), judges relied on cues like hand movements and body behavior to accurately perceive emotions. Wallbott (1998) extended this research, finding specific body movements and postures indicative of distinct emotions through ANOVA analysis. This supports Darwin's perspective on how body language reflects different emotional states.

How do humans encode body language? Body movement is highly representative, but from an application point of view, it is only available in videos, while static body language is a more global representation. Coulson (2004) investigates how judges can recognize emotions from static body postures, focusing on the exact anatomical cues that convey each emotion. They used computer-generated meshes to manipulate postures according to descriptions of emotional body language, and participants had to classify the emotions displayed in these mannequins. The experiments revealed that anger, happiness, and sadness were most accurately recognized from posture, while they rarely identified disgust. Viewpoint also affected the recognition since the same posture viewed from different angles was attributed to different emotions. However, as the body possesses multiple degrees of freedom, the posture is encoded based on various variables in an unconscious manner, meaning it is difficult to describe precisely which cues are used to describe emotion. By manipulating specific postural variables such as head bend, chest bend, abdomen twist, etc., the study was able to begin specifying the anatomical features that convey emotion.

It is clear, however, that dynamic body language is much more representative, as motion is a strong factor regarding body descriptions. A possible way to assess body language is through gait, which is the description of the way that someone walks. Previous research has investigated how gait-related parameters could be used to describe social aspects, including emotion. Montepare, Goldstein and Clausen (1987) evaluated whether emotions could be identified from a person's gait by studying how subjects would perceive emotions of happiness, sadness, anger, and pride through gaits. The results showed that subjects could identify sadness, anger, and happiness at levels better than chance alone. They also investigated which gait characteristics differentiated the emotions: angry gaits were rated as heavier footed, while sad gaits had fewer arm swings. Extending this research, Roether et al. (2009) focused specifically on describing postural and kinematic features that are important for the perception of emotion in human gait by using motion capture to record the gaits of 25 people displaying anger, happiness, sadness, and fear. They applied machine learning to select the most informative features automatically and validated them in a perception experiment where observers classified and rated the animations of the recorded gaits. Statistical analysis showed a high overlap between features from motor data and those within perceptual judgments, indicating that the features are perceptually relevant. Finally, they show that movement speed strongly influences the perception of emotions in gait, as well as features such as limb flexion for anger and fear and head inclination for sadness. These findings point towards a strong correlation between emotional communication and gait from a nonverbal analysis, thus indicating that this is also a suitable cue for emotion extraction.

2.2 CULTURAL ASPECTS OF EMOTION

A continuous discussion in this field arises when researchers try to map the source of the emotion perception abilities, in which two main arguments are presented: The first suggests that the communication and perception of emotion is a biological response that comes from inherent stimuli that were developed through emotion. One piece of evidence for this argument is given in the work of Ekman (1993). Researchers traveled to a community in New Guinea to evaluate how a group of people who have had only minimal contact with Western people would be able to perceive emotions. This group never had any contact with Western media - movies, magazines, etc. The subjects listened to stories describing emotions and had to select a matching facial expression from photos. This group was able to accurately identify

emotions in pictures that Western subjects also identified correctly, providing strong evidence that particular facial configurations are universally associated with discrete emotions across cultures.

Another piece of evidence, now less focused on facial expressions, is the work by Tracy and Matsumoto (2008). The study examines whether nonverbal expressions associated with pride and shame may be innate biological responses to success and failure. They propose to analyze the spontaneous behavior displayed by athletes from over 30 nations in response to winning or losing judo matches at the Olympics and Paralympics. They show that sighted athletes displayed typical components for pride (e.g., head tilted back, smile, expanded posture) and shame (e.g., slumped shoulders, narrowed chest) at the same rate as congenitally blind athletes from various cultures. This suggests the expressions are likely innate rather than learned, as congenitally blind individuals could not have observed and modeled the behaviors from others.

The second point, however, is that human perception and communication of emotion are cultural and learned through interaction with peers in society. Some evidence for this argument is the work by Mesquita, Boiger and Leersnyder (2017), in which they point out that cultural differences in emotions are purposeful, helping people to meet the criteria of a "good" person in their culture. First, they show a cultural variation in the frequency and intensity of emotions. For example, Americans felt more pride and anger, while the Japanese felt more friendly feelings and shame. Through interaction, language representations, and socialization, cultures afford some emotional experiences over others, promoting emotions aligned with cultural values and discouraging those not aligned. This suggests emotions are not static categories but are actively constructed in situated contexts in culture-dependent ways. The same emotion concept may map onto different configurations of appraisals and tendencies across cultures.

We can see, however, that this is not a binary scenario, but rather how universal emotions are shaped and adapted through culture, even though some views point to predominantly biological arguments (EKMAN; FRIESEN; ELLSWORTH, 1982) or predominantly cultural arguments (HARRÉ, 1986). More fundamentally, these discussions point towards to what extent the variety of emotions are universal or cultural in nature (MESQUITA; FRIJDA, 1992). This is an important discussion related to how emotion recognition datasets are built upon and was a driver for the proposal of a dataset representative for Latin American culture, as we will describe later in Subsection 4.2.1.

3 AN OVERVIEW OF EMOTION RECOGNITION

As we discussed in Chapter 1, we propose a definition for emotion recognition that is not application-specific but rather is based on how humans perceive emotions. Emotion recognition is, then, the task of inferring the perceived emotion of a person through nonverbal cues that can be linked to emotion, mood, and thought. This definition means that we are not asking this person how they are feeling, nor are we interested in explicitly communicating emotion in a verbal manner, but rather we want to teach a model to perceive this emotion, such as we do, in an automatic and interference-free manner.

The concept of nonverbal cues is quite broad, as they can come from various sources and be presented in multiple forms. In this chapter, we will discuss emotion recognition as a task of computer vision – in other words, we will focus on visual cues that can be extracted from images or videos, from all the possible nonverbal cues.

The discussion raised by Ekman (1992) has directed the field of emotion recognition, especially the construction of datasets, regarding what models would recognize. Again, given the cognitive aspect of emotions, this task may not look as straightforward as others, such as object detection or semantic segmentation, regarding its design. What is emotion, and what does this task classify? Ekman raised a theory that some separate emotions differ from one another in important manners, which means that some of the formats at that time to measure emotions were not sufficient anymore (for example, pleasant-unpleasant scales). This set of emotions was referred to as *basic* emotions and is still widely applied in today's literature on affective computing in general.

Based on nine characteristics to distinguish the basic emotions from other affective phenomena, Ekman (1992) proposes that the following emotions meet these criteria and therefore qualify as distinct basic emotions according to this framework: *Anger*, *Fear*, *Sadness*, *Enjoyment*, *Disgust* and *Surprise*.

With this strongly supported by evidence definition of classes, the emotion recognition task can now be seen as a classification task. Ekman also discussed other emotional traits, such as *Embarrassment* and *Excitement*, but some of the characteristics were presented unusually. For example, in *Embarrassment*, a very evident signal is blushing, which is more evident in light-skinned persons and, therefore, could not be considered.

Box 1. Classification

Classification is a set of problems in which our model needs to look at features and then predict which category (also called a class) among a set of options that example belongs.

The simplest form of classification is when there are only two classes, called binary classification. Usually, models do not predict a firm categorical assignment but rather express their predictions in the form of probabilities, assigning a probability for each class in that list, and the magnitude of the probability conveys a notion of uncertainty (ZHANG et al., 2023).

When we have more than two possible classes, the problem is defined as multiclass classification. This is the case of Ekman's basic emotions. Therefore, some examples of classification problems are:

- **Handwritten characters recognition**: $\{0, 1, 2..., a, b, c, ...\}$;
- **Object detection**: {car, truck, motorcycle, ..., chair, sofa, table, ...};
- **Emotion recognition**: {happy, sad, angry, ...}

Although this theory is widely accepted, other models of emotion representation are available in the current literature, specifically models that focus on dimensional theories of emotion. This provides a view that emotions are fundamentally similar, except along dimensions such as valence, arousal, and dominance, as stated by the Valence-Arousal-Dominance (VAD) model (SCHLOSBERG, 1954).

This model places emotions in a three-dimensional space. The Valence (V) axis determines whether an emotion is pleasant or unpleasant to the perceiver, therefore distinguishing between positive and negative emotions. The Arousal (A) axis differentiates between active and passive emotions. Finally, the Dominance (D) axis represents the control and dominance over the nature of emotion. Therefore, each emotion can be represented as a linear combination of those three components (KOŁAKOWSKA; SZWOCH; SZWOCH, 2020).

Box 2. Regression

Regression is a set of problems in which our model needs to make estimations, predicting continuous values based on input features. Unlike classification, where the output is a category, in regression, the output is a numerical value that can range from negative infinity to positive infinity.
The simplest form of regression is called linear regression, in which the relationship between the input features and the target is assumed to be linear. In this case, the model tries to fit a line through the data points to minimize to minimize the prediction error.
More complex forms of regression include polynomial regression, where the relationship between input and output is modeled as a polynomial function, and non-linear regression, which can capture more complex relationships. As in classification, regression models also deal with uncertainty, usually by assuming a probability distribution for the errors in the predictions. Some examples of regression problems are:

- **Real Estate Pricing**: Predicting the market value of properties based on attributes like location, square footage, number of bedrooms, etc.
- **Stock Market Forecasting**: Estimating future stock prices or market indices based on historical data, economic indicators, and other relevant factors.
- **Temperature Prediction**: Projecting future temperatures based on historical weather data, climate patterns, and environmental changes.

According to Ekman's arguments (EKMAN, 1992), dimensional theories fail to recognize evidence that emotions differ meaningfully in expression and physiology. However, is this a question of perspective, as there is evidence to support both claims? From this point of view, emotion recognition can now also be seen as a regression task.

Some datasets contain categorical and continuous annotations, allowing models to output formats for regression and classification. However, as we discuss in Chapter 4, annotating VAD in some scenarios can be difficult, especially if the annotator is not trained in emotion recognition. For example, in a similar case in social robotics, a study shows that robot animations that were originally attributed to valence and arousal levels were perceived differently by the

end-user, indicating a discrepancy between the intended perception and the rated perception (MARMPENA; LIM; DAHL, 2018).

3.1 AN EARLY EMOTION RECOGNITION TAXONOMY

The organization of emotion recognition research is not completely agreeable, especially the definitions between FER and emotion recognition. This section aims to propose a broad view of how this field could be arranged by grouping research topics found in the literature. We lay a foundation for future work in emotion recognition by proposing an early taxonomy for guiding work on future emotion recognition applications. We show this early arrangement of the taxonomy in Figure 2, focusing on nonverbal communication.

Also, considering the theory that connects the emotion recognition task with supervised learning problems, we will overview the different methods available in the literature that are also covered in this taxonomy, presenting relevant references of past groundbreaking works and newer research that is shaping this field today. As this is an overview, we will only briefly discuss each topic.

Vision-based methods

Facial Expression Recognition (FER), as the name suggests, focuses on extracting emotional information from faces. Pipelines for this task usually rely on MTCNN (ZHANG et al., 2016) or another framework to align the facial region and achieve better results. With the cropped area of the face, there are mainly two approaches for FER: using standard backbones such as ResNet-50 (HE et al., 2016a) or EfficientNet-B0 (TAN; LE, 2019) associated with fine-tuning strategies between datasets to achieve good results (ZHOU et al., 2019; KUMAR; RAO; YU, 2020; SAVCHENKO, 2021), or sophisticated models that are usually focused on attention or pyramid-like approaches (MENG et al., 2019; ZHANG et al., 2022; WEN et al., 2023; CHEN et al., 2023; ZHANG et al., 2023).

Multi-cue emotion recognition extends FER by extracting other nonverbal cues that can be captured through vision. Most techniques combine face with context (LEE et al., 2019; LE et al., 2021), while others also add body language (THUSEETHAN; RAJASEGARAR; YEARWOOD, 2022; COSTA et al., 2022). Overall, many other cues can also be considered, depending on their visibility on the dataset and their correlation with emotion. For example, Yang et al.

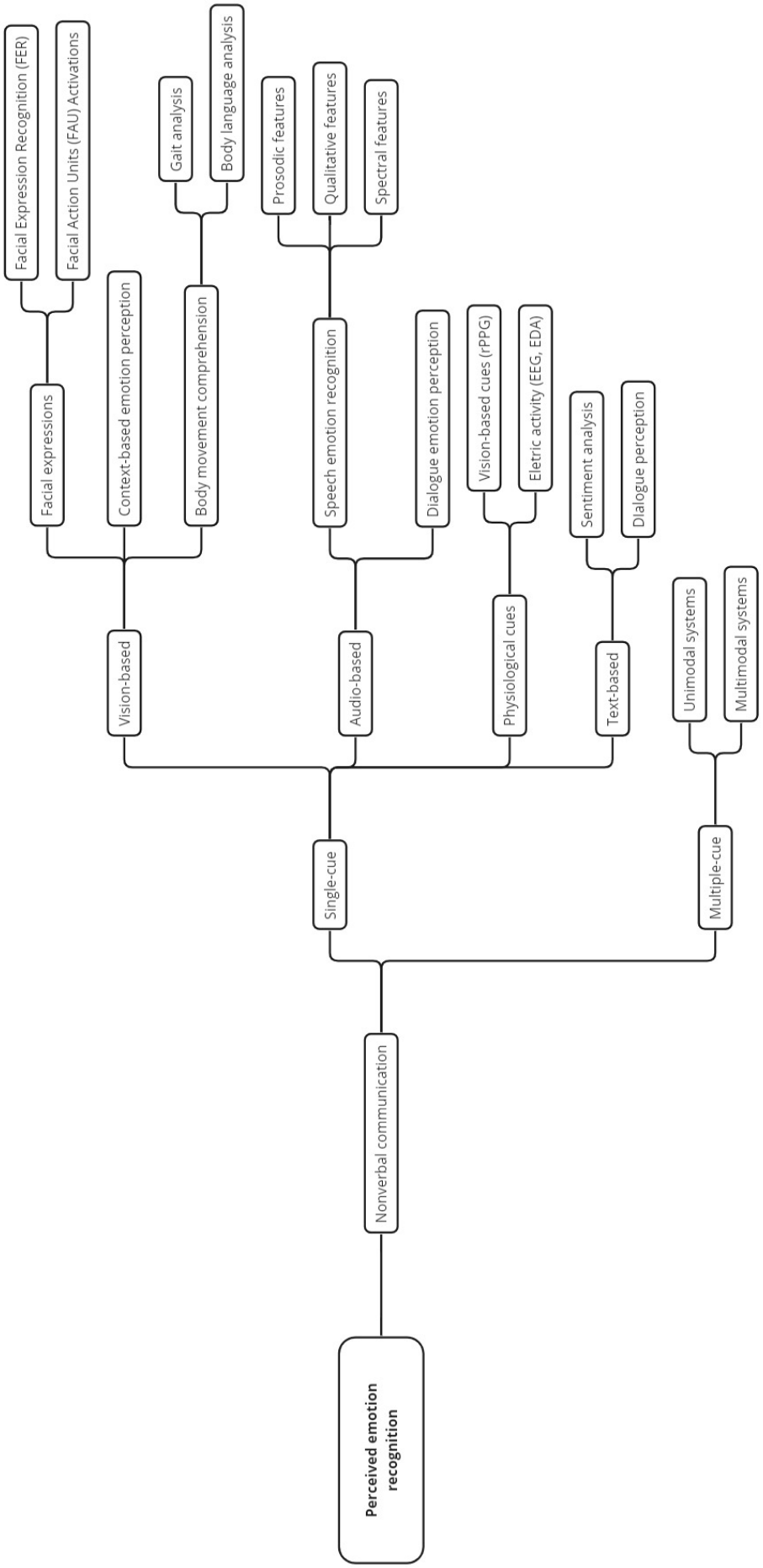


Figure 2 – An early view for an emotion recognition taxonomy, focusing on nonverbal communication.

Source: Author.

(2022) uses five nonverbal cues: face landmarks, body pose, context, agent relationships, and human-object interaction. Mittal et al. (2020) besides using face, context, and body pose, it also uses depth estimation to represent Frege's principle and consider the proximity of agents as factors that affect the emotional state of an agent. There is, however, a trade-off since adding support to other cues increases the computational requirements of the model, limiting their operational application.

Audio-based methods

Speech analysis is another common approach for emotion recognition, as vocal cues also provide strong indicators of a speaker's emotional state. Speech analysis techniques typically involve extracting features from three categories: prosodic, qualitative, and spectral (AL-DUJAILI; EBRAHIMI-MOGHADAM, 2023).

Prosodic features capture characteristics of speech melody, rhythm, and tempo, such as pitch, energy, and duration (CÁMBARA; LUQUE; FARRÚS, 2020). Qualitative features describe the voice quality, such as shimmer and jitter (KERKENI et al., 2018). Finally, spectral features represent the frequency spectrum and give insights into the tonal quality of speech (YANG; HUANG, 2022).

Physiological methods

Remote Photoplethysmography (rPPG) is a contactless technique for measuring heart rate based on subtle skin color changes in the face due to blood flow (CHEN; MCDUFF, 2018). Given how emotion has a direct influence on our physiological signals, elevating heart rate, blood pressure, and breathing rate, rPPG has also been applied to emotion recognition by analyzing these changes (BENEZETH et al., 2018; BRAUN et al., 2023).

Text-based methods

Sentiment analysis, also known as opinion mining, aims to determine whether the sentiment expressed in a text is positive, negative, or neutral. This analysis is typically applied to reviews, social media posts, or any other format where people can express their opinions and feelings. There are three main approaches to sentiment analysis (DANG; MORENO-GARCIA; PRIETA, 2020): lexicon-based approaches, which include dictionary-based and corpus-based methods (DANG; MORENO-GARCIA; PRIETA, 2020; LI et al., 2020; CONSOLI; BARBAGLIA; MANZAN,

2022; SHANG et al., 2023; MACHOVA et al., 2020); Machine Learning (ML)-based techniques, which range from traditional methods such as Support Vector Machines to more robust Deep Learning models through the usage of Word2Vec, which enables words to be mapped to a vector space where similar words have similar representation or TF-IDF, which is a statistical measure of how important a word is to the document (LIU, 2020; KAMYAB; LIU; ADJEISAH, 2021; XU et al., 2020), or hybrid methods that combine both approaches, with lexicon playing a more important role.

Multimodal approaches

Combining different methods to evaluate images and videos with multiple modalities can also yield significant results in some scenarios. Zhang, Pan and Wang (2023), for example, combines vision and text to propose EmotionCLIP, which extracts visual cues and also encodes text descriptions of the scene. Other approaches, such as Chen et al. (2023) and Srivastava, Singh and Tapaswi (2023), also combine vision with text, focused on the description of movies.

Vision and audio can also lead to a significant understanding of the scene since the correlation between speech and facial features can allow for an improved understanding of individuals. Li, Wang and Cui (2023) combines vision, audio, and text by first learning their contributions individually and then fusing these features in a late fashion to achieve predictions.

3.2 POSSIBLE APPLICATIONS

3.2.1 Smart cities, environments, and spaces

The concept of a *smart city* is related to a city that functions intelligently, integrating its infrastructure and services using intelligent devices to monitor and control environments (HANCKE; SILVA; JR, 2012). The human factor is the primary surrounding condition for intelligent services in such scenarios and guides the implementation of a service related to human dynamics (SHAW; SUI, 2018). Emotion recognition, linked with human dynamics and sentiments (SEAGAL; HORNE, 2003), emerges as a key application in this context. This technology could revolutionize city services by adapting them according to the emotional states of citizens.

For example, studies like Meng et al. (2020) have explored the impact of urban soundscapes on facial expressions, indicating how certain city areas or sound environments affect citizen well-being. Such insights could guide decision-makers to identify and improve less pleasant city regions, extending to public and private spaces like buildings, monuments, or stores.

Figure 3 – An abandoned park with overgrown grass, visited only by some people.

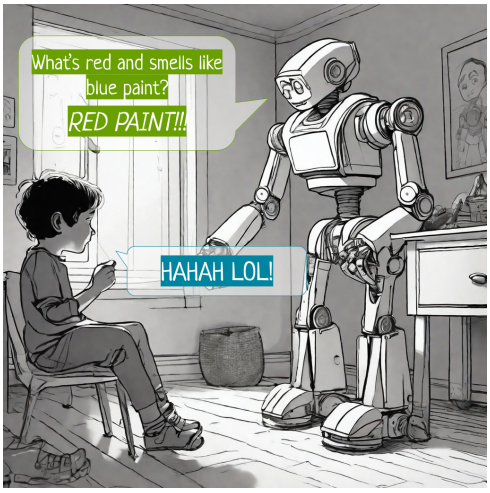
Source: Author.
(Generated using Generative AI, as disclosed in Chapter 7.1)

Urbanism greatly benefits from understanding the emotional responses of citizens towards public spaces. Research such as Wei et al. (2019) has investigated the correlation between facial expressions and the ambiance of public parks, revealing that proximity to the city center can influence visitor emotions. These studies suggest that analyzing emotional responses over time could help urbanists to make informed decisions about space management and interventions. Traditionally, urbanists observe and record community interactions with public spaces (LIMA et al., 2022), but emotion recognition technology offers a more nuanced understanding. This approach could suggest interventions in areas like poorly maintained parks, as illustrated in Figure 3, by gauging the collective emotional response of the community towards these spaces.

3.2.2 Affective and assistive robots or agents

Social robots are emerging as a valuable companion for people with special needs, offering monitoring and interactive capabilities. Researchers are developing these robots to interact empathetically with humans by recognizing and responding to their emotions (CALVO-BARAJAS; PERUGIA; CASTELLANO, 2020). One such example is CuDDler (LIMBU et al., 2013), a robot resembling a baby polar bear. CuDDler can assess a person's emotional state and react in a way that fosters positive feelings, which is especially beneficial for children. We exemplify this interaction in Figure 4.

Figure 4 – A social robot cheering up a child.



Source: Author.
(Generated using Generative AI, as disclosed in Chapter 7.1)

Moreover, affective agents can use emotional cues to evaluate and improve the quality of their interactions with users. For instance, by monitoring a user's emotional state before and after performing a task, an AI agent can gauge whether the interaction alleviated stress or caused confusion. The same can happen in the other direction; modeling emotional behavior in robots or agents can improve interaction quality (LIM; OKUNO, 2015; YADOLLAHI et al., 2021). This feedback loop is crucial in refining Artificial Intelligence (AI) responses, ensuring clearer communication, and fostering a more intuitive user experience.

3.2.3 Emotion tracking and mental health

Systems that track mood and emotion are more common today, mainly due to the recent discussions regarding the importance of mental health. These conditions have increased recently due to the COVID-19 outbreak, which generated outcomes that could last for years. Although clinicians can give mental support through consultations, patients can sometimes feel uncomfortable exposing their feelings verbally (BUSCH et al., 2021). In this context, telemonitoring systems could be expanded to also monitor their moods and emotions over time.

Figure 5 – A teenager with a sad face in a messy room lying in bed for an extended amount of time.

Source: Author.
(Generated using Generative AI, as disclosed in Chapter 7.1)

A study by Gavrilescu and Vizireanu (2019) delves into the potential of recognizing the activation of FAUs to estimate levels of depression, anxiety, and stress. Accurately predicting these metrics requires capturing spontaneous emotional responses, which can provide insightful data for physicians and aid in diagnosis. To enhance this, such systems could monitor the frequency of sudden emotional shifts within a user's home environment, thereby offering detailed insights to healthcare providers. Integrating an off-the-shelf activity recognition model can also link emotional states with corresponding activities. This holistic approach could lead to comprehensive behavioral insights, like identifying when a patient feels sadness and spends an extended period in bed, as depicted in Figure 5. Such detailed monitoring can significantly enhance the understanding and treatment of mental health conditions.

3.2.4 Driver behavior and transportation

Daily commuting is a significant part of our daily routine, often linked with negative emotions like anger (UNDERWOOD et al., 1999) and anxiety (FAIRCLOUGH; TATTERSALL; HOUSTON, 2006). These emotions can adversely affect driving performance and overall well-being (DING et al., 2014; ZEPF et al., 2020). When drivers are angry or anxious, their decision-making skills may be compromised, leading to aggressive behaviors such as tailgating, speeding, and sudden lane changes (MOULOUA; BRILL; SHIRKEY, 2007; ROSEBOROUGH; WICKENS; WIESENTHAL, 2021). This kind of angry driving is not just a personal issue; it poses a serious threat to public safety on the roads. Addressing this problem is crucial for promoting safer and more peaceful driving conditions for everyone.

At the same time, driving while angry might lead to escalation in certain scenarios, which is the main cause of road rage. According to Forbes[1], road rage shootings have suffered a 135% increase from 2018 to 2022, leading to over 400 people injured in the US.

[1] Available at <https://www.forbes.com/advisor/car-insurance/state-rankings-confrontational-drivers/>

Besides the categorical classification of emotion (XIAO et al., 2022), it is also possible to monitor affective states such as stress, fatigue, or distraction (HAOUIJ et al., 2018) that can be used to plan interactions suggesting, for example, a brief pause or the switch of the driver, as we exemplify in Figure 6.

Figure 6 – An assistive agent inside a car suggesting a break for a nervous driver.

Source: Author.
(Generated using Generative AI, as disclosed in Chapter 7.1)

Emotion recognition can also significantly enhance the sensing capabilities of autonomous and semi-autonomous vehicles. Beyond just detecting pedestrians, this technology can also discern their emotional states, adding a vital layer of context to the vehicle's decision-making process. For instance, if pedestrians are angry or distracted, they might not notice an approaching vehicle and could make sudden, unpredictable movements that increase the risk of accidents. By understanding these emotional cues, autonomous vehicles can adjust their actions accordingly, creating a safer environment for pedestrians and drivers.

Table 3 – References for other applications of emotion recognition.

Market	Applications
Recommender systems	Recommendation for tourist spots (SANTAMARIA-GRANADOS; MENDOZA-MORENO; RAMIREZ-GONZALEZ, 2020), music (SAMUVEL; PERUMAL; ELANGOVAN, 2020; DHARSINI et al., 2020), movies (SOLEYMANI; PANTIC; PUN, 2011; ZHANG, 2020), and other multimedia content (MARIAPPAN; SUK; PRABHAKARAN, 2012)
Human resources	Candidate recruitment (KHOSLA; CHU; NGUYEN, 2016; GORBOVA et al., 2017; ADEPU; BOGA; SAIRAM, 2020), workplace quality (BOYD; ANDALIBI, 2023)
Security	Deception detection (ZLOTEANU, 2017; CURTIS, 2021), unusual behavior (CHANDRAN; BINU, 2021), crowd analysis (VELTMEIJER; GERRITSEN; HINDRIKS, 2021)
Education	Student engagement (IMANI; MONTAZER, 2019; GUPTA; KUMAR; TEKCHANDANI, 2023; DHALL et al., 2023)

3.2.5 Other applications

While the sections above have delved into specific and detailed applications of emotion recognition, a broader spectrum of uses can also be cited. To provide a comprehensive overview, we have compiled these additional applications in Table 3, whey they are presented in a summarized format for easy understanding and comparison.

4 DATASETS FOR EMOTION RECOGNITION

This chapter provides an overview of existing vision-based emotion recognition datasets, as also highlighting limitations of current available data. Some portion of this content has been adapted from our previous publication "*A survey on datasets for emotion recognition from vision: limitations and in-the-wild applicability*" (COSTA et al., 2023a), and portions of the text, figures, and tables from the original work have been incorporated in this chapter.

Deep Learning (DL) models are highly dependent on data availability and data quality, and the dataset used for training or validating these models will have a direct impact on their capabilities. Especially for emotion recognition, the presence of multiple cues on the samples will allow these models to extract more robust representations regarding emotion. Although there are several datasets for this task, each dataset has its particularities and limitations, which could severely harm the capacity of execution in real scenarios. Also, there are limitations regarding sample bias, in which samples present in the dataset are not representative of the real world, and recall bias, which is caused by the way multiple annotators are usually handled in these scenarios.

In this chapter we survey datasets currently used for benchmarking techniques for emotion recognition. Specifically, we focus on vision-based datasets containing images or videos for evaluation. Therefore, we do not evaluate datasets focusing on speech tonality, for example. Some of the contributions of our work are:

- We survey and list the datasets currently employed for benchmarking in the state-of-the-art (Section 4.1);
- We explore the difference in the annotations of these datasets and how they could impact training and evaluating techniques (Subsection 4.1.1);
- We discuss annotations described using continuous models, such as the Valence-Arousal-Dominance (VAD) model, and how they can harm the ability of a model to understand emotion (Subsection 4.1.2);
- We investigate the presence of nonverbal cues beyond facial expression in the datasets of the state-of-the-art and propose experiments regarding their representativeness, visibility, and data quality (Subsection 4.1.3);
- We discuss possible application scenarios for emotion recognition extracted from pub-

lished works and how each dataset can impact positively and negatively according to its features (Subsection 4.1.5).

This survey also differs from other published surveys in emotion recognition, such as (SAXENA; KHANNA; GUPTA, 2020; ZEPF et al., 2020; CANAL et al., 2022; VELTMEIJER; GERRITSEN; HINDRIKS, 2021; KHAN et al., 2023) since we focus on listing, understanding, and discussing datasets and the permissions and limitations they bring to techniques, while surveys usually focus on listing techniques, their limitations, specifications, and results on benchmarks. To the best of the authors' knowledge, this work is the first survey that focuses on datasets instead of techniques.

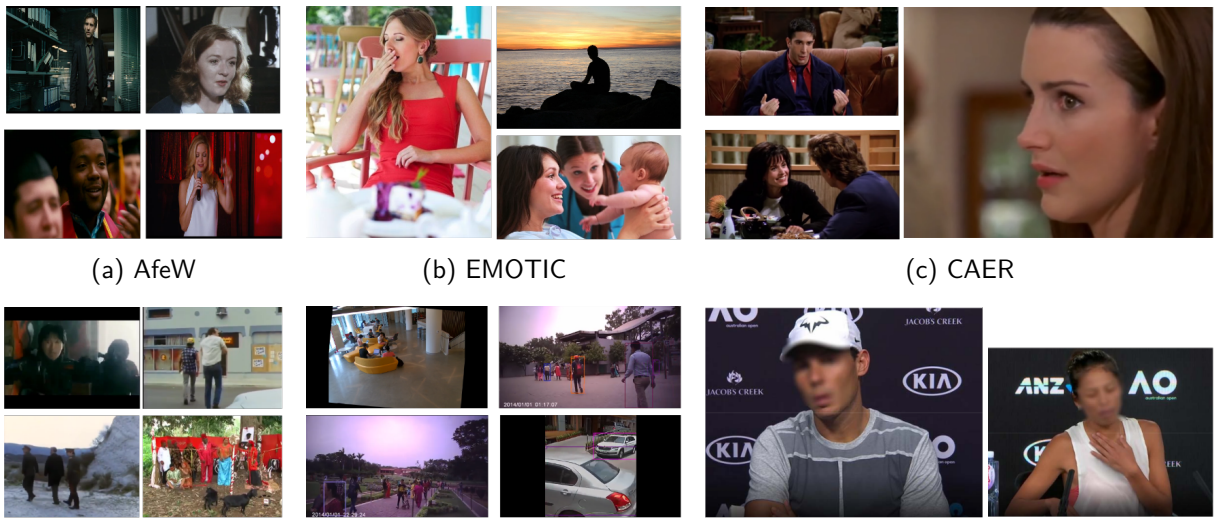
4.1 DATASETS IN THE STATE OF THE ART

In this section, we will introduce the datasets currently employed as benchmarks for techniques in the state of the art. Rather than providing an exhaustive overview of past datasets for emotion recognition, we focus on the recent efforts currently being employed and evaluated in the state of the art. We show samples from this dataset in Figure 7 and give an overview of these datasets in Table 5. As discussed previously, due to the importance of context and other nonverbal cues for perceiving emotion, we will focus our efforts on datasets that explore such features or have them available, discarding, for instance, datasets for Facial Expression Recognition (FER) in this evaluation. Therefore, datasets such as AffectNet (MOLLAHOSSEINI; HASANI; MAHOOR, 2017), the Facial Expression Recognition 2013 (FER2013) dataset (GOODFELLOW et al., 2013), and the Extended Cohn-Kanade dataset (CK+) dataset (LUCEY et al., 2010) will not be considered in this evaluation.

The Acted Facial Expressions in the Wild Dataset (AfeW) (DHALL et al., 2011) addresses the limitation imposed by the lack of data from real-world scenarios, challenging the dominance of lab-posed images. It comprises scenes from 54 movies, focusing on spontaneous expressions in realistic environments. Expression-related keywords extracted from subtitles guide the annotation of expressions. Additionally, the static subset SfeW and the extended AfeW-VA dataset (KOSSAIFI et al., 2017), which introduces a continuous emotion model, extends this resource. We show examples of this dataset in Figure 7a and discuss their annotations in Subsection 4.1.2.

The Emotions in Context (EMOTIC) dataset (KOSTI et al., 2017a; KOSTI et al., 2019) is

Figure 7 – Datasets currently used for the task of emotion recognition. Samples were extracted directly from the dataset, except for **(f)**, which was extracted from their arXiv manuscript with a CC-BY-SA license since the download link for the dataset is currently offline. Although some datasets explicitly show faces, such as **(a)** and **(c)**, others, such as **(b)** and **(d)** have samples with severe occlusion, given their focus on other nonverbal cues.



(a) AfeW (b) EMOTIC (c) CAER

(d) BoLD (e) GroupWalk (f) iMiGUE

Source: Author.

a benchmark widely employed in the current literature for emotion recognition designed to capture subjects in, unconstrained environments, including contexts. It is based on established datasets such as Common Objects in Context (COCO). Annotated using Amazon Mechanical Turk, the dataset underwent multiple rounds of annotation, with up to five annotators per image in test and validation sets. The 2017 version of EMOTIC consists of 18,316 images with 23,788 subjects. The updated 2019 version (KOSTI et al., 2019), which is the focus of this work and illustrated in Figure 7b, includes 23,571 images with 34,320 subjects, maintaining the same gender ratio but slightly altering age distribution.

The Context-Aware Emotion Recognition (CAER) dataset (LEE et al., 2019) dataset was proposed to solve limitations perceived by the authors on other datasets and also on EMOTIC. According to the authors, although EMOTIC contains contextual information, they work on a different aspect and propose a large-scale dataset for context-aware emotion recognition with various context information. We propose an experiment to validate this affirmation later in this work. Another limitation, compared to EMOTIC, is that emotions are highly dynamic, and using images could be a limiting factor for techniques. Therefore, the authors propose the CAER dataset, which contains videos, and the Context-Aware Emotion Recognition (Static) (CAER-S) dataset, for static images, the most commonly used benchmark between these two.

Another main differing factor between CAER and EMOTIC is the sample source. While EMOTIC focuses on reusing images from other popular datasets and complementing them with images from the web, CAER focuses on extracting images from TV shows. This yields questions regarding the expressivity of the emotions in the scenes. The psychology literature tackles the problem of posed emotions, and evidence suggests that they are at least an approximation of what is actually felt. Studies show that we can diminish the impact of these problems by using different actors in the experiments that are unaware of the task - in this case, the actors were unaware that this footage would be used for the emotion recognition task. Therefore we can consider that the negative impact is low (ZUCKERMAN et al., 1976; WALLBOTT; SCHERER, 1986; WALLBOTT, 1998).

The Body Language Dataset (BoLD) (LUO et al., 2020), distinct from EMOTIC and CAER, emphasizes body language as an important nonverbal cue, relatively unexplored in current research (COSTA et al., 2022; CHEN et al., 2022). It consists of crowdsourced emotional data from videos, validated using action recognition methods and Laban Movement Analysis features. Its construction involves three stages: selecting video clips and their durations, annotating poses (using pose estimation and tracking), and identifying perceived emotions. Sourced from the AVA dataset (GU et al., 2018), each character in a clip is uniquely identified and tracked for emotional annotation, with close-up clips omitted for clearer body visibility. The dataset comprises 26,164 video clips with 48,037 instances (characters with landmark tracking), each annotated by five participants on 20 samples. This dataset also has a diversity aspect, including underrepresented ethnic groups such as *Hispanic or Latino* and *Native Hawaiian or Other Pacific Islander*. Despite its potential for in-the-wild emotion recognition applications, the dataset is not yet widely used in the literature.

The ***GroupWalk*** dataset (MITTAL et al., 2020) comprises videos recorded with stationary cameras in 8 real-world scenarios. A total of 10 annotators annotated 3,544 agents with visible faces across all videos. This dataset is still to be widely used in the current literature, given its focus on a very specific nonverbal cue.

The Micro-Gesture Understanding and Emotion Analysis (iMiGUE) dataset (LIU et al., 2021a) is designed to study nonverbal body gestures, especially micro gestures, in the context of perceived emotion. It consists of 359 videos from post-match press conferences featuring professional athletes, offering a total of 2,092 minutes of footage. These videos, capturing athletes' reactions immediately after matches, provide natural, unposed emotional expressions. To ensure privacy and focus on gestures, the dataset is identity-free with masked biometric data

such as faces and voices. It also emphasizes ethnic diversity and gender balance. Annotations in the dataset, done by five annotators, include various aspects like body, head, hand, body+hand, and head+hand gestures, primarily from Grand Slam tournament interviews, with 258 instances of wins and 101 losses.

Recently, datasets recorded in the wild have gained attention for allowing evaluations in uncontrollable scenarios, such as laboratory environments. The AfeW benchmark (DHALL et al., 2011) was one of the first public databases used for this evaluation scenario and was employed as the main dataset for evaluation at the Emotion Recognition in the Wild (EmotiW) (DHALL et al., 2013) challenges. Subsequently, datasets such as EMOTIC (KOSTI et al., 2017a; KOSTI et al., 2019) and CAER (LEE et al., 2019) have been published to explore the participation of context for emotion recognition deeply. The latter are the datasets most commonly used today in the state of the art for evaluation.

Table 5 – Overview of the datasets commonly used in the state-of-the-art. Each row contains a different dataset. VAD or VA are continuous annotations models. For the demographics, M and F stand for Male and Female, respectively, while A, T, and C stand for Adult, Teenager, and Children.

Dataset	Annotations	Annotators	Samples	Demographics
EMOTIC	Categorical (26 classes) and continuous (VAD)	AMT[1]	18,316 images	23,788 people; 66%M and 34%F; 78% A, 11% T, and 11% C
EMOTIC[2]	Categorical (26 classes) and continuous (VAD)	AMT	23,571 images	34,20 people - 66% M and 34% F; 83% A, 7% T, and 10% C
CAER/CAER-S	Categorical (6 classes + neutral)	6 annotators	13,201 video clips (70,000 images)	Unknown
iMiGUE	Categorical (2 classes)	5 trained annotators	359 video clips	72 adults - 50% M and 50% F
AfeW	Categorical (6 classes + neutral)	2 annotators	1,426 video clips	330 subjects aged from 1 to 70 years
AfeW-VA	Continuous (VA)	2 trained annotators	600 video clips	240 subjects aged from 8 to 76 years; 52% female
BoLD	Categorical (26 classes) and continuous (VAD)	AMT	26,164 video clips	48,037 instances - 71% M and 29% F; 90.7% A, 6.9% T and 2.5% C
GroupWalk	Categorical (3 classes + neutral)	10 annotators	45 video clips	3,544 people

[1] Amazon Mechanical Turk. [2] A second version of EMOTIC was published in 2019.

Source: Author

Figure 8 – Examples from the CAER-S dataset with more than one person in the image. The annotation given for the image is displayed as the sub-caption.

(a) Angry (b) Disgust (c) Happy (d) Sad

Source: Author.

4.1.1 Emotion categories and annotations

Techniques related to affective computing usually focus on some variations of the well-known Ekman's (EKMAN, 1992) basic emotions, which are *Anger*, *Fear*, *Sadness*, *Enjoyment*, *Disgust* and *Surprise*. The same applies to datasets related to emotion recognition and parent fields of study, which usually add the *Neutral* emotion.

This set of categorical annotations is the case for the CAER, in which videos and images are annotated using Ekman's basic emotions. A limitation, however, in the format of annotation used in both CAER and CAER-S is that a single annotation is given for the video or image. This means that on images with multiple people, only one label is given, and no bounding box or identifiable information is provided for whom that annotation was made. This harms techniques such as CAER-Net (LEE et al., 2019), GLAMOR-Net (LE et al., 2022), and other techniques trained on this dataset because the authors usually use the first face detected on the image, as discussed in their research papers. EmotionRAM (COSTA et al., 2022) employs a face selector algorithm that searches for the leading performer on the scene based on the assumption that the annotation would be from them but is also limited to images with difficult scenarios. We show examples of images with more than one person framed in Figure 8.

The authors proposed a different approach for EMOTIC by defining an extended list of 26 emotional categories also containing Ekman's basic emotions, leading to 20 novel emotional states for comprehension, which was also employed on BoLD. To define these emotional categories, the authors proposed an approach based on word connections (affiliations and relevance of words) and inter-dependence of words (psychological and affective meaning) to form word groupings. However, the main difference is that both EMOTIC and BoLD do not contain the *Neutral* category, arguing that, generally, at least one category can be applicable,

even with low intensity.

However, the lack of a neutral annotation group could lead to neutral images being sampled into opposite groups and therefore harm the learning process of a network. For example, research on social cognition points out that humans perceive emotionally neutral faces depending on visible traits; positive traits are correlated with happiness, and traits involving dominance and threat are correlated with anger (SAID; SEBE; TODOROV, 2009; MONTEPARE; DOBISH, 2003). Therefore, a person that appears to be emotionally stable would be associated with *Happy*, while a person that appears to be aggressive would be associated with *Angry*. In this scenario, given a neutral image, and more specifically, one with a neutral face, the action of the annotator would likely depend on their perception of the personality traits of the people present in the image instead of the perceived emotion.

4.1.1.1 Assessing annotators agreement

Could a high number of classes impact the agreement between annotators? We hypothesize that the various possibilities of classes could lead to uncertainty for annotation, leading to disagreement among annotators. Therefore, as an example, we investigate the agreement among annotators on the EMOTIC dataset.

The authors also propose a study to assess the level of agreement among annotators. They employed a quantitative metric known as Fleiss' Kappa measure (FLEISS, 1971), which evaluates the reliability of agreement among a fixed number of raters in assigning categorical ratings. They showed that more than 50% of images have $\kappa > 0.30$, indicating that in these cases, the annotations are better than at random. As stated in Fleiss' work, it is reasonable to interpret the absence of agreement among raters as their inability to distinguish subjects - in this case, different emotions. As is also proposed in Fleiss' work, the authors did not evaluate the agreement by category.

We proposed an experiment to overview the agreement among annotators. We empirically chose to focus on the test set, given how this is how current state-of-the-art approaches evaluate their results. For this experiment, given a set of images of the test set $I = \{I_1, ..., I_n\}$ containing a set of persons $P = \{P_1, ..., P_n\}$, we loop each P_i for each I_i and store their annotations if the number of annotations for that person is higher than one. We then count the co-occurrence of each emotion for that sample in pairs and store this information in a co-occurrence matrix. For example, if the annotations for $I_1; P_1$ are *Happy, Engagement* from

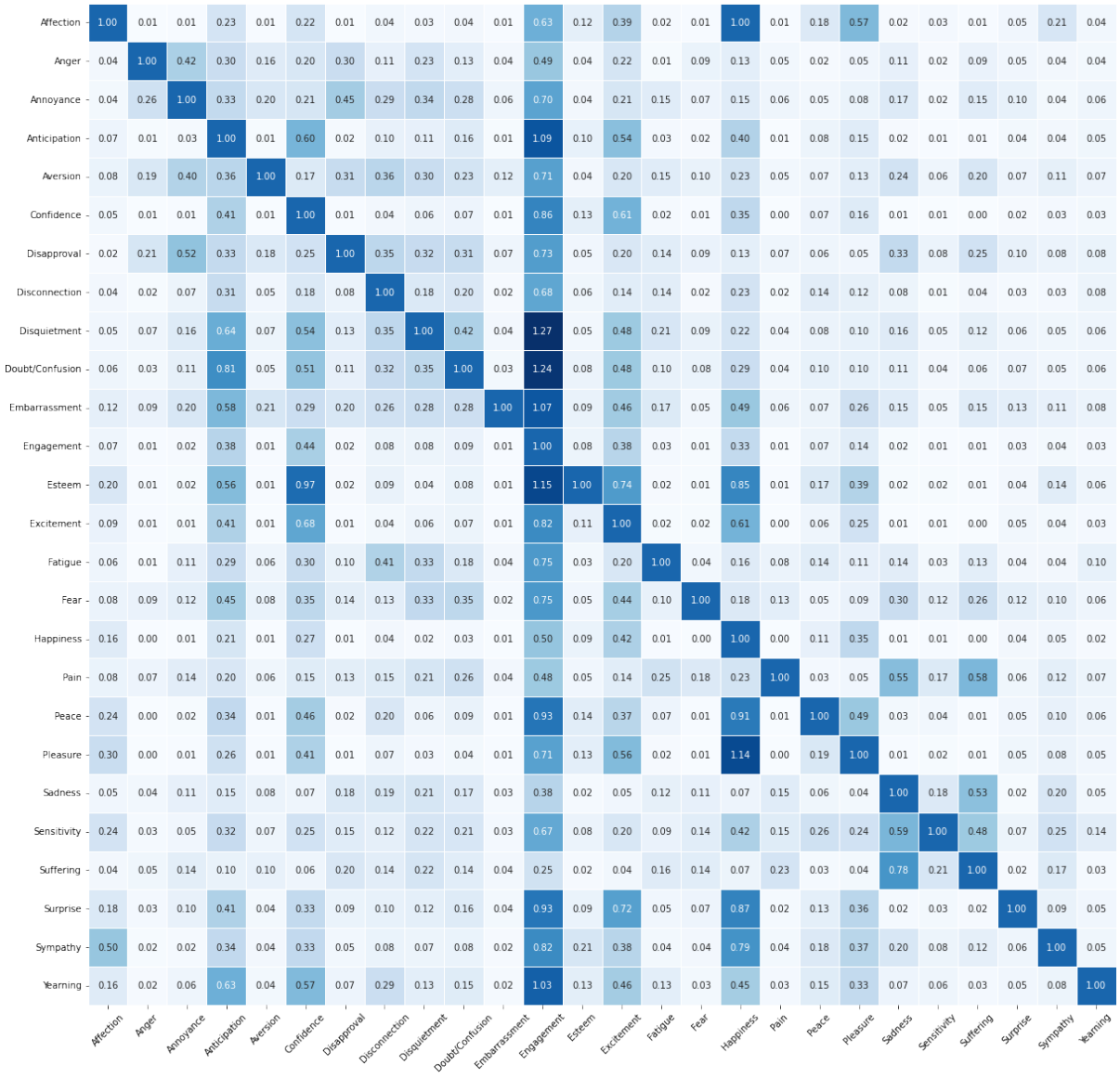
annotator 1 and *Peace* from annotator 2, we would increment the co-occurrences of *(Happy, Engagement)*, *(Happy, Peace)*, *(Engagement, Happy)*, *(Engagement, Peace)*, *(Peace, Happy)* and *(Peace, Engagement)* by one. Therefore, we construct a database containing the co-occurrence of the annotations given by all annotators for each image and person. We display a visualization of this data in Figure 9 as a confusion matrix to allow an easier visualization normalized by emotion. This means that, for a pair of emotions, the value shown indicates the co-occurrence of those emotions. If the value is higher than one, the annotators agreed more with this specific pair instead of agreeing on the emotion. If the value is less than one, they agreed less.

From this overview, we can see a general agreement between classes. For example, the co-occurrence between *Affection* and *Happiness* is 0.99 (1,879 samples), indicating a strong agreement between annotations, as is between *Suffering* and *Sadness* with 0.77 agreement (426 samples). We can also see, however, the odd distribution of *Engagement*, which, by the author's definition, is *paying attention to something; absorbed into something; curious; interested* (KOSTI et al., 2019). *Sensitivity*, which is defined as *feeling of being physically or emotionally wounded; feeling delicate of vulnerable* shares co-occurrences with *Happiness* (0.42; 101 samples) and *Sadness* (0.59; 142 samples), even with these two having opposite definitions. Finally, we can also see outliers in this data composed of directly opposite emotions. For example, *Disconnection* and *Engagement* (0.68; 1,365 samples), *Fear* and *Confidence* (0.35; 104 samples), *Fear* and *Excitement* (0.44; 131 samples), *Pain* and *Happiness* (0.23; 50 samples). A similar experiment was performed by the authors and is published in their reference paper but using a different approach than ours. In our approach, we look directly at the co-occurrence of each pair of emotions and compute the average based on the number of samples, which revealed insights that were not present in their evaluation. These insights should be used by scientists and engineers when developing their applications to be aware of limitations present in the dataset.

4.1.2 Continuous annotations

Some datasets also annotate samples with continuous dimensions instead of only discrete categories. For example, the VAD model (MEHRABIAN, 1980) is usually employed for emotion recognition. In this model, emotions are placed in a three-dimensional space: Valence, Arousal, and Dominance. The Valence (V) axis determines whether an emotion is pleasant or unpleasant

Figure 9 – An overview of the annotations on EMOTIC's test set and the concordance of the annotators. Each cell contains the co-occurrence between the pair of emotions.

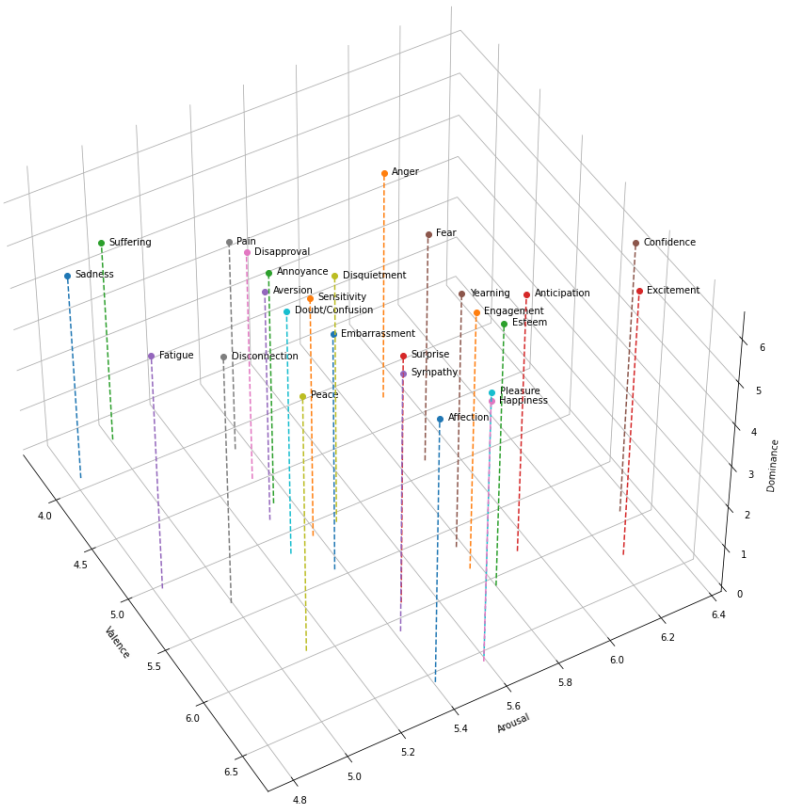


Source: Author.

to the perceiver, therefore distinguishing between positive and negative emotions. The Arousal (A) axis differentiates between active and passive emotions. Finally, the Dominance (D) axis represents the control and dominance over the nature of emotion. Therefore, each emotion can be represented as a linear combination of those three components (KOŁAKOWSKA; SZWOCH; SZWOCH, 2020).

The AFEW-VA dataset (KOSSAIFI et al., 2017; DHALL et al., 2012) contains annotations of valence and arousal for 600 clips, which were also, in part, used in AFEW to create an expanded dataset. The values for each axis of the model range between [-10, 10]. However, as exposed in their work, a significant part of the annotations lies around the neutral value for valence and

Figure 10 – Visualization of the continuous annotations for EMOTIC.



Source: Author.

arousal, indicating that a significant part of the dataset contains neutral expressions. A possible reason for this behavior is that the entire dataset was annotated by just two individuals who, although certified on the facial action units coding system, could be suffering from *annotator burnout* (PANDEY et al., 2022).

The EMOTIC dataset also contains annotations using the VAD model ranging from [1, 10]. However, unlike AFEW-VA, the authors chose to rely on crowdsourced annotations powered by the AMT platform with restrictions to discard annotations that were, in their opinion, not compatible with their discrete annotations. These control images would appear once for every 18 images shown to the annotator. However, as reported in their work, the metrics related to annotation consistency point to disagreement in some cases. For example, the standard deviation for the dominance dimension is 2.12, which can be considered high given that the values range from [1, 10]. For valence and arousal, the standard deviation is 1.41 and 0.70, respectively. Finally, the score distribution for each dimension is higher than AFEW-VA's, indicating that the perceived emotions are more diverse from the annotators' point of view. We extract the mean for each annotated emotion's valence, arousal, and dominance axis on

EMOTIC, and we show this visualization in Figure 10. This figure shows that, as expected, we have a cluster of positive and negative emotions. *Sadness* and *Suffering*, for example, have almost the same VAD mean. At the same time, *Pleasure* and *Happiness* are also close. This visualization also points out annotations that confuse themselves in this form of annotation, such as *Peace* and *Fear*.

Finally, the community does not agree that the VAD model is a good approach to representing emotions. It is challenging to represent emotional categories with numbers (KOŁAKOWSKA; SZWOCH; SZWOCH, 2020), given that this is different from how humans naturally perceive emotions, even more when the emotion needs to be divided into different categories. The representation is also deeply intimate and changes from person to person and culture to culture. In datasets with continuous and categorical annotations, such as EMOTIC, one might employ weights between each annotation format to control their participation in calculating loss, for example.

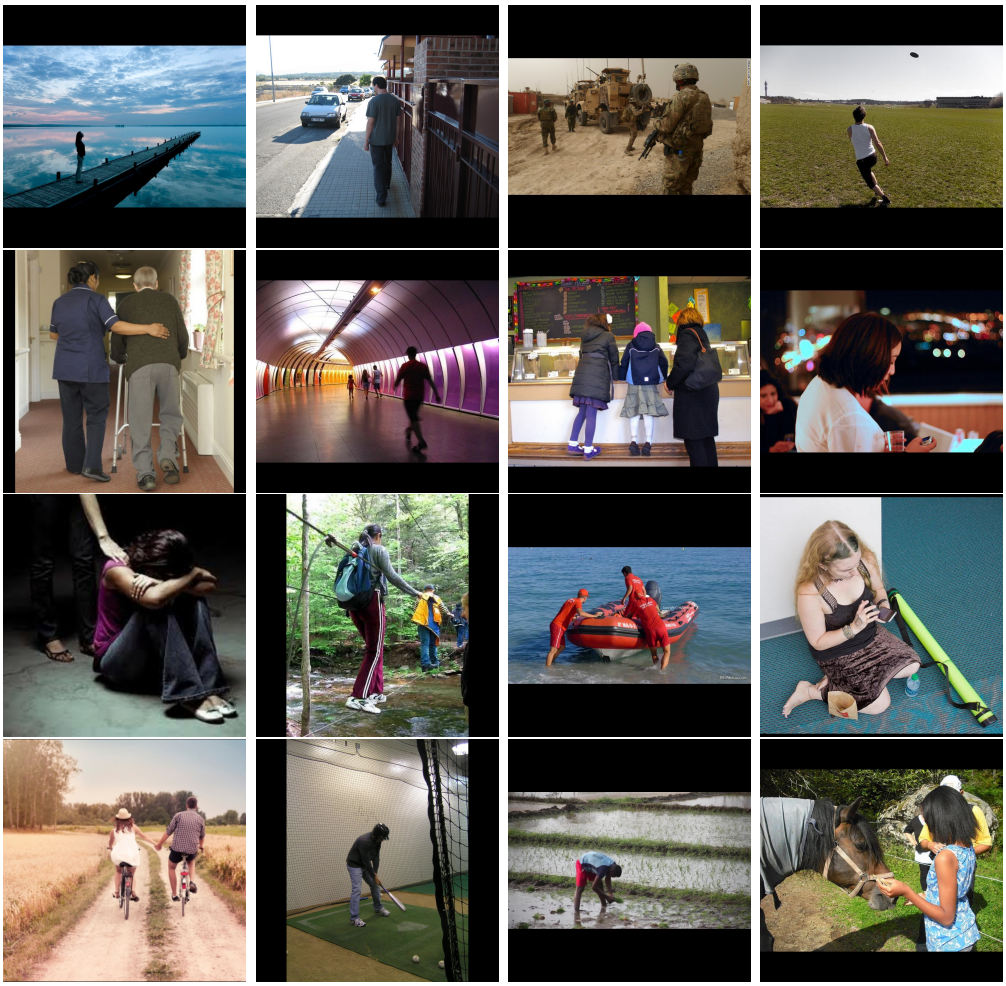
4.1.3 Presence of nonverbal cues

Except for AfeW and AfeW-VA datasets, all other databases surveyed in this work were designed for the presence of nonverbal cues other than facial expressions in their samples. For example, EMOTIC and CAER/CAER-S focus on context; BoLD focuses on body language; GroupWalk focuses on gaits, and iMiGUE focus on microexpressions on the body, hands, face,

Table 7 – Presence and visibility of nonverbal cues in the datasets currently employed at the state-of-the-art. For each dataset (rows), we classify the cues (columns) as ***M**issing*, ***S**omewhat **P**resent*, ***P**resent* or ***A**nnotated* for when the cue is not only visible, but annotated by humans.

Dataset name	**Cue**			
	Facial expressions	Context	Body language	Others
EMOTIC	SP	P	SP	N/A
CAER	P	P	SP	N/A
iMiGUE	P	M	A	Microgestures
AfeW	P	SP	SP	N/A
BoLD	P	P	A	N/A
GroupWalk	SP	SP	SP	N/A

Figure 11 – Examples from the EMOTIC dataset (KOSTI et al., 2017a; KOSTI et al., 2019) with images with severe facial occlusion. For this dataset, it is expected that techniques can extract information from context. Therefore, this is not an issue but rather a characteristic. Each column contains samples from a subset. The images were padded to allow better visualization.



(a) emodb (b) ade20k (c) framesdb (d) mscoco

Source: Author.

and the combination of them. However, it is possible to extract other nonverbal cues from these datasets, even if these were not the focus of their design. We list the presence and visibility of nonverbal cues in Table 7.

Given how facial expressions are severely significant for emotion recognition, it is important to have good crops of the face available, which is the case for most of the samples from these datasets. For EMOTIC, however, it is more common to see images with severe facial occlusion, as shown in Figure 11. For the other datasets, since the data was obtained from TV clips or movies, it is more common for people to be aligned with the camera. For GroupWalk, given how security cameras and handheld devices are also used to record the clips, it is common that sometimes the faces will be far from the camera, making it difficult to explore this cue.

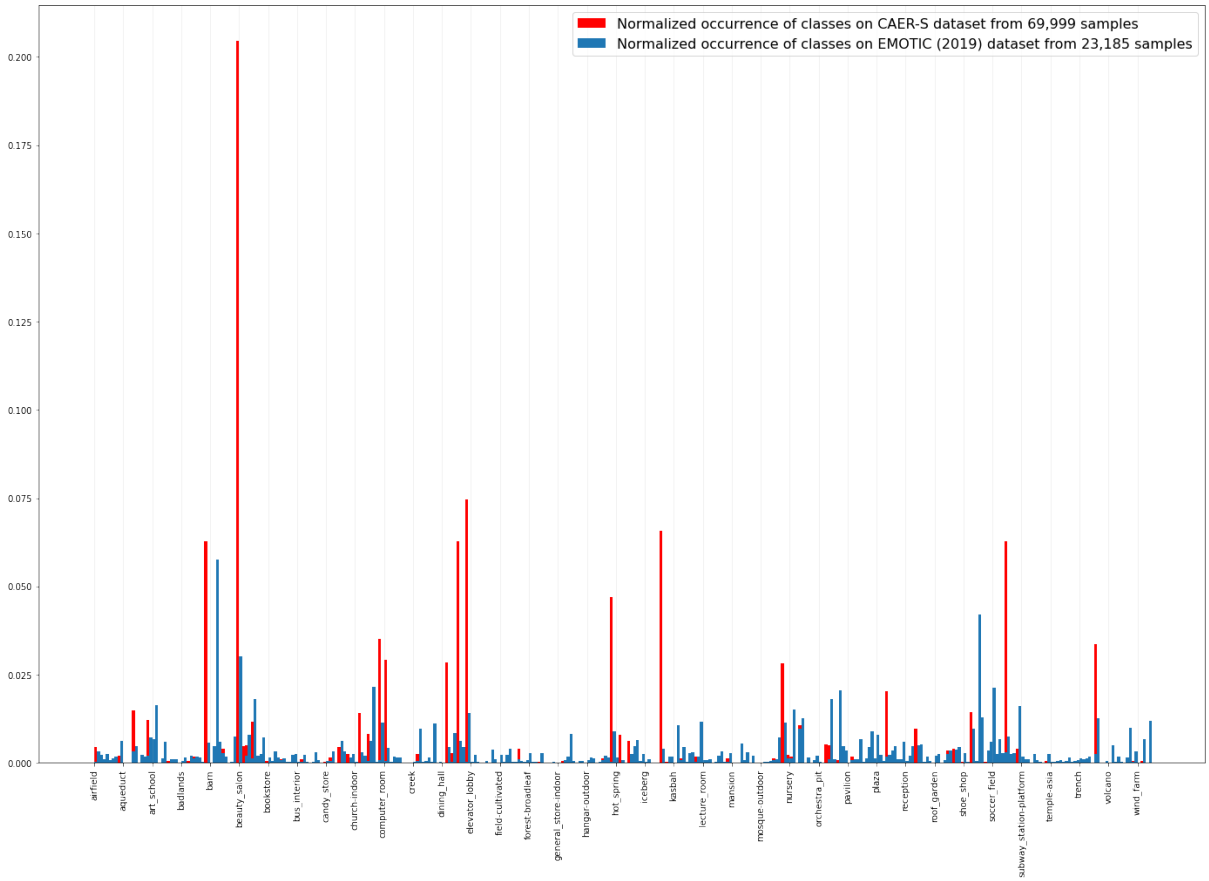
4.1.3.1 Context variability

Context information is also available on most samples of the datasets selected for evaluation in this survey. Even for datasets such as AfeW and AfeW-VA, which are commonly used for FER, background, and scene information is available and can be used to leverage context. However, when proposing the CAER dataset, the authors argue that the list of datasets until the date of publication (including EMOTIC) did not contain a dataset with multiple context information. When exploring CAER-S, we notice that even though the source of the samples is 79 TV shows, a significant number of frames contain repeated background information, which does not occur on EMOTIC at the same significance due to the source of the data.

To validate the variability of background information (context) on EMOTIC and CAER-S, we propose an experiment using open-source code for scene recognition. For this, we use the work by (LÓPEZ-CIFUENTES et al., 2020), which extracts features from images for classification. Given an image I, we feed it to the pipeline of the proposed network to extract the scene label s, which describes at some level the context from the input image I_c. Please notice that although this technique is compatible with the state of the art with a sufficient Top@1 score, our intent is not to quantify the different context information to disclose how many images have a specific background but rather to understand if these images are different enough to have different classifications among the dataset.

First, we modified the data loaders proposed by the authors to load images from the CAER-S and EMOTIC datasets separately. After this, we employ the same transforms used in their evaluation to extract comparable information. We decided empirically to use the model pre-trained on the Places 365 dataset (ZHOU et al., 2017) given its high data variability. After this, we sampled each image and extracted the Top@1 classification of it, storing it in a file for further processing. We expose the results of this experiment in Figure 12. This graph shows that the classifications given by (LÓPEZ-CIFUENTES et al., 2020) are more grouped on CAER-S than on EMOTIC, implying that the images on this dataset are less diverse. Please notice that although the datasets contain a different number of samples, our investigation is regarding data distribution. A dataset with images containing diverse backgrounds would be more distributed among this graph, as it happens with EMOTIC, independent of the number of samples present in it. Finally, this experiment complements individual observation on the dataset, indicating that CAER and, subsequently, CAER-S do not have a higher context variability than EMOTIC.

Figure 12 – A bar chart with the occurrence of the classifications of each image from CAER-S (red) and EMOTIC (blue) with the scene recognition approach proposed by López-Cifuentes et al. (2020). The difference in the height of each bar indicates the difference in the number of samples of each dataset. A dataset could be considered balanced if the samples were distributed equally among the classes. For EMOTIC, we can see that this behavior is more visible than on CAER-S.



Source: Author.

4.1.3.2 Body keypoints visibility

Another essential cue that is present in these datasets is the body. Even though only specific datasets such as iMiGUE have annotated body language (see Table 7), scientists and engineers can extract insights from this cue by using different types of approaches. Costa et al. (2022), for example, employed a body encoding stream that used a simple model proposed by Xiao, Wu and Wei (2018) until its last convolutional layer, allowing the network to correlate the internal representations of this model and emotion. After that, more robust approaches were proposed, which focused on an activity recognition pipeline, such as Mittal et al. (2020) and Chen et al. (2022). We propose an experiment to assess the visibility of the body keypoints on images from three datasets: EMOTIC and CAER, which are the main datasets used today in the literature, and also BoLD, given its body language focus.

In this experiment, we loop through every image of the datasets except for BoLD, in which we choose a single random frame from every video. This different approach is motivated by some reasons, such as (a) there is low variability between frames in each video of BoLD and (b) the high number of recorded frames would make this experiment unfeasible due to the high processing time. Please notice that although BoLD has ground-truth body keypoints annotation, we chose not to use them in our evaluation to keep a comparable basis regarding body keypoint visibility.

We use You Only Look Once (YOLO) (REDMON; FARHADI, 2018) for each image or video frame to detect the people present on the scene. EMOTIC has ground-truth annotations regarding this aspect. Again, we chose not to use them in our evaluation to keep a comparable basis among datasets. For each detected person, we feed the cropped region to MediaPipe (LUGARESI et al., 2019) to extract the detected pose. We empirically chose MediaPipe for this experiment due to its capability of also predicting keypoint visibility. Next, we sum the visibility of each of the 33 keypoints for each dataset. Finally, we plot these points using a spatial distribution representing a neutral pose body. For the drawing, we normalized the keypoint visibility regarding the maximum value stored to allow better comparison between points in the same and different datasets. We show the results of this experiment in Figure 13.

From the results of this experiment, we can notice that CAER-S is the dataset with less visibility regarding the lower body parts. Given how the dataset is built upon video clips from TV shows, it is common to have the lead performer centered on the screen and only show the lower body parts when relevant to the story. BoLD, although being focused on body language, also has less visibility for the lower body parts when compared to EMOTIC. Please notice that they provide body annotations and that this result may differ when their annotation is used. Finally, EMOTIC has a well-distributed visibility of the whole body, which is justified by the multiple camera placements present on the dataset. All of them have good visibility for the face, hands, and arms.

4.1.4 Discussion

Scientists and engineers currently use all datasets listed in this work to develop emotion recognition models and pipelines. However, in which scenarios techniques trained using these datasets could be deployed? Models, no matter how deep or robust, are still limited to the quality and availability of data. Therefore, based on features extracted from these datasets, we

Figure 13 – Visibility of each individual body joint as keypoints for the datasets evaluated in this experiment. From top to bottom, we have face, arms and hands, legs and feet, following the 33 pose landmarks defined by Mediapipe and available here. The size and heatmap color of each sampled point represents the visibility of the keypoint.

(a) EMOTIC (b) CAER-S (c) BoLD

Source: Author.

propose a discussion regarding application scenarios and gaps that we could expect in these approaches. Finally, before working on the construction and deployment of the technology, one should compare the requirements with the details described in each dataset above. Researchers should also know that these datasets were not built toward a specific requirement. The best practice is to have a dataset to fine-tune techniques for each specific need. We discuss below the possible application scenarios and limitations for each dataset.

4.1.5 Application scenarios

In this section, we will discuss how datasets can be used to train models for the application scenarios we defined previously in Section 3.2. To avoid repetition, the discussion and motivation of each application scenario will be omitted, focusing on the aspect of the dataset.

For smart environments and spaces (Subsection 3.2.1), existing surveillance cameras, often top-down, can be utilized, focusing less on facial expressions and more on other cues due to limited visibility. GroupWalk and BoLD can be recommended for this scenario. In retail or private sectors, cameras with a frontal view can capture clear facial expressions, and models trained on AfeW or CAER could benefit from this data.

For affective and assistive robotics (Subsection 3.2.2), models for this task could be trained using EMOTIC to allow context encoding and semantics representation (WU et al., 2022). Of course, one cannot expect a child or a person that must be accompanied to look at the robot all of the time, so techniques must be robust to facial occlusion.

For emotion tracking and mental health (Subsection 3.2.3), EMOTIC could also be employed here, given its high context availability and people who are not always facing the camera,

to track mood from photos from the gallery and cameras placed in the person's house. For example, the techniques could benefit from a high-level context description to correlate house objects with emotions.

4.1.6 Limitations

Specific tasks. The datasets iMiGUE and GroupWalk, discussed in this work, are designed for specific tasks, with each having distinct limitations. iMiGUE, aimed at emotion extraction from interviews, is constrained by its binary emotion representation (positive or negative) and identity protection measures, like hidden facial expressions, limiting its application in broader scenarios like job interview comfort assessment or customer care quality evaluation. GroupWalk, although offering camera placement variety for robustness, suffers from sample quality issues, such as video distortions from stabilization software, noise, low image quality, and bounding box annotations directly on images. These flaws potentially decrease model accuracy, despite known benefits of data augmentation techniques like random crops and rotations. The literature still lacks datasets for other specific tasks, such as emotion recognition from security camera viewpoints, a challenge compounded by the limitations of using 2D data.

Labels. Another clear limitation in this field is related to the labeling of datasets, as revealed in our survey. EMOTIC and BoLD, which uses 26 categories, do not include a neutrality class, leading to issues like correlation of visible traits with emotions and disagreement among annotators. Contrarily, CAER and CAER-S use a 7-category system including neutrality, but lack the structured annotations of EMOTIC, such as individual emotions in scenes with multiple people. There's also inconsistency among datasets regarding emotion classes (Table 5), limiting applicability in specific scenarios. While transfer learning could theoretically bridge these gaps, challenges like overfitting, catastrophic forgetting, and resource investment arise, as discussed in (ZOPH et al., 2020), (LI; ZHANG, 2021), and (CHEN et al., 2019; XU et al., 2020).

Additionally, the lack of ontologies for emotion representation complicates even further this matter. Existing emotion ontologies, such as MFOEM (HASTINGS et al., 2011), are not currently utilized in datasets. This result in vague categorizations. For example, Surprise, according to MFOEM, may include a positive or negative modifier based on event evaluation (and even context), a nuance missing in datasets, as we exemplify in Figure 14. Adopting such ontologies could unify the diverse label sets and provide literature-based ground-truth descriptions, enhancing the efficacy and applicability of emotion recognition datasets.

Figure 14 – An illustration of different types of surprise. In **(a)**, the person is feeling positively surprised due to a surprise birthday party. In **(b)**, the person is feeling negatively surprised due to a message their received in their phone.

(a)

(b)

Source: Author.
(Generated using Generative AI, as disclosed in Chapter 7.1)

Cultural representation. As discussed previously in Section 2.2, the perception and communication of emotion is highly mutable due to culture. Most of these datasets were collected from videos of movies or TV shows in early 2000's, which were not very representative. During our research, we performed some assessment related to this impact in data and have noticed that some models do not work as well on people from Brazil[1]. Although BoLD does contain a "Hispanic or latino" ethnical category in their dataset, it comprises only 8.41% of their data (LUO et al., 2020).

[1] We do not have references for this discussion as we are still formalizing a study on this topic.

4.2 A DATASET FOR EMOTION RECOGNITION ON LATIN AMERICAN CULTURES

In light of the complexities surrounding the cultural specificity of emotions (see Section 2.2), we can raise concerns about how current emotion recognition datasets are built. These datasets, predominantly composed of content from movies and TV shows produced in the United States and Europe, reveal a significant gap in cultural representation. This lack of diversity is not *just* a matter of fairness, as it directly impacts the accuracy and applicability of emotion recognition technologies in underrepresented cultures within these datasets. The lack of cultural representation leads to the risk of biases and misinterpretations in emotion recognition systems when applied to different cultures. Therefore, addressing this imbalance in dataset composition becomes imperative for developing more effective and universally applicable emotion recognition models.

We propose a new dataset for emotion recognition in Latin American cultures, focusing on Brazilian culture. The Emotions in LatAm dataset (EiLA) is a small-scale annotated database containing people in their context. The situational context is important for emotion recognition, as it allows evaluations beyond the typical facial expression recognition approach, allowing techniques to leverage how the context could impact someone's perception of emotion.

4.2.1 The EiLA benchmark

Although most existing datasets, such as EMOTIC and CAER-S, have a sufficient amount of data and cue visibility, the only dataset containing cultural specificity for Latin Americans and/or LatinX people is BoLD (LUO et al., 2020), although containing only 8.41% of data under "Hispanic or Latino." Qualitatively, it is also possible to conclude that this specificity would be equal to or under 8.41% in other datasets. Therefore, our aim was to create a database that contained a structure that was similar to the datasets from the state of the art (see Section 4.1, but with cultural specificity in mind), which would allow the deployment of emotion recognition models on Latin American countries with improved bias robustness. We show samples of our dataset in Figure 15.

Figure 15 – Sample images from EiLA. The dataset contains various different scenarios and people of different skin tones.

Source: Author.

Data collection

It was clear that obtaining completely spontaneous reactions would be difficult, as by the Brazilian General Data Protection Law (*LGPD - Lei Geral de Proteção de Dados*)[2], we are required to notify people before recording them. Therefore, inspired by other datasets in the literature, we have focused on capturing data from publicly available TV shows. After several rounds of brainstorming with other lab members, we have decided to focus on reality shows, given how these are captured in multiple different scenarios and often have different viewpoints of the same person.

Data was collected by manually cropping clips of these TV shows considering the cultural background and skin tone color of the participants. With this approach, we could ensure cultural sensitivity and participation among different emotions.

[2] Available in English at <https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/>

Dataset composition

The EiLA benchmark dataset includes annotations for 15 minutes of video in a dynamic setting, encompassing 4,521 annotated frames. The dataset features 78 distinct participants, evenly split between 39 male and 39 female individuals. In terms of skin tone, 46 participants have a light skin tone, 22 have a dark skin tone, and 10 have a mixed skin tone. The participants' ages range from 18 to 65 years, representing various ethnic groups primarily from Latin America, especially Brazil. Updated versions of EiLA may include increased participation from certain groups. To facilitate access to new versions and related research, a central repository is available at <https://eila-dataset.github.io/>.

Annotation

The annotation was performed independently and blindly by three annotators, who were volunteer students working with emotion recognition at Voxar Labs, Centro de Informática, Universidade Federal de Pernambuco. All students have passed through onboarding processes and have been working with emotion recognition for some time.

We first collected clips from Brazilian TV shows. We manually annotated each person, attributing to them a bounding box with their location and a unique identifier among the video to allow the usage of the dataset on video-based emotion recognition techniques and as a benchmark for person re-identification. Each video was manually annotated using Ekman's basic emotions, which are *Anger*, *Fear*, *Sadness*, *Enjoyment*, *Disgust* and *Surprise*.

Out of the 4,521 frames initially sampled, we have successfully compiled 8,086 annotated samples, each related to the people visible in the scenes. Within this dataset, a subset of 901 samples (approximately 11%) received annotations from three different annotators. Two annotators annotated a further 1,426 samples (representing about 17%). The majority, comprising 5,723 samples, which account for roughly 70% of the total, were annotated by a single individual. This distribution highlights the varying levels of annotator engagement across our dataset. Following EMOTIC's proposal, the sets with two or three annotators could be used to define a test or validation set. In contrast, the larger sample of one annotator could become the training set. Overall, the dataset has an equal division of perceived gender (50% male and 50% female).

Table 9 – Analysis of concordance in the annotated samples.

Dataset	κ **score**
EMOTIC (KOSTI et al., 2017a)	0.31
EMOTIC (KOSTI et al., 2019)	0.30
EiLA (2 annotators)	0.44
EiLA (3 annotators)	0.42

Annotation consistency

We verify the concordance of the annotations for the sets with two or three annotators to check the consistency among different people. To measure this agreement quantitatively, we compute the *Fleiss' Kappa Score* (κ) using the two and three annotations in each set.

We show the result of the consistency in Table 9. Both versions of EMOTIC (KOSTI et al., 2017a; KOSTI et al., 2019) have a κ score close to 0.30. This indicates that the annotators have a higher concordance of the samples in our dataset, which might be related to the fact that they are researchers in emotion recognition and not crowdsourced personnel from AMT.

5 COLLECTING AND PROCESSING AFFECTIVE FEATURES

Revisiting the (early) emotion recognition taxonomy we proposed in Section 3.1, we will proceed to discuss our core technical contributions within the vision-based approaches, comprehensively discussing the construction of our models and frameworks falling under this classification.

5.1 HIGH-LEVEL CONTEXT REPRESENTATION FOR EMOTION RECOGNITION

In this section, we discuss our results of a high-level context representation approach for emotion recognition. This discussion is built upon the results published in our paper with the same title as the section (COSTA et al., 2023b)[1] on the LatinX in AI workshop at CVPR 2023. Portions of the text, figures, and tables from the original work have been incorporated in this section.

Among the diverse representations of human behavior, emotion recognition has been a research topic of interest in the last few years. As we discussed previously in this thesis (Chapter 2), an intelligent system that is able to perceive emotions needs to be able to capture and process nonverbal cues. In unrestricted in-the-wild scenarios, our emotions are influenced by information in the situational context. For example, a person sitting on a beach, enjoying the sun and the sea during their vacation, is more likely to experience positive sentiments such as joy and happiness. In contrast, a person stuck in a traffic jam filled with noise pollution could be inclined towards more negative sentiments such as frustration, anger, or stress. Therefore, environmental stimuli should also be taken into consideration when analyzing emotion.

Researchers have been proposing approaches that take into consideration contextual information for a while in works such as EMOTIC (KOSTI et al., 2017b), CAER-Net (LEE et al., 2019), EmotiCon (MITTAL et al., 2020) GLAMOR-Net (LE et al., 2021), and EmotionRAM (COSTA et al., 2022), each proposing new approaches on how to leverage context and extract its representations from images on different datasets. We hypothesize that, although working on these low-level representational features could and has led to significant results in the past, generating semantic, high-level descriptions could be more assertive to unseen data, leading

[1] Available at <https://openaccess.thecvf.com/content/CVPR2023W/LatinX/html/de_Lima_Costa_High-Level_Context_Representation_for_Emotion_Recognition_in_Images_CVPRW_2023_paper.html>

to better results in test sets and when deployed to solve real-world problems.

In some scenarios, high-level representations of emotions can be a valuable aid for decision-makers to make informed choices. For example, consider a city planner who needs to decide which public parks in the city need to be improved or renovated first. To make this decision, the planner needs to know how people feel when they are in these spaces, but they may not need to know each experience to make this decision. Instead, the high-level representation could be provided as an overview of the emotions associated with that context and how people act towards it. It could be easily compared without needing to act on top of a significant amount of data. Approaches such as this one have several advantages, such as resource-saving.

In this work, we propose an approach for the extraction of high-level context representations of images for the task of emotion recognition. We show in our experiments that these highly representative descriptions of context are capable of yielding results comparable to the state-of-the-art of emotion recognition on the EMOTIC dataset (KOSTI et al., 2019) by itself and could easily be placed into a complete emotion recognition pipeline as a context encoding module to lead to significant improvements in accuracy. We also show that our proposal can perform a fast inference, a desirable feature for low-consumption edge devices that could be deployed in the wild. The contributions of this work are as follows:

- We propose a novel framework that builds a high-level representation of extracted context descriptors from images (Subsection 5.1.2) and employs Graph Convolutional Neural Networks (GCNs) to classify these representations into emotional categories (Subsubsection 5.1.2.5).
- We benchmark on the well-known EMOTIC dataset, achieving a comparable accuracy with the state-of-the-art (Subsection 5.1.4).
- We discuss how our model's low computational power requirements make diverse applications possible to solve real-world problems (Subsection 5.1.4).

5.1.1 Related works

Extending the techniques based on Facial Expression Recognition (FER), researchers have been investigating how adding other nonverbal cues could improve the pipeline, and a cue that is commonly investigated is context. For example, EMOTIC (the model, not the dataset) (KOSTI et al., 2019) proposed a baseline for this approach, in which both the person and the

context in which they are placed would be considered. In the case of facial occlusion, for example, context would also be contributing to emotion perception. CAER-Net (LEE et al., 2019) and GLAMOR-Net (LE et al., 2021) follow the same path, however, employing different forms of how to weigh context contributions and, therefore, how important contextual information should be in each scenario.

However, all of these techniques mentioned above have the same limitation by design: the lack of definition of what should be considered as context. For example, the approach proposed by Le et al. (2021) considers detecting the face of a person, completely occluding it with a black rectangle, and using this new image as a representation of context. However, the other body parts are still visible, as are the body parts of other people in the scene, and this image would be fed to a context encoding stream that is designed to extract features from the scene automatically. However, are these encoding streams capable of doing this task without prior knowledge?

Other approaches, such as EmotiCon (MITTAL et al., 2020), proposed that it is necessary to use multiple independent and specialist streams to generate representations that can be correlated to emotion, given how context is highly descriptive. Specifically for EmotiCon, as an example, the authors propose the usage of the following context streams: (1) multimodal context, with facial landmarks and body keypoints; (2) situational context, extracted by processing the background image with the person occluded by using a pedestrian tracking method; and (3) socio-dynamic context, which computes proximity features using depth maps.

In a more recent approach, Chen et al. (2023) proposed models combining different representations from context. For example, they use a deep network for each person on the scene to calculate their social relations between intimate, not intimate, and no relation. They also propose a deep reasoning module for using multiple context representations that are extracted locally and globally and involve scene recognition and body pose estimation, among other modules.

However, humans perceive context differently (BARRETT; KENSINGER, 2010; BARRETT; MESQUITA; GENDRON, 2011). The literature suggests that humans encode context naturally by using our internal representations of meaning in the image. Therefore, it is not natural for us to take calculated steps to understand context. Instead, our brain automatically classifies these stimuli as positive, neutral, or negative based on our previous knowledge of that information (PASTOR et al., 2008). Our approach differs from the techniques mentioned earlier due to the more straightforward approach for context, in which we try to mimic the context representation

of humans based on our best knowledge of the literature on nonverbal communication and behavioral psychology.

5.1.2 Methodology

In this section, we describe our approach for extracting high-level representations from context. Given how humans describe and understand context in images, we propose extracting high-level descriptions of images to correlate them with semantic features. This approach mimics how humans correlate semantic descriptions with emotions to improve interoperability.

5.1.2.1 High-level descriptions

Given an image as input, we first want to extract high-level image descriptions. We employ ExpansionNet-v2 (HU; CAVICCHIOLI; CAPOTONDI, 2022), an image captioning model based on the Swin-Transformer architecture (LIU et al., 2021b). We first traverse through the EMOTIC dataset, and for each sample, we input the image to ExpansionNet-v2 for captioning generation.

We then process the raw caption to generate a refined caption. First, we perform the removal of stop words from the caption. Stop words are common words in a language, such as articles, prepositions, and pronouns, but do not have any semantic meaning. Therefore, maintaining these words would only elevate the complexity, given their high frequency in the English language, and by removing them, we have a more representative corpus. We use spaCy[2], a public library for natural language processing. We also remove common nouns such as *man*, *woman*, *girl*, and *boy* as to not allow the model to create generalization between gender and emotions, as this could induce to gender-related biases. We also applied lemmatization to reduce each word to its root form. The remaining words are called *valid words* and will be used in the following steps to generate data representations. Finally, we show examples of images, their original captions, and their processed captions in Figure 16.

5.1.2.2 Co-occurrence mining

The second step involves the generation of co-occurrence matrices that will represent patterns of labels within the dataset, which will be employed in the future through conditional

[2] Available at <https://spacy.io/>

Figure 16 – Examples of images from EMOTIC with their raw captions and processed captions.

Raw caption: A bride and groom cutting a wedding cake
Processed caption: bride groom cut wedding cake

(a)

Raw caption: A man riding a wave on a surfboard in the ocean
Processed caption: ride wave surfboard ocean

(b)

Raw caption: A man and a girl sitting at a table with food
Processed caption: sit table food

(c)

Source: Author.

probability. After preprocessing the captions in the dataset, we store this information and count the occurrence of each emotion, and the valid words of each caption, resulting in a matrix $M_c \in \mathbb{N}^{W \times C}$, where W is the number of valid words from the corpus and C is the number of emotion categories in the dataset. Therefore, $M_{c_{ij}}$ denotes the number of times that emotion C_j occurred when the valid word W_i also occurred. We call this matrix the emotion co-occurrence matrix.

Based on the same assumption, we also generate a co-occurrence matrix based on the co-occurrence of valid words. Given a window of size s, we slide this window to capture the co-occurrence of the valid words, resulting in a matrix $M_w \in \mathbb{N}^{W \times W}$. Therefore, $M_{W_{ij}}$ denotes the number of times that the valid word W_i appeared together with the valid word W_j.

5.1.2.3 Semantic descriptions

For each valid word W, we extract semantic representations that can be correlated with emotion. Given how ExpansionNet-v2 is a model to generate captions in a generic context, extracting the semantic representations of the word will lead to better representations of affective meaning in that caption.

For extracting semantic descriptions, we employ SenticNet (CAMBRIA; HUSSAIN, 2015), a knowledge base for semantics, sentics, and polarity associated with natural language concepts. We query each valid word, and we extract the following attributes: the two mood tags associated with the concept; the pleasantness sentic, which represents the perception of pleasantness or unpleasantness of the word; the polarity value, which represents the overall sentiment of the word and finally, the semantically-related concepts.

Except for the valid word not existing on the SenticNet knowledge base, we use WordNet (MILLER, 1995; MILLER, 1998) to search for synonyms. The advantage of this approach is that the words in WordNet are grouped using synsets, which are sets of synonyms with similar concepts or meanings. Therefore, by querying a word in WordNet, its synonym will have a significant relationship and most surely have the same meaning. For each possible synonym, we rank the list according to the similarity with the valid word and iterate through the list, selecting the first synonym present in SenticNet. In the rare case that the valid word is not present on SenticNet and neither are its synonyms, we drop the valid word from the caption and proceed to the next step without it.

5.1.2.4 Graph generation

With this prior knowledge (e.g., co-occurrence and semantic representations), we can capture relationships between valid words and emotions and also between themselves. Given how they are a particularly effective method of describing structured data, we choose to model these representations using graphs. Although some of the knowledge is learned prior, the definition and construction of graphs are done as needed and in real-time. This allows this technique to generate representations from unseen data.

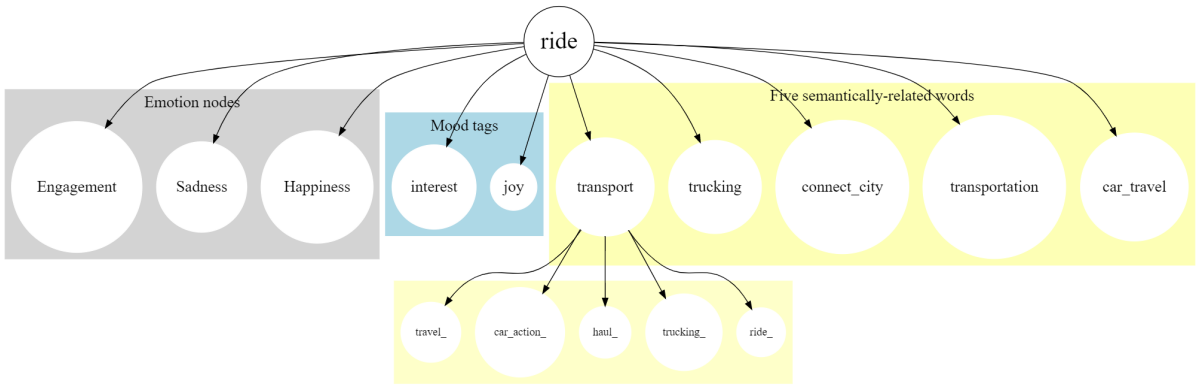
We use Deep Graph Library[3], a framework-agnostic library for generating and manipulating graphs. We start by constructing an empty graph $G = (V, E)$, in which V is a set of nodes and E is a set of edges. In this case, $V = E = \{\emptyset\}$. For each valid word W, we start by adding a new node V_{W_i} to the before empty set of nodes V of the graph. We use GloVe (PENNINGTON; SOCHER; MANNING, 2014) to fetch the valid word embedding and use this representation as the feature $X \in \mathbb{R}^{50}$ for node V_{W_i}. If the valid word is absent on GloVe, we randomly sample this embedding from a uniform distribution [-0.01, 0.01]. We save this representation for future use in case this valid word reappears.

Next, we add a node V_C for each emotion category C in the dataset. For EMOTIC, since we have 26 possible emotions, we add 26 nodes and place edges $e = (V_W, V_{C_i})$ between the valid word and each emotion. We define the weight w_e according to the equation below:

$$w_e = P(C_i|W) = \frac{M_{C_{W,i}}}{\sum M_{C_W}}, \tag{5.1}$$

[3] Available at <www.dgl.ai>

Figure 17 – Example of the generated graph. For brevity and visualization, we consider only one valid word in this scenario. For each valid word node, we create nodes for emotions (also reduced for brevity), mood tags, and semantically-related words. For each semantically-related word, we also query SenticNet and extract their semantically-related words, as shown in the "transport" node.



Source: Author.

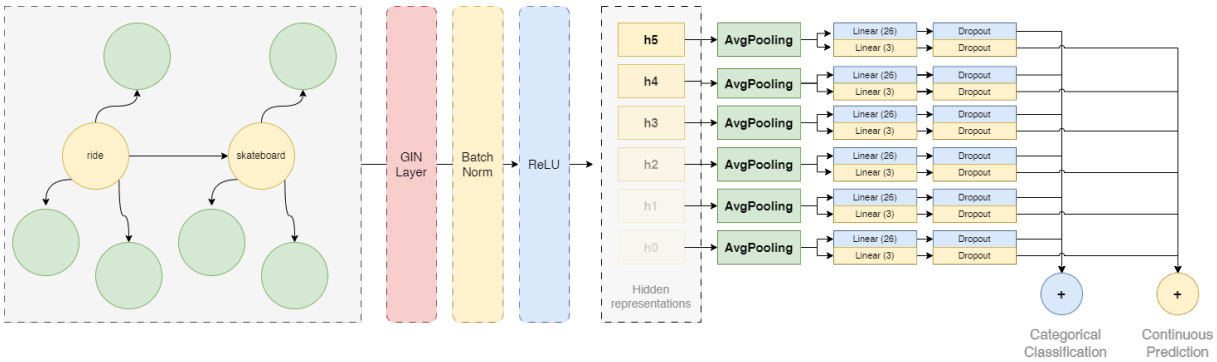
where the edge weight w_e between the valid word node V_W and the i_{th} emotional category C_i is $P(C_i|W)$, which is given by the co-occurrence between the valid word W and emotion category C_i divided by the sum of the co-occurrence between the valid word W and all possible emotions C, extracted from the co-occurrence matrix M.

Next, we add nodes related to the sentic semantic description of the word. First, we add two nodes relative to the mood tags extracted from SenticNet, and we set the weights of the edge between the valid word node and them to be the pleasantness value of the word. Next, we add five nodes relative to the five semantically-related words available from querying SenticNet and add edges using the polarity value as weight.

For each of the five semantically-related words to the valid word, we also query SenticNet and extract their five semantically-related words. The polarity value of the word in the first level gives the edge between these connections. We hypothesize that by adding another level of semantic relationships, we will be able to extract even deeper representations of context.

We perform the same process for each valid word W in the caption. After the nodes for all valid words are created, we add edges between these nodes. The weight of each edge is given by the co-occurrence of the words $M_{W_{i,j}}$, divided by the total number of times the word appeared. Finally, we show an example of the graph with reduced information for brevity in Figure 17.

Figure 18 – Our proposed architecture for high-level context representation. Given an input graph that is generated using previously learned knowledge and image captioning, we use an adaptation of GIN (XU et al., 2018) to classify it among a set of emotions. Given how EMOTIC also has annotations using a continuous model, we adapt the pipeline to generate two predictions, which are considered to calculate the loss.



Source: Author.

5.1.2.5 Deep GCN for Emotion Recognition

Given the construction of graphs to represent context, we use a deep graph convolutional neural network for graph classification and, consequently, for emotion recognition. Given a set of graphs $G_1, ..., G_N$ and a set of emotion categories $C \in \mathbb{R}^{26}$, we aim to classify each graph according to an emotional category. For this task, we propose adapting Graph Isomorphism Networks (GIN) (XU et al., 2018), chosen due to its simple architecture, which could lead to reasonable inference rates in low-energy, low-consumption devices.

First, given a graph as input, we store this graph's features directly in the hidden representations stack as $h0$. After this, we loop through a GIN convolutional block containing a GIN layer, batch normalization, and ReLU. We iterate over this block five times in this approach, generating representations $h1$ to $h5$. Finally, we iterate through the hidden representations, average pooling these features and reducing their dimensionality. In parallel, we keep a stream for the categorical classification, which outputs classification labels C, and another stream for continuous predictions for a VAD model (check Subsection 4.1.2 for more details).

Therefore, we learn categorical labels and continuous values during training. We define our loss as a weighted combination of the individual losses of each output. Given a prediction $\hat{y} = (\hat{y}_{cat}, \hat{y}_{cont})$ in which $\hat{y}_{cat} \in \mathbb{R}^C$ and $\hat{y}_{cont} \in \mathbb{R}^3$, we define the loss in this prediction as $L = \lambda_{cat} L_{cat} + \lambda_{cont} L_{cont}$, where L_{cat} and L_{cont} represents the loss of each individual prediction. For L_{cat} implement a weighted euclidean loss as used in EMOTIC (KOSTI et al., 2019), which is defined as follows:

$$L_{2_{cat}}(\hat{y}_{cat}) = \sum_{i=1}^{26} w_i(\hat{y}_{cat_i} - y_{cat_i})^2, \tag{5.2}$$

in which $\hat{y}_{cat_i}$ is the prediction for the i_{th} category and y_{cat_i} is its ground-truth label. The weight w_i is defined as $w_i = \frac{1}{ln(c+p_i)}$, where p_i is the probability of the i_{th} category and c is a parameter to control the range of valid values. We also employ a L2 loss for L_{cont}, defined as:

$$L_{2_{cont}}(\hat{y}_{cont}) = \sum_{j=1}^{3}(\hat{y}_{cont_j} - y_{cont_j})^2. \tag{5.3}$$

Finally, we train our model on the EMOTIC dataset using the abovementioned features. We use the default PyTorch data loader and implement access to the list of graphs to feed the model during execution. We show an overview of our model in Figure 18.

5.1.3 Experiments

Dataset

We perform our experiments on the Emotions in Context (EMOTIC) dataset (KOSTI et al., 2019), using the 2019 version, to allow direct comparison with the state-of-the-art (check Section 4.1 for more details).

Comparison with the state-of-the-art

We compare our results with other techniques of state-of-the-art, namely EMOTIC(KOSTI et al., 2019), Zhang's work (ZHANG; LIANG; MA, 2019), EmotiCon (MITTAL et al., 2020), DRM (CHEN et al., 2023), LEKG (CHEN et al., 2023), which are two variations of Chen's method (CHEN et al., 2023), and Yang's work (YANG et al., 2022). However, as we later describe in Table 11, these techniques are often built on top of a combination of multiple nonverbal cues.

Validation metrics

Besides the quantitative evaluation using the Mean Average Precision (mAP) metric, as is done in the current literature (KOSTI et al., 2017b; KOSTI et al., 2019; ZHANG; LIANG; MA, 2019; MITTAL et al., 2020; CHEN et al., 2023), we present some examples to perform a brief qualitative evaluation of the predictions.

Table 11 – Quantitative evaluation of our approach compared with state-of-the-art models on EMOTIC dataset.

Technique	mAP	# of nonverbal cues
EMOTIC (KOSTI et al., 2019)	27.38	2 (body and context)
Zhang, Liang and Ma (2019)	28.42	1 (two-stream context analysis)
EmotiCON Mittal et al. (2020)	35.48	4 (face, body pose, context, depth)
DRM (CHEN et al., 2023)	26.48	5 (three body descriptors and two context descriptors)
LEKG (CHEN et al., 2023)	29.47	2 (scene recognition and global context)
Yang et al. (2022)	**37.73**	**5 (face landmark, body pose, context, agent relationships, and human-object interaction)**
Ours	30.02	1 (single-stream context)

Implementation details

We train our model from scratch, learning the parameters using Adadelta (ZEILER, 2012). After an empirical comparison of multiple values on the validation set, the batch size is set to 16. We use a learning rate of 0.001 and a weight decay of 0.0004.

Regarding the experimentation environment, we train and validate our model on a desktop computer running Ubuntu 20.04 LTS with an Intel i7-4790K with 32 GB of RAM and an NVIDIA RTX 2080 Ti with 12GB of VRAM. For training and experimenting with our model, we use PyTorch 1.12 with CUDA 11.3 and CuDNN 8.3.2. For the experimentations regarding inference time, we also compare it with a consumer-grade notebook with Windows 10 Pro, 16 GB of RAM, and an NVIDIA GeForce GTX 1060M with 6GB of VRAM.

5.1.4 Results and discussion

We compare our results with different approaches in Table 11. Our proposed method outperforms other graph-related methods, such as the work by Zhang, Liang and Ma (2019), and also DRM and LEKG, which are two variations of Chen et al. (2023). We did not compare our method with Chen's TEKG model because it has a local approach that could not be extended to real-world problems by itself. Although EmotiCon (MITTAL et al., 2020) reported

Table 13 – Ablation study of the proposed method.

Method	**mAP**
GIN (XU et al., 2018) (AvgPooling)	0.3002
GIN (XU et al., 2018) (SumPooling)	0.1715
Simple GCN	0.2505

a higher mAP than our method by 5.46 mAP, according to Chen et al. (2023), their result of 35.48 mAP is not reproducible, reporting a score of 26.87 mAP, in which our method can perform by 3.15 mAP. Additionally, EmotiCon uses four nonverbal cues, while we only employ one. Yang et al. (2022) reports the highest mAP in this dataset, with a result of 37.73 mAP, which is 7.71 mAP higher than our result, by also employing five nonverbal cues. Therefore, we demonstrate that our model is competitive with the state-of-the-art, even with just one cue. While Zhang, Liang and Ma (2019) also use only one cue, they process it at two levels and can be considered two contextual cues.

The number of nonverbal cues employed is directly related to the inference time of the model, a question that we wanted to tackle. Emotion recognition models should be easily deployed on edge when thinking about real-world situations. This would allow for multiple data capture and processing points without increasing spending too much or requiring high energy consumption. These are two current barriers imposed when deploying deep learning models in the wild. However, in cases where multiple cues are needed, our model could act as a context encoding stream, even with a convolutional neural network, to extract descriptions and contribute to the overall perception of emotion.

We conducted ablation studies on our model to assess its performance under different configurations. Our findings, present in Table 13, indicate that using sum pooling instead of the current pooling approach for GIN results in inferior performance. Furthermore, we also compared the performance of our model with that of using a simple GCN consisting of two GCN blocks, ReLU activations, and a classifier. Our model outperformed this simple GCN, which resulted in a lower mAP.

We also evaluate our model with a qualitative analysis, as shown in Figure 19. For each model on the EMOTIC test set, we feed this image to the proposed pipeline, generating categorical predictions for the image. Although the model can also generate results using the VAD model, it is difficult for humans to understand and compare these values since this is unnatural for us. Therefore, we choose to use only categorical values for the qualitative

evaluation. This experiment shows, as expected that our method looks for cues in context for emotion prediction. In Figure 19a and Figure 19b, the model could predict all categories present on the ground truth. In this case, given how the context is representative, our model can act well and give an overview of the emotion in that scene. In opposite cases, such as Figure 19c, the context is not representative, and the network cannot predict any correct emotion class. In Figure 19d, the context is related to a set of emotions, for example, positive emotions, but the perceived emotion of the person is actually negative. In this example, when looking at the person, we can perceive an emotion related to tiredness, which is confirmed by the ground truth *Fatigue*. However, since the model does not look at face or body language from context, it perceives the wrong emotion. Finally, for Figure 19e and Figure 19f, the context is very generic, but the model can extract cues from it and classify correctly, at least on some level.

Figure 19 – Qualitative results of our model on the EMOTIC dataset. For each image, we have the ground-truth emotion as annotated in the dataset and the prediction of the network.

Source: Author.

Finally, we test our model on different environments to assess the computational power required and the inference time. We execute the entire testing pipeline by setting a batch size of 1 for individual predictions and evaluate on both environments described in Subsection 5.1.3. We store each individual prediction into a list and then compute the minimum and average values. The minimum value indicates the sample in which the inference was faster, while the average value indicates the average inference time for the model. In a moderate deep learning machine, the inference of our model took 4.0264ms as a minimum, and 4.1546ms on average, leading to ≈ 248 fps and ≈ 240 fps, respectively. For a consumer-grade notebook, the inference of our model took 8.9597ms as a minimum, and 10.3898ms on average, leading to

≈ 111 fps and ≈ 96 fps, respectively. Finally, on the same consumer-grade notebook without using CUDA, the inference of the model took 13.0021ms as a minimum and 17.7365ms on average, leading to ≈ 77 fps and ≈ 56 fps, respectively, on an Intel Core i7-7700HQ @ 2.80GHz CPU.

We do not compare our inference time with the other techniques we evaluated above since neither has official open-source implementations. However, we may infer that models such as EmotiCon, which uses various other deep learning models to extract and process specific cues, would take longer than ours for execution.

5.2 AN ANALYSIS OF GAIT FOR EMOTION RECOGNITION

This section is built upon the insights and results gained from the work done by Maria Luisa Lima in her undergraduate final project, which was co-advised by me (LIMA, 2023)[4]. The results from that exploration have laid the groundwork for the analysis presented here.

A person's gait is the description of the way they walk. Multiple researchers have used gait for studies related to human behavior recognition, such as fall risk assessment (BAUTMANS et al., 2011), detection (CHEN; LIN, 2010), and prediction (LIANG et al., 2019), as well as the prediction of illnesses such as Parkinson's disease (BIASE et al., 2020), dementia (ARDLE et al., 2020), and other neurodegenerative diseases (DENTAMARO; IMPEDOVO; PIRLO, 2020). These analyses are possible through the study of quantitative gait-related parameters.

Research in behavioral psychology indicates that gait-related parameters are not merely physical attributes but are also reflective of social aspects. Studies have shown our innate ability to discern individuals through gait patterns, including self-recognition (BEARDSWORTH; BUCKNER, 1981) and identifying close friends (CUTTING; KOZLOWSKI, 1977). These findings highlight gait's role as a unique behavioral marker containing social information. We have also discussed in Subsection 2.1.1 how gait can be viewed as a dynamic body language representation that highlights motion in an unconstrained scenario. Expanding upon these insights, we hypothesize that gait analysis can also be effectively utilized in emotion recognition.

5.2.1 Related works

Early works for gait-based analysis for emotion recognition were based on extracting features from gait and using algorithms to compare motions with databases. Venture et al. (2014) investigated using motion capture data and computational modeling for this approach. First, four actors performed walks displaying emotions and extracted joint angles from inverse kinematics models to analyze motion capture data relative to lower torso, waist, and head movements. They applied Principal Component Analysis (PCA) and verified that these emotions could be distinguished through these extracted features. Finally, they implemented a similarity index algorithm that compares test motions to the previously gathered database to identify the conveyed emotion. Although they achieved a significant result of 80% in their evaluation, these features might have higher significance in lab evaluations and might not be

[4] Available in Brazilian Portuguese at <https://repositorio.ufpe.br/handle/123456789/52705>

able to perform well in other scenarios.

Another work by Daoudi et al. (2017) proposed a more robust geometry-based approach. It represented the dynamics of skeleton joints over time using covariance matrices, which were mapped to the Riemannian manifold of symmetric positive definite matrices. This allowed the paper to exploit the geometric properties of this manifold for classifying emotions. However, covariance matrices may not fully capture body motion's temporal evolution and dynamics, imposing a limited sequence modeling.

The natural move here was related to how to allow models to learn the spatiotemporal relations of joints. Randhavane et al. (2019b) presented a new approach focused on recurrent neural networks, thus allowing a better spatiotemporal relationship. They combined affective features, such as the angles between joints and stride length, with deep features that were learned using a Long Short-Term Memory architecture.

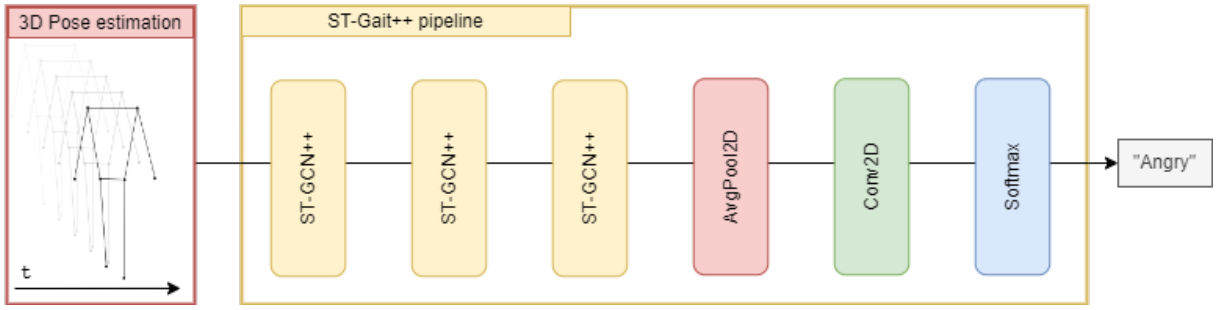
However, the advances on GCNs make a more robust way of learning these relationships possible. Bhattacharya et al. (2020) proposed using such architecture to extract features from videos and classify the emotions implicitly. Using the ST-GCN (YAN; XIONG; LIN, 2018), the joints of the skeletons are directly encoded into the architecture, allowing for a higher representation of the gait.

Still, while ST-GCNs provide an effective approach, there are some limitations that this work aims to address. First, the representational capacity of the base ST-GCN is predefined rather than learned, which was sufficient for its originally intended application of activity recognition. However, for perceiving emotional cues through nonverbal behaviors like gait, these cues are often more subtle than the movements used for activity recognition. Therefore, not learning the topology may limit the ability to capture these subtle movement patterns that are indicative of different emotions.

5.2.2 Methodology

Given a video $V \in \mathbb{R}^{N \times H \times W \times 3}$ containing N frames, H height and W width, we want to infer an emotion y perceived from the gait of a person in this video. We first extract a set of 3D body keypoints $\mathbf{K} \in \mathbb{R}^{16 \times 3}$, in which $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_{16}$, each $\mathbf{k}_i$ represents the location of a body joint in space.

Figure 20 – Our proposed architecture for this method.



Source: Author.

5.2.2.1 Graph generation

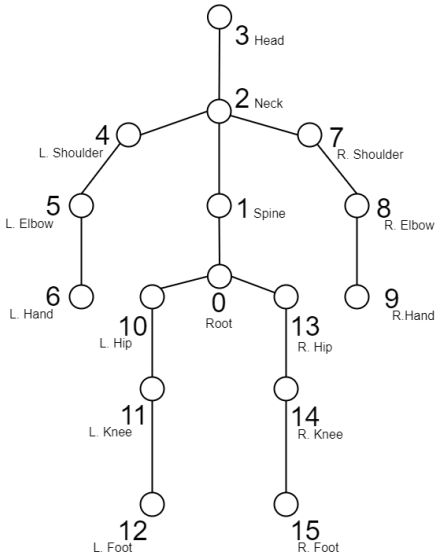
One of the possible ways to represent a skeleton is through a graph. Each body joint, such as the right shoulder and right elbow can be seen as an edge, and the bone that connects these two joints can be represented as a vertex. This is a clear indicative on why GCNs are a good candidate for processing these types of data. Therefore, given K, we start populating a graph $G = (V, E)$ with the location of the joints and their connections, which will be used as input for a GCN.

5.2.2.2 ST-Gait++ for gait processing

We use this graph as input for our gait processing model that is built upon ST-GCN++ (DUAN et al., 2022), a Graph Convolutional Neural Network (GCN) that contains several features for improved spatiotemporal analysis, such as a predefined joint topology in which the weights are learned during training without any sparse constraints or the usage of a temporal convolutional network with branches of different kernel sizes and dilations for improved fine-grained temporal patterns and movements. Inspired on the literature from behavioral psychology (ROETHER et al., 2009), this architecture allows for a complete encoding of affective features related to gait. We show our proposed model in Figure 20.

In the main model, each input gait is fed to a set of 3 ST-GCN++ layers with sizes 32, 64, and 64. We average pool the output of these ST-GCN++ layers in the spatial dimension, as also adding support to translation invariance to the system. Finally, we add a pointwise convolution with a 1×1 kernel to reduce the dimensionality, acting as a linear bottleneck. A pointwise convolution was chosen here due to its parameter sharing capabilities, as well as the

Figure 21 – Composition of joints present in E-Gait.



Source: Author.

preservation of the spatial information, which would be lost when using simple fully connected layers. Finally, we apply a Softmax layer to achieve the classification.

Previous research in gait analysis for emotion recognition has demonstrated that some affective features for gaits can provide meaningful information for emotion perception, improving the accuracy of models (CRENN et al., 2016; BHATTACHARYA et al., 2020). These features include information about posture, such as the angle and distance between joints, and movement features, such as speed and acceleration of individual joints in the gait. These features are calculated during data loading and we append them into the pipeline, following the previously established baseline.

5.2.3 Experiments

Dataset

We have used the E-Gait dataset (BHATTACHARYA et al., 2020) for our experimentation. Given the specificity of this dataset, we have decided not to introduce it in Chapter 4, but rather discuss it here. From qualitative analysis, we have noticed that the synthetic data present in the dataset does not contain naturally-flowing gaits. Therefore, we have decided to use only real data in this evaluation.

Each sample of the dataset is a tensor of shape $T \times V$, in which T is the total time of the gait, and V is the total number of coordinates (16 body joints with 3 dimensions each,

Table 14 – Quantitative analysis of methods for gait-based emotion recognition on the E-Gait dataset.

Methods	**Acc. (%)**
Venture et al. (2014)	30.8
Daoudi et al. (2017)	42.5
Li et al. (2016)	53.7
Crenn et al. (2016)	66.2
Randhavane et al. (2019a)	80.7
Narayanan et al. (2020)	82.4
Bhattacharya et al. (2020)	82.1
Bhattacharya et al. (2020) (our implementation)	83.3
ST-Gait++ (Ours)	**87.5**

leading to $V = 48$. We show in Figure 21 an overview of the composition of the joints.

Implementation details

We implemented ST-Gait++ using PyTorch (PASZKE et al., 2017) 1.7.0. In order to obtain a comparative baseline, we also implemented STEP (BHATTACHARYA et al., 2020), and we trained on the same E-Gait that is publicly available.

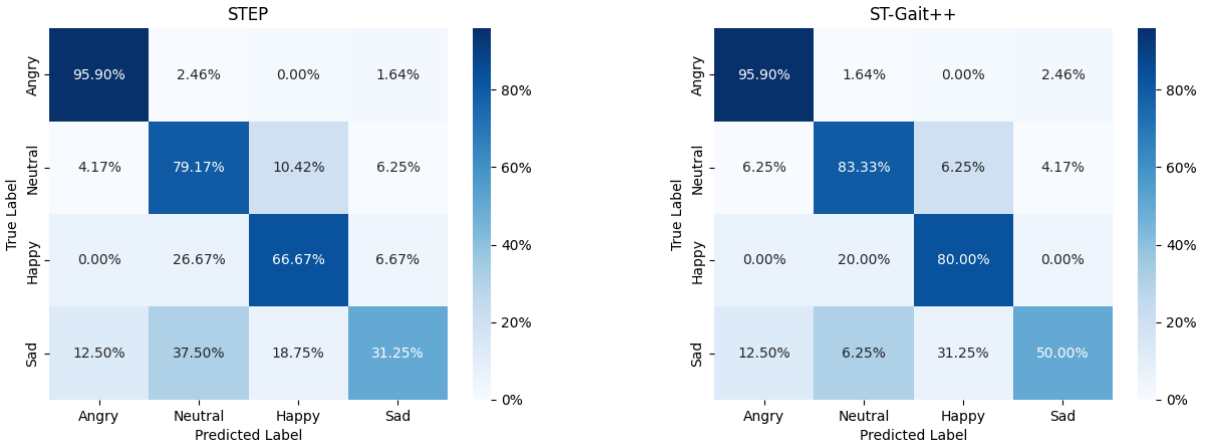
We have trained ST-Gait++ using Adam (KINGMA; BA, 2014) with a learning rate of 0.01 and a weight decay of 3×10^{-4}.

5.2.4 Results and discussion

We compare our quantitative results with different approaches in Table 14. As we can see, our proposed method outperforms other GCN-related methods, such as Bhattacharya et al. (2020). It highly performs other temporal methods (RANDHAVANE et al., 2019a), as well as feature-based or geometric methods. We also evaluate the individual performance of the evaluated classes, and we show in Figure 22 that our method is less ambiguous than STEP (BHATTACHARYA et al., 2020). For *Neutral* and *Happy*, which are the most ambiguous emotional classes (this will be discussed in depth in Subsection 5.3.3), we see that our model has increased accuracy when compared to STEP. Overall, we have increased accuracy in all classes, except for *Angry*, which was maintained the same.

Besides the accuracy increase, our model was also able to converge faster, highlighting several improvements, such as fewer requirements for computational resources or training time,

Figure 22 – Confusion matrices for **(a)** STEP and **(b)** ST-Gait++.



(a) (b)

Source: Author.

increased possibilities for scaling, and better generalization of data. ST-Gait++ converged on epoch #127, while STEP converged on epoch #462.

As discussed previously, we have used the E-Gait dataset for our experimentation. Although this is a well-known, well-accepted dataset in the state of the art, there are several problems in its utilization, such as:

- **Availability**. According to the dataset authors, the current available E-Gait dataset was modified after the original paper was published, and that version is not available anymore[5]. The previous version of the dataset consisted of 4,227 real data and 1,000 synthetic data, and the currently available version has 2,177 real data and 4,000 synthetic data.

- **Lack of original data**. The E-Gait dataset contains only skeletons that were already collected and preprocessed. This means that we do not have access to the original videos, which makes the qualitative evaluation process difficult since we are restricted to skeleton views. This also makes it difficult to adapt a model trained on this dataset to a real-world problem, as the preprocessing function is unknown.

- **Lack of variability**. A significant portion of E-Gait data is imported from the Edinburgh Locomotion Mocap Dataset (ELMD) (KLEINSMITH; BIANCHI-BERTHOUZE, 2012),

[5] Source: <https://web.archive.org/web/20231228142857/https://github.com/UttaranB127/STEP/issues/11>

a dataset recorded using a sole male actor. This severely impacts the variability of the dataset since gender and cultural variability are significantly related to gait perception.

Some of these limitations might have a direct impact on the generalization capabilities of this model, which are yet to be evaluated in-the-wild to assess possible biases and fairness issues. This study will not be possible with the current data, however, since we do not have the mapping of the preprocessing functions used in this dataset. Finally, although the data availability might be different from the other references in Table 14, we may argue that our version of the dataset has close to 50% less real data than others published previously, and if that we had access to that data the overall accuracy could be significantly improved.

5.3 MULTIPLE CUE PROCESSING IN STATIC DOMAINS

In this section, we discuss our results for a multiple cue processing method in images. This discussion is built upon the results published in our preprint entitled "A Fast Multiple Cue Fusing Approach for Human Emotion Recognition" (COSTA et al., 2022) [6]. Portions of the text, figures, and tables from the original work have been incorporated in this section.

Nowadays, given the current advances in smart cities and smart environments, there is a focus on building more attractive and lively spaces. For this, physical sensors and Internet-of-Things frameworks continuously gather data related to weather, pollution, and traffic. However, intelligent systems that can analyze human behavior are essential to gaining a deeper understanding of citizens and users of that area, depicting actions and emotions that could lead to the generation of key insights for policymakers, urbanists, and other interested parties. Examples are the works by Meng et al. (2020) and Wei et al. (2019) that explored how urban noise and the proximity of public parks to city centers impacted the expressions of citizens, and the work by Sajjad et al. (2019) that analyzed emotional shifts in groups that could indicate hostile situations and would prompt security services.

However, these studies highlight two main limitations. First, the current state of research is mainly focused on facial expression recognition, motivated by the number of discriminative features on the human face. Secondly, the amount of computational power required to run inference on current state-of-the-art models makes it unfeasible to expand these systems on a city scale, as the requirements for hardware would grow significantly.

In this work, we tackle these two limitations by investigating the use of multiple cues that correspond to nonverbal communication. We have developed three individual streams that gather and process data from face expressions, context, and static body language, as shown in Figure 23, tackling the first limitation on most current systems. With this, our goal is to recognize emotion in unconstrained scenarios, allowing for different applications in the context of smart cities and smart environments. Our approach for Emotion Recognition on Adaptive Multi-cues (EmotionRAM) comprises image preprocessing and face localization techniques, which improve the generalization ability on the CAER-S dataset (LEE et al., 2019), coupled with the usage of self-calibrated convolutions and body keypoints detection.

Although we are not the first to propose extending facial expressions and leveraging other nonverbal cues, these proposals add a significant aspect to the discussion: the trade-off be-

[6] Available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4255748>

Figure 23 – Given the sample images, we present the information that our proposed multi-cue learning framework learns from to recognize human emotions from images: **(a)** facial expressions, **(b)** context, and **(c)** body pose.

Source: Author.

tween accuracy and computational power. The deployment of deep learning systems has raised significant concerns related to global warming, scarcity of resources, and energy consumption. Also, given the high computational requirement, a significant amount of computational power would need to be applied to recognize emotions on a large scale. Tackling the second limitation, we have developed a model with simplicity by design that is deeply inspired by how humans perceive and process visual features to classify emotions. With this approach, we can maintain a significant accuracy while being faster than the state-of-the-art.

- A novel learning framework that relies on multiple nonverbal cues for emotion recognition, focusing on reproducibility and based on psychological aspects of emotion (Subsection 5.3.2)
- A benchmark on the well-known CAER-S dataset (LEE et al., 2019), in which we ranked second by a difference of 0.12% in accuracy while being more than nine times faster than the first ranked approach (Section 5.4).

5.3.1 Related works

Context awareness

Extending these works, Kosti et al. (2017b) proposed a dataset and a baseline that also leverages context when predicting emotion. By extracting features relative to context, they fuse the prediction of two encoding streams to predict emotion. Lee et al. (2019) and Le et al. (2021) propose direct improvements over this technique, using an attention module that focuses on salient parts of the scene to boost context comprehension. Also, they overlap the face of the person with a black rectangle to force the context encoding to search for cues in the background. By design, these architectures place body language inside context by allowing a single encoding stream to decode background information and body language. Therefore, neither approach is trained with specific knowledge of body pose.

Body pose

Based on the concept that body expression can help the perception of emotions, Randhavane et al. (2019a) propose the analysis of gait to classify emotions. Given an RGB video of an individual walking, the authors use 3D pose estimation techniques to create a set of 3D poses, which are investigated spatiotemporally. Although this technique uses information from body language, it requires the user to be walking to extract features such as the swing of the arms and posture, therefore imposing constraints on the user and limiting the range of applicability. A more recent approach by Bhattacharya et al. (2020) improves the previously mentioned technique by 14% on the accuracy metric. However, it is also limited to the users walking.

Other techniques are more focused on extracting body pose and body language as a cue for emotion recognition. Wu, Zhang and Ning (2019) propose three encoding streams in their work, in which face, body, and context would be processed individually in an architecture that employs DenseXception blocks (CHOLLET, 2017), which would then be fused to extract the emotional category of the image. A more recent approach was proposed in EmoSeC (THUSEETHAN; RAJASEGARAR; YEARWOOD, 2022), which, besides a stream that looks for non-target subjects, also proposes three encoding streams that individually process face, body, and context. However, a limitation of these techniques is that they proposed a single design of an encoding stream and reused it in every cue extraction pipeline. For example, in EmoSeC,

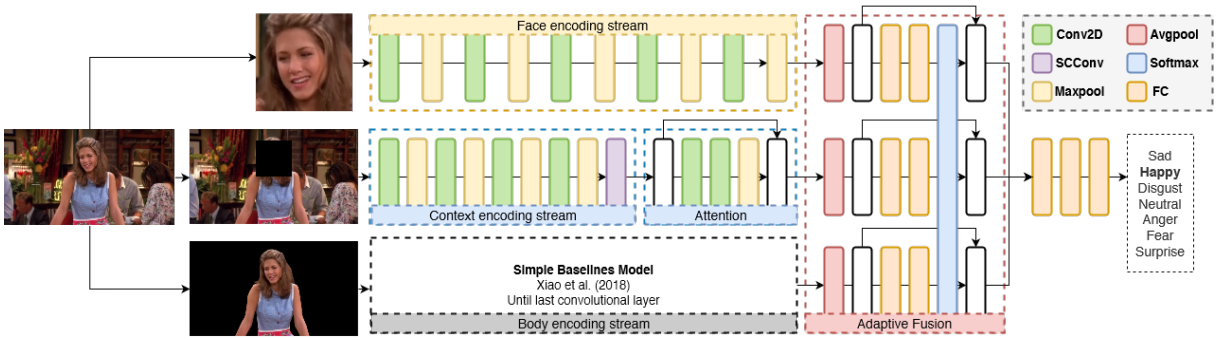
they propose the DeepNet architecture and employ DeepNet-F, DeepNet-B, DeepNet-R, and DeepNet-I for face, body, non-target subject, and image, respectively. The limitation of this design is that each cue is presented differently, and the same process to extract emotional meaning from the face will not necessarily work well for context or body language.

Finally, other works follow the same approach we propose in this work related to body pose (MITTAL et al., 2020; YANG et al., 2022). However, by extracting the 2D locations of each keypoint, the model needs to be able to generalize between coordinates in a space and emotion. Considering that datasets usually have diverse viewpoints, people, and occlusion, it is difficult to understand how these models can create these correlations. For example, Yang et al. (2022) describe that three independent dense layers with GeLU activations (HENDRYCKS; GIMPEL, 2016) are used to extract the features from this space. More recent work by Chen et al. (2023) extracts the 2D representations of pose and inputs it to STEP (BHATTACHARYA et al., 2020), setting the temporal aspect of the spatiotemporal analysis to one, forcing the model to work on static images. It is also unclear how STEP could generalize spatiotemporal configurations with no temporal analysis since this is a restriction imposed by the technique, as is also the alignment of the person to the camera. Therefore, although these previous works also use specialist systems to understand pose, the way that these are processed is not well known or evaluated and, from a generalistic point of view, could lead to limitations. Coulson (2004) proposed an experiment regarding how humans attribute emotion to static body postures, and their findings are that humans model this cognitive behavior by looking at different variables simultaneously. Still, the description of these variables is unknown. Therefore, if one's approach is to design a system inspired by how humans perceive emotions, such representations could also be limited from this aspect.

Comparison with the state-of-the-art

When comparing our work with the state-of-the-art, we notice that these works also employ multiple encoding streams to process nonverbal cues in parallel. However, we may summarize how they process these nonverbal cues in two main groups: first, techniques that employ generalistic encoding streams and, therefore, are unable to learn deep correlations to these cues (KOSTI et al., 2017b; KOSTI et al., 2019; THUSEETHAN; RAJASEGARAR; YEARWOOD, 2022; COULSON, 2004); second, models that are limited in their multiple cue extraction, and therefore have limited capability for processing emotion recognition in diverse scenarios (LEE et al., 2019;

Figure 24 – Our proposed architecture for multi-cue emotion recognition. Given an input image in an unconstrained scenario, we use an off-the-shelf face detector algorithm (KING, 2009) to get the localization of the face on the image. First, we crop the face and use it as input for the *face encoding stream*, responsible for extracting features from the face. Next, we fill the cropped region with a black rectangle and use this new image as input for the *context encoding stream*. Since the facial crop is occluded, this stream is "forced" to search for features from other image regions during training (i.e., the background context). Finally, we apply a segmentation technique (HE et al., 2017) to remove background noise and persons not acting directly on the scene. We use this segmentation mask as input for an off-the-shelf human keypoint extractor (XIAO; WU; WEI, 2018). The features extracted from these three streams are fused adaptively, allowing the emotion classification.



Source: Author.

LE et al., 2021; GAO et al., 2021; ZHAO; LIU; WANG, 2021; ZHAO; LIU; ZHOU, 2021). Our proposed approach extracts and processes each cue individually using encoding streams that were based on specialist models or approaches validated from the human behavior literature while also extracting information from multiple cues at the same time. Each of the proposed cues in this work has a strong motivation from the behavioral sciences and nonverbal communication literature, as will be described in the next sections.

5.3.2 Methodology

Given an image I, we aim to infer an emotion y among a set of K emotion labels by using a convolutional neural network model. The proposed network architecture extracts features of three streams: *face encoding stream*, *context encoding stream* and *body encoding stream*. The proposed method can infer emotion from multiple non-verbal cues by combining these features in an *adaptive fusion network*. In Figure 24, we present the proposed architecture, and each module is detailed below.

5.3.2.1 *Face encoding stream*

Given that in some images we have multiple faces present, we implement a *face selector algorithm* that selects the leading performer's face based on their placement on the scene. Given a set of detected faces $F = \{F_1, ..., F_n\}$, we rank each candidate F_c and select a face for input F_s based on its bounding box area, which points to if the person is on background or foreground, and the 2D position (x, y) of the bounding box centroid on the scene.

We crop the bounding-box region of F_s and use it as input to the *face encoding stream*, as shown in Figure 24. This module consists of five convolutional layers with 3x3 kernels with sizes 32, 64, 128, 256, and 256, followed by batch normalization (BN) and rectified linear unit (ReLU) activation and four max-pooling layers with a kernel size of 2. We spatially average the final feature layer using an average-pooling layer. Although the design of this encoding stream is similar to others available in the literature (LEE et al., 2019), this design is based on how convolutional blocks can generate feature maps that could be used for prediction, therefore arriving from a much-generalized point of view. Based on the nonverbal communication literature, eye movements to specific facial features are crucial in determining emotion perception (ADOLPHS et al., 2005; AVIEZER et al., 2008b). By employing simple convolutional blocks, we hypothesize that this same behavior can be replicated at some level by the proposed architecture, which will be downsampling the input data for feature extraction and, therefore, selecting information.

5.3.2.2 *Context encoding stream*

Extracting emotional information from context is problematic due to the high variability of context information. Moreover, many essential details may be hidden in the scene, motivating a robust encoding stream for context. Therefore, proposing approaches to enrich the representations extracted on the *context encoding stream* could improve the results by allowing more representative features to classify emotion. Based on the nonverbal communication literature, at the same time that context is automatically encoded by perceivers when the perceiver is required to judge the emotion, they will attempt to use whatever contextual information is available. We design this stream to try and replicate this behavior by "looking" at context and searching for contextual information that can be useful for judging emotion (BARRETT; KENSINGER, 2010). To achieve this, we propose the following design, which, again, is very sim-

ilar to other works in the literature due to its simplistic nature but adapted to be a powerful feature extraction module. This module consists of four convolutional layers with 3×3 kernels with sizes 32, 64, 128, and 256, followed by batch normalization layers (IOFFE; SZEGEDY, 2015), ReLU activations (NAIR; HINTON, 2010), and four max-pooling layers with a kernel size of 2×2. Finally, we add an adaptive self-calibrated convolution (LIU et al., 2020) with kernel size 3×3 and a ReLU activation layer.

Adaptive self-calibrated convolutions

We propose the usage of self-calibrated convolutions (LIU et al., 2020) to allow output features to be enriched. Those modules provide internal communications of the convolutional layer. It can generate more discriminative representations and improve the overall quality of the extracted features. The adaptive self-calibrated convolution module receives an input with channels size C and outputs a features map with channels size C'; a restriction when using the original self-calibrated convolutions we overcame. By altering the last convolution on this block, we allowed experimentation with encoder networks that vary the output channel size.

Attention inference module

Given the enriched outputs from the self-calibrated convolution in the *context encoding stream*, an attention inference module is learned in a non-supervised way, allowing this stream to focus on the salient regions of the context. Given the feature map $F \in \mathbb{R}^{256\times8\times14}$ that was outputted by the *context encoding stream*, the attention inference module gradually reduces the channel dimensions from 256 to 128 using convolutional layers and then further to 1 using a fully connected (FC) layer, resulting into an attention map $A \in \mathbb{R}^{1\times8\times14}$.

To apply the attention map A, we normalize it using the Softmax operation along the spatial dimensions, obtaining attention weights α. These weights signify the relative importance of each spatial location in the feature map. Finally, we enhance the original context feature map F by applying element-wise multiplication with the attention weights α, generating an enhanced feature map F' given by $F' = F \odot \alpha$, accentuating the informative regions of the background. We show visualizations of this output in Section 5.4. We chose this attention module because its effectiveness was already demonstrated in other works (LEE et al., 2019; LE et al., 2021).

Figure 25 – The proposed approach to deal with the cluttered background problem. We regress a mask around the main performer to occlude the background information (left column) using Mask R-CNN (HE et al., 2017) and force the keypoints detection model to focus on the foreground information. If more than one mask is regressed on the image, we use the face crop region to select the correct mask. The output is the image with only the main performer's body present (right column).

Source: Author.

5.3.2.3 Body encoding stream

Following our proposal to investigate body pose as a nonverbal communication input, we employ a *body encoding stream* based on an approach proposed by Xiao, Wu and Wei (2018) known as Simple Baselines. Instead of following the entire body pose pipeline, we extract the features learned up to the last convolutional layer of the network. Our intuition with this proposal is to allow the following layers to learn the correlation between features and emotions instead of leveraging the 2D annotations, based on a work by Coulson (2004) that states that perceivers use multiple diverse variables when trying to judge emotion from static body postures. Therefore, differently from other approaches that encode the 2D annotations of pose directly into their pipeline (CHEN et al., 2023), we use this model as a feature extractor and learn the correlations between this previously learned internal knowledge with emotion.

Given that in some images we can have more than one person, especially as an application prerequisite, we used Mask R-CNN (HE et al., 2017) to create segmentation masks of the people present on the scene. With the segmentation masks, we can separate a person of interest in the scene from a cluttered background and extract only their body pose. Given a set of masks $M = \{M_1, ..., M_n\}$, we calculate the overlap between the mask and the face selected for input on the *face encoding stream* F_s and rank each mask according to the number

of pixels overlapped. This approach prevents two different persons from being considered for two different encoding streams. Figure 25 presents an example of how we apply segmentation masks for cluttered backgrounds. We generated annotations of the masks for the training and evaluation procedures to eliminate the need to predict new masks for each batch. The *body encoding stream* receives the full, uncropped image, with the removed background, as we show in Figure 24 and Figure 25.

5.3.2.4 Adaptive fusion networks

The direct concatenation of multiple features often fails to provide adequate performance. In this work, we propose the fusion of features from *face encoding stream* (X_F), *context encoding stream* (X_C) and *body encoding stream* (X_B) using a fusion network with an attention model to infer the weights of each feature.

Given a set of features output from each stream $X = \{X_F, X_C, X_B\}$, the model learns a set of attention weights $\lambda = \{\lambda_F, \lambda_C, \lambda_B\}$ in which we apply a Softmax function to restrict $\lambda_F + \lambda_C + \lambda_B = 1$ and multiply the weights relative to the contribution of each stream with the features learned individually by each stream, as

$$X_A = \Pi(X_F \odot \lambda_F, X_C \odot \lambda_C, X_B \odot \lambda_B). \tag{5.4}$$

given Π as the concatenation operator. We then employ three Fully Connected (FC) layers that reduce the dimensionality of X_A to our final output y. The first FC layer reduces the dimensionality of X_A from 768 to 512, and the subsequent FC layers from 512 to 128 and from 128 to K.

5.3.2.5 Preprocessing pipeline

Before our training procedure, we perform a preprocessing step to allow some variability between epochs and enable training while keeping important features available. For the input of the *face encoding stream*, we resize the facial crops to a fixed size of 96×96. For the *context encoding stream*, we pad each image according to the shape of the larger image on the dataset, which is 400×712. We also resize the images by three to maintain the aspect ratio and use a random crop with a padding of 5 pixels on all sides to augment our training

dataset. For the images used in the *body encoding stream*, we follow the pipeline proposed by Xiao, Wu and Wei (2018) and resize the image to 256×256.

5.3.3 Experiments

Dataset

We perform our experiments on the CAER-S dataset (LEE et al., 2019) to allow direct comparison with the state-of-the-art. An overview of this dataset is available in Section 4.1.

Baseline works

We compare our proposed multi-cue learning framework against baseline works (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SIMONYAN; ZISSERMAN, 2014; HE et al., 2016b) and state-of-the-art approaches on emotion recognition (GAO et al., 2021; ZHAO; LIU; ZHOU, 2021; LEE et al., 2019; ZHAO; LIU; WANG, 2021; LE et al., 2021). Moreover, we assess the contribution of each proposed module using an ablation study on the considered cues and discuss the contribution of our approaches to deal with the limitations of the CAER-S (LEE et al., 2019) dataset, such as the face selector algorithm.

Validation

Our experiments are comprised of qualitative and quantitative evaluation. For qualitative evaluation, we use the Grad-CAM (SELVARAJU et al., 2017) technique to investigate how the *context encoding stream* searches for cues in the background of the scene. The Grad-CAM allows visualizations of which regions of the input image are more relevant for predictions by using class-specific gradient information to localize these crucial regions. We also evaluate the overall accuracy of the model qualitatively. We use unweighted classification accuracy for our experiments to quantify the model's performance for quantitative evaluation. This metric is standard for the state-of-the-art evaluation (LEE et al., 2019; LE et al., 2021; ZHAO; LIU; WANG, 2021; ZHAO; LIU; ZHOU, 2021), which allows for direct comparison of the models since no pre-trained weights are available for CAER-S on most of the compared techniques.

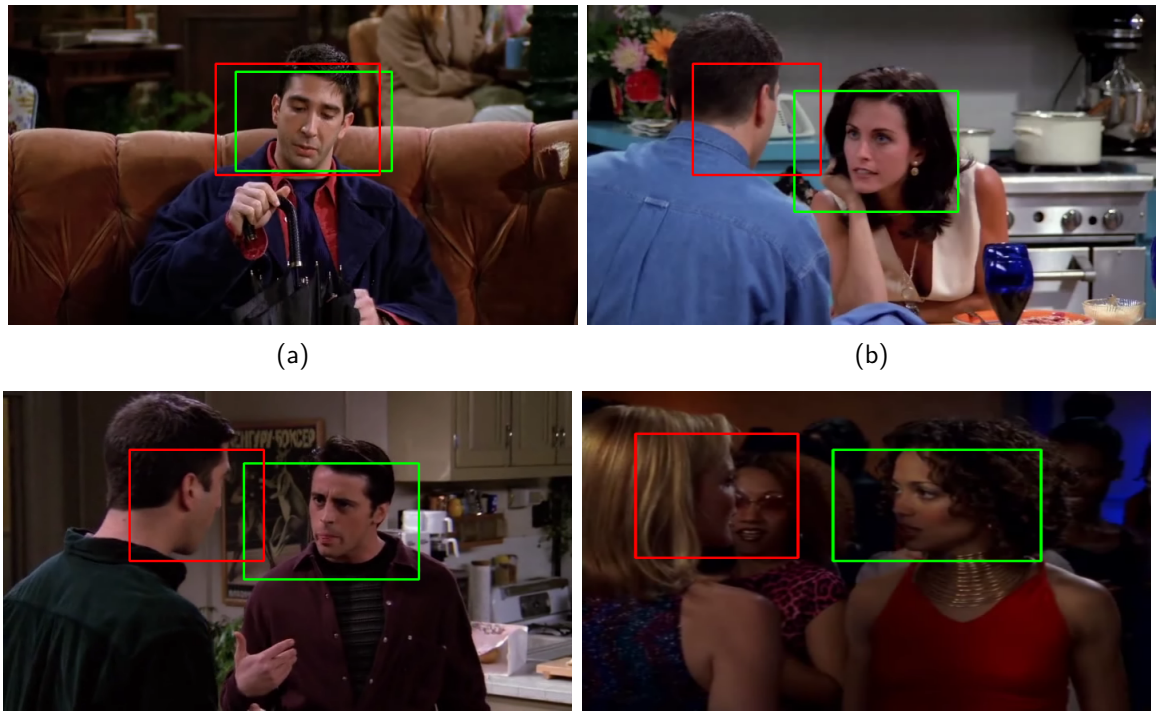
Implementation details

We implemented EmotionRAM using PyTorch (PASZKE et al., 2017) and trained the *face encoding stream* and *context encoding stream* from scratch, while for the *body encoding stream* we load the weights from the pre-trained model from Xiao, Wu and Wei (2018) up to the last convolutional layer, as we describe in Subsubsection 5.3.2.3, with an initial learning rate initialized as 3×10^{-3} and dropped by a factor of 0.5 every 60 epochs using the RMSProp (HINTON; SRIVASTAVA; SWERSKY, 2012) optimizer. Finally, we trained the model using the cross-entropy loss function on a batch size of 32.

Experimentation environment

We performed all experiments on a desktop computer running Ubuntu 20.04.3 LTS with an Intel i7-4790K with 32 GB of RAM and NVIDIA RTX 2080 Ti, with 12GB of RAM using the driver 470.63.01. For training and experimenting with our model, we use PyTorch 1.11

Figure 26 – Comparison of different approaches for face crop. The current approach from the state-of-the-art is to use the dlib toolkit (KING, 2009) and use the first face from the list of predicted faces (red bounding box). Given the same set of faces, we employ the face selector algorithm (green bounding box) to search for the face of the leading performer.



(a) (b)

(c) (d)

Source: Author.

Table 15 – Ablation study of the components proposed in our method.

Methods	Face	Context	Body	Acc. (%)
EmotionRAM (f)	✓			71.36
EmotionRAM (c)		✓		85.58
EmotionRAM (b)			✓	66.67
EmotionRAM (f+c)	✓	✓		88.29
EmotionRAM (f+b)	✓		✓	68.13
EmotionRAM (c+b)		✓	✓	81.97
EmotionRAM (f+c+b)	✓	✓	✓	**89.76**

and CUDA 10.2.89. To experiment with GLAMOR-Net (LE et al., 2021), we created a separate environment that uses Tensorflow 2.1.0 and CUDA 11.6.1 to reproduce their result using the closest available version to Tensorflow 2.0, which is the version they used in their work.

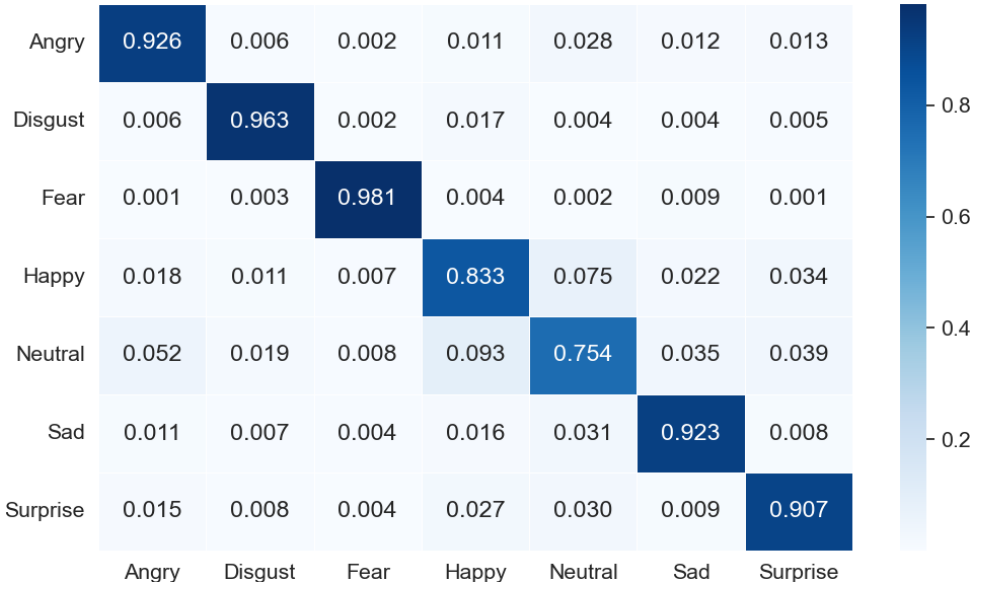
5.4 RESULTS AND DISCUSSION

Face selector

Our first experiment was to evaluate how the dataset structure could impact our accuracy score. As discussed previously, CAER-S does not have per-person annotations, leading to confusion in scenes with multiple actors. The approach currently used in the literature (LEE et al., 2019; LE et al., 2021) is to use an off-the-shelf face detector, such as dlib (KING, 2009), and use the first face detected on the scene as input for the *face encoding stream*. However, we need an algorithm to identify the leading performer in the scene since the scene and context are building up to their action (otherwise, the network would be leaning towards group-level emotion recognition tasks, such as Gupta et al. (2018), Dhall et al. (2018), Dhall et al. (2017), Dhall et al. (2016)).

As we assess in Figure 26, the red bounding box would be selected by other state-of-the-art approaches, while the green bounding box is the face selected by our approach. In cases with only one person on the scene Figure 26a, both approaches work the same; however, in cases with multiple persons, such as Figure 26b and Figure 26c this approach tends to fail and wrongfully select a face from a performer that is not leading the scene, prejudicing the *face encoding stream*. We also show in Figure 26d that this approach also helps to select

Figure 27 – The confusion matrix of the outputs from EmotionRAM (f+c+b) on CAER-S dataset.



Source: Author.

the correct face in crowded scenes. We noticed a slight increase in accuracy when using this approach, which points out the robustness of the *context encoding stream*, indicating that it can leverage emotion from context even with an incorrect face selected.

Ablation study

We evaluated EmotionRAM with ablation studies to investigate the contribution of each component. We varied the combination of components used in each evaluation and logged the accuracy of that model. We expose the results of this study in Table 15. The results show that combining face, context, and body can achieve the best accuracy. Additionally, we can notice that our model powered by face and context only yields a competitive result with other face and context models, such as Le et al. (2021). Besides our own fusion approach employed in this pipeline, we also evaluate two other strategies, which are the direct concatenation of the features and a variation of our approach that sums the contribution of each cue instead of weighing them. These results are available in Table 17.

Table 17 – Comparison of fusion methods for three nonverbal cues (face, context, and body).

Methods	**Acc. (%)**
Adaptive fusion (ours)	**89.76**
Adaptive sum	67.62
Concatenation	69.15

Ambiguity between Neutral and Happy states

In Figure 27, we display a confusion matrix of the outputs from our proposed method. This result provides insights into which classes are the most difficult for the classifier to handle. For example, we can notice from the results that *Happy* and *Neutral* are the most ambiguous classes, leading to difficulties for the classifier to differentiate between them.

However, as tackled by the psychology literature, this behavior can be explained by the structural resemblance of emotionally neutral faces, which tends to resemble happiness by humans (SAID; SEBE; TODOROV, 2009; MONTEPARE; DOBISH, 2003; KNUTSON, 1996), and could also be impacting our results. We show examples of misclassifications between *Happy* and *Neutral* in Figure 28.

Self-calibrated convolutions

We experiment using self-calibrated convolutions (LIU et al., 2020) for an improved feature extraction step. In our experimentation, we found that placing the self-calibrated convolutions on the *face encoding stream* is prejudicial to the model's overall accuracy. The primary reason for this issue is that on some samples of the dataset, the size of the face crops from the leading performer would be too small due to their placement on the scene. Empirically, we chose to swap only the last convolutional layer of the *context encoding stream* to boost the features from context, given that extracting features that could be correlated with emotions is challenging. We also evaluated that swapping only the last convolutional layer yields a higher accuracy score than swapping all dense layers with this self-calibrated layer.

Body encoding stream

Finally, we experiment with using body language as an input to the model. The features learned up to the last convolutional layer are combined on the adaptive fusion module, which learns weights for three inputs. A first experiment focused solely on using the body pose estimation technique as an extra input cue to the model. This raw implementation yielded a low accuracy score compared to the previously implemented improvements.

However, during further investigation, we noticed that multiple actors are present on the scene in many cases, and the body posture of these actors was also considered when leveraging

Figure 28 – Examples of cases in which our model misclassified between *happy* and *neutral* samples.

(a) Incorrectly classified as *Happy* (*Neutral*)

(b) Incorrectly classified as *Neutral* (*Happy*)

Source: Author.

the body pose. Therefore, as previously explained in Subsubsection 5.3.2.3, we use Mask R-CNN (HE et al., 2017) to segment the principal performer's body and isolate it from context, leading to our best result of 89.76%. Although the accuracy increment may be small compared with the approach using only context and face, as we show in Table 15, the pose encoding stream may be decisive in complex cases where the context does not contain useful information. However, further investigation is needed, especially hyperparameter-wise, to understand if any other constraints contribute negatively to this result.

Comparison against state of the art

We compare our results with different approaches in Table 18. The proposed method was 16.25% better when comparing our implementation of the baseline approach - CAER-Net-S (LEE et al., 2019) that obtained an accuracy of 73.51%. We also performed consistently better against traditional deep neural networks approaches for image tasks, such as AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) (47.36%), VGG-Net (SIMONYAN; ZISSERMAN, 2014) (49.89%), and ResNet (HE et al., 2016b) (57.33%), and also to more robust, recent and optimized models that are used for similar tasks, such as DenseNet-121 (HUANG et al., 2017) (81.66%), EfficientNet-B0 (TAN; LE, 2019) (70.52%), Inception-V3 (SZEGEDY et al., 2016) (75.86%) and MobileNet-V3 (HOWARD et al., 2019) (61.57%).

We also compare to GLAMOR-Net (LE et al., 2021) by using their available code[7] and changing the configurations of the data loading procedure to point to the CAER-S dataset on

[7] Available at <https://github.com/minhnhatvt/glamor-net/tree/main>

Table 18 – Quantitative evaluation of EmotionRAM in comparison with baseline methods on the CAER-S dataset.

Methods	**Acc. (%)**
ImageNet-AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012)	47.36
ImageNet-VGG-Net (SIMONYAN; ZISSERMAN, 2014)	49.89
ImageNet-ResNet (HE et al., 2016a)	57.33
DenseNet-121 (HUANG et al., 2017)	81.66
EfficientNet-B0 (TAN; LE, 2019)	70.52
Inception-V3 (SZEGEDY et al., 2016)	75.86
MobileNet-V3 (HOWARD et al., 2019)	61.57
Fine-tuned AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012)	61.73
Fine-tuned VGGNet (SIMONYAN; ZISSERMAN, 2014)	64.85
Fine-tuned ResNet (HE et al., 2016a)	68.46
Fine-tuned DenseNet-121 (HUANG et al., 2017)	74.03
Fine-tuned EfficientNet-B0 (TAN; LE, 2019)	79.76
Fine-tuned Inception-V3 (SZEGEDY et al., 2016)	78.16
Fine-tuned MobileNet-V3 (HOWARD et al., 2019)	62.40
CAER-Net-S (LEE et al., 2019)	73.51
CAER-Net-S (LEE et al., 2019) (our reproduction)	81.50
GRERN (GAO et al., 2021)	81.31
EfficientFace (ZHAO; LIU; ZHOU, 2021)	81.48
MA-Net (ZHAO; LIU; WANG, 2021)	88.42
GLAMOR-Net (LE et al., 2021)	77.90
GLAMOR-Net (LE et al., 2021) (our reproduction)	76.33
GLAMOR-Net (ResNet-18) (LE et al., 2021)	**89.88**
GLAMOR-Net (ResNet-18) (LE et al., 2021) (our reproduction)	83.08
MCF-Net (XU et al., 2022)	75.68
MCF-Net (ResNet-18) (XU et al., 2022)	81.82
CAHFW-Net (ZHOU et al., 2023)	83.75
SMResNet (LIU et al., 2023)	88.52
EmotionRAM (face+context)	88.10
EmotionRAM (face+context+body)	**89.76**

disk, overwriting their default configurations of loading NCAER-S. After training the model using the same configurations available on their paper (and also available on their code), we were not able to achieve their reported performance of 89.88% and obtained an accuracy of 83.08% on the CAER-S dataset. With these new results in mind, our method performs slightly worse than their reported result by a difference of 0.12%, but 6.68% better than the result we could achieve using their available code.

Even though the difference in performance might not seem significant[8], the fact that we

[8] We were unable to compare statistically our results with those published by the authors, since they did not

Table 19 – Inference time of our EmotionRAM framework against GLAMOR-Net.

Method	**Min**.	**Avg**.	**Acc**. (%)
EmotionRAM (f+c+b)	6.1490ms (≈162 fps)	7.0110ms (≈142 fps)	89.76
GLAMOR-Net (LE et al., 2021)	20.6775ms (≈48 fps)	28.7653ms (≈43 fps)	77.90
GLAMOR-Net (LE et al., 2021)*	59.9367ms (≈17 fps)	71.9979ms (≈15 fps)	89.88

* Using ResNet-18 backbone.

rely on an additional cue for describing the scene might allow for a better understanding of the situation. We should also consider that adding a third encoding stream may be helpful in real situations where the face might be occluded or the context is not significant.

Focusing on the application scenario, we compare the computational cost of the top-performing models by measuring the minimum and average inference time, as we show in Table 19.

Since Le et al. (2021) did not report the inference time of their models, we use their available code to assess it on the CAER-S test set. We also report the inference time of their baseline approach because it is considered to be faster than their implementation using the ResNet-18 as a backbone. We want to point out that this model yields a lower accuracy score.

This experiment shows that our model can infer faster than the architectures proposed in Le et al. (2021), while keeping a competitive accuracy score. Furthermore, our model is more suitable for deployment on consumer-grade computers or energy-friendly edge devices, given its lower computational cost.

Qualitative evaluation

After the quantitative evaluation, we proceeded with a visual, qualitative evaluation. In Figure 29, we show examples of correct and incorrect classifications of our model on the CAER-S test set.

Following the qualitative investigation, we also applied a visual investigation based on the Grad-CAM technique (SELVARAJU et al., 2017). In Figure 30, we show a few examples of how the context encoding stream acts towards the correct prediction of the emotion. The qualitative evaluation highlighted some key aspects of our method, which we will now discuss.

publish pre-trained weights for CAER-S dataset.

In Figure 30a, we can see that the *context encoding stream* learned to take into consideration interactions with other people by focusing on the second performer on the scene and started leveraging their emotions, since emotions in a group tend to point towards the same direction, something that is deeply motivated by the field of behavioral psychology. The same idea is reinforced in Figure 30b, in which the *context encoding stream* focuses on the way that a person is holding an infant - by placing them close to their body in a protective manner as if they were in a dangerous situation.

Figure 30c and Figure 30d may be directly compared since this is a classic question regarding context. In Figure 30c, we have a prediction for *Sad*, and the *context encoding stream* focuses on the background of the scene, such as the presence of candles and how the illumination leans towards a darker theme, while in Figure 30d we have a more uplifting scene, with a more relaxed pose on a well-lit stage. In Figure 30e, we have another group interaction in which the *context encoding stream* also considered how the other person was feeling, pointing out a simple conversation with neutral feelings. By investigating the scene, we may notice that they are placed in an uninteresting situation of filing a document, and the network can correctly

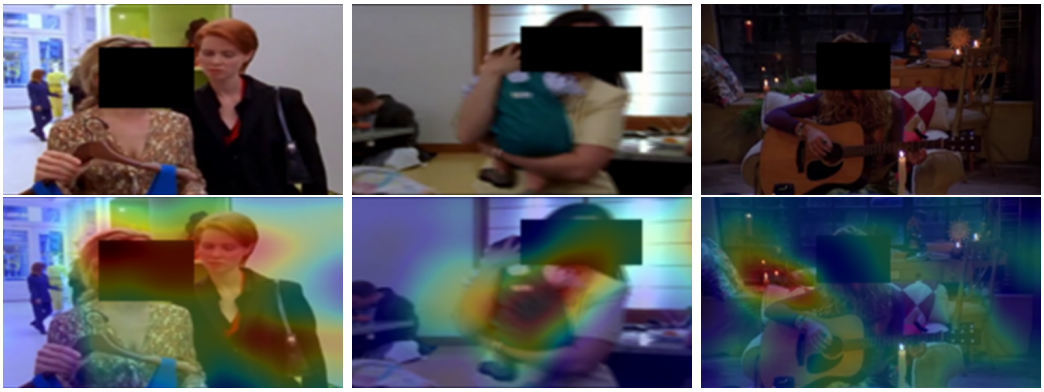
Figure 29 – Qualitative results of our method (EmotionRAM f+c+b) on the CAER-S test set. We present pairs of predictions, in which the top image is a correct prediction, and the bottom image is an incorrect prediction, followed by the correct class between parenthesis.

Source: Author.

predict the emotion for this interaction.

In Figure 30f, we see a case in which person-object interaction is taken into consideration, in which the *context encoding stream* focuses on the object on the performer's hand and on the way that they are holding it. Finally, in Figure 30g, we see a scene in which we do not have an informative context, and the *context encoding stream* focuses on the body language of the user to predict their emotion.

Figure 30 – Visualization of the attention module from the context encoding stream. Since this module is responsible for extracting context information, the main person's facial region on the scene is occluded with a black rectangle. On the top row, we see the image used as input for the module, and on the bottom row, the output from Grad-CAM (SELVARAJU et al., 2017) with respect to the last convolutional layer of the attention inference module.



(a) Disgust (b) Fear (c) Sad

(d) Happy (e) Neutral (f) Surprise

(g) Angry

Source: Author.

6 DISCUSSION

This chapter will overview the results and research goals defined in this thesis, pointing out our specific contributions in this work.

6.1 CONCEPTUAL DEFINITIONS FOR EMOTION RECOGNITION

Two of our research goals - RG1 and RG2 - strongly depended on a conceptual analysis of emotion recognition supported by findings in the behavioral psychology literature. In Chapter 2, we reviewed how humans perceive and communicate nonverbal cues related to emotion in a natural manner.

We gathered references to how humans process each of these cues based on the three nonverbal cues we selected to follow in this research, namely (1) facial expressions, (2) situational context, and (3) body language. For (1) facial expressions, the literature suggests that humans perform eye shifts to specific facial features (ADOLPHS et al., 2005; AVIEZER et al., 2008b); other references indicate that these eye shifts are to FAUs (MARTINEZ, 2017). We have modeled this behavior by placing a face encoding stream based mostly on convolutional blocks since the model will learn these important areas by reducing the spatial parameters of this representation. In the case of (2) situational context, these influences are perceived automatically, and people will routinely encode the context when asked to make a more specific inference about someone else's emotions (BARRETT; LINDQUIST; GENDRON, 2007; BARRETT; KENSINGER, 2010; BARRETT; MESQUITA; GENDRON, 2011; AVIEZER et al., 2008b). We have replicated this behavior by "looking" at the scene and searching for contextual cues, boosting the representations learned by using self-calibrated convolutions and attention blocks for our multi-cue approach as described in Section 5.3, and by generating high-level descriptions of the scene that are augmented with sentic representations for our high-level approach as described in Section 5.1. Finally, for (3) body language, there are different views based on static or dynamic domains: in the first case, Coulson (2004) describes that when humans are asked to judge emotions from static body postures, anger, happiness, and sadness were mostly accurate by looking at some features such as head bend and chest bend, but that overall this judge would use multiple variables that would be encoded in an unconscious manner. We have replicated this behavior in our multi-cue approach by using a pose extraction pipeline, but

instead of generating the 2D positions of the joints, we extracted the features up to the last convolutional layer and allowed the next layers to learn representations. In the second case, it is clear that dynamic body language is much more representative. Roether et al. (2009) evaluated that many features, such as movement speed and limb flexion are encoded when perceiving emotions in gait, and we have modeled this by using a spatio-temporal approach that has a body topology with weights associated with each node and edge.

Given the emotion theories proposed and validated by researchers in that field (as discussed in Chapter 3), we can now model the emotion recognition task. In summary, this task could be modeled as classification or regression, depending on which theory to follow. With that, we now had the information necessary for proposing approaches that are inspired by how humans perceive emotions: first, we could now design deep learning models that were inspired by the process of how humans perceive emotions, adding cognitive aspects to a mostly perceptive task; second, we now could model the deep learning task of emotion recognition based on how humans classify emotion based on such perception.

Finally, the models and frameworks we propose in this thesis can be extended and adapted to work on multiple domains. For example, our multi-cue approach is powered by an adaptive fusion module (as we show in Subsubsection 5.3.2.4) that could receive different descriptions, such as captions or semantic descriptions for context.

6.2 PERFORMANCE OF IMPLEMENTED MODELS

Our third research goal - RG3 - was related to developing frameworks for emotion recognition, adding two requirements. First, they need to be optimized so that deployment would be possible in developing countries, tackling the many limitations that are imposed in these markets related to the difficulty of accessing high-performing hardware and expenses related to deployment and scalability. Moving against the tide, our approaches were designed to allow high inferencing rates even on low-power, low-consumption devices.

Second, our frameworks need to be compatible with the state of the art concerning the currently employed metrics. As the reader might have noticed, from the three quantitative analyses we have performed from our models (high-level context in Subsection 5.1.4, gait analysis in Subsection 5.2.4, and multi-cue emotion recognition in Section 5.4), only our result from gait analysis surpasses the current state-of-the-art. This is not a limitation of this work, as it was never our goal with this research; therefore, we invite the reader to have a different

Table 20 – Overview of the results in this thesis, comparing quantitative results and inferencing time. Negative results indicate that the method performs worse than the state-of-the-art. Inference times are measured in a single RTX 2080 Ti.

Subtask	Result	Performance	Highlight
Multi-cue	2-nd place (-0.12% difference)	$\approx$ 142 fps	Performs $\approx$ 9x faster than 1-st place
Context	3-rd place (-7.71 mAP difference)	$\approx$ 240 fps	Surpasses many techniques using five cues while using only one cue (context)
Gait	1-st place (4.2% difference)	Similar	Requires 72% less epochs for converging

view from these results related to the trade-off between accuracy and inferencing time.

For example, our model for high-level context analysis achieves an inferencing rate of $\approx$ 96 fps – in other words, it can predict the emotion perception from the context of $\approx$ 96 images per second, all in a consumer-grade notebook. This model could easily be applied to large-scale emotion recognition while keeping a lower cost. The same applies to our multi-cue model, which performs 0.12% worse in accuracy when compared to Le et al. (2021) but can perform $\approx$ 9x faster. While our model for gait analysis has a comparable framerate with other models, it also has a significant improvement related to computational resources requirements: training this model takes 72% fewer epochs than the previous state of the art, proposed by Bhattacharya et al. (2020). We overview these results in Table 20.

6.2.1 Limitations in Application Scenarios

In addition to the qualitative and quantitative evaluations in the proposed datasets, we have deployed these models in real-world scenarios, both in controlled lab environments and uncontrolled settings. This subsection discusses some of the conclusions drawn from these deployments.

Firstly, our high-level context model (detailed in Section 5.1) constructs a knowledge graph that facilitates deployment in various unseen scenarios. This approach aligns with the principle of Zero-shot Learning (ZSL) by using auxiliary information to relate seen and unseen classes. However, as the name suggests, it is highly context-dependent, and in some cases, individuals

may experience different emotions than those predicted by the model.

Similarly, our multi-cue emotion recognition method (discussed in Section 5.3) learns from three cues—face, context, and body. Nevertheless, blank background contexts, such as white walls, can lead to incorrect predictions. Another limitation of deploying this model into production is its heavy reliance on face detection algorithms. Inaccurate face detection can significantly affect the model's performance.

Finally, our gait analysis method (explained in Section 5.2) requires a recording of the person walking towards the camera. Although this issue could potentially be addressed through normalization techniques, it remains an area for future work. In summary, while these models have clear limitations, similar challenges are also present in other state-of-the-art models and traditional methods for emotion recognition.

6.2.2 Training and evaluating on EiLA dataset

Finally, another result of this thesis is the EiLA benchmark. Although the number of images is lower than other benchmarks such as EMOTIC or CAER-S, we propose to use this data **(a)** as a fine-tuning dataset to mitigate bias and **(b)** validate how models perform on people with different cultures. As we show in Subsection 4.2.1, EiLA has a higher concordance than EMOTIC, which highlights the quality of the annotation process of this data. EiLA is yet to be made publicly available and used to train emotion recognition models, which has not been done before in this work because it is a consequence of our efforts *after* our endeavors using existing datasets and analyzing them.

6.2.3 Biases

The proposed models may be biased towards the data that we have used for training, i.e. data collected from specific cultural backgrounds. Therefore, we expect them to not work well on people from different cultures, such as Latin Americans, due to the limitations imposed by cultural representations, but also context (please see Subsection 4.1.6 for more details). We expect that EiLA will support training and evaluating state-of-the-art models, reducing bias and allowing for applications to be deployed in Latin America with a fairness aspect associated with them.

A possible exception to these bias-associated factors is our model for high-level context

representation, since it is able to extract global representations (meaning, it is not dependent on the training set to learn them). A comprehensive study on bias, fairness, and accountability is planned as a future work for this thesis.

Another possibility related to biases in our research lies in using the E-Gait dataset for gait analysis. Many samples were imported from the Edinburgh Locomotion Mocap Dataset (ELMD) (KLEINSMITH; BIANCHI-BERTHOUZE, 2012), which features a single male performer recording their gait. It has been shown that gait depends on multiple aspects, such as gender, body type, or height. For this case, this evaluation shows proof that our proposed model can extract these spatiotemporal features; however, recording more data with different subjects is essential to allow this research to be deployed in real-world scenarios.

7 CONCLUSION

In this doctoral thesis, we have evaluated novel approaches to enabling intelligent systems to perceive emotions through the processing and encoding of nonverbal cues present in images and videos. Through a comprehensive review of the relevant behavioral psychology literature, we discuss how human emotion perception occurs naturally through biological mechanisms, and argue that drawing inspiration from these processes can help guide the design of deep learning models for this task.

We proposed three framework designs, each grounded in evaluations and findings from behavioral psychology research. The first, EmotionRAM, uses facial expressions, situational context, and body language to classify emotions in static image domains. Its design is based on three pillars mirroring how humans leverage gaze movements to read facial features, automatically encode context, and interpret body postures for emotion perception. The second, ST-Gait++, applies to dynamic domains utilizing body pose and gait cues, inspired by research highlighting gait as a meaningful behavioral marker of emotion. Finally, we propose a high-level context-focused design using only environmental features.

We also evaluate the role of datasets in model success and limitations, highlighting a significant lack of cultural representation in current benchmarks. To address this, we introduce a new dataset, EiLA, focused on capturing the cultural aspects of emotion through Latin American images and videos across different contexts.

In summary, our quantitative evaluations demonstrate that our proposed psychologically-inspired models achieve state-of-the-art or better performance on benchmark tasks while being optimized for low-resource deployment, evidencing that lighter deep learning models can indeed perform high-quality emotion recognition when drawing from the principles of human perception.

7.1 FUTURE WORKS

With the recent developments in emotion recognition research, we believe that vision-language models may allow a deeper evaluation of nonverbal cues, especially related to fairness, accountability, and explainability. As future works, we are building upon extracting and describing cues using Vision-Language Models (VLMs), especially for describing body language

and behavior. With this in mind, we also plan on expanding the research related to how culture can impact emotion perception in humans, and propose models that leverage these types of cues to increase recognition levels among diverse cultural groups.

Another interest lies in shifting from generalistic to applied emotion recognition systems, which can solve task-related problems, such as driver behavior recognition and mental health. We also plan on improving EiLA benchmark, adding more samples with a deeper investigation related to biases, as well as deploying our frameworks to solve real problems in society and evaluate their technology readiness levels.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, we have used **StableDiffusion-XL** (PODELL et al., 2023) to generate images from Section 3.2. Given how the license of this model attributes the authorship to us, we are not required to ask for permission to use these images in this manuscript.

RESEARCH ACTIVITIES

PUBLICATIONS

- Costa, W., Macêdo, D., Zanchettin, C., Talavera, E., Figueiredo, L. S., & Teichrieb, V. (2024). **A fast multiple cue fusing approach for human emotion recognition**. Available at SSRN 4255748. (Submitted to Image and Vision Computing, a Q1 journal with i.f. of 4.7 - under review)
- Costa, W., Talavera, E., Oliveira, R., Figueiredo, L., Teixeira, J. M., Lima, J. P., & Teichrieb, V. (2023). **A Survey on Datasets for Emotion Recognition from Vision: Limitations and In-the-Wild Applicability**. Applied Sciences, 13(9), 5697. (Published in a Q2 journal with i.f. of 2.7)
- de Lima Costa, W., Talavera, E., Figueiredo, L. S., & Teichrieb, V. (2023). **High-Level Context Representation for Emotion Recognition in Images**. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 326-334). (CVPR Workshops)

RESEARCH & DEVELOPMENT PROJECTS

Six years participating in R&D projects at Voxar Labs, four of them leading projects with national and global partners in the industry. Writing grant requests and submitting public notices to secure funding for research projects.

- **Shorts** (partnership with HP) - Project leader (2024 - present)
- **Activity Monitoring** (partnership with HP) - Project leader (2024 - present)
- **Perception** (partnership with Volkswagen and EyeFlow) - Project leader (2023 - present)
- **Patient Sitting** (partnership with HP) - Project leader (2023 - 2023)
- **Multi-camera Telehealth** (partnership with HP) - Project leader (2022 - 2022)
- **UX²R: User eXperience on eXtended Realities** (partnership with Samsung) - Research scientist (2019 - 2021)

- **Image and video description** (partnership with Samsung) - Research scientist (2018 - 2018)

EDUCATION

One and a half years of teaching experience as a volunteer teacher assistant, developing and teaching practical and theoretical classes for the Digital Image Processing course (for B.Sc. and M.Sc. students) at the Universidade Federal de Pernambuco. I co-advised seven final projects for B.Sc and M.Sc. students among students from Brazil (with Prof. Veronica Teichrieb and Prof. Lucas Figueiredo) and in The Netherlands (with Prof. Estefania Talavera).

OTHER RESEARCH ACTIVITIES

- A Best Doctoral Consortium award in the LatinX in AI workshop of the ICCV 2021;
- Reviewing for journals such as the IEEE Access and the Journal on Interactive Systems;
- Reviewing for conferences such as ISMAR, SIBGRAPI, SVR, and the LatinX in AI Workshop at NeurIPS;
- Active member of AI communities such as the LatinX in AI and Google Research@São Paulo;
- Mentor for the Google CSR LatAm program.

REFERENCES

ADEPU, Y.; BOGA, V. R.; SAIRAM, U. Interviewee performance analyzer using facial emotion recognition and speech fluency recognition. In: IEEE. *2020 IEEE International Conference for Innovation in Technology (INOCON)*. [S.l.], 2020. p. 1–5.

ADOLPHS, R.; GOSSELIN, F.; BUCHANAN, T. W.; TRANEL, D.; SCHYNS, P.; DAMASIO, A. R. A mechanism for impaired fear recognition after amygdala damage. *Nature*, Nature Publishing Group UK London, v. 433, n. 7021, p. 68–72, 2005.

AL-DUJAILI, M. J.; EBRAHIMI-MOGHADAM, A. Speech emotion recognition: a comprehensive survey. *Wireless Personal Communications*, Springer, v. 129, n. 4, p. 2525–2561, 2023.

ARDLE, R. M.; DIN, S. D.; GALNA, B.; THOMAS, A.; ROCHESTER, L. Differentiating dementia disease subtypes with gait analysis: feasibility of wearable sensors? *Gait & posture*, Elsevier, v. 76, p. 372–376, 2020.

AVIEZER, H.; HASSIN, R. R.; RYAN, J.; GRADY, C.; SUSSKIND, J.; ANDERSON, A.; MOSCOVITCH, M.; BENTIN, S. Angry, disgusted, or afraid? studies on the malleability of emotion perception. *Psychological science*, SAGE Publications Sage CA: Los Angeles, CA, v. 19, n. 7, p. 724–732, 2008.

AVIEZER, H.; HASSIN, R. R.; RYAN, J.; GRADY, C.; SUSSKIND, J.; ANDERSON, A.; MOSCOVITCH, M.; BENTIN, S. Angry, disgusted, or afraid? studies on the malleability of emotion perception. *Psychological science*, SAGE Publications Sage CA: Los Angeles, CA, v. 19, n. 7, p. 724–732, 2008.

BARRETT, L. F.; KENSINGER, E. A. Context is routinely encoded during emotion perception. *Psychological science*, Sage Publications Sage CA: Los Angeles, CA, v. 21, n. 4, p. 595–599, 2010.

BARRETT, L. F.; LINDQUIST, K. A.; GENDRON, M. Language as context for the perception of emotion. *Trends in cognitive sciences*, Elsevier, v. 11, n. 8, p. 327–332, 2007.

BARRETT, L. F.; MESQUITA, B.; GENDRON, M. Context in emotion perception. *Current Directions in Psychological Science*, Sage Publications Sage CA: Los Angeles, CA, v. 20, n. 5, p. 286–290, 2011.

BAUTMANS, I.; JANSEN, B.; KEYMOLEN, B. V.; METS, T. Reliability and clinical correlates of 3d-accelerometry based gait analysis outcomes according to age and fall-risk. *Gait & posture*, Elsevier, v. 33, n. 3, p. 366–372, 2011.

BEARDSWORTH, T.; BUCKNER, T. The ability to recognize oneself from a video recording of one's movements without seeing one's body. *Bulletin of the Psychonomic Society*, Springer, v. 18, n. 1, p. 19–22, 1981.

BENEZETH, Y.; LI, P.; MACWAN, R.; NAKAMURA, K.; GOMEZ, R.; YANG, F. Remote heart rate variability for emotional state monitoring. In: IEEE. *2018 IEEE EMBS international conference on biomedical & health informatics (BHI)*. [S.l.], 2018. p. 153–156.

BHATTACHARYA, U.; MITTAL, T.; CHANDRA, R.; RANDHAVANE, T.; BERA, A.; MANOCHA, D. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. *AAAI Conference on Artificial Intelligence*, p. 1342–1350, 2020.

BIASE, L. D.; SANTO, A. D.; CAMINITI, M. L.; LISO, A. D.; SHAH, S. A.; RICCI, L.; LAZZARO, V. D. Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, MDPI, v. 20, n. 12, p. 3529, 2020.

BIRDWHISTELL, R. L. *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. [S.l.]: Department of State, Foreign Service Institute, 1952.

BOYD, K. L.; ANDALIBI, N. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work. *Proceedings of the ACM on Human-Computer Interaction*, ACM New York, NY, USA, v. 7, n. CSCW1, p. 1–37, 2023.

BRAUN, B.; MCDUFF, D.; BALTRUSAITIS, T.; HOLZ, C. Video-based sympathetic arousal assessment via peripheral blood flow estimation. *Biomedical Optics Express*, Optica Publishing Group, v. 14, n. 12, p. 6607–6628, 2023.

BUCK, R. Motivation, emotion and cognition: A developmental-interactionist view. *International review of studies on emotion*, Wiley Chichester, v. 1, p. 101–142, 1991.

BUSCH, A. B.; SUGARMAN, D. E.; HORVITZ, L. E.; GREENFIELD, S. F. Telemedicine for treating mental health and substance use disorders: reflections since the pandemic. *Neuropsychopharmacology*, Nature Publishing Group, v. 46, n. 6, p. 1068–1070, 2021.

CALVO-BARAJAS, N.; PERUGIA, G.; CASTELLANO, G. The effects of robot's facial expressions on children's first impressions of trustworthiness. *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, p. 165–171, 2020.

CÁMBARA, G.; LUQUE, J.; FARRÚS, M. Convolutional speech recognition with pitch and voice quality features. *arXiv preprint arXiv:2009.01309*, 2020.

CAMBRIA, E.; HUSSAIN, A. *Sentic Computing: a common-sense-based framework for concept-level sentiment analysis*. [S.l.]: Springer Cham, 2015.

CANAL, F. Z.; MÜLLER, T. R.; MATIAS, J. C.; SCOTTON, G. G.; JUNIOR, A. R. de S.; POZZEBON, E.; SOBIERANSKI, A. C. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, Elsevier, v. 582, p. 593–617, 2022.

CHANDRAN, P. S.; BINU, A. Facial emotion recognition system for unusual behaviour identification and alert generation. In: SPRINGER. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020*. [S.l.], 2021. p. 789–803.

CHEN, F.; SHAO, J.; ZHU, S.; SHEN, H. T. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 10761–10770.

CHEN, J.; YANG, T.; HUANG, Z.; WANG, K.; LIU, M.; LYU, C. Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition. *Applied Intelligence*, Springer, p. 1–17, 2022.

CHEN, J.; YANG, T.; HUANG, Z.; WANG, K.; LIU, M.; LYU, C. Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition. *Applied Intelligence*, Springer, v. 53, n. 4, p. 4201–4217, 2023.

CHEN, W.; MCDUFF, D. Deepphys: Video-based physiological measurement using convolutional attention networks. In: *Proceedings of the european conference on computer vision (ECCV)*. [S.l.: s.n.], 2018. p. 349–365.

CHEN, X.; WANG, S.; FU, B.; LONG, M.; WANG, J. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *Advances in Neural Information Processing Systems*, v. 32, 2019.

CHEN, Y.; LI, J.; SHAN, S.; WANG, M.; HONG, R. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *arXiv preprint arXiv:2312.05447*, 2023.

CHEN, Y.-C.; LIN, Y.-W. Indoor rfid gait monitoring system for fall detection. In: IEEE. *2010 2nd International Symposium on Aware Computing*. [S.l.], 2010. p. 207–212.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1251–1258, 2017.

CONSOLI, S.; BARBAGLIA, L.; MANZAN, S. Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems*, Elsevier, v. 247, p. 108781, 2022.

COSTA, W.; MACÊDO, D.; ZANCHETTIN, C.; TALAVERA, E.; FIGUEIREDO, L. S.; TEICHRIEB, V. A fast multiple cue fusing approach for human emotion recognition. *SSRN preprint 4255748*, 2022.

COSTA, W.; TALAVERA, E.; OLIVEIRA, R.; FIGUEIREDO, L.; TEIXEIRA, J. M.; LIMA, J. P.; TEICHRIEB, V. A survey on datasets for emotion recognition from vision: Limitations and in-the-wild applicability. *Applied Sciences*, MDPI, v. 13, n. 9, p. 5697, 2023.

COSTA, W. de L.; TALAVERA, E.; FIGUEIREDO, L. S.; TEICHRIEB, V. High-level context representation for emotion recognition in images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 326–334.

COULSON, M. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, Springer, v. 28, p. 117–139, 2004.

CRENN, A.; KHAN, R. A.; MEYER, A.; BOUAKAZ, S. Body expression recognition from animated 3d skeleton. In: IEEE. *2016 International Conference on 3D Imaging (IC3D)*. [S.l.], 2016. p. 1–7.

CURTIS, D. A. Deception detection and emotion recognition: Investigating face software. *Psychotherapy Research*, Taylor & Francis, v. 31, n. 6, p. 802–816, 2021.

CUTTING, J. E.; KOZLOWSKI, L. T. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, Springer, v. 9, p. 353–356, 1977.

DANG, N. C.; MORENO-GARCIA, M. N.; PRIETA, F. De la. Sentiment analysis based on deep learning: A comparative study. *Electronics*, MDPI, v. 9, n. 3, p. 483, 2020.

DAOUDI, M.; BERRETTI, S.; PALA, P.; DELEVOYE, Y.; BIMBO, A. D. Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices. In: SPRINGER. *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part I 19*. [S.l.], 2017. p. 550–560.

DARWIN, C. The expression of the emotions in man and animals. *London, UK: John Marry*, 1872.

DENTAMARO, V.; IMPEDOVO, D.; PIRLO, G. Gait analysis for early neurodegenerative diseases classification through the kinematic theory of rapid human movements. *IEEE Access*, IEEE, v. 8, p. 193966–193980, 2020.

DHALL, A.; GOECKE, R.; GHOSH, S.; JOSHI, J.; HOEY, J.; GEDEON, T. From individual to group-level emotion recognition: Emotiw 5.0. *ACM International Conference on Multimodal Interaction*, p. 524–528, 2017.

DHALL, A.; GOECKE, R.; JOSHI, J.; WAGNER, M.; GEDEON, T. Emotion recognition in the wild challenge 2013. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. [S.l.: s.n.], 2013. p. 509–516.

DHALL, A.; GOECKE, R.; JOSHI, J.; HOEY, J.; GEDEON, T. Emotiw 2016: Video and group-level emotion recognition challenges. *ACM International Conference on Multimodal Interaction*, p. 427–432, 2016.

DHALL, A.; GOECKE, R.; LUCEY, S.; GEDEON, T. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, v. 2, p. 1, 2011.

DHALL, A.; GOECKE, R.; LUCEY, S.; GEDEON, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, IEEE Computer Society, v. 19, n. 03, p. 34–41, 2012.

DHALL, A.; KAUR, A.; GOECKE, R.; GEDEON, T. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. *ACM International Conference on Multimodal Interaction*, p. 653–656, 2018.

DHALL, A.; SINGH, M.; GOECKE, R.; GEDEON, T.; ZENG, D.; WANG, Y.; IKEDA, K. Emotiw 2023: Emotion recognition in the wild challenge. In: *Proceedings of the 25th International Conference on Multimodal Interaction*. [S.l.: s.n.], 2023. p. 746–749.

DHARSINI, S. V.; BALAJI, B.; HARI, K. K. et al. Music recommendation system based on facial emotion recognition. *Journal of Computational and Theoretical Nanoscience*, American Scientific Publishers, v. 17, n. 4, p. 1662–1665, 2020.

DING, D.; GEBEL, K.; PHONGSAVAN, P.; BAUMAN, A. E.; MEROM, D. Driving: a road to unhealthy lifestyles and poor health outcomes. *PloS one*, Public Library of Science San Francisco, USA, v. 9, n. 6, p. e94602, 2014.

DUAN, H.; WANG, J.; CHEN, K.; LIN, D. Pyskl: Towards good practices for skeleton action recognition. In: *Proceedings of the 30th ACM International Conference on Multimedia*. [S.l.: s.n.], 2022. p. 7351–7354.

EKMAN, P. An argument for basic emotions. *Cognition & emotion*, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992.

EKMAN, P. Facial expression and emotion. *American psychologist*, American Psychological Association, v. 48, n. 4, p. 384, 1993.

EKMAN, P.; FRIESEN, W. V.; ELLSWORTH, P. What are the similarities and differences in facial behavior across cultures. *Emotion in the human face*, Cambridge University Press Cambridge, v. 2, p. 128–44, 1982.

EKMAN, P.; O'SULLIVAN, M.; FRIESEN, W. V.; SCHERER, K. R. Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior*, Springer, v. 15, n. 2, p. 125–135, 1991.

EKMAN, P.; ROPER, G.; HAGER, J. C. Deliberate facial movement. *Child development*, JSTOR, p. 886–891, 1980.

FAIRCLOUGH, S. H.; TATTERSALL, A. J.; HOUSTON, K. Anxiety and performance in the british driving test. *Transportation Research Part F: Traffic Psychology and Behaviour*, Elsevier, v. 9, n. 1, p. 43–52, 2006.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, American Psychological Association, v. 76, n. 5, p. 378, 1971.

FRIESEN, E.; EKMAN, P. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, v. 3, n. 2, p. 5, 1978.

GAO, Q.; ZENG, H.; LI, G.; TONG, T. Graph reasoning-based emotion recognition network. *IEEE Access*, p. 6488–6497, 2021.

GAVRILESCU, M.; VIZIREANU, N. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 19, n. 17, p. 3693, 2019.

GOODFELLOW, I. J.; ERHAN, D.; CARRIER, P. L.; COURVILLE, A.; MIRZA, M.; HAMNER, B.; CUKIERSKI, W.; TANG, Y.; THALER, D.; LEE, D.-H. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. [S.l.], 2013. p. 117–124.

GORBOVA, J.; LUSI, I.; LITVIN, A.; ANBARJAFARI, G. Automated screening of job candidate based on multimodal video processing. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. [S.l.: s.n.], 2017. p. 29–35.

GU, C.; SUN, C.; ROSS, D. A.; VONDRICK, C.; PANTOFARU, C.; LI, Y.; VIJAYANARASIMHAN, S.; TODERICI, G.; RICCO, S.; SUKTHANKAR, R. et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 6047–6056.

GUPTA, A.; AGRAWAL, D.; CHAUHAN, H.; DOLZ, J.; PEDERSOLI, M. An attention model for group-level emotion recognition. *ACM International Conference on Multimodal Interaction*, p. 611–615, 2018.

GUPTA, S.; KUMAR, P.; TEKCHANDANI, R. K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia Tools and Applications*, Springer, v. 82, n. 8, p. 11365–11394, 2023.

HALFORD, G. S.; HINE, T. J. Fundamental differences between perception and cognition aside from cognitive penetrability. *Behavioral and Brain Sciences*, Cambridge University Press, v. 39, 2016.

HANCKE, G. P.; SILVA, B. de Carvalho e; JR, G. P. H. The role of advanced sensing in smart cities. *Sensors*, MDPI, v. 13, n. 1, p. 393–425, 2012.

HAOUIJ, N. E.; POGGI, J.-M.; SEVESTRE-GHALILA, S.; GHOZI, R.; JAÏDANE, M. Affectiveroad system and database to assess driver's attention. In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. [S.l.: s.n.], 2018. p. 800–803.

HARRÉ, R. The social construction of emotions. 1986.

HASTINGS, J.; CEUSTERS, W.; SMITH, B.; MULLIGAN, K. Dispositions and processes in the emotion ontology. 2011.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. Mask r-cnn. *IEEE International Conference on Computer Vision*, p. 2961–2969, 2017.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, p. 770–778, 2016.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Identity mappings in deep residual networks. *European Conference on Computer Vision*, p. 630–645, 2016.

HENDRYCKS, D.; GIMPEL, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

HINTON, G.; SRIVASTAVA, N.; SWERSKY, K. *Neural networks for machine learning lecture 6a overview of mini-batch gradient descent*. 2012.

HJORTSJÖ, C.-H. *Man's face and mimic language*. [S.l.]: Studentlitteratur Lund, Sweden, 1970.

HOWARD, A.; SANDLER, M.; CHU, G.; CHEN, L.-C.; CHEN, B. et al. Searching for mobilenetv3. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2019. p. 1314–1324.

HU, J. C.; CAVICCHIOLI, R.; CAPOTONDI, A. Expansionnet v2: Block static expansion in fast end to end training for image captioning. *arXiv preprint arXiv:2208.06551*, 2022.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4700–4708.

IMANI, M.; MONTAZER, G. A. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, Elsevier, v. 147, p. 102423, 2019.

IOFFE, S.; SZEGEDY, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. [S.l.]: arXiv, 2015.

JACOB, H.; KREIFELTS, B.; NIZIELSKI, S.; SCHÜTZ, A.; WILDGRUBER, D. Effects of emotional intelligence on the impression of irony created by the mismatch between verbal and nonverbal cues. *PloS one*, Public Library of Science San Francisco, CA USA, v. 11, n. 10, p. e0163211, 2016.

KAMYAB, M.; LIU, G.; ADJEISAH, M. Attention-based cnn and bi-lstm model based on tf-idf and glove word embedding for sentiment analysis. *Applied Sciences*, MDPI, v. 11, n. 23, p. 11255, 2021.

KERKENI, L.; SERRESTOU, Y.; MBARKI, M.; RAOOF, K.; MAHJOUB, M. A. Speech emotion recognition: Methods and cases study. *ICAART (2)*, v. 20, 2018.

KHAN, M. A. R.; ROSTOV, M.; RAHMAN, J. S.; AHMED, K. A.; HOSSAIN, M. Z. Assessing the applicability of machine learning models for robotic emotion monitoring: A survey. *Applied Sciences*, Multidisciplinary Digital Publishing Institute, v. 13, n. 1, p. 387, 2023.

KHOSLA, R.; CHU, M.-T.; NGUYEN, K. Human-robot interaction modelling for recruitment and retention of employees. In: SPRINGER. *HCI in Business, Government, and Organizations: Information Systems: Third International Conference, HCIBGO 2016, Held as Part of HCI International 2016, Toronto, Canada, July 17-22, 2016, Proceedings, Part II 3*. [S.l.], 2016. p. 302–312.

KING, D. E. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, v. 10, p. 1755–1758, 2009.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

KLEINSMITH, A.; BIANCHI-BERTHOUZE, N. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, IEEE, v. 4, n. 1, p. 15–33, 2012.

KNUTSON, B. Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, Springer, v. 20, n. 3, p. 165–182, 1996.

KOŁAKOWSKA, A.; SZWOCH, W.; SZWOCH, M. A review of emotion recognition methods based on data acquired via smartphone sensors. *Sensors*, MDPI, v. 20, n. 21, p. 6367, 2020.

KOSSAIFI, J.; TZIMIROPOULOS, G.; TODOROVIC, S.; PANTIC, M. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, Elsevier, v. 65, p. 23–36, 2017.

KOSTI, R.; ALVAREZ, J. M.; RECASENS, A.; LAPEDRIZA, A. Emotic: Emotions in context dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2017. p. 61–69.

KOSTI, R.; ALVAREZ, J. M.; RECASENS, A.; LAPEDRIZA, A. Emotion recognition in context. *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1667–1675, 2017.

KOSTI, R.; ALVAREZ, J. M.; RECASENS, A.; LAPEDRIZA, A. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, v. 42, n. 11, p. 2755–2766, 2019.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, p. 1097–1105, 2012.

KUMAR, V.; RAO, S.; YU, L. Noisy student training using body language dataset improves facial expression recognition. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2020. p. 756–773.

LE, N.; NGUYEN, K.; NGUYEN, A.; LE, B. Global-local attention for emotion recognition. *Neural Computing and Applications*, p. 1–15, 2021.

LE, N.; NGUYEN, K.; NGUYEN, A.; LE, B. Global-local attention for emotion recognition. *Neural Computing and Applications*, Springer, v. 34, n. 24, p. 21625–21639, 2022.

LEE, J.; KIM, S.; KIM, S.; PARK, J.; SOHN, K. Context-aware emotion recognition networks. *IEEE International Conference on Computer Vision*, p. 10143–10152, 2019.

LHOMMET, M.; MARSELLA, S. C. Expressing emotion through posture. *The Oxford handbook of affective computing*, Oxford Univ. Press, v. 273, 2014.

LI, B.; ZHU, C.; LI, S.; ZHU, T. Identifying emotions from non-contact gaits information based on microsoft kinects. *IEEE Transactions on Affective Computing*, IEEE, v. 9, n. 4, p. 585–591, 2016.

LI, D.; ZHANG, H. Improved regularization and robustness for fine-tuning in neural networks. *Advances in Neural Information Processing Systems*, v. 34, p. 27249–27262, 2021.

LI, W.; ZHU, L.; SHI, Y.; GUO, K.; CAMBRIA, E. User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models. *Applied Soft Computing*, Elsevier, v. 94, p. 106435, 2020.

LI, Y.; WANG, Y.; CUI, Z. Decoupled multimodal distilling for emotion recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 6631–6640.

LIANG, S.; LIU, Y.; LI, G.; ZHAO, G. Elderly fall risk prediction with plantar center of force using convlstm algorithm. In: IEEE. *2019 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. [S.l.], 2019. p. 36–41.

LIM, A.; OKUNO, H. G. Developing robot emotions through interaction with caregivers. In: *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics*. [S.l.]: IGI Global, 2015. p. 316–337.

LIMA, J. P.; ROBERTO, R.; FIGUEIREDO, L.; SIMÕES, F.; THOMAS, D.; UCHIYAMA, H.; TEICHRIEB, V. 3d pedestrian localization using multiple cameras: a generalizable approach. *Machine Vision and Applications*, Springer, v. 33, n. 4, p. 61, 2022.

LIMA, M. L. L. d. *Um estudo de reconhecimento de emoções baseado em linguagem corporal e marcha*. Master's Thesis (B.S. thesis), 2023.

LIMBU, D. K.; ANTHONY, W. C. Y.; ADRIAN, T. H. J.; DUNG, T. A.; KEE, T. Y.; DAT, T. H.; ALVIN, W. H. Y.; TERENCE, N. W. Z.; RIDONG, J.; JUN, L. Affective social interaction with cuddler robot. *IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, p. 179–184, 2013.

LIU, B. Text sentiment analysis based on cbow model and deep learning in big data environment. *Journal of ambient intelligence and humanized computing*, Springer, v. 11, p. 451–458, 2020.

LIU, J.; HU, M.; WANG, Y.; HUANG, Z.; JIANG, J. Symmetric multi-scale residual network ensemble with weighted evidence fusion strategy for facial expression recognition. *Symmetry*, MDPI, v. 15, n. 6, p. 1228, 2023.

LIU, J.-J.; HOU, Q.; CHENG, M.-M.; WANG, C.; FENG, J. Improving convolutional networks with self-calibrated convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, p. 10096–10105, 2020.

LIU, X.; SHI, H.; CHEN, H.; YU, Z.; LI, X.; ZHAO, G. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 10631–10642.

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: *IEEE/CVF International Conference on Computer Vision (CVPR)*. [S.l.: s.n.], 2021. p. 10012–10022.

LÓPEZ-CIFUENTES, A.; ESCUDERO-VIÑOLO, M.; BESCÓS, J.; GARCÍA-MARTÍN, Á. Semantic-aware scene recognition. *Pattern Recognition*, Elsevier, p. 107256, 2020.

LUCEY, P.; COHN, J. F.; KANADE, T.; SARAGIH, J.; AMBADAR, Z.; MATTHEWS, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE. *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. [S.l.], 2010. p. 94–101.

LUGARESI, C.; TANG, J.; NASH, H.; MCCLANAHAN, C.; UBOWEJA, E.; HAYS, M.; ZHANG, F.; CHANG, C.-L.; YONG, M.; LEE, J.; CHANG, W.-T.; HUA, W.; GEORG, M.; GRUNDMANN, M. Mediapipe: A framework for perceiving and processing reality. In: *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*. [s.n.], 2019. Available at: <https://mixedreality.cs.cornell.edu/s/NewTitle%5FMay1%5FMediaPipe%5FCVPR%5FCV4ARVR%5FWorkshop%5F2019.pdf>.

LUO, Y.; YE, J.; ADAMS, R. B.; LI, J.; NEWMAN, M. G.; WANG, J. Z. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International journal of computer vision*, Springer, v. 128, n. 1, p. 1–25, 2020.

MACHOVA, K.; MIKULA, M.; GAO, X.; MACH, M. Lexicon-based sentiment analysis using the particle swarm optimization. *Electronics*, MDPI, v. 9, n. 8, p. 1317, 2020.

MARIAPPAN, M. B.; SUK, M.; PRABHAKARAN, B. Facefetch: A user emotion driven multimedia content recommendation system based on facial expression recognition. In: IEEE. *2012 IEEE International Symposium on Multimedia*. [S.l.], 2012. p. 84–87.

MARMPENA, M.; LIM, A.; DAHL, T. S. How does the robot feel? perception of valence and arousal in emotional body language. *Paladyn, Journal of Behavioral Robotics*, Sciendo, v. 9, n. 1, p. 168–182, 2018.

MARTINEZ, A. M. Visual perception of facial expressions of emotion. *Current opinion in psychology*, Elsevier, v. 17, p. 27–33, 2017.

MEHRABIAN, A. Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. 1980.

MENG, D.; PENG, X.; WANG, K.; QIAO, Y. Frame attention networks for facial expression recognition in videos. In: IEEE. *2019 IEEE international conference on image processing (ICIP)*. [S.l.], 2019. p. 3866–3870.

MENG, Q.; HU, X.; KANG, J.; WU, Y. On the effectiveness of facial expression recognition for evaluation of urban sound perception. *Science of The Total Environment*, Elsevier, v. 710, p. 135484, 2020.

MESQUITA, B.; BOIGER, M.; LEERSNYDER, J. D. Doing emotions: The role of culture in everyday emotions. *European Review of Social Psychology*, Taylor & Francis, v. 28, n. 1, p. 95–133, 2017.

MESQUITA, B.; FRIJDA, N. H. Cultural variations in emotions: a review. *Psychological bulletin*, American Psychological Association, v. 112, n. 2, p. 179, 1992.

MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, ACM New York, NY, USA, v. 38, n. 11, p. 39–41, 1995.

MILLER, G. A. *WordNet: An electronic lexical database*. [S.l.]: MIT press, 1998.

MINSKY, M. *Society of mind*. [S.l.]: Simon and Schuster, 1988.

MITTAL, T.; GUHAN, P.; BHATTACHARYA, U.; CHANDRA, R.; BERA, A.; MANOCHA, D. Emoticon: Context-aware multimodal emotion recognition using frege's principle. June 2020.

MOBBS, D.; WEISKOPF, N.; LAU, H. C.; FEATHERSTONE, E.; DOLAN, R. J.; FRITH, C. D. The kuleshov effect: the influence of contextual framing on emotional attributions. *Social cognitive and affective neuroscience*, Oxford University Press, v. 1, n. 2, p. 95–106, 2006.

MOLLAHOSSEINI, A.; HASANI, B.; MAHOOR, M. H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, IEEE, v. 10, n. 1, p. 18–31, 2017.

MONTEPARE, J. M.; DOBISH, H. The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior*, Springer, v. 27, n. 4, p. 237–254, 2003.

MONTEPARE, J. M.; GOLDSTEIN, S. B.; CLAUSEN, A. The identification of emotions from gait information. *Journal of Nonverbal Behavior*, Springer, v. 11, p. 33–42, 1987.

MOULOUA, M.; BRILL, J. C.; SHIRKEY, E. Gender differences and aggressive driving behavior: A factor analytic study. In: SAGE PUBLICATIONS SAGE CA: LOS ANGELES, CA. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. [S.l.], 2007. v. 51, n. 18, p. 1283–1286.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. *International Conference on Machine Learning*, p. 807–814, 2010.

NARAYANAN, V.; MANOGHAR, B. M.; DORBALA, V. S.; MANOCHA, D.; BERA, A. Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In: IEEE. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. [S.l.], 2020. p. 8200–8207.

NES, A.; SUNDBERG, K.; WATZL, S. The perception/cognition distinction. *Inquiry*, Taylor & Francis, v. 66, n. 2, p. 165–195, 2023.

PANDEY, R.; PUROHIT, H.; CASTILLO, C.; SHALIN, V. L. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, Elsevier, v. 160, p. 102772, 2022.

PASTOR, M. C.; BRADLEY, M. M.; LÖW, A.; VERSACE, F.; MOLTÓ, J.; LANG, P. J. Affective picture perception: emotion, context, and the late positive potential. *Brain research*, Elsevier, v. 1189, p. 145–151, 2008.

PASZKE, A.; GROSS, S.; CHINTALA, S.; CHANAN, G.; YANG, E.; DEVITO, Z.; LIN, Z.; DESMAISON, A.; ANTIGA, L.; LERER, A. *Automatic differentiation in PyTorch*. 2017.

PATEL, D. S. Body language: An effective communication tool. *IUP Journal of English Studies*, p. 2, 2014.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543.

PICARD, R. W. *Affective computing*. [S.l.]: MIT press, 2000.

PODELL, D.; ENGLISH, Z.; LACEY, K.; BLATTMANN, A.; DOCKHORN, T.; MÜLLER, J.; PENNA, J.; ROMBACH, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

RANDHAVANE, T.; BHATTACHARYA, U.; KAPSASKIS, K.; GRAY, K.; BERA, A.; MANOCHA, D. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*, p. 1, 2019.

RANDHAVANE, T.; BHATTACHARYA, U.; KAPSASKIS, K.; GRAY, K.; BERA, A.; MANOCHA, D. Learning perceived emotion using affective and deep features for mental health applications. In: IEEE. *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. [S.l.], 2019. p. 395–399.

RANGANATHAN, H.; CHAKRABORTY, S.; PANCHANATHAN, S. Multimodal emotion recognition using deep learning architectures. In: IEEE. *2016 IEEE winter conference on applications of computer vision (WACV)*. [S.l.], 2016. p. 1–9.

REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

RIGHART, R.; GELDER, B. D. Rapid influence of emotional scenes on encoding of facial expressions: an erp study. *Social cognitive and affective neuroscience*, Oxford University Press, v. 3, n. 3, p. 270–278, 2008.

ROETHER, C. L.; OMLOR, L.; CHRISTENSEN, A.; GIESE, M. A. Critical features for the perception of emotion from gait. *Journal of vision*, The Association for Research in Vision and Ophthalmology, v. 9, n. 6, p. 15–15, 2009.

ROSEBOROUGH, J. E.; WICKENS, C. M.; WIESENTHAL, D. L. Retaliatory aggressive driving: A justice perspective. *Accident Analysis & Prevention*, Elsevier, v. 162, p. 106393, 2021.

ROUAST, P. V.; ADAM, M. T.; CHIONG, R. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, IEEE, v. 12, n. 2, p. 524–543, 2019.

SAID, C. P.; SEBE, N.; TODOROV, A. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, American Psychological Association, v. 9, n. 2, p. 260, 2009.

SAJJAD, M.; NASIR, M.; ULLAH, F. U. M.; MUHAMMAD, K.; SANGAIAH, A. K.; BAIK, S. W. Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services. *Information Sciences*, Elsevier, v. 479, p. 416–431, 2019.

SAMUVEL, D. J.; PERUMAL, B.; ELANGOVAN, M. Music recommendation system based on facial emotion recognition. *3C Tecnologia*, 3Ciencias, p. 261–271, 2020.

SANTAMARIA-GRANADOS, L.; MENDOZA-MORENO, J. F.; RAMIREZ-GONZALEZ, G. Tourist recommender systems based on emotion recognition—a scientometric review. *Future Internet*, MDPI, v. 13, n. 1, p. 2, 2020.

SAVCHENKO, A. V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In: IEEE. *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. [S.l.], 2021. p. 119–124.

SAXENA, A.; KHANNA, A.; GUPTA, D. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, Institute of Electronics and Computer, v. 2, n. 1, p. 53–79, 2020.

SCHLOSBERG, H. Three dimensions of emotion. *Psychological review*, American Psychological Association, v. 61, n. 2, p. 81, 1954.

SEAGAL, S.; HORNE, D. Human dynamics for the 21st century. *Systems Thinker*, PEGASUS COMMUNICATIONS, v. 14, p. 2–6, 2003.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision*, p. 618–626, 2017.

SHANG, L.; XI, H.; HUA, J.; TANG, H.; ZHOU, J. A lexicon enhanced collaborative network for targeted financial sentiment analysis. *Information Processing and Management*, Elsevier, v. 60, n. 2, p. 103187, 2023.

SHAW, S.-L.; SUI, D. *Human dynamics research in smart and connected communities*. [S.l.]: Springer, 2018.

SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [S.l.]: arXiv, 2014.

SOLEYMANI, M.; PANTIC, M.; PUN, T. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, IEEE, v. 3, n. 2, p. 211–223, 2011.

SRIVASTAVA, D.; SINGH, A. K.; TAPASWI, M. How you feelin'? learning emotions and mental states in movie scenes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 2517–2528.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2818–2826.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 6105–6114.

TASSINARY, L. G.; CACIOPPO, J. T. *Unobservable facial actions and emotion*. [S.l.]: SAGE Publications Sage CA: Los Angeles, CA, 1992.

THUSEETHAN, S.; RAJASEGARAR, S.; YEARWOOD, J. Emosec: Emotion recognition from scene context. *Neurocomputing*, Elsevier, v. 492, p. 174–187, 2022.

TRACY, J. L.; MATSUMOTO, D. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 105, n. 33, p. 11655–11660, 2008.

UNDERWOOD, G.; CHAPMAN, P.; WRIGHT, S.; CRUNDALL, D. Anger while driving. *Transportation Research Part F: traffic psychology and behaviour*, Elsevier, v. 2, n. 1, p. 55–68, 1999.

VALLI, A. Natural interaction. *White Paper*, Citeseer, 2007.

VALLI, A. *Alessandro Valli - Notes on Natural Interaction*. [S.l.: s.n.], 2008.

VELTMEIJER, E. A.; GERRITSEN, C.; HINDRIKS, K. V. Automatic emotion recognition for groups: a review. *IEEE Transactions on Affective Computing*, IEEE, v. 14, n. 1, p. 89–107, 2021.

VENTURE, G.; KADONE, H.; ZHANG, T.; GRÈZES, J.; BERTHOZ, A.; HICHEUR, H. Recognizing emotions conveyed by human gait. *International Journal of Social Robotics*, Springer, v. 6, p. 621–632, 2014.

WALLBOTT, H. G. Bodily expression of emotion. *European journal of social psychology*, Wiley Online Library, v. 28, n. 6, p. 879–896, 1998.

WALLBOTT, H. G.; SCHERER, K. R. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, American Psychological Association, v. 51, n. 4, p. 690, 1986.

WEI, H.; HAUER, R. J.; CHEN, X.; HE, X. Facial expressions of visitors in forests along the urbanization gradient: What can we learn from selfies on social networking services? *Forests*, Multidisciplinary Digital Publishing Institute, v. 10, n. 12, p. 1049, 2019.

WEN, Z.; LIN, W.; WANG, T.; XU, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, MDPI, v. 8, n. 2, p. 199, 2023.

WU, J.; ZHANG, Y.; NING, L. The fusion knowledge of face, body and context for emotion recognition. *IEEE International Conference on Multimedia & Expo Workshops*, p. 108–113, 2019.

WU, S.; ZHOU, L.; HU, Z.; LIU, J. Hierarchical context-based emotion recognition with scene graphs. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, 2022.

XIAO, B.; WU, H.; WEI, Y. Simple baselines for human pose estimation and tracking. *European Conference on Computer Vision*, p. 466–481, 2018.

XIAO, H.; LI, W.; ZENG, G.; WU, Y.; XUE, J.; ZHANG, J.; LI, C.; GUO, G. On-road driver emotion recognition using facial expression. *Applied Sciences*, MDPI, v. 12, n. 2, p. 807, 2022.

XU, D.; TIAN, Z.; LAI, R.; KONG, X.; TAN, Z.; SHI, W. Deep learning based emotion analysis of microblog texts. *Information Fusion*, Elsevier, v. 64, p. 1–11, 2020.

XU, H.; KONG, J.; KONG, X.; LI, J.; WANG, J. Mcf-net: Fusion network of facial and scene features for expression recognition in the wild. *Applied Sciences*, MDPI, v. 12, n. 20, p. 10251, 2022.

XU, K.; HU, W.; LESKOVEC, J.; JEGELKA, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

XU, Y.; ZHONG, X.; YEPES, A. J. J.; LAU, J. H. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In: IEEE. *2020 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2020. p. 1–8.

YADOLLAHI, E.; CHANDRA, S.; COUTO, M.; LIM, A.; SANDYGULOVA, A. Children, robots, and virtual agents: Present and future challenges. In: *Interaction design and children*. [S.l.: s.n.], 2021. p. 682–686.

YAN, S.; XIONG, Y.; LIN, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2018. v. 32, n. 1.

YANG, D.; HUANG, S.; WANG, S.; LIU, Y.; ZHAI, P.; SU, L.; LI, M.; ZHANG, L. Emotion recognition for multiple context awareness. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2022. p. 144–162.

YANG, Z.; HUANG, Y. Algorithm for speech emotion recognition classification based on mel-frequency cepstral coefficients and broad learning system. *Evolutionary Intelligence*, Springer, v. 15, n. 4, p. 2485–2494, 2022.

ZEILER, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

ZEPF, S.; HERNANDEZ, J.; SCHMITT, A.; MINKER, W.; PICARD, R. W. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 53, n. 3, p. 1–30, 2020.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. *Dive into Deep Learning*. [S.l.]: Cambridge University Press, 2023. <https://D2L.ai>.

ZHANG, J. Movies and pop songs recommendation system by emotion detection through facial recognition. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2020. v. 1650, n. 3, p. 032076.

ZHANG, K.; ZHANG, Z.; LI, Z.; QIAO, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, IEEE, v. 23, n. 10, p. 1499–1503, 2016.

ZHANG, M.; LIANG, Y.; MA, H. Context-aware affective graph reasoning for emotion recognition. In: IEEE. *2019 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.], 2019. p. 151–156.

ZHANG, S.; PAN, Y.; WANG, J. Z. Learning emotion representations from verbal and nonverbal communication. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 18993–19004.

ZHANG, S.; ZHANG, Y.; ZHANG, Y.; WANG, Y.; SONG, Z. A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, MDPI, v. 12, n. 17, p. 3595, 2023.

ZHANG, Y.; WANG, C.; LING, X.; DENG, W. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2022. p. 418–434.

ZHAO, Z.; LIU, Q.; WANG, S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, p. 6544–6556, 2021.

ZHAO, Z.; LIU, Q.; ZHOU, F. Robust lightweight facial expression recognition network with label distribution training. *AAAI Conference on Artificial Intelligence*, p. 3510–3519, 2021.

ZHOU, B.; LAPEDRIZA, A.; KHOSLA, A.; OLIVA, A.; TORRALBA, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 40, n. 6, p. 1452–1464, 2017.

ZHOU, H.; MENG, D.; ZHANG, Y.; PENG, X.; DU, J.; WANG, K.; QIAO, Y. Exploring emotion features and fusion strategies for audio-video emotion recognition. In: *2019 International conference on multimodal interaction*. [S.l.: s.n.], 2019. p. 562–566.

ZHOU, S.; WU, X.; JIANG, F.; HUANG, Q.; HUANG, C. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. *International Journal of Environmental Research and Public Health*, MDPI, v. 20, n. 2, p. 1400, 2023.

ZLOTEANU, M. *Emotions and deception detection*. Phd Thesis (PhD Thesis) — UCL (University College London), 2017.

ZOPH, B.; GHIASI, G.; LIN, T.-Y.; CUI, Y.; LIU, H.; CUBUK, E. D.; LE, Q. Rethinking pre-training and self-training. *Advances in neural information processing systems*, v. 33, p. 3833–3845, 2020.

ZUCKERMAN, M.; HALL, J. A.; DEFRANK, R. S.; ROSENTHAL, R. Encoding and decoding of spontaneous and posed facial expressions. *Journal of Personality and Social Psychology*, American Psychological Association, v. 34, n. 5, p. 966, 1976.