



UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE



UNIVERSITÉ DE TUNIS EL MANAR

FACULTÉ DES SCIENCES MATHÉMATIQUES, PHYSIQUES ET NATURELLES DE TUNIS

ÉCOLE DOCTORALE DE MATHÉMATIQUES, INFORMATIQUE, SCIENCES ET TECHNOLOGIES DE LA MATIÈRE

UNIVERSITÀ DEGLI STUDI DI FIRENZE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE (DINFO)

THESIS

Recognition of facial emotions from videos

Submitted in fulfilment of the requirements
for the degree of

DOCTOR OF COMPUTER SCIENCE

PhD IN SMART COMPUTING

Presented by

Hajer GUERDELLI

Defended on 14/09/2025

Before the jury composed of:

Ezzeddine ZAGROUBA	Professor at ISI	President
Khaled NOURI	Professor at ENICarthage	Reviewer
Roberto VEZZANI	Professor at University of Modena	Reviewer
Leonardo GALTERI	Professor at Pegaso Telematic University	Examiner
Walid BARHOUMI	Professor at ENICarthage	Thesis Director
Stefano BERRETTI	Professor at University of Florence	Thesis Director



UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE



UNIVERSITÉ DE TUNIS EL MANAR

FACULTÉ DES SCIENCES MATHÉMATIQUES, PHYSIQUES ET NATURELLES DE TUNIS

ÉCOLE DOCTORALE DE MATHÉMATIQUES, INFORMATIQUE, SCIENCES ET TECHNOLOGIES DE LA MATIÈRE

UNIVERSITÀ DEGLI STUDI DI FIRENZE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE (DINFO)

THÈSE

Reconnaissance des émotions faciales à partir de vidéos

En vue de l'obtention du

DOCTORAT EN INFORMATIQUE

PhD IN SMART COMPUTING

Présentée par

Hajer GUERDELLI

Soutenue le 14/09/2025

Devant le jury composé de:

Ezzeddine ZAGROUBA

Khaled NOURI

Roberto VEZZANI

Leonardo GALTERI

Walid BARHOUMI

Stefano BERRETTI

Professeur à ISI

Professeur à ENICarthage

Professeur à Université de Modena

Professeur à Pegaso Telematic Université

Professeur à ENICarthage

Professeur à Université de Florence

Président

Rapporteur

Rapporteur

Examineur

Dir. de thèse

Dir. de thèse

SUMMARY

This thesis addresses the challenge of facial emotion recognition from videos, a critical task with wide-ranging applications in many fields such as human-computer interaction and healthcare. While traditional methods often focus on static and isolated emotions, this work captures the dynamic nature of emotions as they evolve in real-world interactions. The multi-emotion approach introduced in this thesis stems from the need to move beyond the six basic emotions. By incorporating insights from Russell’s circumplex model and Parrott’s taxonomy, this research proposes a fresh approach to emotion classification, expanding beyond the basic emotion framework. This broader taxonomy enables the recognition of a wider and more nuanced range of emotional states, significantly enhancing emotion recognition capabilities in complex real-world environments. To implement this refined approach, the thesis introduces the IMEmo dataset, annotated with 16 distinct emotions. This dataset is specifically designed to capture emotional transitions during video sequences, offering a more comprehensive understanding of emotional dynamics. IMEmo represents a significant contribution by enabling detailed analysis of how emotions evolve within interactions. Building on this, the thesis also proposes an emotion prediction model based on Long Short-Term Memory (LSTM) networks. The model leverages past emotional states in an interaction to predict future emotional shifts, providing a more accurate and realistic forecast of short-term emotional transitions. Extensive experiments conducted on the IMEmo dataset demonstrate the model’s effectiveness in predicting these changes over time. Together, the IMEmo dataset and the LSTM-based prediction model form the core contributions of this thesis, advancing both the recognition and prediction of emotions in dynamic, real-world settings. This research not only broadens the emotional spectrum analyzed but also contributes to advancing the understanding of dynamic emotion prediction in video sequences.

RESUME

Cette thèse aborde le défi de la reconnaissance des émotions faciales à partir de vidéos, une tâche cruciale avec des applications variées dans des domaines tels que l'interaction homme-machine et les soins de santé. Alors que les méthodes traditionnelles se concentrent souvent sur des émotions statiques et isolées, ce travail met en lumière la nature dynamique des émotions qui évoluent au cours des interactions réelles. L'approche multi-émotion introduite dans cette thèse découle de la nécessité de dépasser le cadre limité des six émotions de base. En intégrant des idées issues du modèle circumplex de Russell et de la taxonomie de Parrott, cette recherche propose une nouvelle approche de classification des émotions, élargissant ainsi la reconnaissance à un éventail plus large et plus nuancé d'états émotionnels. Cette taxonomie améliore de manière significative la reconnaissance des émotions dans des environnements complexes et réels. Pour mettre en œuvre cette approche affinée, la thèse présente la base de données IMEmo, annotée avec 16 émotions distinctes. IMEmo est spécifiquement conçu pour capturer les transitions émotionnelles dans les vidéos, offrant ainsi une compréhension plus approfondie de la dynamique des émotions. IMEmo représente une contribution importante pour l'analyse détaillée de l'évolution des émotions au sein des interactions. Ainsi, la thèse propose un modèle de prédiction des émotions basé sur les réseaux de mémoire à long court terme (LSTM). Le modèle s'appuie sur les états émotionnels passés dans une interaction pour prédire les changements émotionnels futurs, offrant une prévision plus réaliste et précise des transitions émotionnelles à court terme. Des expériences approfondies menées sur IMEmo démontrent l'efficacité du modèle dans la prédiction de ces changements au fil du temps. En effet, la base de données IMEmo et le modèle de prédiction constituent les contributions centrales de cette thèse, faisant progresser à la fois la reconnaissance et la prédiction des émotions dans des environnements dynamiques et réels. Ceci élargit non seulement le spectre émotionnel analysé, mais contribue également à améliorer la compréhension de la prédiction dynamique des émotions dans des vidéos.

Table of Contents

List of Tables	vii
List of Figures	ix
General Introduction	1
1 The Limits of Six: A New Look at Emotion Recognition	5
1.1 Introduction	5
1.2 Motivation and Background	6
1.3 Facial Macro- and Micro-Expressions	11
1.3.1 Macro-Expressions	12
1.3.1.1 Macro-Expression Datasets	12
1.3.1.2 Spontaneous Datasets	13
1.3.1.3 Spontaneous and Posed Datasets	23
1.3.1.4 In-the-Wild Datasets	25
1.3.1.5 Other Categorizations of Macro-Expression Datasets	29
1.3.2 Micro-Expressions	35
1.3.2.1 Micro-Expression Datasets	35
1.3.2.2 Spontaneous Datasets	36
1.3.2.3 In-the-Wild Datasets	38
1.3.2.4 Other Categorizations of Micro-Expression Datasets	38
1.3.3 Applications	51
1.4 Human Emotion Recognition and Dataset Challenges	52
1.4.1 Proposed Emotion Taxonomy	52
1.4.1.1 Russell vs. Parrott	53
1.4.1.2 Terms Selection and Mapping	55
1.4.1.3 Terms Taxonomy	56
1.4.2 Experiments	56

1.4.2.1	Relabeling the AffectNet Dataset	56
1.4.2.2	Results	60
1.5	Conclusion	62
2	Emotions in Interpersonal Relationships	63
2.1	Introduction	63
2.2	Interpersonal Relation Recognition	64
2.2.1	Social Relation Traits	64
2.2.2	Objective relations	67
2.2.2.1	Kinship Recognition	69
2.2.2.2	Asian Social Psychology Theory	70
2.2.2.3	Bugental's Social Psychology Theory	71
2.2.3	Subjective Relations	78
2.2.3.1	Kiesler's Circle	81
2.2.3.2	Fiske's Theory of Relational Models	83
2.2.3.3	Others	84
2.3	Prediction of Emotions	86
2.4	Conclusion	87
3	The Florence Interpersonal Relations with Multi-Emotions Dataset	89
3.1	Introduction	89
3.2	The Interpersonal Relations with Multi-Emotions (IMEmo) Dataset	90
3.2.1	Main Dataset Statistics	91
3.2.2	Data Pre-processing	93
3.2.3	Annotations	93
3.2.3.1	Clip-level Annotations	94
3.2.3.2	Shot-level Annotations	94
3.2.3.3	Potential Investigation Scenarios	105
3.2.4	Applications of the IMEmo Dataset	106
3.3	Experiments and Results	108
3.3.1	Reference Architecture and Compared Datasets	108
3.3.2	Emotion Recognition Results	110
3.3.3	Interpersonal recognition results	113

3.4 Conclusion	114
4 Emotion Prediction in Interpersonal Relations	116
4.1 Introduction	116
4.2 Proposed Emotion Prediction Method	117
4.2.1 Proposed Prediction Pipeline	117
4.2.1.1 Feature Extraction (video Encoder)	119
4.2.1.2 Temporal Dependencies	119
4.2.1.3 Prediction Module	119
4.2.2 Application of the Proposed Model: Emotion Prediction from the IMEmo Dataset	120
4.3 Experiments and Results	121
4.3.1 Prediction Performance	121
4.3.2 Dependency Analysis	122
4.3.3 Discussion	125
4.3.4 Ablation Study	127
4.4 Conclusion	129
General Conclusion	130
Publications	139
References	140

List of Tables

1.1	Proposed categorization of macro- and micro-expression datasets.	11
1.2	Classification of macro-expression datasets according to their content.	32
1.3	Classification of macro-expression datasets according to the number of subjects.	34
1.4	Classification of micro-expression datasets according to the number of subjects.	40
1.5	Macro-expressions datasets.	41
1.6	Micro-expressions datasets.	49
1.7	Parrott's distribution of emotions	54
1.8	Our selected emotional terms.	55
1.9	The distribution of images after relabeling	58
2.1	Objective relations datasets: Publicly available audiovisual human-human (face-to-face) non-acted interaction datasets.	68
2.2	Objective relations: Methods and architectures.	74
2.4	Subjective relation datasets: Publicly available audiovisual human-human (face-to-face) non-acted interaction datasets.	79
2.5	Objective relations: Methods and architectures	80
3.1	Available interpersonal relation recognition datasets in comparison to ours.	92
3.2	Objective relation categories included in our IMEmo dataset.	95
3.3	Distribution of emotion annotations by number of frames and by number of video clips.	98
3.4	Distribution of emotion annotations according to the division of 16 emotions, 6 emotions and sentiment.	99
3.5	Distribution of annotations according to the number of annotators.	105

3.6	Recognition accuracy obtained after training and testing the CNN in [5], DAN [170] and DDAMFN [183] on two video datasets, our IMEmo and MELD, and on the image datasets AffectNet and FER-2013. Results are reported for the sentiment positive and negative (plus neutral), the 6 basic emotions (plus neutral) and, on IMEmo, for the proposed annotations with 16 emotion terms (plus neutral). This latter test was not possible for MELD, AffectNet and FER-2013 because their data were annotated only with 6 emotions (plus neutral), respectively.	113
3.7	Accuracies of relation recognition by domain / relation.	114
4.1	Prediction performance: Prediction accuracy obtained after training and testing the proposed model on the IMEmo dataset.	122
4.2	Dependency analysis: Prediction accuracy obtained after training and testing the proposed model on the IMEmo dataset.	124
4.3	Prediction performance in percentage: Evaluation of the contribution given by looking in the past to the segments of one subjects at a time (<i>Same subject</i> and <i>Different subject</i> cases) or to the segments of both the interacting subjects (<i>Both subjects</i> case).	128
4.4	The Definition of Terms Denoting Emotions in Terms of Pleasure, Arousal, and Dominance as proposed by Russell	133

List of Figures

1.1 The valence–arousal continuous emotional space.	9
1.2 Structure of the BP4D+ dataset.	14
1.3 Mapping between emotion definitions.	55
1.4 First level and second level of the proposed classification.	57
1.5 Example of relabeling according to the proposed approach.	59
1.6 Difference between AffectNet and our re-labeling.	59
1.7 Impact of class unbalance on AffectNet dataset	62
2.1 Our proposed organization of the literature on interpersonal relations. . .	66
2.2 The deep supervised feature selection framework	72
2.3 Visualization of the Kiesler’s model using the interpersonal circle.	81
2.4 The model for interpersonal relation prediction	83
3.1 Statistics of clips duration (a), and age plus gender distribution in IMEmo	
(b).	91
3.2 Overview of the annotations included in the IMEmo dataset.	93
3.3 Plots of the emotions frequency at clip level for a few selected clips. . . .	97
3.4 Emotion frequencies for the mapping of 6 emotions and sentiment.	100
3.5 Our video clip annotation website	102
3.6 Annotations of the shots of clip number 94	103
3.7 Frequencies of the 16 emotion terms plus neutral in the used training and	
validation sets.	111
3.8 Some sample frames from our IMEmo dataset with extreme cases for pose,	
illumination, and occlusion.	112
4.1 Our proposed model predicts the emotion of the next segment in a video. .	118
4.2 Example of emotional transitions between two individuals within the IMEmo	
dataset.	120

4.3	Confusion matrix for: (a) 6 emotions, (b) 16 emotions, and (c) sentiment.	126
4.4	Prediction process of the ground-truth and predicted labels.	127

General Introduction

In recent years, the study of human emotions, particularly through facial expressions and body language, has gained significant momentum, largely driven by advancements in artificial intelligence and machine learning. The ability to automatically interpret human emotions has found broad applications, ranging from enhancing human-computer interaction to improving diagnostic tools in healthcare. In fact, in today's increasingly digital world, accurate emotion prediction has become crucial in many fields such as video surveillance, personalized marketing, and social robotics. However, capturing and predicting human emotions presents significant challenges. Indeed, emotions are complex and dynamic, while being influenced by various internal and external factors. While early research primarily focused on recognizing isolated emotions in controlled environments, these studies have limited real-world applicability. In natural settings, emotions often emerge spontaneously and evolve during social interactions. As a result, analyzing emotions in dynamic and interpersonal contexts is essential for a more comprehensive understanding of human affective behavior. In particular, the study of emotions within interpersonal interactions —where emotions are triggered and modulated by exchanges with others— has increasingly become a focus of recent research. Despite advances in recognizing individual emotions, predicting how emotions transition in social contexts remains a significant challenge. This thesis addresses this gap by focusing on short-term emotion prediction in interpersonal interactions.

This thesis makes several significant contributions to the field of emotion recognition and prediction. Initially, we focused on the recognition of individual emotions using datasets of micro and macro expressions, starting from the six basic emotions. However,

we sought to go beyond this limited framework and explore the broader range of emotions described in Russell’s work, while exploring 151 distinct emotions. We experimented with various classes of these emotions to capture more nuanced and complex emotional expressions. Furthermore, as we delved deeper into the study of emotions, we shifted our focus to emotions in interpersonal relations. Our began with an extensive survey of existing datasets in this domain. However, none of the available datasets fully captured the complexity of multiple emotional states occurring simultaneously in interpersonal interactions. Thus, one of the major contributions of this thesis is the creation of the Interpersonal Relations with Multi-Emotions (IMEmo) dataset, a novel dataset specifically designed to capture the complex dynamics of emotional exchanges within interpersonal interactions. This dataset is unique in that it is the first, to our knowledge, that is annotated with 16 distinct emotions, providing a comprehensive understanding of both individual emotional states and how they interact within a dyadic context. Additionally, IMEmo is annotated with critical demographic information such as age, gender, and the type of relationship between the interacting individuals, providing a richer context for emotional analysis. Extracted from commercial movies, this dataset captures natural emotional exchanges, offering a realistic representation of how emotions are triggered and modulated in real-world interpersonal dynamics. To ensure the reliability and consistency of data, the scenes in IMEmo are carefully curated to focus on the emotional expressions of individuals captured in frontal or near-frontal views. For dyadic interactions, each individual is recorded separately rather than simultaneously, ensuring detailed analysis of facial expressions and emotional transitions. These design choices enhance the dataset’s applicability for emotion prediction tasks and maintain uniformity across all samples. The dataset’s segmentation into video shots allows for a detailed temporal analysis of emotional transitions as interactions unfold. Thus, IMEmo’s comprehensive annotations and its focus on realistic, emotionally charged interactions make it a valuable resource for advancing the study of emotion prediction in interpersonal contexts.

Building on this dataset, the second key contribution of this thesis is the development of a novel emotion prediction model. Leveraging Long Short-Term Memory (LSTM) architectures, the model is designed to capture the temporal progression of emotions within video sequences. By integrating visual cues, particularly facial expressions, the

model predicts future emotional states, utilizing the temporal dependencies within the IMEmo dataset. Our analysis revealed a significant temporal correlation between emotional states, which forms the basis for effective emotion prediction. This correlation enables the model to excel at understanding how emotions transition in response to interactions, thanks to the rich emotional and demographic annotations provided by IMEmo. The ability to model these dynamics makes the approach well-suited for real-time emotion prediction in social settings. In fact, through extensive experimentation on the IMEmo dataset, we have demonstrated the effectiveness of our model in predicting short-term emotional transitions, offering a more accurate and nuanced understanding of emotional dynamics within interpersonal interactions. Our model significantly improves the ability to forecast how emotions evolve between individuals, highlighting its potential to advance the field of emotion recognition. Thus, this research has practical applications in enhancing interactive systems where the accurate prediction and understanding of human emotions are crucial for improving responsiveness and the overall quality of human-computer interactions.

The rest of this manuscript is structured as follows. Chapter 1 provides a review of the state of the art in emotion recognition and prediction. It explores datasets related to both micro and macro expressions, while examining the methods proposed for emotion recognition. The second part of the chapter presents our work on the classification of 151 emotions, inspired by Russell’s circumplex model of affect and Parrott’s taxonomy of emotional terms. We explored a fine-grained emotion classification approach that goes beyond the traditional six basic emotions, in order to leverage the valence-arousal (VA) space to classify a wide range of emotions. We highlight the challenges we encountered, including class imbalance and inconsistencies between VA values and categorical emotion labels in existing datasets. This exploration provides valuable insights into the complexity of emotion recognition and sets the foundation for the interpersonal emotion analysis. Chapter 2 shifts the focus to emotions within interpersonal interactions. This chapter begins by surveying existing datasets and approaches and identifies the limitations in capturing the complexity of emotional exchanges between individuals. It also sets the foundation for the development of the IMEmo dataset, which captures these interpersonal emotional dynamics. Then, Chapter 3 presents the first contribution of this work: the de-

velopment of the IMEmo dataset. This chapter details the methodology used for creating and annotating the dataset, and includes experiments evaluating the dataset’s robustness in various emotional recognition tasks, highlighting its advantages for modeling interpersonal emotional dynamics. Chapter 4 introduces the second contribution, which focuses on predicting future emotions in interactions between two individuals. Based on DAN and LSTM, the proposed model leverages the past emotional states of both individuals to capture temporal dependencies and accurately forecast their future emotional states. The chapter includes detailed experimental results, demonstrating the effectiveness of this approach in predicting interpersonal emotional states. The manuscript concludes with a general conclusion that synthesizes the key findings of this research, highlights the contributions made to the field of emotion prediction, and discusses potential future directions for further improving emotion recognition in interpersonal contexts.

Chapter 1

The Limits of Six: A New Look at Emotion Recognition

The object of this chapter is to provide a comprehensive overview of the field of emotion recognition, focusing on both macro- and micro-expressions as key components in understanding human emotional states. The chapter explores Facial Emotion Recognition (FER) systems, highlighting advancements in deep learning and computer vision that have enhanced their performance. It particularly emphasizes the study of datasets that distinguish between macro-expressions—longer-lasting, visible emotions—and micro-expressions, which are brief and subtle, revealing hidden feelings. Moreover, the chapter examines the limitations of current emotion recognition systems, particularly the need for more refined classification methods. To address these challenges, a new taxonomy for emotion classification is proposed. This new approach aims to capture a broader spectrum of emotions, offering a more nuanced understanding that can improve the performance and applicability of FER systems across diverse real-world settings.

1.1 Introduction

In recent years, the automatic recognition of human emotions through facial expressions has become a central focus within the field of affective computing and computer vision. This growing interest is driven by the widespread use of digital devices that capture vast amounts of visual and contextual data, including images, videos, and 3D information. Coupled with advancements in deep learning and artificial intelligence, these datasets have enabled the development of facial expression recognition systems that significantly surpass earlier approaches based on hand-crafted features [54]. Within this broad domain,

one of the most intriguing areas of research lies in the analysis of macro- and micro-expressions, which are critical to understanding both conscious and unconscious human emotions. These expressions provide valuable cues for a wide range of applications, including security, healthcare, and human-computer interaction [151]. In fact, macro-expressions are larger, more visible facial movements, typically lasting longer and often reflecting voluntary or conscious emotions. Recognizing these expressions is relatively well-explored, with numerous datasets and models available for capturing these emotional signals [91]. In contrast, micro-expressions are subtle, involuntary facial movements that occur in response to concealed or suppressed emotions. Their brief duration, usually a fraction of a second, poses a unique challenge for current FER systems, as they require specialized algorithms to detect these fleeting emotional cues [161]. These two types of expressions, while both vital to emotion recognition, represent different layers of human emotional complexity, each with its own set of technical and theoretical challenges.

The second major aspect of this chapter deals with the broader field of Human Emotion Recognition (HER), where challenges surrounding datasets and the classification of emotions remain persistent hurdles. Current HER systems often face limitations due to the diversity and complexity of human emotional states, which are not easily captured by existing datasets. Emotions are inherently subjective, context-dependent, and dynamic, making their accurate recognition an ongoing challenge [54]. Additionally, the lack of comprehensive, standardized datasets that cover the full spectrum of human emotions has hindered progress in this field. To address these challenges, this chapter proposes a new taxonomy for the classification of emotions, aiming to provide a more nuanced and structured approach to categorizing emotional states. By rethinking the way emotions are classified and organized, this new taxonomy aims to advance the development of more accurate and reliable HER systems.

1.2 Motivation and Background

Though facial expressions are innate in humans expressiveness, their relation to emotions is less delineated, so it is first relevant to explore, in some detail, the relation and differences between facial expressions and human emotions. An emotion is a complex and intense psycho-physiological experience of an individual's state of mind when react-

ing to biochemical (internal) and environmental influences (external) [120]. For humans, emotions fundamentally include “physiological behavior, expressive behaviors and consciousness” [109]. Paul Ekman theorized that some basic human emotions are innate and shared by everyone, and that they are accompanied across cultures by universal facial expressions. Therefore, according to Ekman, an emotion causes facial expressions. However, some researchers have found that reproducing the facial expressions causes the corresponding emotion [24]. By making a happy face, for example, we begin to feel happiness. Therefore, facial expressions also cause emotions. This suggests to us that emotions and facial expressions are strongly correlated. However, according to the study of [62], the face does not always tell us the truth about emotions. At one extreme, the work of [8] has found no clear link between facial movements and internal emotional states. It is worth mentioning that almost all the studied datasets are based on the assumption that the same emotion causes nearly the same facial expressions. Indeed, within the framework of posed expressions the subject is asked to express him/herself for a given emotion. Differently, for in-the-wild applications and spontaneous expressions, the ground truths are made based on the opinions of experts who assign an emotion to the subject’s face according to his/her facial expressions. We can, therefore, state that FER datasets are benchmarks for the facial expression classification rather than the emotion recognition.

The state-of-the-art methods have progressed up to a level that has made them perfectly accurate for the first datasets, which were of reduced size while being acquired in *posed* conditions. This performance saturation on standard benchmarks [28] has induced the collection of more challenging datasets. In this respect, one trend is represented by the shift from *posed* to *spontaneous* and *in-the-wild* capturing conditions. In particular:

- *Posed* datasets are typically acquired by asking the subjects to show one of the six basic expressions as defined by Ekman [35]. In most of the cases, experienced actors are enrolled, and capturing takes place in constrained laboratory conditions;
- *Spontaneous* datasets include expressions that are stimulated by the participants. This can be the result of watching a video or a face-to-face interaction. Participants are aware that they are monitored, but emotions are shown in a natural way, rather than acted. In most of the cases, the acquisition context is a constrained one;
- *In-the-wild* datasets relax any acquisition constraint, and expressive subjects are

filmed in real-world scenarios. This is obtained by analyzing facial expressions in images and videos in movies, talk-shows, interviews, etc.

Furthermore, newly proposed facial expression datasets try to fill the gap between data and algorithms. They provide the amount of variegated data that allow facial expression recognition solutions based on neural networks to learn effective internal weights. In the recently introduced datasets, the effort of providing more annotations than those given in the oldest ones is also evidently noticed. These annotations include more expressions than the six ones categorized in the Ekman’s model, and also some emotional states. For instance, additional states have been included as follows: the neutral state in the iSAFE [148], AFEW [29], and FER-2013 [50] datasets, the neutral state and the pain state in Hi4D-ADSIP [95], the neutral state and the contempt one in BAUM-2 [38], and the embarrassment and the pain emotions in BP4D-Spontaneous [184]. Moreover, emotions have been categorized into nine categories in FER-Wild [106], thirteen emotional and mental states in BAUM-1 [179], and twenty-three categories of emotion in EmotioNet [11].

In addition, there are also datasets that propose more continuous models based on the *valence* of the expression (either positive or negative) and its intensity (called *arousal*). According to this model, as proposed by Russel [139], expressions are regarded as continuously distributed in a 2D chart, where the horizontal axis is labeled with valence values from *displeasure* to *pleasure*, while arousal passes are sorted from low- to high-activation along the vertical axis (the center of the circle represents a neutral valence and a medium level of arousal [135]). Using this diagram, several combinations are possible (see Figure 1.1), according to the different quadrants:

- *First quadrant*—emotional states go from pleased (high valence, medium arousal) to excited (about neutral valence, high arousal);
- *Second quadrant*—high arousal with about neutral valence here indicates an alarmed state, while high-negative valence and medium arousal bring to a frustrated state;
- *Third quadrant*—in this quadrant, high-negative valence and medium arousal indicate sad/depressed condition, while the status with low arousal and about neutral valence corresponds to a tired state;
- *Fourth quadrant*—finally, in this quadrant for low arousal and about neutral valence a calm/sleepy state is valence and medium arousal.

The states reported above are just given to exemplify the representations in the continuous space obtained by moving on the boundary circle of the 2D chart, while many other emotional states can be defined by moving inside the circle.

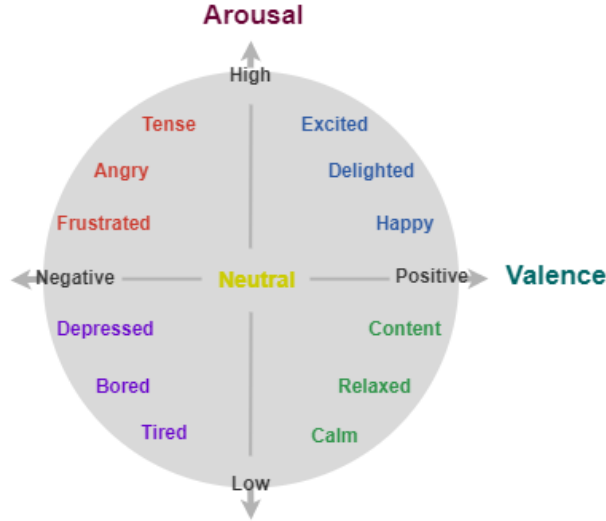


Figure 1.1: The valence–arousal continuous emotional space.

In the above overview, we have implicitly referred to *macro-expression* datasets that are normally indicated just as facial expression datasets, while omitting the “macro-” prefix. Actually, research on facial expressions can be also conducted by analyzing *micro-expressions*. These are unconscious reactions to emotional states that last only a fraction of the time of macro-expressions (less than half a second). This poses a series of additional challenges. Though micro-expressions can be categorized in the same way as the macro ones, they are subtle and difficult to observe with a normal camera working at 25 frames per second. Furthermore, since they are unconscious reactions, micro-expressions have proven to be difficult to hide and also to act. This makes capturing data for micro-expression datasets, as well as annotating them in a reliable way, quite complicated. In particular, this would require the adoption of cameras with high frame rates (up to 100/200 frames per second). It is also worth mentioning that spontaneous and in-the-wild emotions induced by specific video clips are very challenging to label. There are mainly two procedures used to label uncontrolled emotions. For the first one, as used to label the DISFA and the MMI datasets, the data are annotated based on the Facial Action Coding System (FACS), which is a coding of facial muscle actions. The second procedure uses self-reported data of subjects as the real emotion labels as performed, for example, to label the USTC-NVIE dataset. Therefore, many challenges are intrinsic to the process

of facial expression dataset labeling. First, different emotions can act on the same facial muscles, such as glaring and raising the chin, which are often spotted for both disgust and fear emotions. Second, unlike posed expressions, spontaneous emotions display may vary from one person to another, which makes their annotation more difficult. Third, relying on self-reported data makes the dataset annotation subjective. Finally, the absence of a protocol to unify these procedures can be an obstacle to conduct deeper investigations to determine their influence on emotion detection.

All the above considered, we provide in the first section of this chapter an overview of the existing datasets for facial expression analysis by categorizing them as oriented to *macro*- and *micro-expression* recognition. Actually, most of the existing datasets fall into the first category, mainly because macro-expressions are easier to collect and detect than micro-expressions. In particular, we are not aware of works that have summarized, in a systematic way, the existing datasets for both macro- and micro-expression recognition. Due to the increasing number of datasets, and their different characteristics, we believe that this review can provide researchers with a useful guide for orienting the choice of the adequate datasets for training and testing their models. In fact, for both macro- and micro-expressions datasets, we have identified several features that characterize and make specific each dataset. The main distinction we used is between *spontaneous* and *posed* datasets, on the one hand, and datasets acquired *in-the-wild*, on the other hand. In fact, these result in different acquisition requirements, protocols and labeling that ultimately open the way for investigating different challenges. Thus, in this section, we will follow the above categorization in presenting the existing datasets, but we do not refer to posed datasets. The reason for this is that most of the posed datasets were collected in the early stage of the research on macro-expression recognition. Therefore, compared to the benchmarks used in the current literature, such datasets have a small size with the saturated performance shown by the state-of-the-art methods. We enclose eighty datasets, covering both publicly available and not publicly available ones in order to provide a comprehensive overview. We described each dataset, and categorized it based on particular characteristics such as number of subjects, age, frame per second, ethnicity and amount of data. Table [1.1](#) summarizes the proposed classification of macro- and micro-expression datasets.

Table 1.1: Proposed categorization of macro- and micro-expression datasets.

Category	Datasets
Macro-Expression Datasets	
Spontaneous	EB+, TAVER, RAVDESS, GFT, SEWA, BP4D+ (MMSE), BioVid Emo, 4D CCDB, MAHNOB Mimicry, OPEN-EmoRec-II, AVEC'14, BP4D-Spontaneous, DISFA, RECOLA, AVEC'13, CCDB, DynEmo, DEAP, SEMAINE, MAHNOB-HCI, UNBC-McMaster, CAM3D, B3D(AC), CK+, AvID, AVIC, DD, SAL, HUMAINE, EmoTABOO, ENTERFACE, UT-Dallas, RU-FACS, MIT, UA-UIUC, AAI, Smile dataset, iSAFE, ISED
In-the-wild	RAF-DB, Aff-Wild2, AM-FED+, AffectNet, AFEW-VA, Aff-Wild, EmotioNet, FER-Wild, Vinereactor, CHEAVD, HAPPEI, AM-FED, FER-2013, AFEW, Belfast induced, SFEW, VAM-faces, FreeTalk, EmoTV, BAUM-2
Micro-Expression Datasets	
Spontaneous	SAMM, CAS(ME)2, Silesian deception, CASME II, CASME, SMIC-E, SMIC, Canal9, YorkDDT
In-the-wild	MEVIEW

1.3 Facial Macro- and Micro-Expressions

Facial emotion recognition is a critical field of study that focuses on identifying and interpreting human emotions through the analysis of facial expressions. By leveraging advanced algorithms and machine learning techniques, FER systems aim to recognize a variety of emotions such as happiness, sadness, anger, fear, surprise, and disgust. The ability to accurately detect these emotions has wide-ranging applications, like human-computer interaction, psychological research, security, and entertainment. As FER technology continues to evolve, it is increasingly capable of analyzing not only static images but also dynamic facial expressions captured in real-time videos, making it a powerful tool for understanding human emotional states in diverse and complex environments.

1.3.1 Macro-Expressions

Macro-expressions are the most commonly recognized facial expressions, typically associated with the basic emotions of happiness, sadness, anger, surprise, fear, and disgust. These expressions are characterized by noticeable facial movements and are usually visible to the naked eye, lasting between half a second and four seconds. Because they are deliberate and often consciously controlled, macro-expressions provide clear and easily interpretable cues about a person’s emotional state. In the context of FER, macro-expressions have been extensively studied due to their prominent and prolonged nature, which allows for more straightforward analysis and classification compared to more subtle facial cues. Understanding macro-expressions is foundational to the development of FER systems, as they represent the initial step in recognizing emotions based on visual data.

1.3.1.1 Macro-Expression Datasets

A macro-expression dataset is intended as a collection of images or videos of subjects that exhibit a facial expression as a consequence of an emotional state. There are also collections of static and dynamic three-dimensional (3D) scans of the face that capture the same range of emotions as for the 2D counterparts. A common trend that can be observed in these datasets is that of capturing facial expressions that go one step further than the strict categorization provided by the Ekman’s theory [36, 34]. In fact, while Ekman proposed the expression categorization into six universal categories (i.e., angry, disgust, fear, happy, sad and surprise) there is now the conviction that, despite this basic categorization being useful for a high-level view, it is too schematic to span the broad spectrum of human facial expressions. Therefore, other insights have made their way, with the *circumplex* model [139] being one of the most impactful. Furthermore, in addition to the type of the collected data (either images, videos or 3D scans), the capturing conditions and the expression model, the existing macro-expression datasets can be characterized according to several other features. In particular, we have identified the following features:

- *Number of subjects*: Existing datasets vary between four and thousands of subjects. The number of different individuals is particularly relevant for methods that need large quantities of data to learn models capable of generalizing to unseen identities;

- *Age*: Enrolled subjects vary from infants to young children and elderly people;
- *Frames per second (FPS)*: This can vary depending on the application context. For instance, to study the facial expression dynamics, a high FPS can help, whereas low FPS is often adopted for samples captured in real-life conditions;
- *Ethnicity*: Variability in terms of ethnic groups can be relevant and is typically a desired feature in collecting expression datasets;
- *Amount of data*: Number of images, videos or video frames.

Moreover, datasets are usually accompanied with annotations that are essential for training, testing and validating methods. These annotations are particularly relevant for videos where, depending on the fact the annotations are given at frame or video level, analysis at different granularity can be performed. This has a considerable impact depending on whether the datasets include posed, spontaneous or in-the-wild capturing, and on the expression model, either based on the six basic expressions or the circumplex model. In fact, while providing the six expression labels for posed and spontaneous datasets is an easy task, some more difficulties are experienced when the circumplex model is adopted. For in-the-wild capturing, ground-truth annotations are provided offline, and require experienced annotators. This introduces a lot of work from human annotators, which is costly and time-consuming. Sometimes, this human effort is alleviated by resorting to some form of *Mechanical Turk* that distributes the load to low-experienced and low-cost workers. However, the resulting annotations can show a diminished accuracy being originated by averaging annotations across several mechanical workers.

1.3.1.2 Spontaneous Datasets

In this section, we focus on spontaneous macro-expression datasets. These datasets are listed in Table [1.5](#).

EB+ (An expanded version of BP4D+): The EB+ [\[39\]](#) dataset is an expanded version of BP4D+ [\[185\]](#). It contains videos from a total of 200 subjects: 140 subjects from BP4D+, plus 60 additional subjects associated with five to eight tasks that involve inductions of varied emotions of a participant interacting with an experimenter. The

emotions are inducted when the participants interact with the experimenter. A certified FACS coders team annotated the dataset manually.

BP4D+ (Multimodal Spontaneous Emotion): Those tasks in EB+ are minutely explained in the BP4D+ or MultiModal Spontaneous Emotion (MMSE) dataset. This dataset is collected for human behavior analysis, and it illustrates 140 participants from different ethnic origins. The collected data included thermal (infrared) sensing, high-resolution 2D videos, high-resolution 3D dynamic imaging and contact physiological sensors that included respiration, heart rate, electrical conductivity of the skin and blood pressure. BP4D+ (see Figure 1.2) presents ten different emotion categories (happiness or amusement, surprise, sadness, startle or surprise, skeptical, embarrassment, fear or nervous, physical pain, angry and disgust) recorded per person according to the ten tasks that each person experienced. More specifically, these tasks include: listen to a funny joke, watch 3D avatar of participants, listen to 911 emergency phone calls, experience a sudden burst of sound, response to true or false question, improvise a silly song, dart game, submerge hands into ice water, complained for a poor performance and smell a smelly odor. BP4D+ has a larger scale and variability for images than BP4D Spontaneous [184].

- *Tasks* : ten tasks for spontaneous emotion elicitation and the activity in those tasks.
- *Emotion* : the target Emotion.
- *Data sequences & Metadata* : a variety of sensors of the face.

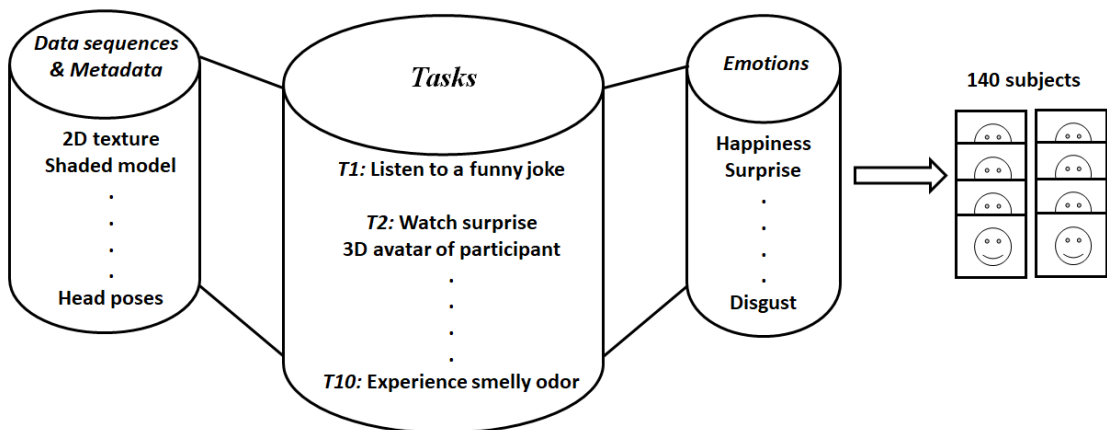


Figure 1.2: Structure of the BP4D+ dataset.

BP4D (Binghamton-Pittsburgh 3D DynAMic Spontaneous Facial Expression Data-base): BP4D Spontaneous [184] contains 41 participants from four different

ethnic origins (Asian, African-American, Hispanic, and Euro-American). It presents eight emotions (happiness or amusement, sadness, surprise or startle, embarrassment, fear or nervous, pain, anger or upset and disgust) derived through a combination of interviews, planned activities, film watching, cold pressor test, social challenge and olfactory stimulation. The facial expressions in the dataset had been annotated using the FACS.

iSAFE (Indian Semi-Acted Facial Expression Database): iSAFE [146] contains 44 volunteers from Indo-Aryan and Dravidian, 395 clips and seven emotions (happy, sad, fear, surprise, angry, neutral, disgust) captured with a camera behind a laptop, where the volunteers were asked to watch a few stimulant videos. The facial expressions were manually self-annotated by a user-interface portal and cross annotated by an annotator.

TAVER (Tri-modal Arousal-Valence Emotion Recognition database): TAVER [79] contains 17 subjects from one ethnic origin (Korean). The emotion (arousal–valence) was elicited through embarrassing and stressing people by asking them questions in a different language (English) than their own (Korean). The participants self-report feeling uncomfortable for the interviews with another language. Six human operators annotated the video sequence, with three annotators for each video sequence for more accuracy.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): The RAVDESS [86] dataset contains 24 participants from different ethnic groups (Caucasian, East-Asian, and Mixed (East-Asian Caucasian, and Black-Canadian First nations Caucasian)). The emotional elicitation in RAVDESS dataset is done through the true performance of emotion by actors. Actors were told to induce the desired state and provide genuine expressions of emotion.

GFT (Group Formation Task): GFT [49] contains 96 participants and 172,800 frames from a larger study on the impact of alcohol on group formation processes. In this study, participants affirmed that they could comfortably drink at least three drinks in 30 min. They were seated around a circular table in an observation room where they were asked to consume a beverage and to discuss any topics except their level of intoxication.

SEWA (Automatic Sentiment Analysis in the wild): SEWA [77] contains 398 participants of different nationality (British, German, Hungarian, Greek, Serbian, and Chinese), and 1990 audio-visual recording clips were collected during the experiment, comprised of 1600 min of audio-visual data of people’s reaction to adverts and 1057 min of video-chat recordings. To stimulate the emotions, the participants were asked to watch four advertisements, each being around 60 s long. These adverts had been chosen to elicit mental states including amusement, empathy, liking and boredom. In a second part, the participants were divided into pairs based on their cultural background, age and gender (for natural interaction, each pair was required to know each other personally in advance). After watching the fourth advertisement, the two participants were asked to discuss, for three minutes on average, the advertisement they had just watched. The subtle changes in the participant’s emotional state (valence, arousal, and liking/disliking) were annotated by human operators from the same cultural background of the recorded subjects (five for each). The SEWA dataset contains annotations for facial landmarks, acoustic low-level descriptors, hand gestures, head gestures, facial action units, verbal and vocal cues, continuously valued valence, arousal and liking/disliking, template behaviors, episodes of agreement/disagreement and mimicry episodes.

BioVid Emo (psychophysiological signals with video signals for discrete basic emotions): The BioVid Emo [180] dataset combines psycho-physiological signals with video signals for discrete basic emotions that were effectively elicited by film clips from 86 participants. The psycho-physiological signals that have been considered in this study are: skin conductance level, electrocardiogram, trapezius electromyogram and four video signals. Five discrete emotions (amusement, sadness, anger, disgust and fear) were elicited by 15 standardized film clips.

ISED (Indian Spontaneous Expression Database): ISED [60] contains 50 Indian subjects and 428 videos. Emotions were induced among the participants by using emotional videos and simultaneously their self-ratings were collected for each experienced emotion (sadness, surprise, happiness, and disgust).

4D CCDB (4D Cardiff Conversation Database): 4D CCDB [160] contains four

participants recording 17 conversations, which have been fully annotated for a speaker and listener activity: conversational facial expressions, head motion, and verbal/non-verbal utterances. The annotation tracks included were: front channel, backchannel, agree, disagree, utterance (verbal/non-verbal), happy (smile or laugh), interesting-backchannel, surprise-positive, surprise-negative, thinking, confusion, head nodding, head shake, head tilt and other.

MAHNOB Mimicry (The mahnob mimicry database: A database of naturalistic human interactions): MAHNOB Mimicry [12] contains 60 subjects from staff and students at Imperial College London (Europe or the Near-East). The subjects were recorded over 54 sessions of dyadic interactions between 12 confederates and their 48 counterparts, being engaged either in a socio-political discussion or negotiating a tenancy agreement.

OPEN-EmoRec-II (A Multimodal Corpus of Human-Computer Interaction): OPEN-EmoRec-II [137] has been designed in order to induce emotional responses in HCI users during two different parts of a HCI-experiment. It contains 30 subjects involving video, audio, physiology (SCL, respiration, BVP, EMG Corrugator supercilii, EMG Zygomaticus Major) and facial reaction annotations.

AVEC'14 (Audio-Visual Emotion recognition Challenge (AVEC 2014)): AVEC'14 [158] contains 84 German subjects with 300 audio-visualse. The challenge has two goals logically organized as sub-challenges: to predict the continuous values of the affective dimensions valence, arousal and dominance at each moment in time; and to predict the value of a single self-reported severity of depression indicator for each recording.

DISFA (A Spontaneous Facial Action Intensity Database): DISFA [96] contains 27 subjects from different ethnicities (Asian, Euro American, Hispanic, and African American) and 130,000 annotations. Participants viewed a four-minute video clip intended to elicit spontaneous Action Units (AUs) in response to videos intended to elicit a range of facial expressions of emotions.

RECOLA (REmote COLlaborative and Affective interactions = Multi-modal Corpus of Remote Collaborative and Affective Interactions (in French: RECOLA)): RECOLA [131] contains 46 subjects of different nationality (French, Italian, German and Portuguese). It is based on a study focusing on emotion perception during remote collaboration, where participants were asked to perform individual and group tasks.

AVEC'13 (Audio-Visual Emotion recognition Challenge (AVEC 2013)): AVEC'13 [159] contains 292 German subjects and 340 audio-visuals. Subjects performed a human-computer interaction task, while being recorded by a webcam and a microphone.

CCDb (Cardiff conversation database): The CCDb [7] 2D audiovisual dataset contains natural conversations between pairs of people. All 16 participants were fully fluent in the English language. It includes 30 audio-visuals.

DynEmo (Dynamic and spontaneous emotional facial expression database): The DynEmo [154] dataset contains 358 Caucasian participants filmed in natural but standardized conditions. The participants were enrolled into ten tasks to display a subjective affective state rated by both the expresser (self-reported after the emotion inducing tasks, using dimensionally, action readiness and emotional labels items) as well as the observers (continuous annotations).

DEAP (A Database for Emotion Analysis Using Physiological Signals): DEAP [73] contains 32 mostly European students and 40 videos. Participants watched music videos and rated them on a discrete nine-point scale for valence, arousal and dominance.

SEMAINE: SEMAINE [100] contains 24 undergraduate and postgraduate students between 22 and 60 years old. It consists of 130,695 frames of typical session duration for Solid SAL (Sensitive Artificial Listener) and semi-automatic SAL. In these sessions, participants were asked to change character when they got bored, annoyed or felt they had nothing more to say to the character.

MAHNOB-HCI (multimodal database for affect recognition and implicit tagging): MAHNOB-HCI [150] illustrates 27 participants from different educational backgrounds, from undergraduate students to postdoctoral fellows, with different English proficiency from intermediate to native speakers. Participants were shown fragments of movies and pictures, while monitoring them with six video cameras, a head-worn microphone, an eye gaze tracker, as well as physiological sensors measuring ECG, EEG (32 channels), respiration amplitude, and skin temperature.

UNBC-McMaster (McMaster University and University of Northern British Columbia (UNBC)–Painful data: The UNBC-McMaster shoulder pain expression archive database): The UNBC-McMaster (UNBC Shoulder Pain Archive (SP)) [90] dataset contains 25 participants who were self-identified as having a problem with shoulder pain. It contains physical pain/temporal expressions/spontaneous facial expressions relating to genuine pain, while discriminating 48,398 frames/200 video sequences.

CAM3D (3D corpus of spontaneous complex mental states): CAM3D [93] contains 16 participants from different ethnic backgrounds (Caucasian, Asian and Middle Eastern). It involves 108 videos, where the use of hand-over-face gestures as a novel affects cues for automatic inference of cognitive mental states.

B3D(AC) (A 3-D Audio-Visual Corpus of Affective Communication): The B3D(AC) [40] audio-visual corpus dataset contains 14 participants native English speakers and 1109 sequences. The annotation of the speech signal includes: transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals.

CK+ (Extended Cohn-Kanade Dataset): CK+ [89] contains 593 sequences, where the 123 participants have performed series of 23 facial displays. It involves seven emotion categories.

AvID (Audiovisual speaker identification and emotion detection for secure communications): AvID [46] contains 15 subjects, recorded while they describe neutral photographs, play a game of Tetris, describe the game of Tetris and solve cognitive tasks. A one-hour video is captured for each subject, discriminating four class emotions (neutral, relaxed, moderately aroused and highly aroused).

AVIC (Audiovisual Interest Corpus): AVIC [144] contains 21 participants from Asian and European ethnic groups, while involving 324 episodes that consist of spontaneous as well as conversational speech demonstrating "theoretical" effectiveness of the approach.

DD (Detecting depression from facial actions and vocal prosody): The DD dataset [22] illustrates 57 participants from a clinical trial for treatment of depression. Trials were conducted using the Hamilton Rating Scale for Depression (HRS-D), which is a criterion measure for assessing the severity of depression. Participant facial behavior was registered in response to the first three of 17 questions in the HRS-D, such that the questions concerned core features of depression: depressed mood, guilt, and suicidal thoughts.

SAL (The Sensitive Artificial Listener): The SAL [30] dataset is based on the observation that it is possible for two people to have a conversation in which one pays little or no attention to the meaning of what the other says and chooses responses on the basis of superficial cues. SAL provides a context in which sustained emotionally colored human-machine interaction seems to be achievable. It identifies the four users' emotional state itself during sessions of 30 min for each user, using evidence from faces, upper body, voice, and key words. The range of emotions is wide, but they are not very intense.

HUMAINE (The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data): HUMAINE [31] contains 50 clips selected to cover material showing emotion in action and interaction spanning a broad emotional space (positive and negative, active and passive), selected from the following corpora: the Belfast Naturalistic dataset (in English, naturalistic, ten clips),

the Castaway Reality Television dataset (in English, naturalistic, ten clips), Sensitive Artificial Listener (in English, induced, 12 clips), Sensitive Artificial Listener (in Hebrew, induced, one clip), Activity/Spaghetti dataset (in English, induced, seven clips), Green Persuasive dataset (in English, induced, four clips), EmoTABOO (in French, induced, two clips), DRIVAWORK corpus (in German, induced, one clip), and GEMEP corpus (in French, acted, one clip).

EmoTABOO (Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics: consisting in letting pairs of people play the game “Taboo”): EmoTABOO [178] is a French dataset containing ten audiovisual clips collected during game playing. People were playing at Taboo, a game in which one person has to explain to the other using gestures and body movement a ‘taboo’ concept or word. It involves multimodal interactions between two people and provides an emotional content, with a range of emotions including embarrassment, amusement, etc.

ENTERFACE: ENTERFACE [142] includes acquisitions for three multimodal emotion detection modalities: the first modality is given by brain signals via fNIRS and contains 16 participants; the second modality includes face videos of five participants; and the third modality captures the scalp EEG signals of 16 participants. EEG and fNIRS provided an “internal” look at the emotion generation processes, while video sequences gave an “external” look on the “same” phenomenon.

UT-Dallas (University of Texas at Dallas): UT-Dallas [116] contains 1540 video clips of 284 people of Caucasian descent walking and conversing. During filming, the subject watched a ten-minute video, which contained scenes from various movies and television programs intended to elicit different emotions in order to capture emotions such as happiness, sadness and disgust.

RU-FACS (Rochester/UCSD Facial Action Coding System): RU-FACS [9] contains 100 subjects that attempted to convince an interviewer he or she is telling the truth. Interviewers were current and former members of the police and FBI.

MIT (The MIT Media Laboratory, Cambridge MA, USA): MIT [61] contains over 25,000 frames scored of 17 drivers that gave their consent to having video and the physiological signals recorded during the drive.

UA-UIUC (University of Illinois at Urbana-Champaign): UA-UIUC [145] contains 28 subjects and one video clip for each subject. First, the subjects could not know that they were being tested for their emotional state. Second, subjects were interviewed after the test to find out their true emotional state for each expression.

AAI (Adult Attachment Interview): The AAI [134] dataset contains 60 subjects from different ethnic groups (European American and Chinese American). The subjects were interviewed and asked to describe the childhood experience. It contains one audio-visual for each subject.

Smile dataset (Dynamics Of Facial Expression: Normative Characteristics And Individual Differences): The Smile dataset [143] contains 195 spontaneous smiles of 95 subjects. Videos were collected throughout a session that included baselines (seated with eyes open) and viewing of film clips.

Overall, the investigated datasets including spontaneous macro-expressions are the majority with 39 instances. The number of subjects included in such datasets ranges from less than 50 to more than 500. The typical number of subjects is not related with other features, like age range or ethnic diversity or even the amount of data. For instance, the TAVER dataset includes 17 subjects, with an age range between 21 and 38 years and only one ethnicity (Korean); the DISFA dataset comprises 27 subjects with an age ranging between 18 and 50 years and four ethnicities (Asian, Euro American, Hispanic, and African American). A large number of subjects does not necessarily correspond to more diversity. For example, the DynEmo dataset with 358 subjects has an age that ranges between 25 and 65 years, and only one ethnicity (Caucasian). That being said, the SEWA dataset with 398 subjects, has an age ranging between 18 and 65 years, and six ethnicities (British, German, Hungarian, Greek, Serbian, and Chinese), and it

contains annotations for facial landmarks, acoustic low-level descriptors, hand gestures, head gestures, facial action units, verbal and vocal cues, continuously valued valence, arousal and liking/disliking (toward an advertisement), template behaviors, episodes of agreement/disagreement and mimicry episodes. Finally, each dataset handles a different class of emotions, the six basic emotions and neutral (iSAFE) or the six basic emotions and embarrassment and pain (BP4D-Spontaneous), four emotions (ISED) or even one emotion (smile dataset). Some other datasets represent emotions in form of valence and arousal (DEAP, AVEC'14).

1.3.1.3 Spontaneous and Posed Datasets

We consider herein the spontaneous and the posed datasets due to the fact that we are interested in the spontaneous part of it.

4DFAB (4D Facial Expression Database for Biometric Applications): The 4DFAB [20] dataset includes six posed expressions, spontaneous expressions (anger, disgust, fear, happiness, sadness and surprise), and nine words utterances (puppy, baby, mushroom, password, ice cream, bubble, Cardiff, bob, rope). It contains recordings of 180 subjects captured in four different sessions spanning over five years. This dataset encloses 4D videos of subjects displaying both spontaneous and posed facial behaviors.

BAUM-1 (Bahcesehir University Multimodal Affective Database-1): BAUM-1 [179] contains 31 Turkish subjects and 1,184 multimodal facial video clips. The expressed emotional and mental states consist of happiness, anger, sadness, disgust, fear, surprise, boredom, contempt, confusion, neutral, thinking, concentrating, and bothered.

MAHNOB Laughter (The MAHNOB Laughter database): MAHNOB Laughter [125] contains 22 subjects from 12 different countries and of different origins recorded in 180 sessions. There are 563 laughter episodes, 849 speech utterances, 51 posed laughs, 67 speech–laughs episodes and 167 other vocalizations annotated in the dataset.

PICS-Stirling ESRC 3D Face (Psychological Image Collection at Stirling-ESRC project 3D Face Database): PICS-Stirling ESRC 3D Face [1] contains 99

subjects, a number of 2D images, video sequences as well as 3D face scans. Seven different expression variations were captured.

Hi4D-ADSIP (High Resolution 4 Dimensional Database from the Applied Digital Signal and Image Processing Research Centre): Hi4D-ADSIP [95] contains 80 subjects from undergraduate students from the Performing Arts Department at the University as well as undergraduate students, postgraduate students and members of staff from other departments. It involves 3,360 images/sequences and consists of seven basic facial expressions and further seven facial articulations.

USTC-NVIE (University of Science and Technology of China (USTC)-Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference): USTC-NVIE [165] contains 215 subjects, 236 images and six basic expressions. Two kind of facial expressions were recorded: spontaneous expressions induced by the film clips and posed ones obtained by asking the subjects to perform some series of expressions in front of the cameras.

MMI-V (Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database): MMI-V [157] contains 25 subjects from different ethnic groups (European, South American, and Asian) recorded in one hour and 32 min of data and 392 segments. Part IV of the dataset was annotated for the six basic emotions and facial muscle actions. Part V of the dataset was annotated for voiced and unvoiced laughter. There are Part IV and Part V because MMI-V dataset was added to the MMI [122] facial expression dataset.

MMI (The acronym MMI comes from M&M Initiative where the Ms are the initials of the two main authors. Although other colleagues joined the development efforts of the main authors, the acronym remained in use): The MMI dataset contains 19 subjects from different ethnic groups (European, Asian, or South American), 740 static images sequence of frontal and side view and 848 videos.

AVLC (The AVLaughterCycle Database): AVLC [156] contains 24 subjects

from different nationality and ethnic groups (Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA, and South Korea) and 1000 spontaneous laughs elicited by a funny movie and 27 acted laughs.

IEMOCAP (The Interactive Emotional Dyadic Motion Capture): IEMOCAP [15] contains 120 actors (fluent English speakers) recorded in 12 h of audiovisual data, including video, speech, motion capture of faces and text transcriptions. The actors performed selected emotional scripts and also improvised spontaneous spoken communication scenarios to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state).

AMI (Augmented Multi-party Interaction): The AMI [19] dataset contains a multi-modal set of data consisting of 100 h of meeting recordings, where some of them are naturally occurring, and some others are elicited. In this latter case, a particular scenario is used where the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day.

Although we did not discuss posed expressions, we included spontaneous and posed macro-expressions in this thesis with 11 datasets. In these categories, the 4DFAB dataset presents an interesting age range that covers infants and elders from 5 to 75 years. Furthermore, the USTC-NVIE dataset presents the highest number of subjects with 215 students. Although MAHNOB Laughter dataset contains an important ethnicity variation (12 different countries and of different origins), its average age is between 27 and 28 years.

1.3.1.4 In-the-Wild Datasets

For in-the-wild datasets, the human–human interaction results in a spontaneous expression, so that the emotional content and the experimental conditions are uncontrolled.

RAF-DB (Real-world Affective Faces Database): RAF-DB [81] includes thousands of subjects with 30,000 facial images collected from Flickr.

Aff-Wild2 (Extending the Aff-Wild Database for Affect Recognition): The Aff-Wild2 dataset contains videos downloaded from YouTube with 258 subjects from infants and young children to elderly people [74]. It illustrates various ethnicity groups (Caucasian, Hispanic or Latino, Asian, black, and African American), different professions (e.g., actors, athletes, politicians, journalists); as well as changes in head pose, illumination conditions, occlusions and emotions.

AM-FED+ (An Extended Dataset Affectiva-MIT Facial Expression Dataset): In the AM-FED+ [98] dataset, 416 participants from around the world (their locations are not known) were recruited to watch video advertisements. It contains 1044 videos of naturalistic facial responses to online media content recorded over the Internet.

AffectNet (Affect from the InterNet): AffectNet [104] contains more than 1,000,000 facial images from the Internet of more than 450,000 participants, presenting valence and arousal in eight emotion categories.

AFEW-VA (Database for valence and arousal estimation in-the-wild): The AFEW-VA dataset [76] contains 240 movie actors in a range of age between 8 and 76 years and 600 video clips.

Aff-Wild (Affectiva-MIT Facial Expression Dataset): Within the Aff-Wild dataset [177], more than 500 videos were collected from YouTube, while capturing subjects displaying a number of spontaneous emotions. The data were tagged using emotion-related keywords such as feeling, anger, hysteria, sorrow, fear, pain, surprise, joy, sadness, disgust, love, wrath, contempt, etc.

EmotioNet (Annotating a million face images in the wild): EmotioNet [11] contains one million images of facial expressions downloaded from the Internet, categorized within one of the 23 basic and compound emotion categories. The images have been annotated either with emotion category or with corresponding AUs.

FER-Wild (Facial Expression Recognition from World Wild Web): FER-

Wild [106] contains 24,000 Web images from Google, Bing, Yahoo, Baidu and Yandex. These image were categorized in nine categories (no-face, six basic expressions: happy, sad, surprise, fear, disgust, anger, neutral, none, and uncertain). The 'no-face' category is defined in the following cases: there is no face in the image, there is a watermark on the face, the bounding box was not on the face or did not cover the majority of the face, the face is a drawing, animation, painted, or printed on something else, and the face is distorted beyond a natural or normal shape. The 'no-face' is defined even if an expression could be inferred. The 'none' category is defined when the images do not present the six basic emotions or neutral (such as sleepy, bored, tired, etc.). The 'uncertain' category is defined when the annotators are unsure of the facial expressions.

Vinereactor (Reactions for vine videos): Vinereactor [72] contains 222 mechanical tuckers works filmed with a webcam watching 200 random vine videos from the comedy vine.co channel to get their reactions.

CHEAVD (Chinese Natural Emotional Audio–Visual Database): CHEAVD [84] is extracted from 34 films, two TV series and four other television shows presenting 26 non-prototypical emotional states, including the six basic ones, from 238 speakers.

HAPPEI (HAPpy PEople Images): HAPPEI [27] contains 4,886 images downloaded from Flickr of 8500 faces, manually annotated by four human labelers. The emotions have been categorized according to group level happiness intensity (neutral, small smile, large smile, small laugh, large laugh and thrilled).

AM-FED (Affectiva-MIT Facial Expression Dataset): AM-FED [99] contains 242 facial videos of 242 webcam videos recorded in real world conditions of viewers, from a range of ages and ethnicities, while watching films. It is labeled for frame-by-frame labels for the presence of ten symmetrical FACS action units, four asymmetric (unilateral) FACS action units, two head movements, smile, general expressiveness, feature tracker fails and gender.

FER-2013 (Facial Expression Recognition 2013 dataset): FER-2013 [50] con-

tains 35,685 facial expressions from images queried from the web. Images were categorized based on the emotion shown in the facial expressions (happiness, neutral, sadness, anger, surprise, disgust, fear).

SFEW (Static Facial Expressions in the Wild): SFEW [28] is an extracted dataset (by selecting frames) from the AFEW [29] dataset.

AFEW (Acted Facial Expressions in the Wild): AFEW contains 330 subjects from fifty-four movie DVDs, 1426 sequences, seven emotions (anger, disgust, fear, sadness, happiness, neutral, and surprise) and 1747 numbers of expressions.

Belfast induced (Belfast Natural Induced Emotion Database): The Belfast induced dataset [149] is divided into three tasks: Set 1 tasks contains 114 subjects from undergraduate students and encloses 570 audio-visuels. It is developed as stimuli for research into the individual differences that might influence human abilities to encode and decode emotional signals. Set 2 tasks contains 82 subjects from undergraduate students and postgraduate students or employed professionals, and encloses 650 audio-visuels. It is developed to allow comparison of these new tasks with more traditional film elicitors that had previously been validated for their ability to induce discrete emotions. Set 3 tasks contains 60 subjects from three different ethnic groups (Peru, Northern Ireland) and encloses 180 audio-visuels. It contains variants of the disgust and fear (both active/social) tasks and the amusement (passive/social) task from Set 1. The emotions were self reported by the participants.

VAM-faces ("Vera Am Mittag"—German TV talk show): The VAM-faces [53] dataset consists of 12 h of audio-visual recordings of the German TV talk show "Vera am Mittag", which were segmented into broadcasts, dialogue acts and utterances. It contains 20 speakers and a set of 1,867 images (93.6 images per speaker on average).

FreeTalk (Tools and Resources for Visualising Conversational-Speech Interaction): The FreeTalk [18] dataset contains four subjects from different countries having a conversation in English. It consists of two 90-minute multiparty conversations,

and the naturalness of the dialogues is further indicated by the topics of the conversation.

EmoTV (emotional video corpus: TV interviews (monologue)): The EmoTV [2] dataset is a corpus of 51 video clips recorded from French TV channels containing interviews. It contains 48 subjects interviewed with a range of 14 emotions classes (anger, despair, doubt, disgust, exaltation, fear, irritation, joy, neutral, pain, sadness, serenity, surprise, and worry).

BAUM-2 (a multilingual audio-visual affective face database): BAUM-2 [37] contains 286 subjects from 122 movies and TV-series result 1,047 video clips in two languages (Turkish, English). It involves eight emotions (anger, happiness, sadness, disgust, surprise, fear, contempt and neutral). The dataset also provides a set of annotations such as subject age, approximate head pose, emotion labels and intensity scores of emotions.

Overall, the twenty investigated in-the-wild macro-expressions datasets have the highest number of subjects, reaching thousands of subjects in the RAF-DB dataset, the highest diversity of emotions with 23 categories of emotions in EmotioNet, the maximum number of subjects with participants from around the world in the AM-FED+ dataset.

1.3.1.5 Other Categorizations of Macro-Expression Datasets

In the following, we propose other categorizations for the spontaneous and in-the-wild datasets. One way is that of considering the different ways the data have been collected:

- In *spontaneous* datasets, unlike posed datasets, where participants are asked to perform an emotion, subjects' emotions are stimulated. For example, in [146], face expressions were captured when volunteers were asked to watch a few stimulant videos. In a similar way, in [150], participants were shown fragments of movies and pictures. In [60], emotional videos were used for each emotion, and in the dataset investigated in [184], combined interviews, planned activities, film watching, cold pressor, test/social challenge and Olfactory stimulation were explored. In [100], participants were told to change character when they got bored, annoyed or felt

they had nothing more to say to the character. The dataset proposed in [144] collected conversational speech, and the work in [30] had been based on a conversation between two people in which one pays little or no attention to the meaning of what the other says and chooses responses on the basis of superficial cues. In [22], participants were from a clinical trial for treatment of depression, however, in [86], the participant has a dialogue script with vignettes for each emotional category. In [159], subjects had performed a human-computer interaction task, similarly to the work of [7], where natural conversations between pairs of people were investigated. In [134], subjects were interviewed and asked to describe the childhood experience, and in [9], subjects tried to convince the interviewers they were telling the truth. In [46], subjects had described neutral photographs, played a game of Tetris, described the game of Tetris and solved cognitive tasks. Differently, in [61], a driver was recorded during the drive, and [31] presented an interaction from TV chat shows and religious programs and discussions between old acquaintances. In [178], participants were playing a game in which one person has to explain to the other using gestures and body movement a ‘taboo’ concept or word.

- Within the framework of *in-the-wild* datasets, the collected data come from movies [37, 29], films, TV plays, interviews and talk shows [2, 84, 53], videos downloaded from Youtube [74], images and videos from the Internet [104, 11, 68] as well as from Flickr [81, 27].

Most of the datasets have classified emotions into the six basic categories (angry, disgust, fear, happy, sad, surprise) [122, 165, 157, 28], with some datasets adding the neutral one [146, 29, 50]. There are also datasets that further extended the basic six plus neutral expression model with one additional expression, like pain [95], or contempt [37]. Other datasets added more expressions, like happiness or amusement / sadness / surprise or startle / embarrassment / fear or nervous / pain / anger or upset / disgust [184]. Actually, a variety of expressions can be found in the existing datasets over those indicated above. For example, there are twenty-three categories of emotion in [11] according to [32]; nine categories (no-face, six basic expressions, neutral, none, and uncertain) in [106]; thirteen emotional and mental states are included in [179], where the six basic emotions plus boredom and contempt are complemented with some mental states (i.e., confusion, neutral, thinking, concentrating, and bothered); four emotions (sadness, surprise, happiness, and

disgust) are given in [60]; with only one emotion (smile) being included in [99, 143]. The Valence and Arousal expression model was instead followed in [104, 73, 177, 158]. We note some datasets that also included AU annotations. For instance, the EmotioNet [11] and DISFA [96] datasets have 12 AUs annotations, and in the CASME [176] dataset, AUs are coded by two coders based on Ekman’s study. Table 1.2 groups the datasets according to the different ways emotions are categorized.

It is worth mentioning that some datasets contain 3D scans of expressive faces. For example, 4DFAB [20] contains 3D faces (over 1,800,000 3D meshes), and PICS-Stirling ESRC 3D Face Database [1] presents 3D face scans along with 2D images and video sequences. Likewise, CAM3D [93] is a 3D multimodal corpus dataset, and B3D(Ac) [40] dataset presents facial expressions in dynamic 3-D face geometries. Likewise, BP4D+ [185] contains high-resolution 3D dynamic imaging with a variety of sensors of the face, 4D CCDB [160] is a 4D (3D Video) audio-visual dataset, BP4D-Spontaneous [184] is a 3D video dataset of spontaneous facial expressions, and Hi4D-ADSIP [95] presents a comprehensive 3D dynamic facial articulation dataset.

In what follows, we propose some other categorizations for macro-expression datasets:

- **Number of subjects:** Table 1.3 presents a classification of macro-expression datasets according to the number of subjects. Most of the datasets contain less than 50 subjects, with just few datasets containing more than 500 subjects. The number of subjects can reach more than thousands, if the expressions are spontaneous or in-the-wild.
- **Age variation:** There are many age ranges in macro-expression datasets. Most of the datasets include subjects in a relatively small range (from 18 to 30 years), namely TAVER, RAVDESS, GFT, MAHNOB Mimicry, BP4D-Spontaneous, MAHNOB Laughter, DEAP, USTC-NVIE, MMI-V, AvID, AVIC, ENTERFACE, UT-Dallas, RU-FACS, UA-UIUC, AAI, iSAFE, and ISED. Some other datasets have a moderate range (18-60), including EB+, SEWA, BP4D+ (MMSE), BAUM-1, BioVid Emo, 4D CCDB, AVEC’14, DISFA, AVEC’13 AViD-Corpus, CCDB, DynEmo, SEMAINE, MAHNOB-HCI, Hi4D-ADSIP, CAM3D, B3D(AC), CK+, VAM-faces, and MM. Few datasets contain children, including CHEAVD, 4DFAB, BAUM-2, AFEW-VA, AFEW, and Aff-Wild2. However, child facial expressions were mixed within adult expression samples without differentiating them based on age. On the

Table 1.2: Classification of macro-expression datasets according to their content.

Expression Representation	Macro-Expression Datasets
Six basic expressions	MMI, USTC-NVIE, MMI-V, SFEW
Six basic expressions + neutral	iSAFE, AFEW, FER-2013
Six basic expressions + neutral, pain	Hi4D-ADSIP
Six basic expressions + neutral, contempt	BAUM-2
Six basic expressions (happiness or amusement, sadness, surprise or startle, fear or nervous, anger or upset, disgust) + embarrassment, pain	BP4D-Spontaneous
23 categories of emotion	EmotioNet
Nine categories of emotions (no-face, six basic expressions, neutral, none, and uncertain)	FER-Wild
13 emotional and mental states (six basic emotions plus boredom and contempt plus mental states, confusion, neutral, thinking, concentrating, and bothered)	BAUM-1
Four emotions (sadness, surprise, happiness, and disgust)	ISED
One emotion (smile)	AM-FED, Smile dataset
Valence-arousal	AffectNet, DEAP, Aff-Wild, AVEC'14

other hand, in the CHEAVD dataset, the participants were divided into six groups of ages, and in the 4DFAB dataset, the age distribution includes five categories, with infants being in the 5-18 category. However, the datasets did not take into

consideration the difference of the facial expressions according to the age.

- **Frame per second (FPS):** In macro-expression analysis, the number of FPS is relevant depending on the application context. In the following datasets, the number of FPS is smaller or equal to 20: TAVER, AM-FED+, and AM-FED. Instead, the number of FPS is greater than 50 for the 4DFAB, 4D CCDB, MAHNOB-HCI, Hi4D-ADSIP, FreeTalk, iSAFE, and ISED datasets. The largest number of FPS, equal to 120, is reached in the IEMOCAP dataset, which makes it a relevant source for studying macro expressions.
- **Ethnicity:** The existing macro-expression datasets contain various ethnicities such as Latino (EB+, 4DFAB, Aff-Wild2, BP4D+, RU-FACS), Hispanic (EB+, 4DFAB, Aff-Wild2, BP4D+, BP4D-Spontaneous, DISFA), White (EB+, BP4D+), African (EB+, Aff-Wild2, BP4D+, BP4D-Spontaneous, DISFA), Asian (EB+, 4DFAB, Aff-Wild2, BP4D+, BP4D-Spontaneous, DISFA, CAM3D, MMI-V, AVIC, MMI, RU-FACS, iSAFE), and Caucasian (4DFAB, Aff-Wild2, RAVDESS, DynEmo, CAM3D, UT-Dallas). However, some datasets contain participants from around the world or randomly selected (RAF-DB, AM-FED+, GFT, AffectNet, AFEW-VA, EmotioNet, AM-FED, AFEW, FreeTalk).
- **Amount of data:** Here, the main distinction is between datasets that include images; like EB+, TAVER, Aff-Wild2, AM-FED+, AFEW-VA, SEWA, Aff-Wild, BAUM-1, BioVid Emo, Vinereactor, CHEAVD, 4D CCDB, OPEN-EmoRec-II, AVEC'14, RECOLA, AM-FED, AVEC'13, CCDB, DynEmo, DEAP, AFEW, Belfast induced, MAHNOB-HCI, UNBC-McMaster, CAM3D, B3D(AC), UT-Dallas, EmoTV, UA-UIUC, and AAI; and those that comprise videos; like RAF-DB, AffectNet, EmotioNet, FER-Wild, HAPPEI, FER-2013, SFEW, USTC-NVIE, iSAFE, and ISED.

Up to this point, we have described and discussed characteristics of macro-expressions related datasets. Research on macro-expression recognition has evolved significantly in the last few years, while reaching saturated performance on lab-controlled, small-sized datasets, and the significant advancement of recognition methods call for new challenges. However, the number of datasets in-the-wild is still limited compared to spontaneous datasets. Indeed, most of the spontaneous datasets contain few subjects unlike in-the-wild ones which contain many more subjects that can reach thousands as in the RAF-DB

Table 1.3: Classification of macro-expression datasets according to the number of subjects.

Number of Subjects	Macro-Expression Datasets
≤ 50	TAVER, RAVDESS, BAUM-1, OPEN-EmoRec-II, BP4D-Spontaneous, DISFA, RECOLA, CCDB, MAHNOB Laughter, DEAP, SEMAINE, MAHNOB-HCI, UNBC-McMaster, CAM3D, B3D(AC), MMI-V, AVLIC, AvID, AVIC, VAM-faces, ENTERFACE, MMI, MIT, EmoTV, UA-UIUC, 4D CCDB, FreeTalk, IEMOCAP, SAL, iSAFE, ISED
$\in [50, 100]$	GFT, SEWA, BioVid Emo, MAHNOB Mimicry, AVEC'14, PICS-Stirling ESRC 3D Face Database, Belfast induced (Set2 and Set3), Hi4D-ADSIP, DD, RU-FACS, AAI, Smile dataset
$\in [100, 250]$	EB+, 4DFAB, AFEW-VA, BP4D+ (MMSE), Vinereactor, CHEAVD, AM-FED, Belfast induced (Set1), USTC-NVIE, CK+
$\in [250, 500]$	SFEW, Aff-Wild2, AM-FED+, BAUM-2, AVEC'13 AViD-Corpus, DynEmo, AFEW, UT-Dallas
≥ 500	RAF-DB, AffectNet, Aff-Wild, EmotioNet, FER-Wild, FER-2013, HAPPEI, HUMAINE

dataset [81]. The variation in ethnicity in spontaneous datasets ranges between one and six different ethnic groups in each dataset, while the captured subjects in in-the-wild datasets are from around the world. The ethnic element is important because the larger the diversity, the more interesting the dataset can be, and this is due to the fact that there are differences in facial expression depending on ethnicity [102]. For instance, in [69], authors have found “a lower mean recognition accuracy of Caucasian faces among African and Japanese subjects than among Europeans and Americans subjects.” The age ranges between infants and elderly, however, few datasets contain children; e.g., CHEAVD, BAUM-2, AFEW-VA, AFEW, Aff-Wild2, and few datasets contain elders,

e.g., EB+, 4DFAB, Aff-Wild2, BAUM-2, and BP4D+(MMSE); the rest of the datasets have an average of 20-30 years. The age variance is important due to the fact that child as elders' facial expressions can actually be different compared to adult expressions. It could be interesting to have datasets with a larger number of subjects in order to have a wider diversity in the ethnicity and age range. This would also help in including more diversity in the way expressions are performed. More in general, increasing the number of emotion categories, going beyond the six basic emotions, is a further direction for the facial expression datasets in the next years.

1.3.2 Micro-Expressions

Micro-Expressions are brief, involuntary facial expressions that typically last only a fraction of a second, often occurring when an individual is attempting to conceal or suppress an emotion. Unlike macro-expressions, which are more easily observable, micro-expressions are subtle and fleeting, making them challenging to detect and analyze. Despite their short duration, micro-expressions can reveal a person's true emotions, even when they are trying to mask them, which makes them a critical area of study in fields such as psychology, security, and interpersonal communication. In facial emotion recognition, accurately capturing and interpreting micro-expressions requires sophisticated algorithms and high-resolution and high-frame-rate video analysis techniques. The study of micro-expressions holds significant potential for applications where understanding genuine emotional responses is crucial, such as in lie detection, negotiation scenarios, and therapeutic settings. As research in this area progresses, the ability to reliably recognize micro-expressions will enhance the sensitivity and accuracy of emotion detection systems.

1.3.2.1 Micro-Expression Datasets

Micro-expressions are shown as the result of an emotional response that activates, in an innate way, both voluntary and involuntary expressions of the face that conflict one with the other. As a result, the individual shows the true expression just for a very short time interval, which is then followed by a false expressive reaction [152]. Overall, studies have shown that this occurs when a part of our brain (i.e., the amygdala) responds to the emotional stimuli experienced by an individual in an appropriate way by showing a facial expression, but then the individual consciously decides to cancel that expression/emotion.

In fact, while macro-expressions lasts from 0.5 to 4 s [152], a micro-expression normally has a duration of less than half of a second [67]. Due to this very short duration, and differently from macro-expressions, micro-expressions cannot be controlled and so they are very difficult to hide. The fact that micro-expressions are expressive reactions that last just some fractions of seconds implies that they are best captured by high-speed cameras [128]. According to the works of Ekman, micro-expressions can be categorized into the seven universal emotions: disgust, anger, fear, sadness, happiness, contempt, and surprise. Ekman himself subsequently expanded this list including a range of positive and negative emotions; namely, amusement, embarrassment, anxiety, contentment, guilt, pleasure, pride, relief, and shame (not all of them are encoded by facial muscles). Herein, we summarize the existing datasets for micro-expression analysis while following a similar organization as the one that we proposed for macro-expression ones. We start by describing the spontaneous datasets, then we discuss the in-the-wild ones. We also put in other evidence relevant features for these datasets. Indeed, comparing the abundance of macro-expression datasets with the limited number of micro-expression ones, it evidently comes the unbalanced proportion, suggesting that the research on micro-expression recognition is less developed than that for macro-expressions.

1.3.2.2 Spontaneous Datasets

We have identified nine datasets in the literature that have been used with a certain frequency for the analysis of micro-expressions. They have been acquired with quite heterogeneous devices and proposing different evaluation protocols.

SAMM (Spontaneous Micro-facial Movement): The SAMM [25] dataset contains 32 participants from 13 different ethnic groups and 159 samples with seven emotions (contempt, disgust, fear, anger, sadness, happiness and surprise).

CAS(ME)2 (Chinese Academy of Sciences Micro-expression–A Database for Spontaneous Macro-expression and Micro-expression Spotting and Recognition): This dataset was proposed in [101]. It contains 22 subjects and 53 samples with four emotions (positive, negative, surprise, and others).

Silesian deception dataset: The Silesian deception dataset [130] includes 101 students of the third year and the fourth year at the Faculty of Automatic Control, Electronics and Computer Science. This comprises 101 samples and frame annotations, including eye closures (eye movements), gaze aversion and micro-tensions.

CASME II (Improved Spontaneous Micro-Expression): CASME II [175] contains 247 samples selected from 3000 elicited facial movements, from 26 participants presenting five classes of emotions (happiness, disgust, surprise, repression and others).

CASME (The Chinese Academy of Sciences Micro-expression): This dataset was introduced in [176]. It includes 7 emotions (tense and repression in addition to the basic emotions), from 35 participants, with only 19 considered as valid. It contains 195 micro-expressions (selected from more than 1500 elicited facial movements) divided into two classes (*A* and *B*). The class *A* includes 100 samples, while 95 samples are comprised by the class *B*. The facial expressions were recorded in two different environmental configurations by using two different cameras: the first one with natural light and 1280×720 resolution, and the second one with two LED lights and a resolution of 640×480 .

SMIC-E (Extended version of SMIC–Spontaneous Micro-expression): SMIC-E [83] contains 32 participants filmed under different conditions and timings. There are 16 participants recorded with a High-Speed (HS) camera (PixeLINK PL-B774U, 640×480) at 100 fps. The longest micro-expression clips have a duration of 50 frames for a total of 167 samples. Other 8 participants were recorded with a normal visual camera (VIS) at 25 fps, in addition to the high-speed camera. In this case, the longest micro-expression clips have 13 frames for a total of 71 samples. The same number of samples were included in the third part of the dataset captured with a near-infrared (NIR) camera at 25 fps in addition to the high-speed camera. In this case, the longest micro-expression clips have 13 frames, such that all of them involve three emotions (positive, negative and surprise).

SMIC (Spontaneous Micro-expression): In the SMIC [126] dataset, six subjects were recorded, with a camera at 100 fps, while watching 16 films. The acquisitions were performed while instructing the participants to suppress their facial expressions whilst

carefully watching the clips. The proposed experiments aim to guess which film clip the subject is watching by looking at her/his face. The acquisitions involve five emotions (disgust, fear, happiness, sadness and surprise) and 77 spontaneous micro-expressions.

Canal9 (A Database of Political Debates for Analysis of Social Interactions): This dataset was recorded by the Canal 9 local TV station and broadcast in Valais, Switzerland [163]. It includes 190 participants, collected during 70 debates for a total of 43 h and ten minutes of material, involving 24 sequences of micro-expressions.

YorkDDT (York Deception Detection Test–University of York): In the YorkDDT dataset [167], the micro-expression sequences of nine subjects were segmented and labeled as truthful/deceptive and/or emotional/non-emotional. This resulted into 18 samples from 20 videos for a Deception Detection Test (DDT) and two emotion classes.

1.3.2.3 In-the-Wild Datasets

We were able to identify just one micro-expression accessible dataset acquired under in-the-wild conditions.

MEVIEW (MicroExpression ViDEos in-the-Wild): The MEVIEW [68] dataset contains 31 videos of 16 subjects from poker games and TV interviews downloaded from the Internet. It includes macro- and micro-expressions while illustrating five emotions.

1.3.2.4 Other Categorizations of Micro-Expression Datasets

We have observed that annotations in micro-expression datasets are classified either with onset (start), apex (peak), offset (end) of the emotion, self-reported, or facial muscle contraction. For example, the frame annotations in SAMM [25], CAS(ME)2 [101], CASME II [175], and CASME [176] are onset, offset, apex; while in the MEVIEW dataset [68], they are onset and offset. Differently, the Silesian deception dataset [130] is annotated with eye closures, gaze aversion and micro-tensions. The annotation of micro-expression video clips in the SMIC-E [83] and the SMIC [126] datasets have been obtained according to participants' self-reported emotions. However, in the Canal9 dataset [163],

the annotations are manual speaker segmentation, role, agreement and disagreement, automatic speaker segmentation, manual shot segmentation, automatic shot segmentation, manual shot classification and manual identification of participants in personal shots. In a similar way to what we presented for macro-expressions, we herein organize the micro-expression datasets according to different categorizations: number of subjects, FPS, amount of data/frames, samples, FACs coded, lights and resolution.

- **Number of subjects:** Table 1.4 presents a classification of micro-expression datasets according to the number of enrolled subjects. We classify the datasets according to the fact they involve less than 50 participants or more than 100 ones.
- **Frame per second and resolution:** Due to the importance of the FPS rate in the detection of micro-expression datasets, we have found that the number of FPS reaches 200 in both the SAMM and the CASME II datasets, which is a higher number than that used in macro-expression datasets. In the following datasets, the number of FPS is equal or greater than 100: Silesian deception, CASME, SMIC-E HS, and SMIC. There are also micro-expression datasets, where the number of FPS is smaller than 50 as for CAS(ME)2, MEVIEW, SMIC-E VIS, SMIC-E NIR, and YorkDDT. To help capture more subtle facial movements, a higher number of FPS and resolution is needed. As best as we know, the highest resolution available for micro-expressions datasets is 2040×1088 pixels as presented by the SMM dataset; and the lowest resolution set, instead, is equal to 320×240 for the YorkDDT dataset. For the rest of the micro-expression datasets, the resolution is set to 640×480 in the CAS(ME)2, Silesian deception, CASME II, CASME, SMIC-E, and SMIC datasets.
- **Amount of data and samples:** Unlike macro-expression datasets, most of the micro-expression datasets contain videos. The major difference between micro- and macro-expressions resides in the number of samples and/or the number of micro-expressions. We classify the datasets according to whether they contain less than 50 samples as in MEVIEW, Canal9 and YorkDDT, between 50 and 100 samples as in CAS(ME)2, SMIC-E VIS, SMIC-E NIR and SMIC, or between 100 and 200 samples as in SAMM, Silesian deception, CASME and SMIC-E HS. The CASME II dataset includes 247 samples.

- **Lights:** Micro-expression datasets propose several lightning conditions. Four lights have been used in both the CASME II and the SMIC-E datasets, while two lights were performed for SAMM and CAS(ME)2 and in the second class of CASME.

Table 1.4: Classification of micro-expression datasets according to the number of subjects.

Number-of Subjects	Micro-Expression Datasets
≤ 50	SAMM, CAS(ME)2, MEVIEW, CASME II, CASME, SMIC-E, SMIC, YorkDDT
≥ 100	RAF-DB, AffectNet, Aff-Wild, EmotioNet, FER-Wild, FER-2013, HAPPEI, HUMAINE

Overall, apart from having only one dataset with in-the-wild expressions, i.e., the MEVIEW dataset [68], there is still work to do for micro-expression datasets. First, the number of subjects is still small, not exceeding 200 subjects. Second, the age range is limited, and most of the subjects are in their twenties. In fact, we did not find children or elders included in the micro-expression datasets in order to have more diversity. The emotion variance is also limited, with only two datasets including seven emotions; i.e., SAMM [25] and CASME [176], while the rest present a number of emotions between two and five. It is worth noting that almost all the micro-expression datasets were created in an indoor environment. It could be interesting to have more in-the-wild micro-expression datasets with a larger number of subjects and a wider age range, so as to include children and elders, and have more emotion variation.

Table 1.5: Macro-expressions datasets. The columns report: the dataset name (Dataset); the number of subjects; the range of subjects' age (Age); the number of frames captured per second (FPS); ethnicity; and the amount of data/frames. In the table cells, a '-' indicates that no information is available, while a '*' following the dataset name indicates that the data is publicly available. *Italicized* names indicate "in the wild" datasets, while **bold** names indicate spontaneous datasets.

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
EB+ [39]	2020	200	18–66	25	Five ethnicities (Latino/Hispanic, White, African American, Asian, and Others)	1216 videos, with 395 K frames in total
iSAFE [146]	2020	44	17–22	60	Two ethnicities (Indo-Aryan and Dravidian (Asian))	395 clips
<i>RAF-DB</i> * [81]	2019	thousands	-	-	The images URLs were collected from Flickr	30,000 facial images
TAVER * [79]	2019	17	21–38	10	One ethnicity (Korean)	17 videos of 1—4 mn
4DFAB [20]	2018	180	5–75	60	Three ethnicities (Caucasian (Europeans and Arabs), Asian (East-Asian and South-Asian) and Hispanic/Latino)	Two million frames. The vertex number of reconstructed 3D meshes ranges from 60k to 75k
<i>Aff-Wild2</i> * [74]	2018	258	infants, young and elderly	30	Five ethnicities (Caucasian, Hispanic or Latino, Asian, black, or African American)	Extending it with 260 more subjects and 1,413,000 new video frames
RAVDESS * [86]	2018	24	21–33	30	(Caucasian, East-Asian, and Mixed (East-Asian Caucasian, and Black-Canadian First nations Caucasian))	7356 recordings composed of 4320 speech recordings and 3036 song recordings
<i>AM-FED+</i> * [98]	2018	416	-	14	Participants from around the world	1,044 videos of naturalistic facial responses to online media content recorded over the Internet
GFT * [49]	2017	96	21–28	-	Participants were randomly selected	172,800 frames

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
<i>AffectNet</i> * [104]	2017	450,000	average age 33.01 years	-	More than 1,000,000 facial images from the Internet	1,000,000 images with facial landmarks. 450,000 images annotated manually
<i>AFEW-VA</i> * [76]	2017	240	8–76	-	Movie actors	600 video clips
SEWA * [77]	2017	398	18–65	20–30	Six ethnicities (British, German, Hungarian, Greek, Serbian, and Chinese)	1990 audio-visual recording clips
BP4D+ (MMSE) [185]	2016	140	18–66	25	Five ethnicities (Latino/Hispanic, White, African American, Asian, and Others)	1.4 million frames. Over 10TB high quality data generated for the research community
<i>Aff-Wild</i> * [177]	2016	500	-	-	-	500 videos from YouTube
<i>EmotioNet</i> * [11]	2016	1,000,000	-	-	One million images of facial expressions downloaded from the Internet	Images queried from web: 100,000 images annotated manually, 900,000 images annotated automatically
<i>FER-Wild</i> * [106]	2016	24,000	-	-	-	24,000 images from web
BAUM-1 * [179]	2016	31	19–65	30	One ethnicity (Turkish)	1184 multimodal facial video clips contain spontaneous facial expressions and speech of 13 emotional and mental states
BioVid Emo * [180]	2016	86	18–65	-	-	15 standardized film clips
<i>Vinereactor</i> * [72]	2016	222	-	web- cam	Mechanical tuckers	6029 video responses from 343 unique mechanical truck workers in response to 200 video stimulus. Total number of 1,380,343 video frames
CHEAVD * [84]	2016	238	11–62	25	-	Extracted from 34 films, two TV series and four other television shows. In the wild

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
ISED * [60]	2016	50	18–22	50	One ethnicity (India)	428 videos
4D CCDB * [160]	2015	4	20–50	60	-	34 audio-visuals
MAHNOB Mimicry * [12]	2015	60	18–34	25	Staff and students at Imperial College London	Over 54 sessions of dyadic interactions between 12 confederates and their 48 counterparts
OPEN-EmoRec-II * [137]	2015	30	Mean age: women 37.5 years; men 51.1 years	-	-	Video, audio, physiology (SCL, respiration, BVP, EMG Corrugator supercilii, EMG Zygomaticus Major) and facial reactions annotations
<i>HAPPEI</i> * [27]	2015	8500 faces	-	-	-	4886 images.
AVEC’14 * [158]	2014	84	18–63	-	German	300 audio-visuals
<i>BAUM-2</i> * [37]	2014	286	5–73	-	two ethnicities (Turkish, English)	1047 video clips
BP4D-Spontaneous * [184]	2013	41	18–29	25	four ethnicities (Asian, African-American, Hispanic, and Euro-American)	368,036 frames
DISFA * [96]	2013	27	18–50	20	four ethnicities (Asian, Euro American, Hispanic, and African-American)	130,000 frames
RECOLA * [131]	2013	46	Mean age: 22 years, stan- dard deviation: three years	-	four ethnicities (French, Italian, German and Portuguese)	27 videos
<i>AM-FED</i> * [99]	2013	242	Range of ages and ethnicities	14	Viewers from a range of ages and ethnicities	168,359 frames/242 facial videos

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
<i>FER-2013</i> * [50]	2013	35,685	-	-	-	Images queried from web
AVEC'13 (AViD- Corpus) * [159]	2013	292	18–63	30	one ethnicity (German)	340 audio-visuals
CCDb * [7]	2013	16	25–56	-	All participants were fully fluent in the English language	30 audio-visuals
<i>MAHNOB Laughter</i> * [125]	2013	22	Average age: 27 and 28 years	25	12 different countries and of different origins.	180 sessions 563 laughter episodes, 849 speech utterances, 51 posed laughs, 67 speech–laughs episodes and 167 other vocalizations annotated in the dataset
DynEmo * [154]	2013	358	25–65	25	One ethnicity (Caucasian)	Two sets of 233 and 125 recordings of EFE of ordinary people
<i>PICS-Stirling ESRC</i> <i>3D Face Database</i> * [1]	2013	99	-	-	-	2D images, video sequences and 3D face scans
DEAP * [73]	2012	32	19–37	-	Mostly European students	40 one-minute long videos shown to subjects
<i>AFEW</i> * [29]	2012	330	1–70	-	Extracted from movies	1426 sequences with length from 300 to 5400 ms. 1747 expressions
SEMAINE * [100]	2012	24	22–60	-	Undergraduate and postgraduate students	130,695 frames
MAHNOB-HCI * [150]	2012	27	19–40	60	Different educational background, from undergraduate students to postdoctoral fellows, with different English proficiency from intermediate to native speakers	756 data sequences

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
Belfast in- duced * [149]	2012	Set1: 114	Undergraduate students	-	undergraduate students	570 audio-visuals
		Set2: 82	Mean age of participants 23.78	-	Undergraduate students, postgraduate stu- dents or employed professionals	650 audio-visuals
		Set3: 60	age of partici- pants 32.54	-	(Peru, Northern Ireland)	180 audio-visuals
Hi4D-ADSIP * [95]	2011	80	18–60	60	Undergraduate students from the Performing Arts Department at the University. Under- graduate students, postgraduate students and members of staff from other departments	3360 images/sequences
UNBC-McMaster (UNBC Shoulder Pain Archive (SP)) * [90]	2011	25	-	-	Participants were self-identified while having a problem with shoulder pain	48398 frames/200 video sequences
CAM3D * [93]	2011	16	24–50	25	Three ethnicities (Caucasian, Asian and Mid- dle Eastern)	108 videos of 12 mental states
SFEW * [28]	2011	95	-	-	-	700 images: 346 images in Set 1 and 354 images in Set 2
B3D(AC) * [40]	2010	14	21–53	25	Native English speakers	1109 sequences, 4.67 s long
USTC-NVIE * [165]	2010	215	17–31	30	Students	236 apex images
CK+ * [89]	2010	123	18–50	-	Three ethnicities (Euro-American, Afro- American and other)	593 sequences

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
<i>MMI-V</i> * [157]	2010	25	20–32	25	Three ethnicities (European, South American, Asian)	1 h and 32 min of data. 392 segments
<i>AVLC</i> * [156]	2010	24	Average ages were respectively 30, 28 and 29 years	25	eleven ethnicities (Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA and South Korea)	1000 spontaneous laughs and 27 acted laughs
<i>AvID</i> * [46]	2009	15	19–37	-	Native Slovenian speakers	Approximately one-hour video for each subject
<i>AVIC</i> [144]	2009	21	≤ 30 and ≥ 40	25	Two ethnicities (Asian and European)	No. episodes 324
Belfast in- duced * [149]	2012	Set1: 114	Undergraduate students Mean age of participants 23.78	-	undergraduate students	570 audio-visuals
		Set2: 82	age of participants 23.78	-	Undergraduate students, postgraduate students or employed professionals	650 audio-visuals
		Set3: 60	age of participants 32.54	-	(Peru, Northern Ireland)	180 audio-visuals
<i>Hi4D-ADSIP</i> * [95]	2011	80	18–60	60	Undergraduate students from the Performing Arts Department at the University. Undergraduate students, postgraduate students and members of staff from other departments	3360 images/sequences
UNBC-McMaster (UNBC Shoulder Pain Archive (SP)) * [90]	2011	25	-	-	Participants were self-identified while having a problem with shoulder pain	48398 frames/200 video sequences
CAM3D * [93]	2011	16	24–50	25	Three ethnicities (Caucasian, Asian and Middle Eastern)	108 videos of 12 mental states

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
<i>SFEW</i> * [28]	2011	95	-	-	-	700 images: 346 images in Set 1 and 354 images in Set 2
B3D(AC) * [40]	2010	14	21–53	25	Native English speakers	1109 sequences, 4.67 s long
<i>USTC-NVIE</i> * [165]	2010	215	17–31	30	Students	236 apex images
CK+ * [89]	2010	123	18–50	-	Three ethnicities (Euro-American, Afro-American and other)	593 sequences
<i>MMI-V</i> * [157]	2010	25	20–32	25	Three ethnicities (European, South American, Asian)	1 h and 32 min of data. 392 segments
<i>AVLC</i> * [156]	2010	24	Average ages were respectively 30, 28 and 29 years	25	eleven ethnicities (Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA and South Korea)	1000 spontaneous laughs and 27 acted laughs
AvID * [46]	2009	15	19–37	-	Native Slovenian speakers	Approximately one-hour video for each subject
AVIC [144]	2009	21	≤ 30 and ≥ 40	25	Two ethnicities (Asian and European)	No. episodes 324
DD [22]	2009	57	-	30	19% non-Caucasian	No. episodes 238
<i>VAM-faces</i> * [53]	2008	20	16–69 (70% ≤ 35)	25	One ethnicity (German)	1867 images (93.6 images per speaker on average)
<i>FreeTalk</i> * [18]	2008	4	-	60	Originating from different countries and each of them speaking a different native language (Finnish, French, Japanese, and English)	No. episodes 300

Table 1.5: *Cont.*

Dataset	Year	Number of Sub- jects	Age	FPS	Ethnicity	Amount of Data/Frames
<i>IEMOCAP</i> * [15]	2008	10	-	120	Actors (fluent English speakers)	Two hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions
<i>SAL</i> * [30]	2008	4	-	-	-	30 min sessions for each user
<i>HUMAINE</i> * [31]	2007	Multiple	-	-	-	50 ‘clips’ from naturalistic and induced data
<i>EmoTABOO</i> * [178]	2007	-	-	-	French dataset	10 clips
<i>AMI</i> [19]	2006	-	-	25	-	A multi-modal data set consisting of 100 h of meeting recordings
<i>ENTERFACE</i> * [142]	2006	16	average age 25	-	-	-
		5	22-38			
		16	average age 25			
<i>RU-FACS</i> [9]	2005	100	18–30	24	Two ethnicities (African-American and Asian or Latino)	400–800 min dataset
<i>MMI</i> * [122]	2005	19	19–62	24	Three ethnicities (European, Asian, or South American)	Subjects portrayed 79 series of facial expressions. Image sequence of frontal and side view are captured. 740 static images/848 videos
<i>UT-Dallas</i> * [116]	2005	284	18–25	29.97	One ethnicity (Caucasians)	1540 standardized clips
<i>MIT</i> [61]	2005	17	-	-	-	Over 25,000 frames were scored
<i>EmoTV</i> * [2]	2005	48	-	-	French	51 video clips
<i>UA-UIUC</i> * [145]	2004	28	Students	-	Students	One video clip for each subject
<i>AAI</i> [134]	2004	60	18–30	-	Two ethnicities (European American and Chinese American)	One audiovisual for each subject
<i>Smile dataset</i> [143]	2001	95	-	30	-	195 spontaneous smiles

Table 1.6: Micro-expressions datasets. Number subjects: Number of subjects. Ages: age range of the subjects. FPS: frames captured per second. - : No Information. Samples: micro-expressions. Content: **Spontaneous** or *in-the-wild*.

Dataset	Year	Number Subjects	Age	FPS	Ethnicity	Data/Frames	of FACs Coded	Samples	Lights	Resolution	Emotions
SAMM [25]	2018	32	average 33.24	200	Thirteen ethnicities (white British and other)	338 micro movements	Yes	159	Two lights as array of LEDs	2040 × 1088	Seven emotions. Macro/Micro
CAS(ME)2 [101]	2018	22	Average 22.59	30	One ethnicity	250 macro, 53 micro	No	53	Two light-emitting diode (LDE) lights	640 × 480	Four emotions. Macro/Micro
MEVIEW [68]	2017	16	-	25	-	31 videos	Yes	31	-	-	Five emotions. Macro/Micro
Silesian deception [130]	2015	101	Students	100	Third and fourth year students	101 videos 1.1 M frames.	Yes	183 micro-tensions	Proper illumination	640 × 480	Macro/Micro
CASME II [175]	2014	26	Average 22.03	200	One ethnicity	Among 3000 facial movements	Yes	247	Four selected LED lamps under umbrella reflectors	640 × 480	Five emotions
CASME [176]	2013	35 (19 valid)	Average 22.03	60	One ethnicity	More than 1,500 elicited facial movements	Yes	195 in Class A, 100 in Class B, 95	Class A: natural light, Class B: room with two LED lights	Class A: 1280 × 720. Class B: 640 × 480	Seven emotions

Table 1.6: *Cont.*

Dataset	Year	Number Subjects	Age	FPS	Ethnicity	of Data/Frames	FACs Coded	Samples	Lights	Resolution	Emotions
SMIC-E: HS VIS NIR [83]	2013	HS: 16	(22–34)	100	Three ethnic- ities (Asians, Caucasians and African)	Longest micro- expression clips: 50 frames	No	164	4 lights at the four upper corners of the room	640×480	3 emotions (positive, negative and surprise)
				25		The longest micro- expression clips: 13 frames			71		
				25		Same as VIS			71		
SMIC [126]	2011	6	-	100	-	1,260,000 frames	No	77	Indoor bunker en- vironment resembling an interrogation room	640×480	Five emo- tions: Micro
Canal9 [163]	2009	190	-	-	-	70 debates for a total of 43 h and 10 min of material	No	24	-	720×576	Political debates recorded by the Canal 9 local Switzer- land TV station
YorkDDT [167]	2009	9	-	25	-	20 videos for a deception detection test (DDT). seven frames	No	18	-	320×240	Two emotion classes

1.3.3 Applications

Datasets vary in the number of participants, head pose, age, video resolution, number of frames, number of subjects, and context. In this section, we comment about the most used datasets in the state-of-the-art and the main contexts of applications according to the most recent works.

Medical Applications Detecting signs of depression, pain, or even diagnosing rare conditions of disease can be identified based on specific features that are derived just by looking at the face, like heart beat, skin texture, and skin temperature. Some datasets include those features to detect emotions, and thus they can be used to detect health signs. For instance, GET [49] is a dataset of group formation (Healthy social drinkers), and the BioVid Emo [180] dataset combines psycho physiological signals with video signals for discrete basic emotions, whereas the OPEN EmoRec II [137] dataset includes physiology annotations (SCL, respiration, BVP, EMG Corrugator supercilii, EMG Zygomaticus Major). Likewise, the MAHNOB-HCI [150] dataset provides physiological sensors measuring ECG, EEG (32 channels), respiration amplitude, and skin temperature. Nevertheless, the UNBC-McMaster or the UNBC Shoulder Pain Archive (SP) [90] dataset has spontaneous facial expressions relating to genuine pain, where participants were self-identified as having a problem with shoulder pain. The DD [22] dataset was created with participants from a clinical trial for treatment of depression. It is worth noting that, in addition to the above mentioned pain and depression datasets, there are studies imaging facial expressions of patients with Alzheimer’s disease [48], Parkinson’s [147], schizophrenia [3], and autism [141]. However, most of the datasets from these studies are protected by privacy laws such as HIPPA laws and hence are not shared publicly.

Smart Driving Applications Driving a car has become like interacting with a social robot. Therefore, the emotional status of drivers is mandatory to build smart driving applications. In the MIT [61] dataset, drivers were recorded while their physiological signals were recorded during the drive (consent to the driving monitoring was collected).

Social Marketing Applications To predict buyers’ practices, commercial applications tend to watch the reactions of people to ads, such as in the AM-FED+ [98] dataset

and in the AM-FED [99] dataset, where subjects were watching amusing super bowl commercials.

Human Computer Interactions Some datasets present the facial expressions when participants are performing a human-computer interaction. For example, the iSAFE [146] dataset is an Indian dataset where volunteers were watching a few stimulant videos, and in the AVEC'13 dataset as well as the AViD-Corpus [159] dataset, subjects were recorded using a webcam and a microphone.

1.4 Human Emotion Recognition and Dataset Challenges

Despite the advancements in FER systems and the development of extensive macro- and micro-expression datasets, there remain significant challenges in the accurate labeling and categorization of emotions. One of the central issues is the reliance on traditional categorical models, such as the six basic emotions, which may oversimplify the complexity of human emotional expression. An alternative approach that leverages Valence-Arousal models offers a more nuanced representation of emotional states, but presents its own set of challenges. In fact, a key issue is the inconsistency between VA values and traditional categorical emotion labels, which leads to difficulties in accurately aligning these two systems. Furthermore, current datasets often exhibit a significant imbalance in emotion categories, with certain emotions being underrepresented. This imbalance complicates the development of robust FER models that can generalize across diverse emotional expressions. To address these challenges, a more fine-grained taxonomy of emotions can be derived from VA values, offering a potential solution for automating the annotation process and reducing the labor-intensive nature of manual dataset labeling. This refined approach aims to bridge the gap between categorical and dimensional models of emotion, enabling more consistent and balanced emotion recognition systems.

1.4.1 Proposed Emotion Taxonomy

In this section, we discuss the selection processes and the taxonomy we chose for our work.

1.4.1.1 Russell vs. Parrott.

Describing an emotional state of a person with the six prototypical emotions, i.e., joy, sadness, disgust, fear, anger and surprise, provides us with her/his expression but does not give us the person's emotions.

Russell In the work of Russell [140], a set of 151 terms (see Appendix 4.4) were provided with the corresponding distributions of *valence*, *arousal* and *dominance* values (this is given as a mean value plus a standard deviation term). The 151 terms not only describe the emotions in a precise way but also include terms that represent emotional interaction between two individuals. Russell defined the emotional state by three independent and bipolar dimensions, pleasure-displeasure, degree of arousal, and dominance-submissiveness. He defined it as a result of a study on 200 subjects and 42 verbal-report emotion scales. In our work, we use the *dimensional* representation of emotions based on VA, without using the dominance dimension. This is motivated by a subsequent work of Russell [139] that introduced the *circumplex* model of emotion, where the dimensions of excitation and valence were distributed in a two-dimensional circular space. Most of facial expression datasets that are labeled with the dimensional representation use these two dimensions.

Parrott In the same context, Parrott [123] defined a taxonomy for emotion-related terms in a tree-like structure (Table 1.7) starting with six primary emotions (i.e., love, joy, surprise, anger, pain and fear), then secondary emotions and tertiary emotions, for a total of 140 terms. Although the Parrott's hierarchy accurately describes the emotion of a person, it does not provide the measure of VA. Hence, one founding idea of our work is that of taking the common terms between Russell's work and Parrott's classification to derive the VA of each emotion.

Table 1.7: Parrott's distribution of emotions

Primary emotion	Secondary emotion	Tertiary emotions
Love	Affection	Adoration, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Desire, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, excitement, thrill, exhilaration
	Contentment	Pleasure
	Pride	Triumph
	Optimism	Eagerness, hope
	Enthrallment	Enthrallment, rapture
	Relief	Relief
Surprise	Surprise	Amazement, astonishment
Anger	Irritability	Aggravation, agitation, annoyance, grouchy, grumpy, crosspatch
	Exasperation	Frustration
	Rage	Anger, outrage, fury, wrath, hostility, ferocity, bitterness, hatred, scorn, spite, vengefulness, dislike, resentment
	Disgust	Revulsion, contempt, loathing
	Envy	Jealousy
Suffering	Suffering	Agony, anguish, hurt
	Sadness	Depression, despair, gloom, glumness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, displeasure
	Shame	Guilt, regret, remorse
	Neglect	Alienation, defeatism, dejection, embarrassment, homesickness, humiliation, insecurity, insult, isolation, loneliness, rejection
	Sympathy	Pity, mono no aware, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, suspense, uneasiness, apprehension (fear), worry, distress, dread

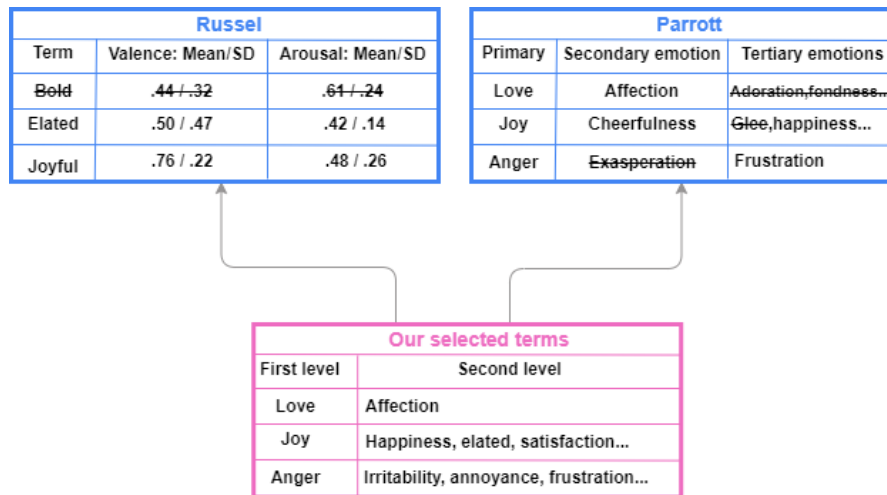


Figure 1.3: Mapping between emotion definitions.

1.4.1.2 Terms Selection and Mapping

To define a larger set of emotional terms with respect to the six basic emotions, we first merged the terms of Russell and Parrott; then, we selected the subset of most similar terms to get a mapping from Parrott's structure to Russell's values for VA. For example, the terms *joy* and *joyful* appear in both taxonomies, and the related (*valence, arousal*) pair given by Russell for joyful is (0.76, 0.48). We can therefore exploit both value mapping and Parrott's hierarchical structure. This process is depicted in Figure 1.3.

Table 1.8: Our selected emotional terms (6 and 32, respectively. at the first and second level of the categorization) plus neutral.

First level emotions	Second level emotions
Love	Affection
Joy	Happiness, Elation, Satisfaction, Excitement, Triumph
Surprise	Astonishment
Anger	Irritability, Annoyance, Frustration, Rage, Hostility, Hatred, Scorn, Disgust, Contempt
Sadness	Depression, Despair, Displeasure, Shame, Guilt, Regret, Rejection, Embarrassment,
	Humiliation, Insecurity, Isolation, Loneliness, Rejection
Fear	Terror, Anxiety, Distress

This process, resulted into 32 terms that we organized as reported in Table 1.8. In

addition, we also included the *neutral* emotion to the selected set of terms. In Parrott’s classification of emotions, 3 positive and 3 negative terms were proposed, respectively, for primary emotion, 11 positive and 14 negative terms for secondary emotion, and 42 positive and 73 negative terms for tertiary emotion, respectively. With this classification of emotions, we can see an imbalance between positive and negative terms. Since our classification is derived from Parrott’s, we can observe the dominance of negative emotions over positive ones in the proposed 32 terms: 3 positive and 3 negative terms, and 7 positive and 25 negative terms for the first and second level of emotions, respectively.

1.4.1.3 Terms Taxonomy

After merging Russell’s and Parrott’s terms, the intersection between the two classifications yields 38 terms (see Table 1.8) from the primary, secondary, and tertiary layers of the Parrott term representation. Taking Parrott’s classification as a guideline for our classification, we divided the 38 terms into two levels (Figure 1.4): the first level includes 6 terms (*love, joy, surprise, anger, sadness, fear*); the second level comprises 32 terms plus the neutral state (*neutral* is positioned in the origin of the VA plane). The choice of having only two levels in the hierarchy, unlike Parrott’s classification, is due to the fact that most of the selected terms come from Parrott’s tertiary layer. In the experiments, we used only the terms in the second level to re-label the dataset because the terms in the first level are independent emotions, intense, long-lasting and irreversible: even in case of reversing them, they need longer period. Contrarily, emotions in the second class are less intense and dependent upon primary emotions.

1.4.2 Experiments

The objective of our evaluation is to demonstrate the applicability of the proposed relabeling to a benchmark dataset. In the following, we considered the AffectNet dataset [104]. In fact, due to the limited number of datasets annotated with both emotions and VA [59], the AffectNet dataset was chosen.

1.4.2.1 Relabeling the AffectNet Dataset

AffectNet [104] is a large FER dataset acquired in the wild, with around 0.4 million images manually annotated for the presence of six categorical facial emotions (i.e., *happy, sad,*

First level		The terms in the first level are independent emotions, intense, long-lasting and irreversible, and even in case to reverse them, they need longer period.	
Love	Affection		
Joy	Happiness		
Surprise	Astonishment		
Anger	Irritability, Annoyance, Frustration, Rage, Hostility, Hatred, Scorn, Disgust, Contempt		
Sadness	Depression, Despair, Displeasure, Shame, Guilt, Regret, Rejection, Embarrassment, Humiliation, Insecurity, Isolation, Loneliness, Rejection		
Fear	Terror, Anxiety, Distress		

The terms in the second class are less intense and dependent upon primary emotions.		Second level	
		Affection	
		Happiness, Elation, Satisfaction, Excitement, Triumph	
	Surprise	Astonishment	
	Anger	Irritability, Annoyance, Frustration, Rage, Hostility, Hatred, Scorn, Disgust, Contempt	
	Sadness	Depression, Despair, Displeasure, Shame, Guilt, Regret, Rejection, Embarrassment, Humiliation, Insecurity, Isolation, Loneliness, Rejection	
	Fear	Terror, Anxiety, Distress	

Figure 1.4: First level and second level of the proposed classification.

surprise, fear, disgust, angry), plus contempt and neutral. Values of VA are also provided. In addition, the three non-emotional labels of *none*, *uncertain*, and *no-face* are used. In particular, the *none* (“none of the eight emotions”) category is an expression/emotion (such as sleepy, bored, tired, seducing, confuse, shame, focused, etc.), that could not be assigned by annotators to any of the six basic emotions, contempt or neutral. However, VA could be assigned to these images. The relabeling process was performed according to the VA annotations in AffectNet. These values were used to link each image with one of the proposed 32 emotional terms plus neutral. The relabeling is as follows:

Validation set - 19 terms were used in the relabeling (Insolent, Anxious, Disgusted, Insecure, Self-satisfied, Frustrated, Astonished, Depressed, lonely, Shamed, Excited, Affectionate, Elated, Hate, Defeated, Hostile, Irritated, Enraged, Happy). For a total of 248 images (6,20% of the relabeled images) no matching terms were found.

Training set - 21 terms were used in the relabeling (Insolent, Affectionate, Self-satisfied, Anxious, Disgusted, Insecure, Shamed, Elated, Depressed, Hate, Excited, Astonished,

Irritated, lonely, Happy, Frustrated, Defeated, Enraged, Regretful, Hostile, Despairing).

In this case, 44,871 images (15,60% of the relabeled images) were not relabeled.

Thus, the number of images and the percentage of each terms are distributed as shown in Table 1.9.

Table 1.9: The distribution of images after relabeling

Emotion	Number of images	Percentage
Anxious	1213	30.33%
Disgusted	876	21.91%
Isolent	547	13.68%
Affectionate	199	4.98%
Insecure	190	4.75%
Self-satisfied	181	4.53%
Shamed	118	2.95%
Elated	88	2.20%
Excited	73	1.83%
Depressed	63	1.58%
Astonished	54	1.35%
Frustrated	44	1.10%
Hate	43	1.08%
Lonely	18	0.45%
Irritated	16	0.40%
Defeated	8	0.20%
Surprised	8	0.20%
Joyful	7	0.18%
Enraged	5	0.13%
Hostile	2	0.05%
Angry	2	0.05%
Contempt	1	0.03%

Figure 1.5 shows the multi-term relabeling process for an example image from AffectNet. With the proposed terms, emotions can be described with more than the basic

terms: for example “sad” has a variety of terms, such as depression, despair, displeasure, shame, guilt, regret, embarrassment that describe the emotion more accurately. More in detail, the image on the left was originally annotated in AffectNet with VA values of, respectively, -0.793651 and -0.373016, and “Sadness” as expression label. With the proposed labeling, we can associate the point in the VA plane with first structure emotion “Sadness” and second structure emotion “Shame”. By applying the proposed terms hierarchy, we can derive a coarser expression characterization with 6 first class emotions (love, joy, surprise, anger, sadness, fear) and 32 second class emotions, that are more descriptive of the main emotion.

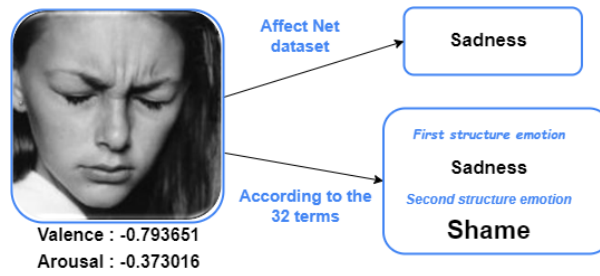


Figure 1.5: Example of relabeling an image from the AffectNet dataset according to the proposed approach.

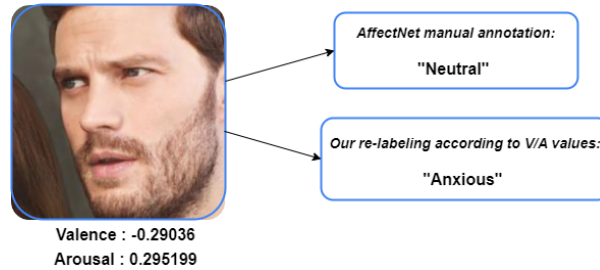


Figure 1.6: Difference between AffectNet and our re-labeling.

It turns out that AffectNet is manually annotated and labeled with a balanced number of classifications for the emotions (500 images for each emotion). But when we re-labeled the AffectNet images based on VA values, we found a large imbalance in the re-labeling results. This is due to the large difference between annotating a dataset according to the VA values and according to manual labeling. An example showing the efficacy of our proposed terms is illustrated in Figure 1.6. In this case, an image with valence -0.29036 and arousal 0.295199 was originally labeled in AffectNet as “Neutral”, though the “Neutral” label is the non-emotional state associated with the origin of the VA plane (0.0 value for both dimensions). The relabeling process applied to this image resulted

in a relabeling with the “Anxious” term. According to Russell’s work, the distributions for VA associated to this term are as follow: mean valence of 0.01 ± 0.45 , and arousal of 0.59 ± 0.31 . Our observation here, is that VA values provide a viable way to expand the categorical emotion representation with 7 (6 basic plus neutral) or 8 (with contempt) terms to a finer grained set of emotional states. Proposing the use of Russell’s mapping and the relation between Russell’s and Parrott’s terms, we contribute a motivated and flexible way to include more terms in the classification of expression datasets fostering the design of more effective and realistic classification solutions.

1.4.2.2 Results

In this section, we aim at evaluating the proposed classification emotion annotations. We are interested in estimating the consistency of our mapping with the manual annotations for the basic 8 emotion classes. With this aim, we used the Distract your Attention Network (DAN) [170]. The DAN model was proposed for facial expression recognition, while we used it as classifier using our proposed annotations as classes to perform emotion recognition. We used 287,651 images for the training set and 3,999 images for the validation set, according to the annotation of VA values provided by AffectNet. DAN includes three key components: a Feature Clustering Network (FCN) extracts robust features by adopting a large-margin learning objective to maximize class separability; a Multi-head cross Attention Network (MAN) instantiates a number of attention heads to simultaneously attend to multiple facial areas and build attention maps on these regions; finally, an Attention Fusion Network (AFN) distracts these attentions to multiple locations before fusing the attention maps to a comprehensive one.

In AffectNet, a total of 291,650 images manually annotated with eight labels were released for public research. We used the validation set including 3,999 images (the test set is not released). Results of our experimentation have shown an accuracy of **27.94%** compared to the original accuracy of **61%** obtained by running DAN as classifier on the original 8 expression labels of AffectNet. Observing the distribution of the 32 emotions over the whole dataset, and the lack of homogeneity between the number of images for each term, an increased difficulty of the classification task is revealed. We believe this result can open the way to new challenges in the task of emotion recognition, especially in the availability of datasets that are annotated not only with more than the 6 basic

emotions but also with the values of VA.

In what follows, we address the problem of class unbalance in facial recognition datasets, and how this problem affects the final performance. In fact, without handling unbalance during data collection results in classifications that are biased toward the majority class with poor accuracy for the minority classes [66]. The class unbalance problem represents one of eight different unbalance problems [113]; it happens when there is a noticeable disparity in the number of examples that belong to the various classes. The unbalance in datasets is not only evident in terms of emotions, but also in the distribution of positive and negative emotions. We refer further to the unbalance in Parrott’s classification regarding positive and negative emotions. AffectNet’s validation set, which includes 500 images for each emotion manually annotated, is unbalanced when considering the VA value of each image. However, AffectNet’s training set is out of balance in terms of VA value. In Figure 1.7, we show the impact of class unbalance of the AffectNet training set on our relabeling: the relabeling results shift towards the majority of classes already distributed in AffectNet: see the over-represented class “happy”, with 13,4915 images, versus the under-represented class “contempt” with 4,250 images. This induces not only unbalance in the data but also inconsistency between the manual labeling and the assigned VA values. This leads us to one outcome of our work: data unbalance results into poor accuracy, even when data annotation is improved with the proposed 32 classes.

Furthermore, regarding the inconsistency between emotion and VA labels, as presented in Section 1.4.2.1 and in Figure 1.6, VA values are attributed regardless of the actual value of the labeled emotion, and this is due to the manual annotation of the dataset (the manual annotation in AffectNet was done by one annotator for each image). For “neutral” the attributed VA values are not only 0.0 but also assume other values such as $(V,A)=(-0.176846,-0.176846)$, $(V,A)=(-0.135501,0.00483933)$. This presents a challenge in relabeling an existing benchmark dataset according to its valence and arousal values, while the emotions are labeled manually.

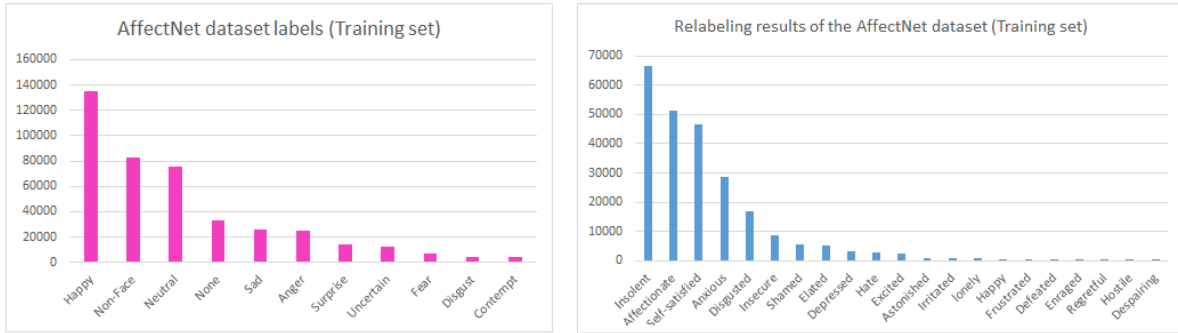


Figure 1.7: Impact of class unbalance on: AffectNet training set (left); our relabeling (right).

1.5 Conclusion

This chapter has provided a comprehensive review of the current research in facial emotion recognition, with a particular emphasis on micro- and macro-expressions, and the broader challenges in human emotion recognition, focusing on the need to move beyond the traditional six basic emotions to a more detailed and nuanced understanding of human emotional states. Current FER systems, while effective in many areas, often struggle to capture the complexity of emotions, which cannot be adequately described by a limited set of categories. Indeed, this chapter has highlighted the necessity of refining emotion classification methods to describe emotions with greater specificity, addressing the subtleties of emotional experience. Additionally, we have explored the limitations of existing datasets, which poses a significant challenge to the development of more accurate FER models. Improving the quality and diversity of emotion datasets is critical to advancing the field and enabling systems to recognize a broader spectrum of emotions.

In the following chapter, we will examine the various interpersonal relations datasets currently available, with a focus on those that capture emotional interactions between individuals. This examination will identify the limitations and potential areas for improvement within these datasets, thereby providing a foundation for the development of more robust and accurate emotion recognition models in future research.

Chapter 2

Emotions in Interpersonal Relationships

This chapter explores the role of emotions in interpersonal relationships, focusing on how emotional exchanges influence communication and behavior between individuals. It reviews key datasets that annotate emotions within both objective (e.g., family, professional) and subjective (e.g., friendly, hostile) relationships. The chapter also discusses psychological theories that inform the classification of these emotions, such as Bugental's and Kiesler's models. By analyzing existing datasets and methods, the chapter highlights gaps in the research and identifies opportunities to develop more comprehensive resources for studying emotional dynamics in social interactions.

2.1 Introduction

In this chapter, we present a comprehensive overview of datasets annotated with emotions in interpersonal relationships, highlighting the importance of emotional dynamics in the context of human interactions. Understanding emotions within interpersonal relationships is crucial, as these emotional exchanges shape how individuals connect, communicate, and behave with one another. Although there has been significant progress in studying individual emotions through various modalities, the exploration of emotions within the broader scope of interpersonal relationships remains a relatively underdeveloped area of research. This chapter aims to bridge that gap by examining how emotions are reflected and categorized within these complex social interactions. We begin by discussing the various types of interpersonal relationships—both objective, such as family

or professional connections, and subjective, which are defined by emotional qualities like warmth, hostility, or trust. We then review key datasets that have been developed to annotate emotions in these contexts, providing insights into the types of emotional categories used, the methodologies for data collection, and the challenges inherent in annotating relationships with emotional nuances. The chapter also explores the psychological foundations that inform the classification of emotions in interpersonal relations, drawing from established theories in social and emotional psychology. These frameworks help explain how emotions not only manifest between individuals but also influence the nature and quality of the relationships themselves. By reviewing these theoretical foundations alongside the available datasets, we aim to offer a clear understanding of how emotions are being captured and categorized within the domain of interpersonal relations. Ultimately, this chapter seeks to provide an organized and detailed perspective on existing annotated datasets and their contribution to the study of emotions in interpersonal relationships. By mapping out the current landscape, we identify gaps, challenges, and opportunities for developing richer and more comprehensive datasets that can better support the analysis and understanding of emotional dynamics in human interactions.

2.2 Interpersonal Relation Recognition

After examining micro and macro expression datasets, which offer significant insights into individual behaviors, it becomes evident that a deeper understanding can be gained by also exploring interpersonal relation datasets. Despite the rapid development of methods for analyzing human behavior, the study of interpersonal relations remains relatively underdeveloped and in its early stages. This area holds immense potential due to its broad social implications and numerous application scenarios, yet it has not received the same level of attention as other domains.

2.2.1 Social Relation Traits

Social relations can be described under different points of view, and involve various aspects. Social relations can be defined according to *objective* terms or based on *subjective* behavior. Among others, this classification was also followed in [91], where authors classified their dataset according to objective and subjective relations [71, 63]. For the

objective relations case, we can think of relationships such as kinship or work roles that do not depend on the particular behavior or emotional state. Differently, subjective relations depend on the perceived emotion, and can be defined using categories such as friendly/hostile, cooperative/submissive to mention some. These relations are directly originated from the particular interaction between people. The large categorization introduced above can be further specialized by considering the different behavioral and social psychology models that were followed by the research in this field. In this thesis, the objective relations comprise Kinship recognition [132, 133, 23], Asian social psychology and the Bugental’s social psychology theory [85, 151, 182]; the subjective relations, instead, include the interpersonal circle theory as defined by Kiesler [91, 186], and the Fiske’s theory of relational models [43]. In order to automatically detect and analyse such different relations using a computer vision/machine learning algorithm, a major problem is that of identifying some distinguishing features that can provide useful clues to the specific task. Clearly, they depend on the different type of relation of interest. For example, to determine a mother/son relation, it could be necessary to rely on some facial traits such as hair or eyes color, skin color, shape of mouth, eyes and nose, age difference. In this case, face recognition features can be useful. If, instead, we want to determine whether two people are in a friendly relation, analyzing expressions or actions is surely more informative. Subjective relations can change for the same class of objective relations though. For example, a mother and a son can either have a good relationships, namely, warm, friendly, involved and demonstrative, or a bad one, namely, strict, harmful, distant and silent. Same as work relations, i.e., boss and employee, it can be dominant, trusting, friendly and assured or equal, mistrusting, hostile and unassured. This makes the estimation of subjective relations significantly more challenging than the objective ones.

Several social psychology theories have been proposed to formalize the above [43, 14, 71, 63, 21, 44, 124, 92]. Among them, we choose those that are used in the interpersonal relation datasets collected in this thesis. Figure 2.1 illustrates our proposed organization of the literature works: we first categorized the datasets and methods according to the fact they address objective or subjective relations. For each category, several theories and relational models were proposed and formalized in the literature, each of which focuses on different yet related aspects. We first divided the datasets according to the five theories that we analyzed. Then, methods were ranked according to the used dataset

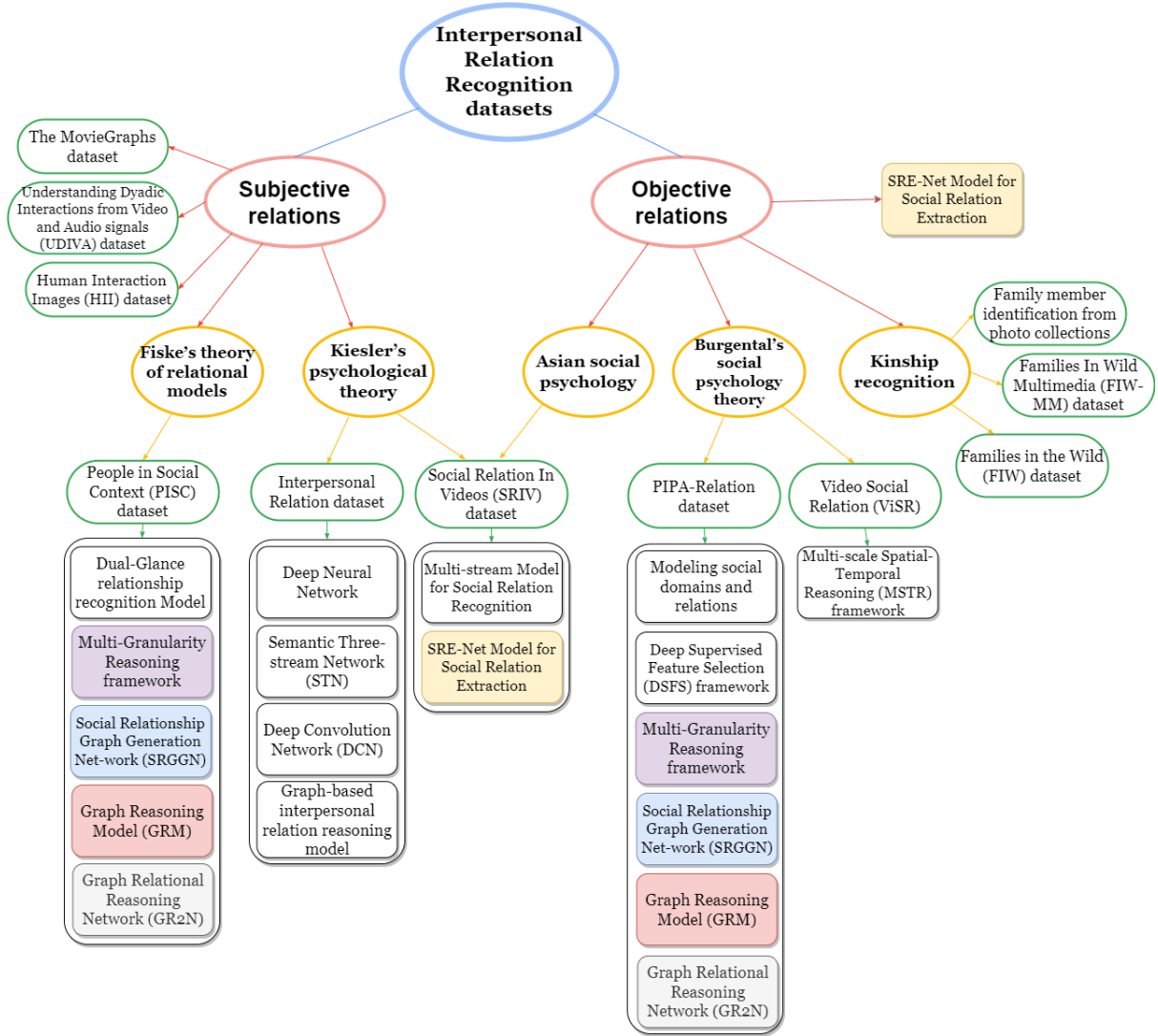


Figure 2.1: Our proposed organization of the literature on interpersonal relations. Yellow circles illustrate the theories we analyzed; green circles indicate the datasets, while the rounded rectangles refer to frameworks and methods. A color coding is used for the frameworks and methods to evidence methods that appear more than once in the figure and are applied to different datasets.

and connected to it in the graph (we used a color-coding for methods that appear more than once). In this way, it is immediate to identify the topic of interest of each approach, and which datasets can be used to address a specific task.

In what follows, we separately describe datasets and methods based on the type of relation they refer to, either objective or subjective.

2.2.2 Objective relations

Objective relations are defined in terms of social relationships that people have with respect to others, such as teacher/student or mother/child. Given the large and different number of such relations that exist in real life, some effort was put in defining meaningful categories. In the literature, the two models most largely employed are the *Bugental's Social Psychology Theory* [14], and the *Asian Social Psychology Theory* [63]. Both develop on the idea of categorizing people relationships as they likely influence their behavior. In fact, in social contexts, people are expected to behave differently depending on the type of relation they have. It is easy to imagine that one can change his/her behavior significantly whether the interaction is with a friend, a brother/sister, or one's boss/supervisor. While the above theories thoroughly consider both kinship as well as other general types of relationships, some literature works instead focus only on the former, others focus on the different levels of family relationship [26]. In the following, we will first briefly present datasets and methods related to kinship recognition in Section 2.2.2.1 and then describe the two theories in Section 2.2.2.2 and Section 2.2.2.3, respectively. Table 2.1 and Table 2.3 summarize the existing datasets for objective relations with their context/size and main annotations, and methods in the literature that were designed for objective relation recognition. For each method, we also reported information about the network architecture and the data used in the training.

Table 2.1: Objective relations datasets: Publicly available audiovisual human-human (face-to-face) non-acted interaction datasets.

Dataset	Year	Annotations	Content/Size	Observations
Families In Wild Multimedia (FIW-MM) [132]	2020	Families IDs, Member IDs, gender information, relationship types	Videos: 550 subjects in 660 videos	Kinship dataset with image, audio and video
Video Social Relation (ViSR) [85]	2019	Eight common social relations	Videos: over 8,000 videos	Limited types of social relations
Social Relation In Videos (SRIV) [91] ¹	2018	Eight subjective relations and eight objective relations	Videos: 3,124 videos	The multi-label strategy makes the relation ambiguous
PIPA-Relation [151]	2017	Head bounding box, identity number and social relations	Images: extend PIPA with 26,915 person	Semantic attributes are collected from both body and head images according to Bugental's theory
Families in the Wild (FIW) [133]	2016	Families IDs, Member IDs, gender information, relationship types	Images: 10,000 family photos	The largest kinship verification dataset
Family member identification from photo collections dataset [23] ²	2015	Identity label shows social role, bounding box indicating location of the face and the body skeleton	Images: photos of 16 families	The volume of the dataset is small

Not-available^{1,2}

2.2.2.1 Kinship Recognition

The objective of kinship recognition is to determine the exact type of kinship (e.g., parent-child) rather than if the subjects are related or not. Furthermore, methods described here only deal with kinship verification and family recognition. Datasets for kinship recognition are differentiated by the type of relation, e.g., Parent-Child [41, 88, 173], Twin Pairs [162], Siblings [13]. In the following, we present the datasets that are labeled with multiple-types of kin relations from pair of people or more images and videos.

Families in the Wild (FIW) The Families in the Wild (FIW) dataset [133] was collected using various search engines (e.g., Google, Bing, Yahoo) and social media outlets (e.g., Pinterest). It contains 10,000 family photos of 1,000 families, each with at least 3 family members. It divided the relationships into three categories with a total of 11 types: the *Parent-child* relation includes Father-Daughter (F-D), Father-Son (F-S), Mother-Daughter (M-D), Mother-Son (M-S); the *Grandparent-grandchild* relation comprises: Grandfather-Granddaughter (GF-GD), Grandfather-Grandson (GF-GS), Grandmother-Granddaughter (GM-GD), Grandmother-Grandson (GM-GS); and finally, the *Siblings* relation includes Sister-Brother (SIBS), Brother-Brother (B-B), Sister-Sister (S-S).

Families In the Wild Multimedia (FIW-MM) The Families In the Wild Multimedia (FIW-MM) dataset [132] is an extended dataset from the FIW dataset with multimedia (MM) data (video, audio, and text captions). It contains 550 subjects in 660 videos, subset of 200 FIW families. In FIW-MM, the data have three types: non-speaking face tracks (visual only), speech segments (audio only), and face tracks of speakers (visual-audio). This dataset is annotated in the same way as the FIW dataset, with families IDs, Member IDs, gender information and relationship types.

Family member identification from photo collections This dataset [23] contains photo collections of 16 different families taken at amusement parks, annotated with multiple labels such as social role, for example: “child1”, “father”, while non-family members are given identities such as “femaleAdult1”, “maleAdult1”, a bounding box around the face and the body skeleton.

2.2.2.2 Asian Social Psychology Theory

The Asian social psychology theory [63] asserts the need to include the context of relationships in any study of social behavior because of the profound impact that relationships have on individual behavior. Interestingly, it classifies relationships into fourteen relations according to the basis of their formation: Kinship (by blood-consanguinity-, marriage, adoption, or godparenthood), Connection by birth, Political authority, Subjugation (by military conquest, slavery, or colonialism), Social class, Office or employment, Residential location, Institutional affiliation, Social connections based on ascription (e.g., inheritance) or achievement, Tutelage apprenticeship or guardianship, Professional consultation, Companionship affection or sexual attraction, Situational temporary or chance encounters. To the best of our knowledge, only the following dataset exists that is collected and annotated according to this categorization.

Social Relation In Videos (SRIV) The Social Relation In Videos (SRIV) dataset [91] was proposed to recognize social relationships from a video, where different behaviors are classified along two dimensions: Subjective Relations (Sub-Relation) and Objective Relations (Obj-Relation). It contains 3,124 videos, collected from movies and TV dramas. The social relations were extracted from movies and television domains, and classified based on two classification approaches: the Kiesler theory (this theory is described in Section 2.2.3.1), and the Asian Social Psychology theory. According to this, the dataset was divided into 16 subclasses: eight subclasses were derived based on the Kiesler theory, and eight subclasses were arranged based on the Asian Social Psychology theory, using from the last one only three types of relationships (work, kinship and other) divided into eight subclasses (supervisor-subordinate, peer, service, parent-daughter, mating, sibling, friendly and hostile).

In the same work [91], authors proposed a multi-stream model for social relation recognition, using a ConvNet architecture. The network used RGB images of videos to learn clues of social relations, such as video scenes and people representations. Then, a multi-stream ConvNet including spatial, temporal and audio features was proposed, where the action features were fused by a temporal segment network, which is used to learn spatial and temporal features. In doing so, Inception with Batch Normalization (BN-Inception) was chosen to achieve a balance between accuracy and efficiency, and

GoogleNet was adopted to learn audio features using audio spectra.

In [187], the authors used the objective relation classes from the SRIV dataset with their SRE-Net model for social relation extraction. To predict the social relation (leader-member, peer, service, parents-offspring, lover, sibling, friend and enemy) between characters, they proposed a MoCNR method (method of character node recognition) to identify the number of people and characters that appear in the video by clustering people’s facial features (keyframe extraction, face detection, face alignment, face feature extraction and clustering). Then, the social relation recognition method was introduced based on scene segmentation: the scene changes in videos are detected and used to split the video into separate clips. The GoogleNet model was used to extract audio features with a C3D model to capture the time sequence information of video. Finally, AdaBoost (Adaptive Boosting), XGBoost, GBDT and LightGBM classifiers were used to accomplish social relation prediction.

2.2.2.3 Bugental’s Social Psychology Theory

The Bugental’s domain-based theory [14] partitions social life relations into 5 domains, namely, Attachment, Reciprocity, Mating, Hierarchical power, and Coalitional groups. Then, from this partition 16 fine-grained categories are derived: father-child, mother-child, grandpa-grandchild, grandma-grandchild, friends, siblings, classmates, lovers/spouses, presenter-audience, teacher-student, trainer-trainee, leader-subordinate, band members, dance team members, sport team members and colleagues. We observe that both kinship and other types of relations are defined in this theory, which makes it widely exploited with several methods addressing the problem of recognizing these classes.

PIPA-Relation The PIPA-Relation dataset [151] is an extended version of the People In Photo Albums (PIPA) dataset [182] with 26,915 person pair annotations for social relations. The dataset is labeled with five social domain according to Bugental’s theory. In fact, the PIPA dataset is collected from Flickr photo albums for the task of person recognition and contains 37,107 photos with 63,188 instances of 2,356 identities, labeled by head bounding box and identity number, where the same person can appear in multiple albums.

In [151], authors experimented with two models: a CNN model trained end-to-end,

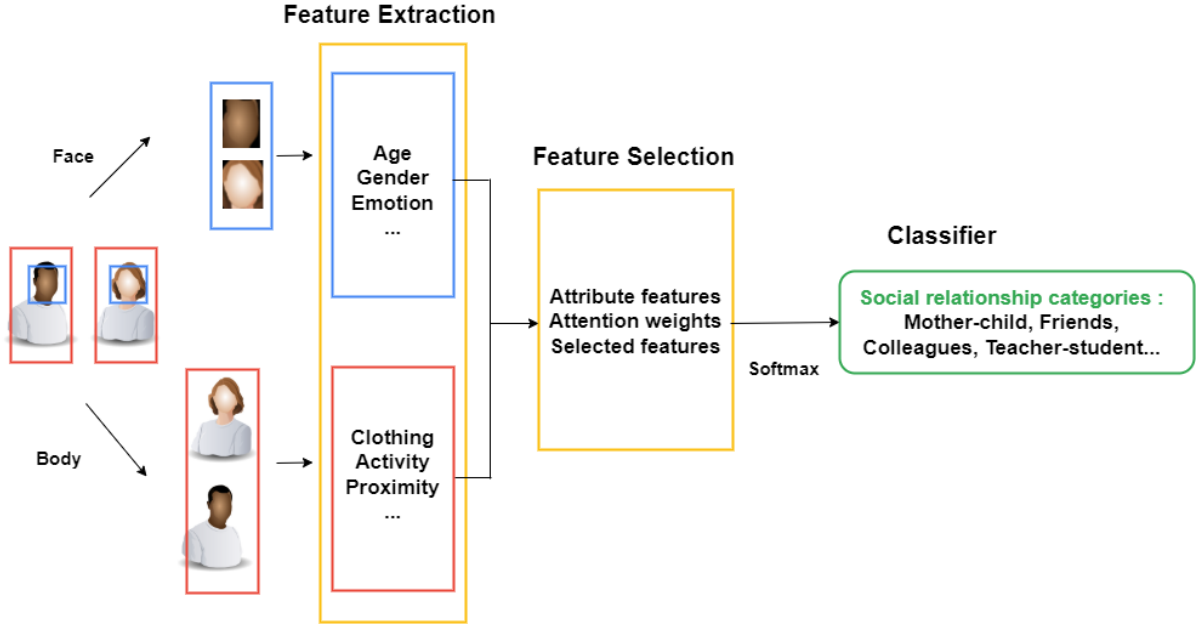


Figure 2.2: The deep supervised feature selection framework proposed in [164].

and a CNN model for semantic attribute recognition derived from the social domain theory. This latter network is capable of identifying attributes such as age and age difference, the latter being motivated by its importance to distinguish relations, gender, head appearance such as straight hair, wavy hair, wearing earring, wearing hat and so on, head pose and facial expressions. Other attributes that are classified are clothing such as hat, tShirt, jeans, actions such as holding hands, high five, hug, and activities such as adjusting, ailing, applauding, arranging, attacking, ballooning, baptizing and so on. Then, the method uses the concatenated feature to learn a linear SVM and categorize the interaction. The results of the experiments are quite preliminary in terms of accuracy, since authors reported a relation recognition accuracy for all the attributes of 57.2%, which is the same for body attributes, and only 44.8% for head attribute.

In [164], the authors adopted 12 face and body attributes including age, gender, appearance, emotion, pose, location & scale, face appearance, face pose, face emotion, body clothing, body proximity and body activity. A Deep Supervised Feature Selection (DSFS) framework is proposed, where the input photo can include more than two subjects that are segmented and used as input to the DSFS framework as pairs (the DSFS framework is illustrated in Figure 2.2). Then, the feature extraction module, included in the DSFS, extracts the 12 attribute's features using a pre-trained Double-Stream CaffeNet, a pre-trained CNN-CRF, or multi-task RNN. To deal with the problem of noises and

redundancies, authors used a feature selection module with two feature selection policies: group feature selection to extract the optimal feature subset based on contributions of attributes, and dimensional feature selection to learn the optimal feature subset at a fine-grained level and so remove most of the redundancy. Then, the final classification was performed by a Softmax classifier used to compute the probability distribution of the input pair on social relationship categories. The accuracy of social relationship recognition obtained in this work was 61.51%. Authors also tried to eliminate some attributes that effected the accuracy; an interesting outcome of this work is that body attributes contribute more than face attributes to recognize social relationships, which is quite in line with the observation that objective relations are independent from the emotional status of the person, and behavioral features such as actions or body proximity are more informative to recognize the social interaction.

Table 2.2: Objective relations: Methods and architectures.

Method	Year	Architecture	Training Set (Dataset)	Observations
Deep Supervised Feature Selection (DSFS) framework [164]	2020	Architecture of the proposed deep supervised feature selection (DSFS) framework	PIPA-relation dataset	The outcome is that body attributes contribute more than face attributes to recognize social relationships
Graph Relational Reasoning Network (GR2N) [82]	2020	Graph Relational Reasoning Network	PIPA-Relation and PISC dataset	Generate a reasonable and consistent social relation graph
Multi-scale Spatial-Temporal Reasoning (MSTR) framework [85]	2019	- Triple Graphs model - Pyramid Graph Convolutional Network (PGCN)	ViSR dataset	A graph network to capture social relations according to long-term and short-term temporal cues in the video
Multi-Granularity Reasoning framework [181]	2019	Deep CNN: - Person-Object Graph (POG) - Person-Pose Graph (PPG)	PIPA-Relation and PISC dataset	Predict social relations by fusing scene information as surrounding objects, but malfunction is observed for images with noisy features

Table 2.3: *Cont.*

Method	Year	Architecture	Training Set (Dataset)	Observations
Social Relationship Graph Generation Network (SRG-GN) [6]	2019	ConvNet architecture	PIPA-Relation and PISC dataset	Relationships are treated independently on the same image, which neglects the logical constraints between social relationships on an image
Graph Reasoning Model [166]	2018	Gated Graph Neural Network (GGNN)	PIPA-Relation and PISC dataset	Building a graph where persons and objects are fully connected, then using GGNN to predict the social relation
SRE-Net Model for Social Relation Extraction [187]	2018	MoCNR algorithm for Character Nodes Recognition	SRIV and LFW dataset	Extracting social relations between characters based on scene to uncover the relations within the same scene
CNN models trained end-to-end and trained for semantic attributes [151]	2017	Double-stream CaffeNet	PIPA and PIPA-Relation dataset	Recognize social relations from a group of semantic attributes

The works in [181, 6, 166, 82] also used the PIPA-Relation and PISC datasets [151] (this latter dataset is introduced in Section 2.2.3.2). Given that the goal is to understand the interaction among people, the common intuition of these methods is that of modeling the interaction with graphs. The multi-Granularity Reasoning framework as proposed in [181] has two branches. The first one used a CNN to learn knowledge about the scenes from the whole image. The other branch focused on regional cues and fine interactions among persons and contextual objects, with three main procedures: a CNN object detection model to crop persons and objects in an image, a human pose estimation method, and a Person-Object Graph (POG) to connect each person with the other person and the objects. They built a Person-Pose Graph (PPG) to model the interaction between two persons. Social relation reasoning was performed on the two graphs by GCNs. The social relation was predicted by the global feature (deep convolutional neural network ResNet-101) from CNN and the reasoning feature from the GCNs. They obtained the accuracy of 64.6% for Friends, 67.8% for Family, 60.5% for Couple, 76.8% for Professional, 34.7% for Commercial, and 70.4% for No Relation on the PISC dataset. An accuracy of 64.4% was reported on the PIPA-relation dataset.

In [6], authors proposed a Social Relationship Graph Generation Network (SRGGN) with two modules: A Multi-Network Convolutional Neural Network (MN-CNN) module for attribute and relationship representations followed by a Social Relationship Graph Inference Network (SRG-IN) module for generating a structured graph representation. The MN-CNN module has two sub-modules, SN1 and SN2, with an input image and a set of bounding box annotations for the people in the image. The input to SN1, the Attribute ConvNet architecture, is the cropping for a single-body image of a person, with the network having layers for each attribute: age, gender, and clothing. The sub-module SN2 is a network of pairwise-relationship ConvNet architectures, with two VGG-16 architectures to compute activity and scene features from the context images of people. The SRG-IN module predicts relationships in an image according to (person1, relation, person2), and the social relationships between people are classified in the form of a social graph. Gated Recurrent Units (GRUs) are used to improve the prediction of relationships between persons, where each relationship in an image gets information from its nearby nodes (person attributes) and also its nearby edges (relationships).

A Graph Reasoning Model (GRM) was finally proposed in [166] to detect the correlations between social relationships and semantic objects in the scene. To this end, a social relationship and object nodes are included. The GRM takes an image and a person pair of interest and extracts features from the regions of the person pair to initialize the relationship node. For the object node, it uses a Faster-RCNN detector to extract their features. Then, to explore the interaction of the persons with the contextual objects, it uses the Gated Graph Neural Network (GGNN), and finally, the graph attention mechanism is used to select the most informative nodes.

In [82], authors created a social relation graph called the Graph Relational Reasoning Network (GR2N). It jointly gathers all relations within a single image by constructing several virtual relation graphs to explicitly model the logical constraints among different types of social relations. They presented each person by a node in a graph, and the edge between each node is the relation between the people. GR2N is created to predict the existence of edges and the type of edges, in order to solve the problem of unknowing topology of the graph at the beginning, then generate a reasonable and consistent social relation graph.

Video Social Relation (ViSR) The Video Social Relation (ViSR) dataset [85] contains 8,000 video clips with more than 200 movies. The ViSR dataset is labeled with eight social relations such as Parent-offspring, Couple, Leader-subordinate, Service, Sibling, Friend, Colleague and Opponent. A Multi-scale Spatial-Temporal Reasoning (MSTR) framework was also proposed in that work. The approach started by cropping the persons and objects from frames with Mask R-CNN. Then, a triple Graphs model was designed: with an Intra-Person Graph (IntraG) for the same person, an Intra-Person Graph (IntraG) to model the spatial and temporal representation of persons and objects, and a Person-Object Graph (POG) to capture the co-existence of persons and contextual objects. Then, a relation reasoning by Pyramid Graph Convolutional Network (PGCN) was used to learn dynamics in varied temporal range.

In geneal, all the datasets and methods described so far are related to the problem of recognizing and classifying objective social relationships, so identifying the type of

connection among individuals. Overall, we can draw some conclusions; first, most of the methods rely more on body, e.g., age difference, gender, clothing, or action features, e.g., proximity, interaction, rather than emotional features, e.g., facial expressions, to discriminate the relationships. This somehow confirms that behavioral patterns are more informative in this scenario. We also observe another common way of addressing the problem is that of using graphs, which makes sense inasmuch as connections among people indeed form sort of graphs, where edges can be useful to verify the consistency of relations and perform reasoning on the graph. As an example, if subject A is the father of subject B , and subject C is the brother of subject B , then subject A must be also the father of subject C . Furthermore, observing the overall performance of the reported methods though, we can conclude that the problem is rather challenging, and that there is still room for large improvements.

2.2.3 Subjective Relations

We defined subjective social relations as those involving emotional states and behaviors, independently from the objective relationship that exists among subjects. Differently from objective relations, these are much more complex, variegated, and challenging both to define and detect. In fact, they depend on both the particular emotional status of the individuals, the person whom he/she is interacting with, and other possibly varying aspects. We identified two major theories related to subjective relations: the *Kiesler's Circle Theory* [71], and the *Fiske's Theory of Relational Models* [43]. Both analyze and categorize social relations in terms of perceived emotional status, even though the latter includes also a partial objective categorization. In the following, we separately describe the two theories and the related datasets and methods that are most closely aligned with them. Table 2.4 and Table 2.5 summarize the datasets, architectures, and methods belonging to this category.

Table 2.4: Subjective relation datasets: Publicly available audiovisual human-human (face-to-face) non-acted interaction datasets.

Dataset	Year	Annotations	Content/Size	Observations	
Understanding Interactions from Video and Audio signals (UDIVA) [118]	Dyadic	2020	Personality scores (self- and peer-reported), sociodemographics, mood, fatigue, relationship type	Videos: 147 subjects in 188 sessions	The self-report annotation is difficult to understand
Social Relation In Videos (SRIV) [91] ¹		2018	Eight subjective relations and eight objective relations	3,124 videos	The multi-label strategy makes the relation ambiguous
MovieGraphs [161]		2018	Interactions between characters, relationships, reasons behind certain interactions, situation, scene, and natural language	7,637 videos	An exhaustively annotated dataset. Some of the information is incomplete due to the discarded parts between scenes
People in Social Context (PISC) [151]		2017	Social relationship	22,670 images	Scene and global contextual cues to predict social relationships
Interpersonal Relation [186]		2016	Faces' bounding boxes and interpersonal relation traits	8,016 images	Interpersonal relationship traits are divided into 16 segments, thus describing a wide range of relations
Human Interaction Images (HII) [153]		2016	Face locations and orientations	1,971 images	Recognizing interactions between people by face information

Table 2.5: Objective relations: Methods and architectures

Method	Year	Architecture	Training Set (Dataset)	Observations
Graph-based interpersonal relation reasoning model [47]	2021	Multi-scale features with the GNNs	Interpersonal relation dataset	Exploiting the interactions between two faces for interpersonal relationship recognition
VGG-FACE model [57]	2019	Siamese-like architecture	Interpersonal relation dataset	Extracting information from whole images to predict group-level emotions
Semantic three-stream network (STN) [174]	2019	Siamese network	Interpersonal relation dataset	Using the semantic information to understand the social relation from the entire image
Multi-stream Model for Social Relation Recognition [91]	2018	Network Architecture (deep learning)	SRIV dataset	Using visual and audio features that represent social relations of people to ameliorate the recognition
Hybrid1365-VGG model [161]	2018	-	MovieGraphs dataset	Multi-labels to summarize and localize each scene, generating explanations for human interactions
Dual-Glance Model [151]	2017	Attentive RCNN	PISC dataset	Bringing attention to regions of persons and objects to predict social relations
Deep convolutional network (DCN) [186]	2016	Siamese-like architecture	Interpersonal relation dataset	Purchasing the inherent correspondences between facial expressions and other heterogeneous attributes



Figure 2.3: Visualization of the Kiesler’s model using the interpersonal circle.

2.2.3.1 Kiesler’s Circle

The Kiesler’s psychological theory [71] has proposed to organize the interpersonal relations into a circle called interpersonal circle or circumplex, where the relations are divided into 16 segments, with 8 pairs: Dominant, Competitive, Trusting, Warm, Friendly, Involved, Demonstrative and Assured/Submissive, Deferent, Mistrusting, Cold, Hostile, Detached, Inhibited and Unassured. Figure 2.3 shows the Kiesler circle, with the 16 segments and 8 pairs.

In the following, we present the interpersonal relation datasets and methods that have defined the social relation traits based on the interpersonal circle proposed by Kiesler.

Interpersonal Relation In [186], the authors built two datasets: the Expression in-the-Wild (ExpW) for facial expression recognition and the Interpersonal Relation dataset. This latter contains 8,016 videos and movies, collected from the Web, which are annotated with interpersonal relation traits. The authors proposed a deep convolution network to capture a rich face representation and a novel attribute propagation method that, despite the different datasets, takes advantage of the inherent correspondences between facial expressions and other heterogeneous attributes. Next, they jointly considered pairwise faces for interpersonal relation prediction by arranging two identical DCNs obtained in a Siamese-like architecture (the interpersonal relation model is illustrated in Figure 2.4). Using the interpersonal relation dataset, they trained the new Siamese network end-to-end to map raw pixels of a pair of face images to relation traits. Then, they performed relation traits reasoning using face representation and additional spatial cues. Thus, the proposed method uses a pre-trained model, which already learned face attributes (facial expression, gender, age, and head pose), and therefore, it predicts interpersonal relations.

They obtained 71% average accuracy across all the relation traits prediction performance (dominant, competitive, trusting, warm, friendly, involved, demonstrative and assured).

In [57], the authors proposed a deep neural network combining a Siamese-like Network with two VGG-Face model branches to learn the social relation between two people and a deep model to extract information (such as layout of objects, posture of people and background of a scene) from whole images to predict group-level emotions. The authors used the same evaluation metric and the same spatial cues as [186]. In particular, the used measures are the balanced accuracy to account for the imbalanced positive and negative attribute samples as evaluation metric, and the two faces' positions, the relative faces' positions and the ratio between the faces' scales as spatial cues. They selected VGG, Inception-v2 and ResNet to be fine-tuned on the social relation dataset. The reason behind this choice, is because of the performance of these networks in the ImageNet Large Scale, Visual Recognition Challenge (ILSVRC) and the information learned from it. This resulted useful to infer high level social relation, such as layout of objects, posture of people and background of a scene.

In [174], the authors proposed a semantic three-stream network (STN) for social relation recognition. They used a Siamese network with a semantic augmentation structure to extract features from a pair of face images to reduce parameters, since they tried to understand social relationships from the entire original image. They mapped each face separately, using a Siamese network to extract features and using a ResNet where convolutional networks share weights. They also proposed a semantic augmentation structure to help the network to better sense objects in the environment, which contains five parts: a size adjustment module, a channel adjustment module, a multi-view module, a feature transition module, and a semantic fusion module.

For interpersonal relation recognition, authors in [47] proposed a graph-based interpersonal relation reasoning model with multi-scale features. Their proposed architecture is divided into a feature extraction, the GNN module, and the classifiers. In the feature extraction stage, a pair of faces is cropped, represented in a bounding box and fed to two VGGFace models. Then, the joint area of every face pair is cropped as an input to another ResNet. The extracted features are represented as five nodes to construct the GNN graph structure. The GNN module based graph reasoning extract hierarchical features from the connected nodes. Finally, they designed eight binary classifiers using

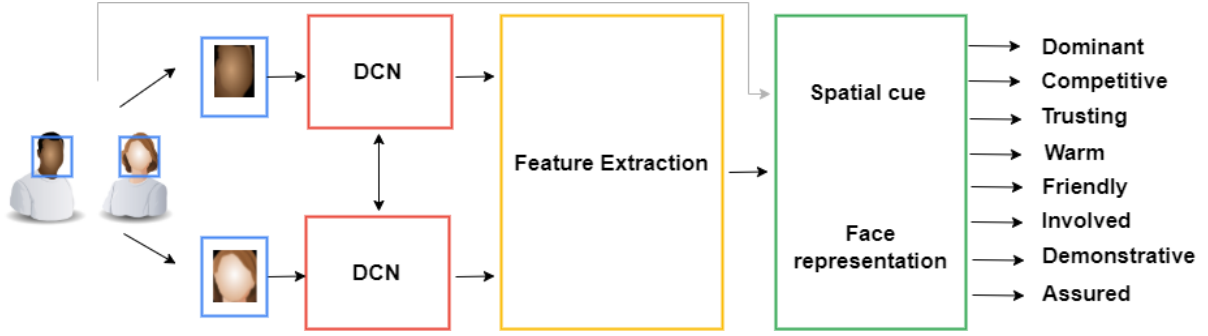


Figure 2.4: The model for interpersonal relation prediction as proposed in [186].

cross-entropy for making a final multinomial prediction.

Social Relation in Videos (SRIV) As mentioned in Section 2.2.2.2, the Social Relation In Videos (SRIV) dataset [91] uses subjective relations and objective relations, from the Asian social psychology theory along with the Kiesler theory, using the eight sub-classes. Other than the method in [91], we are not aware of other approaches reporting results on subjective relations on this dataset. Still, we included it in this discussion as it provides annotation of subjective interactions, and can be of interest for developing approaches addressing this specific task.

2.2.3.2 Fiske’s Theory of Relational Models

The Fiske’s theory of relational models [43] targets the reflection of personal history onto the cognitive individual experiences and differentiates social relations into four parts: Communal sharing, Authority ranking, Equality matching, and Market pricing. This allows dividing the relationships into intimate, non-intimate, no relation, friends, family, couple, professional, commercial. Even though some of these classes are actually objective rather than subjective, e.g., family, couple, other classes are defined based on the personal experience and how this is reflected onto interpersonal relations. Given this, we included this model within subjective relation models.

People in Social Context (PISC) The People in Social Context (PISC) dataset [151] consists of 22,670 images and 76,568 manually annotated labels from 9 types of social relationship (No Relation, Has Relation, Intimate Relation, Non-Intimate Relation, Friends,

Family Members, Couple, Professional, Commercial) and consists of 66 annotated occupation categories. It collects around 40k images containing people from a variety of sources, including Visual Genome, MS-COCO, YFCC100M, Flickr, Instagram, Twitter and commercial search engines (i.e., Google and Bing), using a combination of key words search (i.e., co-worker, people, friends, etc.) and people detector (Faster RCNN).

The proposed Dual-Glance relationship recognition Model [151] has two glances. The first one takes in three inputs, two bounding boxes covering each person and one for the union region insert into three CNNs. In the second glance, they adapted Faster RCNN to process the input image with a Region Proposal Network (RPN) to generate a set of region proposals, the process with a CNN. Next, they allocated attention to each region, and aggregated their outputs to refine the score. Their experiments show 63.2% accuracy on three-relationships (Intimate, Non-Intimate and No relation), and 79.7% on six-relationship (Friends, Family, Couple, Professional, Commercial and No Relation).

2.2.3.3 Others

Here, we separately report datasets for subjective interpersonal relation analysis that do not follow any of the previous models/theories.

Understanding Dyadic Interactions from Video and Audio signals (UDIVA)

The UDIVA dataset [118] includes face-to-face dyadic interaction videos based on free and structured tasks, recorded from different camera positions. It is composed of 90.5h of recordings of dyadic interactions between 147 voluntary participants. It includes socio-demographic (age, gender, ethnicity, occupation, maximum level of education and country of origin), self- and peer-reported personality, internal state (pre- and post-session mood and fatigue), and relationship profiling. Authors propose a novel method for self-reported personality (characterized by the basic Big Five traits (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), or OCEAN, based on self-reported assessments) using R(2+1)D network for the visual chunks and VGGish for audio to extract features, with a visual input (face, local context, and extended context chunks), audio input (raw chunks), and metadata input (both interlocutors' characteristics, and session and dyadic features), to get an OCEAN scores as an output.

The MovieGraphs The MovieGraphs dataset [161] presents 20 relationships such as family (e.g., parent, spouse), friendship/romance (e.g., friend, lover), or work (e.g., boss, coworker). It contains 51 movies and 7,637 video clips annotated with scene label, situation label (topic), description, graph with 8 nodes: character, attributes (age, gender, ethnicity, profession, appearance, mental and emotional states), relationship, interaction (verbal or non-verbal), topic, reason, time stamp. In this work, authors proposed three tasks starting with graph-based situation retrieval to retrieve relevant clips using as nodes character, attribute, relationship, interaction, topic, and reason. Then, interaction ordering is applied, where they used an attention-based decoder RNN. Finally, reason prediction is performed using information about the scene in the form of attributes of each character, their relationship, and an interaction to predict the reason of the interaction.

Human Interaction Images (HII) The Human Interaction Images (HII) dataset [153] includes 10 human interaction classes: boxing-punching, dining, handshaking, highfive, hugging, kicking, kissing, partying, speech and talking. HII contains 1,972 images with 150 images for each class. There can be two, three or more subject sin an images, with a visible facial region for at least one of them. Authors proposed a novel descriptor based on facial regions to recognize the human interactions. The proposed approach started with face detection, then a set of descriptors were used: Histogram of Face Orientations (HFO) for face orientations distribution, Histogram of Face Directions (HFD) for direction frequencies in the images distribution, Distances of Faces (DF) for the location of each face and the distances between them, Circular Histogram of Face Locations (CHFL) for the relative layout of people using a histogram of their locations by fitting a circle to the center of the extracted faces, and Grid Histogram of Face Locations (GHFL) same as CHFL for the spatial layout of the multiple people. Finally, Scene features with GIST descriptors and Bag-of-Words (BoW) were used to detect the scene and deep features with deep learning using convolutional neural networks.

Overall, differently from the objective relations category, the different types of sub-

jective interpersonal relations are way more variegated and complex, making it hard to define clearly separated categories. Despite referring to two main common behavioral models, all the reported datasets slightly differ from each other in terms of data annotation and focus. This makes it complex to gather comprehensive conclusions that are shared among the described approaches. In any case, a very clear difference that can be noted with respect to objective relations is that much more attention and emphasis is put on analyzing faces. Indeed, many of the described approaches use networks that are pre-trained either for face or expression recognition. Still, the less developed literature represents a piece of evidence of the challenging nature of the problem, which is yet far from being solved.

2.3 Prediction of Emotions

Forecasting the future is a topic that has a variety of applications, including action, motion, trajectories or even preference prediction [75, 136, 94, 10]. Understanding temporal dynamics is also fundamental for human action generation [171, 111, 115] or completion [17, 172]. In the literature though, there is a clear absence of works focusing on the problem of *predicting future emotions* in interpersonal interaction videos [55]. In the following, we briefly discuss methods for short-term prediction in generic videos. In particular, recent advancements in short-term video prediction have been significantly influenced by the development of deep predictive coding networks, notably the PredNet architecture in [87]. This hierarchical recurrent neural network, inspired by principles of neuroscience, predicts future frames in a video sequence by forwarding only the deviations to subsequent layers. While PredNet effectively learns robust internal representations of the visual world, its reliance on unsupervised learning limits its applicability in scenarios requiring high precision and detailed contextual understanding.

In video-based person re-identification, the Temporal Complementary Learning Network (TCLNet) in [65] has achieved significant advancements. TCLNet employs Temporal Saliency Erasing (TSE) and Temporal Saliency Boosting (TSB) modules to enhance complementary features from video frames, demonstrating superior performance on major benchmarks. However, its complexity and computational demands can be prohibitive. Its focus on person re-identification may limit its generalizability to other video prediction

tasks.

In [58], authors introduced ICSWM, which integrates physical dynamics within predictive models to enhance interpretability and predictive accuracy. By incorporating physical laws and using a curriculum-based training strategy, ICSWM demonstrated superior localization and prediction of object interactions. Nonetheless, the model’s emphasis on physical dynamics may not capture more abstract or high-level semantic changes in video sequences, which are key issues for our problem. In [78], authors presented an effective approach leveraging multi-level prediction spaces to improve video forecasting. Their method integrates predictions across multiple feature spaces, using recurrent connections and LSTM-based modules, significantly improving performance metrics such as Mean Absolute Error (MAE) and Structural Similarity Index (SSIM). The integration of multi-level prediction spaces adds layers of complexity and computational overhead. The Bidirectional Constraint Network (BCnet) in [64] addresses the challenges of high-dimensionality in video prediction by employing both forward and backward prediction modules. This bidirectional approach effectively reduces motion confusion and enhances prediction accuracy. However, the necessity for bidirectional processing can lead to increased training times and resource requirements, potentially limiting its practical deployment in resource-constrained environments.

2.4 Conclusion

This chapter has provided a detailed exploration of emotions within both individual and interpersonal contexts. In the previous chapter, we have focused on datasets related to individual emotions, typically categorized by the six basic emotions, whereas along this chapter, we have extended that exploration to interpersonal relationships, where emotional dynamics are far more complex. Through our review, we have identified a significant gap in the literature—while many datasets exist for individual emotions, none, to the best of our knowledge, capture a broader range of emotions beyond the basic six in the context of interpersonal relationships.

Moreover, drawing inspiration from Russell’s psychological model of emotions, which advocates for a more nuanced and comprehensive classification, we have recognized the need for a dataset that reflects the diversity of emotions experienced between individuals.

This led to our central contribution: the creation of a new dataset annotated with a wider spectrum of emotions that occur specifically in interpersonal interactions. This dataset is designed to provide deeper insights into the emotional dynamics between two people, moving beyond the basic emotion categories. Thus, the next chapter will be dedicated to introducing this novel dataset, detailing its structure, the range of emotions it covers, and the process of annotation. This new dataset aims to bridge a critical gap in emotion research, enabling future studies to model and analyze the complexity of emotional interactions in a more granular and realistic manner.

Chapter 3

The Florence Interpersonal Relations with Multi-Emotions Dataset

This chapter presents the IMEmo dataset, a novel resource for capturing emotional expressions in dyadic interactions. Unlike traditional datasets, IMEmo offers annotations for 16 distinct emotions, addressing a gap in existing emotion recognition resources. The chapter details the dataset’s creation, structure, and annotation process, highlighting its potential to advance research in multi-emotion recognition and interpersonal dynamics.

3.1 Introduction

In this chapter, we introduce a novel dataset: the Interpersonal Relations with Multi-Emotions (IMEmo) dataset [56], a new and comprehensive dataset designed to capture a wide range of emotional expressions and transitions in interpersonal interactions. The IMEmo dataset consists of video clips where two individuals engage in interactions, reflecting a diverse array of emotional states. Unlike traditional datasets focused on recognizing basic emotions of an individual, our dataset emphasizes the complexity of multi-emotion recognition in dyadic scenarios, where both subjects’ emotions influence each other over time. A key innovation of the IMEmo dataset is the annotation of 16 distinct emotions, inspired by our earlier research on emotion classification. Through this work, we identified the need for a dataset that could capture the full spectrum of emotions ex-

pressed during interpersonal interactions. To our knowledge, no existing dataset provides such detailed annotation of emotions in these social contexts, which makes the IMEmo dataset a significant contribution to the field.

This chapter details the creation of the IMEmo dataset, including its structure, the annotation process for the 16 proposed emotions, and the challenges posed by the nature of multi-emotion prediction in dyadic interactions. Additionally, we address how this dataset fills a critical gap by moving beyond the recognition of basic emotions and focusing on the nuanced emotional exchanges between interacting individuals.

3.2 The Interpersonal Relations with Multi-Emotions (IMEmo) Dataset

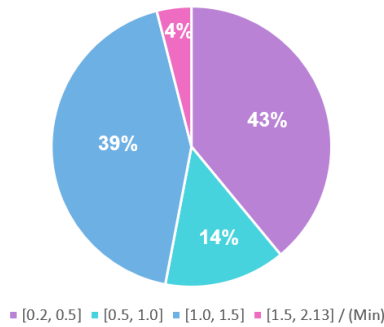
In the last few years, an increasing number of datasets for facial expression analysis have been collected and made publicly available [96, 138, 148, 184]. Most of them include images and/or videos each containing one subject [77, 80, 148]. Annotations include the six categorical emotions, while the dimensional annotation with VA values is reported in some cases [105, 177]. We proposed a new dataset for facial expression analysis that includes video clips extracted from movies, where two subjects interact each with the other showing a variety of emotions. Being it characterized by the collection of **Interpersonal relations with Multi-Emotions**, we refer to our dataset as **IMEmo** for short. We want to move beyond the common scenario where the task is that of recognizing one of the six basic emotions of an individual in an image or video using facial analysis. Differently, we target the problem of recognizing the emotion of two individuals, when interacting with each other.

As shown in the Table 3.1, some datasets entered this investigation direction, but following different objectives. For example, the datasets in [15, 16, 117] include dyadic interactions but are limited in the number of participants, context annotations, language, or purpose. One of the aims of our research is to investigate the potential of emotion recognition and its role in understanding how interaction affects emotion changes. As an additional design choice, we decided to use video clips taken from commercial movies in public use. This is relevant to solve two problems that commonly arise when collecting human face datasets: *(i)* by construction, we use videos where emotions are acted by

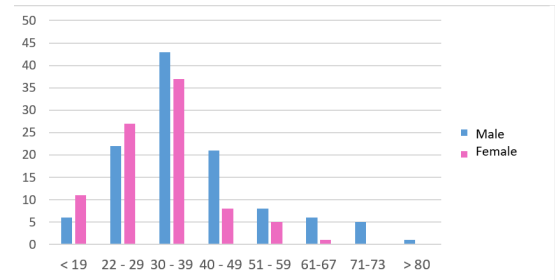
professional actors interacting with each other. This makes unnecessary defining specific laboratory setups, subjects enrollment, and allows a reasonable confidence on the extent at which expressions are reliably executed; *(ii)* using video materials which are in a common domain, we remove concerns about privacy and protection of personal data. The IMemo dataset will be made publicly available to cultivate a collaborative environment, promoting data sharing and cross-disciplinary cooperation. In the remaining part of this section, we describe the main features of our dataset.

3.2.1 Main Dataset Statistics

First, we manually extracted 100 *video clips* from romantic and dramatic genre movies that are accessible in the public Web domain. Each clip includes two persons, who are in an engaged interaction, either talking, laughing, etc. This process involved two steps: first, we identified movies in the two used categories with public and free access on the Web; then, for each movie, we manually selected one clip or more, where an interpersonal relation is present. Overall, the duration of the video clips is of 79.3 minutes, with individual clips lasting from 0.20 to 2.13 minutes. More specifically, 40 videos have duration from 0.20 to 0.50 minutes, 14 videos from 0.51 to 1.00, 45 videos from 1.01 to 1.55, and one video lasts 2.13 minutes (see Figure 3.1a).



(a) Duration of individual clips of IMemo dataset.



(b) Age statistics for females and males by age group.

Figure 3.1: Statistics of clips duration (a), and age plus gender distribution in IMemo (b).

We also targeted a multi-language dataset, with emotion annotations that are independent from the language. In particular, the collected video clips are available in English (80), Indian (7), French (3), Italian (2), and sign language (6), while there are also some

Table 3.1: Available interpersonal relation recognition datasets in comparison to ours.

Dataset	# subjects	Focus	Annotation	Language
IEMOCAP [15]	10	Emotion recognition	Emotions, transcripts, turn-taking	English
CreativeIT [103]	16	Emotion recognition	Transcripts, speech features, emotion	English
MSP-IMPROV [16]	12	Emotion recognition	Turn-taking, emotion	English
UDIVA [117]	147	Social interaction analysis	Audiovisual, heart rate, personality scores, socio-demographics, mood, fatigue, relationship type	Spanish, Catalan, English
MELD [129]	6	Social interaction analysis	Emotions, sentiment	English
IMEmo	163	Emotion recognition, interpersonal relation recognition, emotion prediction	Age, gender, emotions, social relationships, actions	English, Indian, French, Italian, sign language, silent scene

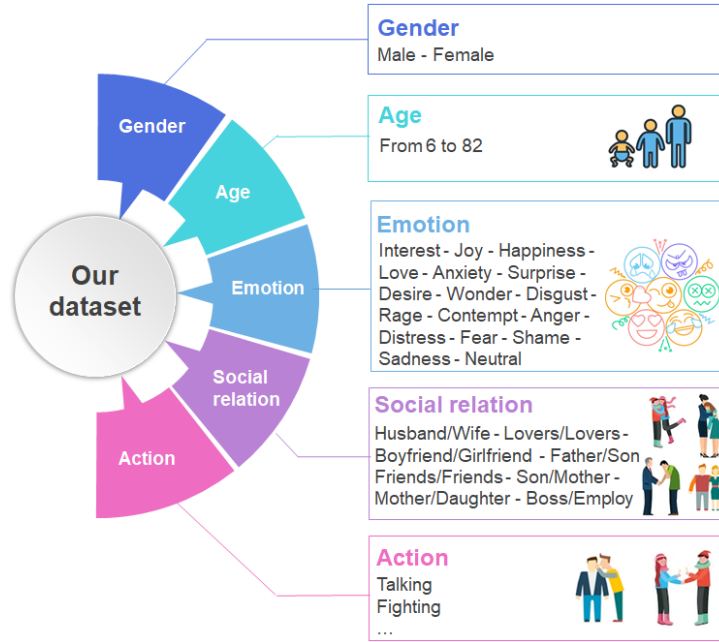


Figure 3.2: Overview of the annotations included in the IMEmo dataset.

silent scenes (2). The diversity of languages aims to highlight that emotions are universal, and that emotions are recognizable despite whether or not the dialogue is understood.

3.2.2 Data Pre-processing

For every video in the dataset, we provide both the raw data and a pre-processed version for the sake of fair comparison across methods that will use our dataset. In particular, we obtained the face crop by using a customized code sourced from GitHub [110], tailored to our specific dataset. This approach allowed for the precise isolation and identification of facial regions within our images, crucial for subsequent analysis. The extracted faces were curated to ensure accuracy and consistency across the dataset. By employing this adapted code, we were able to streamline the process of facial recognition and analysis, contributing to the robustness of our experiments.

3.2.3 Annotations

The IMEmo dataset has been annotated both at *clip* and *shot* level. A summary of the annotations is shown in Figure 3.2.

3.2.3.1 Clip-level Annotations

Clips were manually annotated while extracting information of interest either from the metadata of the related movie, or from public Web sources (e.g., age of the actors, relationship between two characters in a movie, etc.).

Relationship annotation: In general, relationships in human behavior datasets are classified according to *objective terms* or *subjective behavior* [55]. The relationship label in our dataset is extracted from the movie metadata, and it is based on the relationship between the two characters played in the movie. Hence, the relationships in our dataset are classified as *objective relations*, because they do not depend on a specific behavior or emotional state. Examples of objective relations are kinship, e.g., *husband/wife* or working roles, e.g., *boss/employee*, *colleague/colleague*, as summarized in Table 3.2.

Action annotation: In the work of Morshed [107], a taxonomy for human action recognition was proposed. It divides activities into *atomic actions*, *behavior*, *interaction* and *group activities*. Considering that our IMEmo dataset includes interpersonal relation clips, we decided to annotate the actions using the *interaction* category that comprises, for example, *playing*, *hugging*, *dancing*, etc. For the clips annotated so far, we actually used only two categories: *fighting* (20%) and *talking* (80%). Despite the limited range of variations, we decided to keep this annotation for possible future extensions.

3.2.3.2 Shot-level Annotations

As introduced above, a set of shots was manually extracted from each video clip in the dataset based on the camera changes from one character to the other in the clip. With this subdivision, shots can include one of the two interacting characters or both of them in the case the camera view includes the two. Shots have been annotated with the intervention of voluntary users. This represented the most important and time consuming part of our annotation process that allowed us to derive a fine level emotion annotation of the initial clips. In particular, since *shots* have been selected so as to include homogeneous emotion content, the shot level annotation is extended to all the frames included in the shot.

Table 3.2: Objective relation categories included in our IMemo dataset.

Objective relation	Classification	#clips
Working	Boss / Employee	5
	Colleague / Colleague	4
	Teacher / Student	4
	Psychologist / Patient	4
	Bartender / Client	1
	Social worker / Child	2
Kinship	Husband / Wife	17
	Girlfriend / Boyfriend	7
	Lover / Lover	20
	Brother / Brother	1
	Daughter / Father	3
	Daughter / Mother	2
	Son / Father	5
	Son / Mother	2
	GrandFather / Granddaughter	2
	Husband / Mother-in-law	1
Other	Friend / Friend	13
	Criminal / Lawyer	2
	ex-girlfriend / ex-boyfriend	1
	ex-wife / ex-husband	1
	Santa Claus / Kid	1
	Seducer / Victim	1
	Serial killer / FBI trainee	1

Emotion annotation We annotated our dataset with *16 emotions*, which represent the main annotations, the standard *6 basic emotions*, and *sentiment*, as follows:

16 emotions: In order to annotate our dataset with emotions, we started with Ortony’s conclusion about basic emotions [114]: *“The problem is not just that investigators cannot find basic emotions; we do not even have, and probably cannot have, a satisfactory criterion of basicity that is generally acceptable to emotion theorists”*. In that work, 14 groups were identified in the literature to characterize the “basic” emotions:

- Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness [4];
- Anger, disgust, fear, joy, sadness, surprise [33];
- Desire, happiness, interest, surprise, wonder, sorrow [45];
- Rage and terror, anxiety, joy [52];
- Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise [69];
- Fear, grief, love, rage [70];
- Anger, disgust, elation, fear, subjection, tender-emotion, wonder [97];
- Pain, pleasure [108];
- Anger, disgust, anxiety, happiness, sadness [112];
- Expectancy, fear, rage, panic [119];
- Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise [127];
- Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise [155];
- Fear, love, rage [168];
- Happiness, sadness [169].

Overall, these groups with different emotion categories include 81 terms. These categories stand for emotions regarding relationships, tendencies, universal facial expressions, actions, hardwired, bodily involvement, biological processes, density of neural firing and attribution independent. By combining all the previously described emotions, in the IMemo dataset, we decided to work with the “basic” emotions that appear with highest frequency in these groups. For instance, 7 out of the 14 categories include “anger”;

therefore, we selected “anger” as one of the chosen emotions for our terms. Following this approach, we obtained 16 terms: anger, desire, fear, love, sadness, disgust, joy, surprise, happiness, interest, wonder, rage, anxiety, contempt, distress, and shame. In addition, the neutral term was also used. We note that 9 of the terms are connoted by a negative mood (i.e., angry, anxiety, contempt, disgust, distress, fear, rage, sadness, shame), while the remaining 7 emotion terms are considered as positive ones (i.e., desire, happy, interest, joy, love, surprise, wonder). These emotion terms were used to annotate the shots in our dataset, also keeping the subdivision between positive and negative.

Table 3.3 shows the frequency of the 16 emotions of our proposed categorization at both clip and frame levels. Using such a number of emotions, and considering also the unbalanced nature of the data poses additional challenges to the emotion recognition task. As an example, in Figure 3.3 we display a distribution of emotion annotations at clip level for a few selected clips.

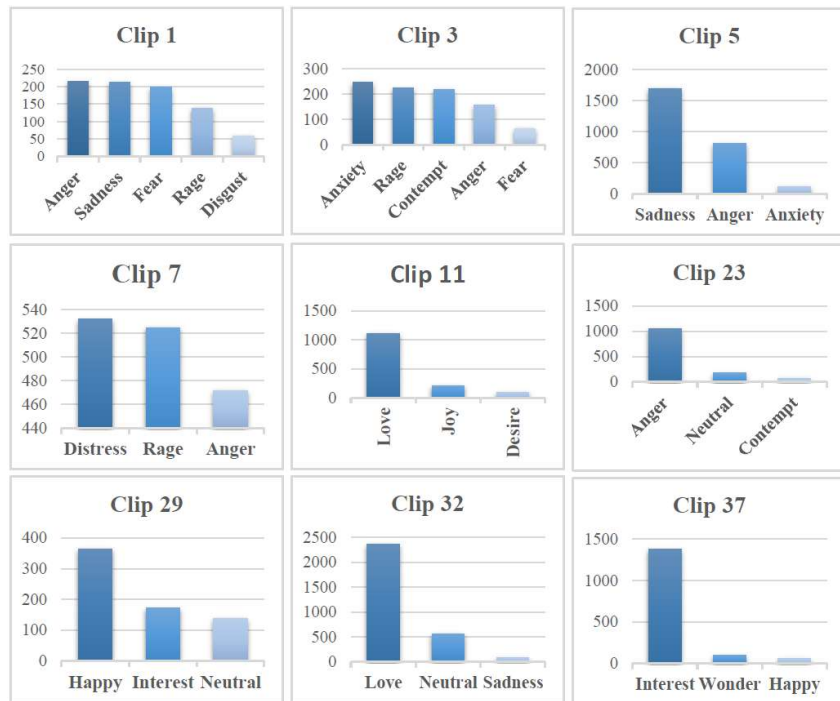


Figure 3.3: Plots of the emotions frequency at clip level for a few selected clips.

6 emotions: Since our dataset was labeled with 16 emotions, which represent a super-set of the Ekman’s six universal emotions (anger, disgust, fear, happy, sad, and surprise), we opted to label each shot also with them. This allowed us to make a more thorough comparison with previous datasets that adopted such emotion annotation. As

Table 3.3: Distribution of emotion annotations by number of frames and by number of video clips.

Emotion	# frames	# video clips
Sadness	37693	60
Anger	25410	47
Interest	19235	51
Neutral	15355	51
Happy	9586	32
Anxiety	7565	41
Love	7216	18
Contempt	6824	26
Rage	5830	18
Fear	4380	15
Wonder	4235	29
Surprise	4020	20
Shame	2385	10
Disgust	2060	13
Distress	1751	13
Desire	1406	8
Joy	463	3

shown in Table 3.4, for each shot, we mapped the 16 emotions back to the 6 basic emotion labels as follows: happy (includes happy, interest, joy, love, and desire), fear (includes fear, distress, and anxiety), surprise (includes surprise and wonder), disgust (includes disgust and contempt), anger (includes anger and rage), and sadness (includes sadness and shame).

Sentiment: During the process of annotating the dataset, sentiment was present for the annotators, not as a choice but as a division of emotions (see the user annotation panel of the Web interface in Figure 3.5c). The categorization of positive, negative, and neutral sentiment classes is given in Table 3.4. In the sentiment classes, we considered

angry, anxiety, contempt, disgust, distress, fear, rage, sadness, shame as negative, while desire, happy, interest, joy, love, surprise, wonder were considered as positive, and neutral as neutral sentiment-bearing class. For the 6 emotions, the division into positive and negative inherits from the 16 ones. This is illustrated in Figure 3.4 that also shows the occurrences for individual emotions.

Table 3.4: Distribution of emotion annotations according to the division of 16 emotions, 6 emotions and sentiment.

16 emotions	6 emotions	Sentiment
Happy	Happy	Positive
Interest		
Love		
Desire		
Joy		
Surprise	Surprise	Negative
Wonder		
Fear	Fear	
Distress		
Anxiety		
Disgust	Disgust	
Contempt		
Anger	Anger	
Rage		
Sadness	Sadness	
Shame		
Neutral	Neutral	Neutral

Valence/Arousal annotation: We provide a valence/arousal (VA) annotation for each clip, based on the work of Russell [140]. In particular, Russell identified 151 fine-

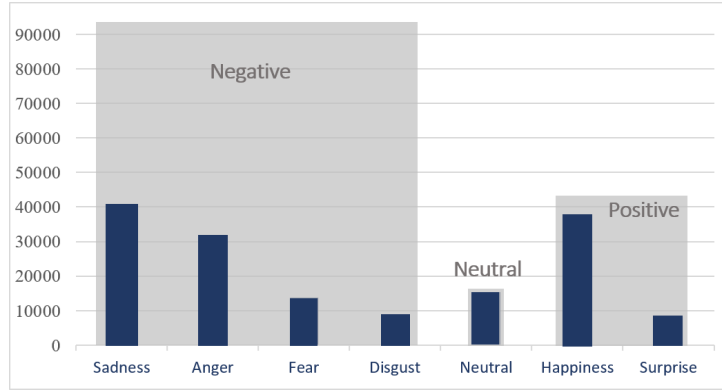


Figure 3.4: Emotion frequencies for the mapping of 6 emotions and sentiment.

grained emotional terms, while also providing a map from each of them and a distribution in the dimensional VA plane. For example, the emotion term “excited” was associated with the values of 0.62 valence and 0.75 arousal, with a dispersion range of 0.25 and 0.2, respectively. Doing this mapping for all the emotion terms resulted in a complete coverage of the VA plane, though the dispersion regions associated to each term largely overlap with each other. In the annotation, we exploited this mapping to automatically provide a VA value for each shot and, in turn, for each frame in the shot. To this end, we considered the emotion term with the highest frequency (majority voting) as the emotion associated to the shot and mapped it to its corresponding point on the VA plane: valence and arousal values for that point were then used for the shot and all its frames.

Annotation website The annotations detailed above were given by human annotators using a Web tool as shown in Figure 3.5. We identify three main steps in the annotation process:

1. A user can register into the annotation website so that he/she can recover his/her annotations at any time. Only anonymous information about the annotator is stored (e.g., age, gender). A brief instruction and explanation page is also presented to the annotator at this stage. Then, the annotator can select a clip from a grid that groups clips with same language or genre (see Figure 3.5a). The user can otherwise select a different genre or language group;
2. Once a clip is selected, the annotator is redirected to a web page where the entire video clip can be watched on the top of the page. The shots—each constituted by

a small number of frames—that compose the clip are displayed as animated GIF images in the bottom part of the page. The user annotation takes place at shot level by following the “annotation” button as shown in Figure 3.5b;

3. The annotation panel for emotion selection is illustrated in Figure 3.5c. In this panel, the 16 fine emotion categories plus the neutral one are presented to the user following a daisy-like visualization. Icons for positive and negative emotions are visualized separately; in both the cases, an icon for the neutral emotion is also shown in the center of the daisy-like visualization.

Annotators The IMEmo dataset was annotated by 20 subjects of different cultures and educational level, with an age range going from 22 to 54 years old. Typically, annotations are carried out by experts, but in our dataset, we found it necessary to involve multiple users. Given the deeply personal nature of emotional perception, we sought agreement across a diverse set of perspectives. Each clip was annotated by at least 3 subjects, with 10 clips as average number of annotations. There were 5 annotators on 63 clips, 4 annotators on 2 clips, and 3 annotators on 35 clips. Each user was requested to annotate 5 clips as a minimum. The user annotation can be performed in different sessions. On average, the time effort lasted about 5 minutes per clip. As an example, the shots composing the video clip #94 and its annotations from 5 users are given in Figure 3.6. In (a), the initial frame for the 8 shots of person 1, and the 7 shots of person 2 are shown. In (b), it is then visualized the annotations from the 5 users at the shot level, separately for person 1 (first 8 cases) and person 2 (second 7 cases). The table included in (c) of the same figure reports the final annotation of the shot for the two characters based on majority voting.

Annotations In Table 3.5, we outline the distribution of annotations based on the number of annotators. Our dataset consists of 100 clips, with 1,232 shots annotated by a minimum of 3 annotators (805 shots), up to a maximum of 5 annotators (411 shots), with 16 shots annotated by 4 annotators. Among shots annotated by 5 annotators, 7.79% displayed unanimous agreement in emotion annotation, while 75.42% were annotated by a majority consensus of 4 or 3 out of 5 annotators. The remaining shots received

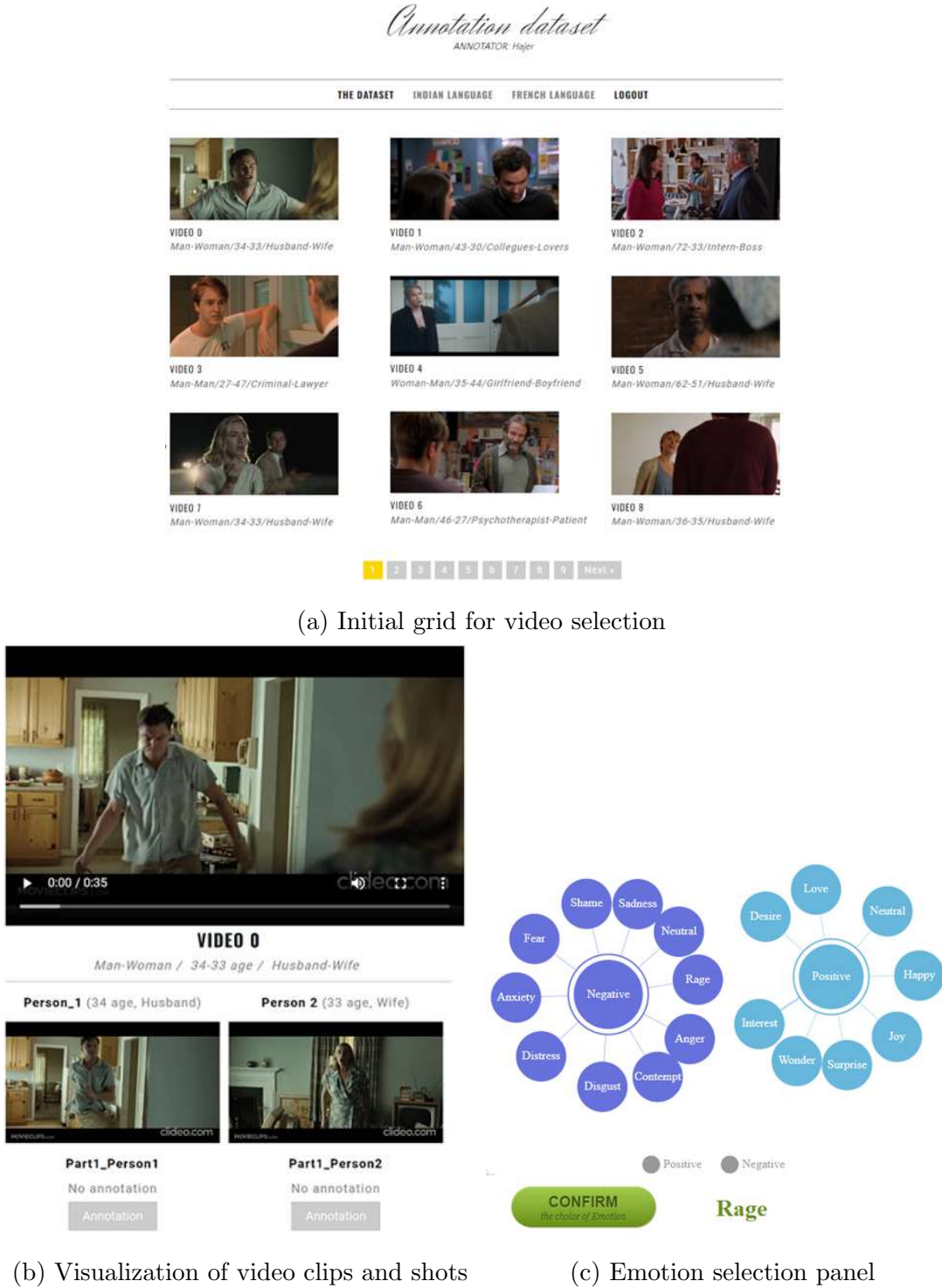
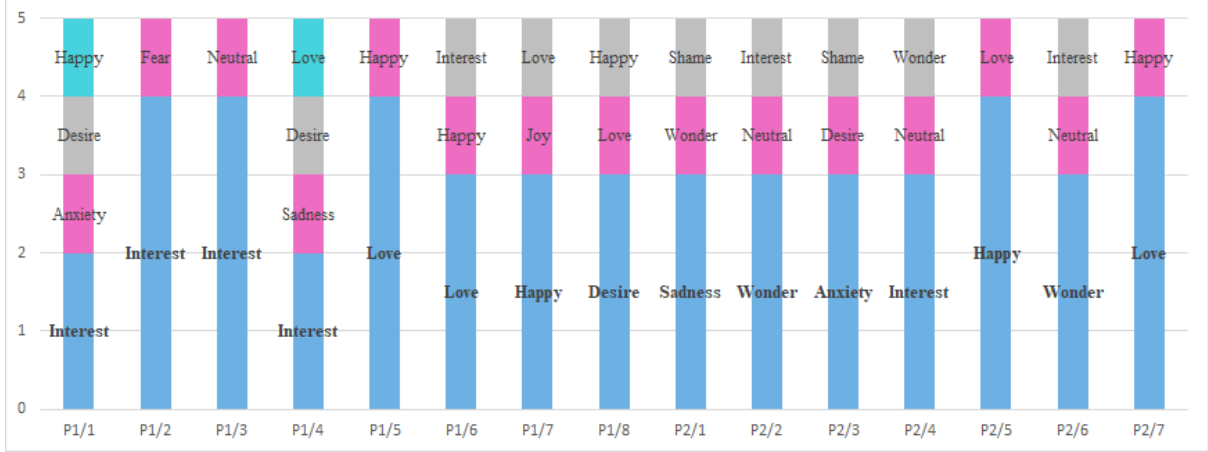


Figure 3.5: Our video clip annotation website: (a) Each user starts the annotation process by selecting a video clip from a grid; (b) The selected clip and its shots are then shown to the human annotator in a same page; (c) The annotation panel showing the 9 negative (plus neutral) and the 7 positive (plus neutral) fine grained emotion categories derived according to the work of Ortony [114] are finally displayed for user selection.



(a) Subdivision of clip #94 into 8 shots for person 1 and 7 shots for person 2.



(b) Annotations of the shots of clip #94 for person 1 (from P1/1 to P1/8) and for person 2 (from P2/1 to P2/7).

(c) Final annotation of the shots.

# Person	Shot1	Shot2	Shot3	Shot4	Shot5	Shot6	Shot7	Shot8
1	Interest	Interest	Interest	Interest	Love	Love	Happy	Desire
2	Sadness	Wonder	Anxiety	Interest	Happy	Wonder	Love	

Figure 3.6: Figure illustrating the process of annotation between different annotators for clip #94. The clip is subdivided into 8 shots for person 1 and 7 shots for person 2. Annotations for each shot are provided, labeled for person 1 (P1/1 to P1/8) and person 2 (P2/1 to P2/7), highlighting the variation in annotations. Additionally, table (c) presents the final annotations of the shots, indicating the most voted emotion in cases of discrepancy between annotators.

annotations from less than half of the annotators.

Regarding shots annotated by 3 annotators, 30.43% exhibited unanimous agreement in emotion annotation among all annotators, constituting 245 out of 805 shots. Additionally, 69.57% of shots were annotated with the same emotion agreed upon by 2 annotators, while a minority of shots had only one annotator presenting a different emotion. Given that the dataset was annotated by multiple annotators, and the selection of the emotion

annotation for each shot was determined by majority agreement, we offer a quantification of the uncertainty associated with each emotion annotation. The uncertainty value μ for each emotion label e in each shot S_j in a clip was computed by assessing the probability p_e of that emotion e , as follows:

$$p_e = \frac{\text{\#annotations for emotion } e}{\text{Total number of annotation}}, \quad (3.1)$$

$$\mu(S_e^j) = - \left(\sum_e p_e \log_2(p_e) \right), \quad (3.2)$$

The overall uncertainty for a clip is instead computed as:

$$\mathcal{E} = \frac{\sum_j \mu(S_j)}{\sum_j S_j}. \quad (3.3)$$

For instance, shot 1 of video 1 was labeled “rage” by four annotators and “anger” by one annotator, so we provide the secondary annotation “anger” as an uncertain annotation since it was chosen by only one annotator. As another example, shot 1 in video 3 was annotated by two annotators with “contempt,” serving as the final and primary annotation. The uncertain annotations include “rage”, “disgust”, and “anger”, with each being chosen by one of the annotators. Finally, in the case of video 4, shot 3, “anger” was annotated unanimously by all annotators, establishing it as the definitive and sole annotation for that shot.

Given that emotions are subjective and can vary between individuals, there exist minor disagreements, particularly between positive and negative emotions. In some cases, two annotators may annotate the same shot with opposing emotions. For instance, in shot 8 of video 6, three annotators labeled it with “sadness,” while the remaining two annotators labeled it with “love” and “interest”. While efforts were made to achieve consensus in labeling, variations in individual interpretation persisted, leading to instances of conflicting annotations. Hence, recognizing the subjective nature of emotion perception, we have included, alongside the primary annotations, supplementary annotations

Table 3.5: Distribution of annotations according to the number of annotators. The columns represent, respectively: Annotators: number of *Annotators* with their percentage; All: percentage of cases where *All* the annotators agreed on the annotation; More half: when *More* than *half* of the annotators agreed on the annotation, and Less half: when *Less* than *half* of the annotators agreed on the annotation; Uncertainty: value of $\mathcal{E}$ as defined in Eq. (3.3)

Number of clips : 100				
Number of shots : 1232				
Number of shots annotated by 5 : 411				
Number of shots annotated by 4 : 16				
Number of shots annotated by 3 : 805				
Annotators	All	More half	Less half	Uncertainty $\mathcal{E}$
5 (35%)	7.79%	75.42%	16.79%	<i>4.00%</i>
4 (2%)	0 %	43.75%	56.25%	<i>0.14%</i>
3 (63%)	30.43%	69.57%	0 %	<i>1.88%</i>

representing alternative interpretations. These additional annotations serve to provide researchers with a comprehensive dataset that accommodates the diverse range of emotional responses.

Furthermore, it is worth noting that the annotation of emotions does not always match exactly with what the annotator perceives as the actor’s emotion; occasionally, the annotator’s emotion imitates the actor’s feelings, which is why for some clips, we can find annotators describing it as sad, while others describe it as happy, even though these are completely different emotions. This is because the annotators’ emotions often took place and mimic the actor’s emotions.

3.2.3.3 Potential Investigation Scenarios

We designed our dataset while having in mind a set of investigations and analysis that are not possible with datasets that are currently available in the literature. In particular, we believe that our dataset can be potentially used in two main scenarios (but more can be defined as well):

- *Emotion recognition*: this is actually the most “traditional” way our dataset can be employed with. In the simplest setting, emotions of the two interacting individuals could be considered and analyzed independently. Compared to existing datasets like [77, 121, 165], we provide an annotation with a larger number of emotion categories at shot level. In addition, this analysis can be performed while two subjects are interacting in a variety of different contexts and with different objective relations. This is also a unique feature of our collected data;
- *Estimation of the interpersonal relation*: looking at the emotion of the two individuals in a video clip, the idea is that of estimating an emotion label of their interpersonal relation. This is similar to the scenario proposed in other interpersonal relation datasets [117, 161].

3.2.4 Applications of the IMEmo Dataset

The IMEmo dataset, with its comprehensive annotations of 16 distinct emotions, demographic information, and interpersonal interaction dynamics, provides several unique opportunities for researchers and practitioners across various domains. In fact, the IMEmo dataset opens doors to numerous practical applications that extend beyond academic research, impacting fields such as healthcare, security, entertainment, and education. By enabling a better understanding of emotional transitions in real-world interactions, this dataset can serve as a foundational resource for building emotionally intelligent systems that enhance human experiences across various domains. Below are potential applications and future research avenues that could benefit from this dataset.

Human-Computer Interaction (HCI) and Social Robotics: The IMEmo dataset can be used to train advanced emotion recognition systems in social robots, making them more responsive to human emotional states. For instance, robots can be programmed to adapt their behavior based on the emotional exchanges between users during a conversation, creating more intuitive and empathetic interactions. For example, a virtual assistant or a service robot in healthcare settings could monitor emotional transitions in patients to detect distress, frustration, or sadness and offer appropriate support or interventions (e.g., contacting healthcare providers or suggesting relaxation techniques).

Mental Health Monitoring: IMEmo’s detailed emotional annotations could be employed to develop emotion-based diagnostic tools for mental health practitioners. The dataset’s ability to track emotional transitions over time in real-world interactions allows for the development of models that recognize patterns related to mental health issues like depression, anxiety, or emotional dysregulation.

Improving Customer Experience in Marketing and Retail: Retail environments and marketing campaigns increasingly rely on understanding customer emotions to personalize user experiences. The IMEmo dataset can aid in developing systems that track customer reactions during their interactions with sales representatives or when viewing advertisements.

Enhanced Video Surveillance Systems: In security and surveillance, understanding emotions can play a critical role in identifying suspicious or abnormal behavior. By leveraging IMEmo’s dynamic emotion tracking, security systems can be enhanced to recognize changes in emotional states that may indicate potential security threats, such as fear or anger, in crowded environments.

Interpersonal Communication Research: The IMEmo dataset provides a rich resource for researchers studying how emotions evolve during interpersonal interactions. By analyzing the emotional exchanges between individuals, researchers can develop new theories on how emotions influence communication dynamics, conflict resolution, or team collaboration.

Development of Emotionally Intelligent AI in Entertainment: In the entertainment industry, AI systems that can recognize and predict emotions in real time are increasingly being used to enhance user engagement in gaming and virtual environments.

IMemo’s detailed annotations can be used to train these systems to create more emotionally responsive characters in video games or interactive movies.

Educational Tools for Emotional Intelligence: Emotional intelligence training tools, particularly for children or individuals with social challenges (e.g., individuals with autism), can benefit from datasets like IMemo. These tools can help users better understand emotional cues and responses in social interactions. For example, an educational app could analyze a user’s interactions and provide feedback on recognizing and responding to emotional changes in others, using video examples from IMemo to illustrate common emotional transitions in conversations.

3.3 Experiments and Results

Our objective here is to perform a baseline frame-by-frame analysis of our data and verify how challenging it is estimating emotions of individual subjects in our shots (i.e., first application scenario as proposed in Section 3.2.3.3). In doing so, we are also interested in comparing results obtained for our dataset with respect to those scored in the same task for other benchmarks that are in common use in the facial expression recognition community. Details are given below.

3.3.1 Reference Architecture and Compared Datasets

In our experiments, we deployed a variety of diverse neural network architectures for emotion classification:

Convolutional Neural Network (CNN): As a first step, we identified a CNN-based emotion recognition architecture, which is sufficiently simple and versatile, being robust at the same time, to be trained for different datasets and number of emotion classes. Our choice was for the CNN proposed in [5].

Distract your Attention Network (DAN): DAN [170] includes three key components: a Feature Clustering Network (FCN) extracts robust features by adopting a large-margin

learning objective to maximize class separability; a Multi-head cross Attention Network (MAN) instantiates a number of attention heads to simultaneously attend to multiple facial areas and build attention maps on these regions; finally, an Attention Fusion Network (AFN) distracts these attentions to multiple locations before fusing the attention maps to a comprehensive one.

Dual-Direction Attention Mixed Feature Network (DDAMFN): an innovative model designed for facial expression recognition (FER). DDAMFN [183] integrates the Mixed Feature Network (MFN) as its backbone, leveraging mixed-size kernels for robust feature extraction, and the Dual-Direction Attention Network (DDAN) to capture long range dependencies effectively.

In addition to our IMEmo dataset, the other benchmarks that we considered in the evaluation are the MELD [129], AffectNet [105] and FER-2013 [51] datasets.

IMEmo: The IMEmo dataset was collected from 67 movies including a total number of 163 actors. In the training set we used 139 actors and in the validation set 33, with only 9 identities shared between the two sets. The shots for these actors contained all the 16 expressions of our categorical representation. There are 155,414 frames in total in the 100 videos clips of our dataset. The dataset was split into a training set, which contained 127,644 frames and represented 83% of the data, and a validation set, comprising 27,770 frames representing 17% of the data. The emotion frequencies based on our emotion annotations for the training and validation sets are provided in Figure 3.7

MELD: Contains 13,000 utterances from 1,433 dialogues from the TV-series Friends. Annotated by three annotators with emotion (joy, sadness, fear, anger, surprise, and disgust plus neutral and non-Neutral), and sentiment (positive, negative and neutral) labels, and encompasses audio, visual, and textual modalities.

AffectNet: A total of 450,000 images collected from Internet were annotated by 12 experts to label the face using both the categorical (discrete) and dimensional (continuous using valence/arousal) models. Each image was annotated by one annotator. Eight discrete categories were defined including the six basic, plus *neutral*, and *contempt*; in

addition to these, the *none*, *uncertain*, and *non-face* classes were considered. The “none of the eight emotions” category includes the expression/emotions such as sleepy, bored, tired, seducing, confused, ashamed, focused, etc., that could not be assigned by annotators to any of the six basic emotions, contempt or neutral. If the annotators were uncertain about any of the facial expressions, images were tagged as uncertain. The dataset is imbalanced with about 46% of happy images and 25% of images showing a neutral expression. The other categories ranged from $\sim 9\%$ for sad and anger to $\sim 2\%$ for disgust and contempt. A subset of 36,000 images were annotated by two annotators and reserved for the test set. Validation was performed on a randomly selected set of five hundred samples of each category, while the remaining images were used for training. The average image resolution of faces is 425×425 with standard deviation of 349×349 pixels.

FER-2013: The database was introduced in the ICML 2013 Challenges in Representation Learning [51]. The database was created by collecting Internet images that capture the six basic expressions plus the neutral one. Images were resized to 48×48 pixels and converted to gray-scale. Human annotators rejected incorrectly labeled images, corrected the cropping if necessary, and filtered out some duplicate images. The resulting database contains 35,887 images most of which are a set in natural environments. The *disgust* expression has the fewest images (~ 600), while other labels have nearly 5,000 samples each. The training set consists of 28,709 examples. The test set, consists of 3,589 examples.

3.3.2 Emotion Recognition Results

In a first experiment, we compared the classification accuracy for our dataset with that of MELD, AffectNet, and FER-2013 for both sentiment and the six basic emotions.

The networks, namely CNN [5], DAN [170], and DDAMFN [183], were trained separately on each dataset using the suggested set for training and validation. The results are reported in Table 3.6, reveal notable insights into emotion classification across diverse datasets. While the emotion recognition accuracy scores achieved by the image datasets AffectNet and FER-2013 are comparable, the values drop considerably in video datasets, including both the MELD dataset and our own dataset. Specifically, when examining

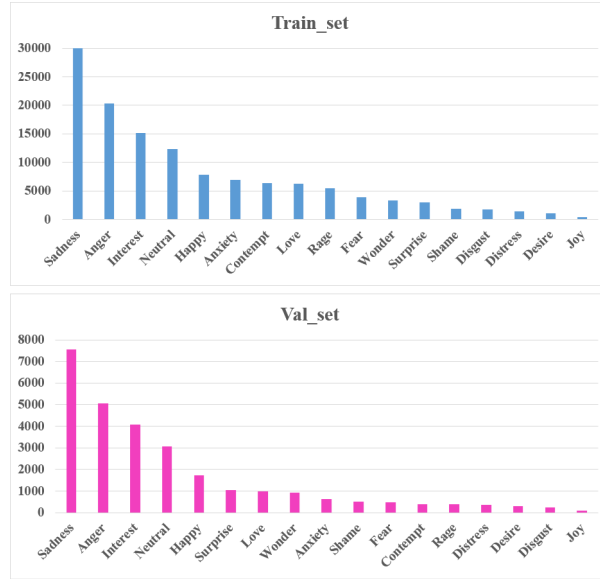


Figure 3.7: Frequencies of the 16 emotion terms plus neutral in the used training and validation sets.

the performance on the MELD dataset, we find a close alignment with our dataset. For instance, sentiment using the CNN network attains an accuracy of 59% for our dataset compared to MELD’s 43%. Similarly, for the classification of the six basic emotions, our dataset achieves 32% accuracy, while MELD achieves 39%. Further comparisons with image datasets AffectNet and Fer2013 underscore significant differences in performance. AffectNet achieves accuracy of 79% in positive and negative sentiment, consistent with Fer2013, but outperforms Fer2013 in classifying the six basic emotions, achieving 73% accuracy compared to Fer2013’s 66% accuracy using the CNN network. Similarly, with the DAN network, AffectNet achieves 65% accuracy for sentiment, while Fer2013 achieves 77%. Our dataset attains 50% accuracy for sentiment classification and MELD achieves 43%. However, for the classification of the six basic emotions, AffectNet demonstrates better performance with 65% accuracy compared to Fer2013’s 56%. Our dataset achieves 25% accuracy for the classification of the six basic emotions and MELD achieves 21%. Lastly, employing the DDAMFN network, AffectNet achieves a sentiment accuracy of 67%, while Fer2013 achieves 78%. Our dataset reaches 47% accuracy for sentiment classification and MELD 38%. For the classification of the six basic emotions, both datasets demonstrate comparable performance, with AffectNet and Fer2013 achieving 63% accuracy. Our dataset achieves 34% accuracy and MELD achieves 18% respectively for the classification of the six basic emotions.

This can be explained by the imbalanced distribution of the classes, as well as by the quality of the images in terms of resolution, illumination condition, pose of the head, movement, etc. This can be observed in Figure 3.8 for some frames in our dataset. This also evidences that the emotion classification difficulty on our data is greater than that of state-of-the-art benchmarks. Furthermore, our findings highlight a notable disparity in emotion classification difficulty between video datasets and image datasets, indicating that the temporal dimension inherent in video data and the domain shift add a layer of complexity beyond that encountered in state-of-the-art benchmarks based on static images. A similar finding was reported in [42] for the face recognition task.

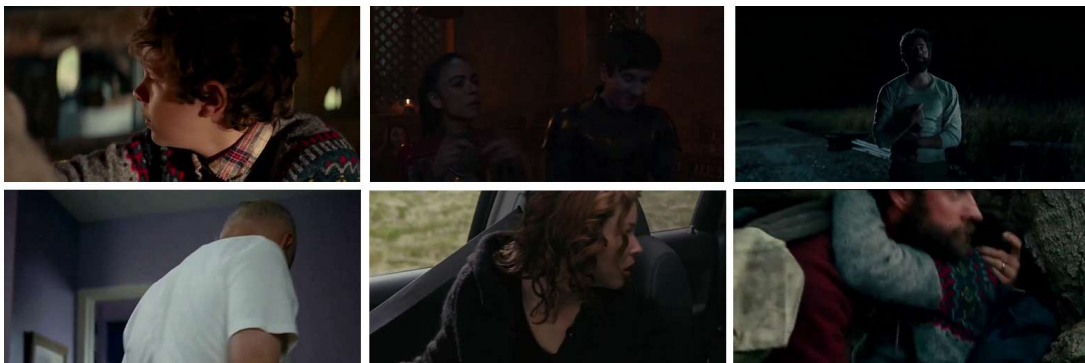


Figure 3.8: Some sample frames from our IMEmo dataset with extreme cases for pose, illumination, and occlusion.

For our dataset, we also performed a classification experiment using all the 16 classes (plus neutral) of the proposed emotion annotation (see Section 3.2.3.2). This experiment was not possible on MELD, AffectNet and FER-2013 since those datasets are annotated only with 6 and 7 emotion classes (plus neutral). As a first evidence, we note that the accuracy obtained for 6 emotion classes is greater than that scored with 16 categories. Though this behavior is somewhat expected, due to the increased difficulty of the problem, it reveals a drop in the ability of separating fine expressions with respect to coarser ones.

Finally, we also present a classification experiment aiming to separate positive from negative emotions (see the last row in Table 3.6). For IMEmo, as reported in Section 3.2.3.2, 9 out of the 16 labels used in our annotation are labeled by a negative mood (i.e., angry, anxiety, contempt, disgust, distress, fear, rage, sadness, shame), while the remaining 7 emotion terms are considered as positive ones (i.e., desire, happy, interest, joy, love, surprise, wonder). For AffectNet the positive/negative sets comprise {happy, surprise} and {angry, contempt, disgust, fear, sad}, respectively; for FER-2013 the sets are

the same with the exclusion of “contempt” from the negative one. The MELD dataset has been labeled with seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. Additionally, MELD provides sentiment annotations (positive, negative, and neutral) for each video. Also in this case, results show the bar for emotion recognition methods on our proposed benchmark is posed to a higher level.

Table 3.6: Recognition accuracy obtained after training and testing the CNN in [5], DAN [170] and DDAMFN [183] on two video datasets, our IMEmo and MELD, and on the image datasets AffectNet and FER-2013. Results are reported for the sentiment positive and negative (plus neutral), the 6 basic emotions (plus neutral) and, on IMEmo, for the proposed annotations with 16 emotion terms (plus neutral). This latter test was not possible for MELD, AffectNet and FER-2013 because their data were annotated only with 6 emotions (plus neutral), respectively.

Network	emotions	Video dataset		Image dataset	
		IMEmo	MELD	AffectNet	FER-2013
CNN	Positive / negative + <i>neutral</i>	59%	43%	79%	79%
	6 + <i>neutral</i>	32%	39%	73%	66%
	16 + <i>neutral</i>	25%	–	–	–
DAN	Positive / negative + <i>neutral</i>	50%	43%	65%	77%
	6 + <i>neutral</i>	25%	21%	65%	56%
	16 + <i>neutral</i>	21%	–	–	–
DDAMFN	Positive / negative + <i>neutral</i>	47%	38%	67%	78%
	6 + <i>neutral</i>	34%	18%	63%	63%
	16 + <i>neutral</i>	24%	–	–	–

3.3.3 Interpersonal recognition results

We follow the work in [151], that constructed a large-scale People in Photo Album (PIPA) dataset, and employed a two-stream model for social domain/relation recognition. The foundation of the relationships stems from Bugental’s domain-based theory [55], which categorizes social relationships into five distinct domains. From this theoretical framework, sixteen social relationships are derived, and the PIPA dataset is expanded to include

26,915 relationship annotations between pairs of individuals.

PIPA dataset: Sourced from Flickr photo albums, serves the purpose of person recognition [182]. Comprising 37,107 photos, it encompasses 63,188 instances featuring 2,356 distinct identities. Each annotated individual is accompanied by both a head bounding box and an identity ID. This diversity allows for individuals to appear in multiple social scenarios and interactions.

Table 3.7 shows the results of relationship recognition experiments for head attributes in relations, and both head and body attributes in domains. The results reported for PIPA dataset are from photo albums that present the two or more people in the same image, with the same lighting and background. In this case, we obtained results of 37.8% for our dataset and 67.7% for PIPA. However, when considering only the image of the head, unaffected by previous conditions, we achieved 32.3% for relations compared to PIPA’s 44.8%, and 62.6% for our dataset compared to PIPA’s 59.4% for domains.

Table 3.7: Accuracies of relation recognition by domain / relation.

Attributes	IMEmo	PIPA
Body attributes	37.8%	67.7%
Head attributes	62.6%	59.4%
Head attributes (relation)	32.3%	44.8%

3.4 Conclusion

In this chapter, we have introduced a novel dataset, the IMEmo dataset, which includes rich annotations at both the clip and shot levels, emphasizing interpersonal interactions and multi-emotion recognition. The IMEmo dataset provides a unique contribution to the field by offering a diverse, multi-lingual, and carefully curated collection of movie clips, annotated with a comprehensive range of emotions. This dataset forms the foundation for exploring complex emotional dynamics and relationships in video-based emotion recognition. Through its comprehensive structure and annotations, the IMEmo dataset addresses a critical need in the field, offering new opportunities to advance the study of multi-emotion recognition and interpersonal dynamics. By providing this dataset, we

aim to contribute to the development of more sophisticated models and methodologies that can better understand and predict emotional states in social contexts, ultimately fostering deeper insights into human emotional behavior. Additionally, this thesis proposes a second major contribution: a predictive model that forecasts the future emotional states between two individuals based on their previous interactions. This model, aimed at predicting the evolution of emotions in interpersonal contexts, will be presented in the next chapter.

Chapter 4

Emotion Prediction in Interpersonal Relations

In this chapter, we investigate the proposed relevant method for predicting emotions in video sequences during interpersonal interactions by capturing the temporal dynamics of emotional changes. Our approach combines visual features extracted using a Distract Your Attention (DAN) network with emotion embeddings to model emotional transitions over time. A Long Short-Term Memory (LSTM) network is employed to capture the temporal dependencies, predicting the emotion of the next video shot based on previous ones. Experimental results demonstrate the effectiveness of the model in predicting both sentiments and specific emotions, although challenges remain in forecasting fine-grained emotional states. This method contributes to advancing emotion prediction by considering the evolving nature of emotions during social interactions.

4.1 Introduction

Emotion plays a critical role in social interactions, guiding communication and shaping relationships. In recent years, significant advancements have been made in the automatic analysis of emotions, particularly in human-computer interaction and affective computing. However, accurately predicting emotional shifts during interpersonal relations remains a challenging task due to the dynamic nature of emotions. Emotions evolve over time, especially in response to social cues and interactions between individuals. Understanding these emotional transitions is crucial for improving the accuracy and effectiveness of emotion prediction models. Within this framework, we introduce in this chapter an effective

approach for predicting emotions in video sequences, focusing on social interactions between individuals. The core idea is to predict the emotion displayed in the next video shot by analyzing the temporal dynamics of emotional changes in previous shots. By leveraging neural networks designed to capture both visual and emotional features, the proposed method aims to improve the accuracy of emotion prediction in the context of interpersonal relations. This chapter outlines the architecture, methodology, and experimental results of this emotion prediction model, which offers promising advancements for emotion recognition in dynamic social scenarios.

4.2 Proposed Emotion Prediction Method

The proposed method defined a protocol that allows a short-term emotions prediction. Given a video clip with N segments, our idea is that of letting a prediction network to look at the segments from $k - m$ to $k - 1$, in order to predict the most probable emotion in segment k (i.e., m represents the number of segments the network looks at in the past). In addition, we designed a neural network architecture that analyzes the video segments and integrates visual features and emotion embeddings to model emotional progression over time. Specifically, we have used a Distract Your Attention Network (DAN) [170] in order to extract the visual features, which are then enhanced with emotion feature vectors derived from corresponding emotion labels. These combined features are thereafter provided to a long short-term memory network that captures the temporal dependencies and emotional context across the video. The final output of the LSTM is classified into one of 16 distinct emotions, six basic emotions and sentiment, predicting the emotion of the next shot.

4.2.1 Proposed Prediction Pipeline

First, we clarify the proposed scenario and the problem we want to answer to, which by itself is a contribution of this study and something new in the literature of automatic analysis of human emotions. Indeed, our main goal is to predict the emotion of the subject in the next shot of a video clip (let us refer to her/his identity as subject B), by using the emotions of both subject A and subject B throughout the previous shots of the video. The main assumption here is that, during social interactions, emotions can dynamically

evolve and change.

More formally, each video clip C is regarded as a concatenation of N shots (video segments) $C = s_1^A \cup s_2^B \cup \dots \cup s_{N-1}^A \cup s_N^B$. Each shot s_i^j includes one subject (j refers to A or B alternatively), with an emotion that does not change during the shot itself. Instead, the emotion may change passing from one shot to the subsequent one for both subjects. The goal is predicting the emotion of the subject in shot s_k^j , looking at the previous $s_{k-m}^j \cup s_{k-m+1}^j \cup \dots \cup s_{k-1}^j$ shots, being m the number of shots we look at back in time (*memory*).

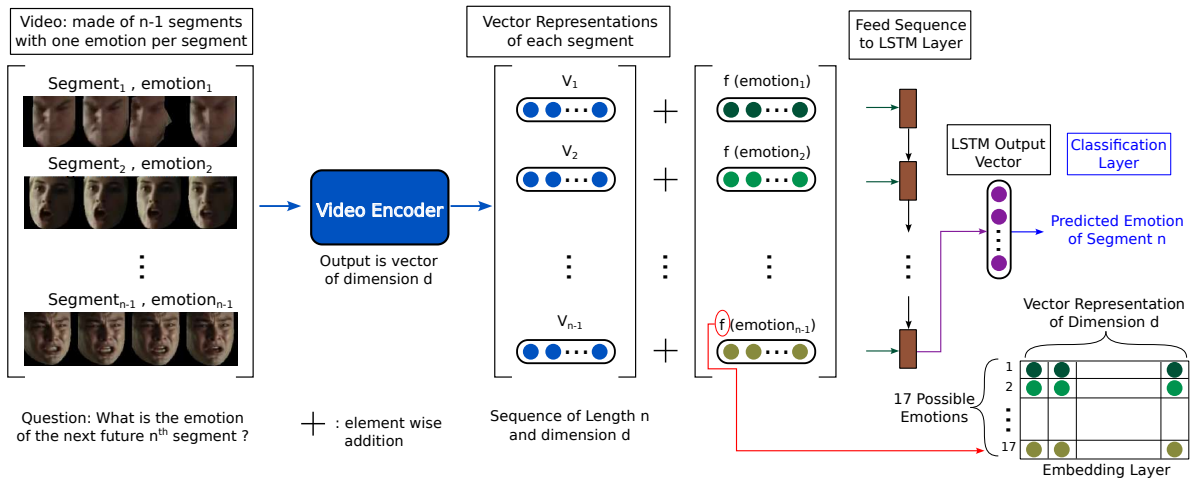


Figure 4.1: Our proposed model predicts the emotion of the next segment in a video. The video is divided into n segments, each tagged with an emotion. These segments are encoded into vector representations by a DAN network. An embedding layer ensures each segment's emotion is accurately represented in the vector space of dimension d . The sequence of vectors, with the vector representing the segment added to the emotion from the embedding layer, is then fed into an LSTM layer. The LSTM's output vector is processed by a classification layer to predict the emotion of the n -th segment, choosing from 17 possible emotions.

To decide about such estimation, we propose the learning architecture illustrated in Figure 4.1, where sequences of video segments, each corresponding to an emotional state, are processed through two main components: the *Feature Extraction* module, which is presented by video Encoder, extracts features from the video segments, and the *Temporal Dependency* module, which accounts for temporal dependencies between shots. The

proposed architecture is detailed in the following subsections.

4.2.1.1 Feature Extraction (video Encoder)

This module is realized by the distract your attention network (DAN) [170] that has been designed for facial expression recognition and comprises three main components. First, the Feature Clustering Network (FCN) extracts robust features from a batch of face images using a large-margin learning objective, which enhances class separability. Second, the Multi-head Cross Attention Network (MAN) employs multiple attention heads to simultaneously focus on various facial regions, generating attention maps for these areas. Finally, the Attention Fusion Network (AFN) distributes attention across multiple locations, then fuses these attention maps into a single comprehensive map. Each shot is represented by a segment that gathers all the frames of that shot. The DAN network processes each shot, or segment, as a whole, encoding the visual features of each segment into vectors (V_{k-m} to V_{k-1}) the visual features of each segment comprised in the *memory* of length m we decide to use. Simultaneously, each emotion label is passed through an embedding layer to generate emotion feature vectors ($f(e_{k-m})$ to $f(e_{k-1})$). These emotion feature vectors are then added ($V_{k-m} + f(e_{k-m})$) with their respective visual vectors to form combined feature representations $\hat{V}_i$ for each segment.

4.2.1.2 Temporal Dependencies

The LSTM network is designed to model the temporal dependencies between frames, enabling it to predict short-term emotional transitions. The LSTM captures how emotions like happiness or surprise change dynamically over time, and the network is trained to predict future emotional states based on past frames. Specifically, the LSTM units retain information about past frames, which allows the model to forecast the emotional state of the next segment accurately. The combined vectors $\hat{V}_{k-m}, \dots, \hat{V}_{k-1}$ are then fed into a Long Short-Term Memory (LSTM) network. The LSTM processes these sequences and outputs a final representation, given by the vector that represents the next shot, V_k .

4.2.1.3 Prediction Module

A final classification layer takes V_k and predicts the emotion for segment k , capturing the temporal dynamics and context of the emotional expressions across the video. The

Person A	Segment 1	Segment 3	Segment 5	Segment 7	Segment 9	Segment 11
	 <i>Interest</i>	 <i>Interest</i>	 <i>Happy</i>	 <i>Interest</i>	 <i>Interest</i>	 <i>Joy</i>
Person B	Segment 2	Segment 4	Segment 6	Segment 8	Segment 10	
	 <i>Interest</i>	 <i>Wonder</i>	 <i>Surprise</i>	 <i>Surprise</i>	 <i>Joy</i>	

Figure 4.2: Example of emotional transitions between two individuals within the IMEmo dataset.

predicted emotion for segment k is derived from this classification process, ensuring a comprehensive analysis of the sequential emotional states, ultimately giving one of the possible emotions according to the used set (either 6 or 16 plus neutral, or positive/negative).

4.2.2 Application of the Proposed Model: Emotion Prediction from the IMEmo Dataset

To illustrate our proposed emotion prediction method, we consider herein a sequence of emotional exchanges between Person A and Person B from the IMEmo dataset, as shown in Figure 4.2. The dataset captures their dynamic emotional interactions across multiple video segments, providing a rich resource for modeling and predicting emotional transitions.

Step 1: Feature Extraction from Video Segments Each video segment, as depicted in the figure (Segments 1 to 11), is processed to extract key facial features from both participants. In Segment 1, Person A displays an expression labeled as “Interest,” while Person B shows the same emotional state in Segment 2. The feature extraction step employs the DAN (Deep Attention Network) to encode these facial expressions into high-dimensional feature vectors, capturing the critical facial landmarks and movements necessary for emotion recognition.

Step 2: Temporal Dependency Modeling with LSTM Once the features are extracted from each segment, they are passed through the LSTM (Long Short-Term

Memory) network. The LSTM captures the temporal dependencies between consecutive segments, modeling how emotional states evolve over time. In this example, Person A’s emotional state shifts from “Interest” (Segments 1, 3, 7, 9) to “Happy” (Segment 5) and eventually to “Joy” (Segment 11). Similarly, Person B’s emotions transition from “Interest” (Segment 2), to “Wonder” (Segment 4), “Surprise” (Segments 6, 8), and finally to “Joy” (Segment 10).

The LSTM retains information from previous segments, allowing the model to predict the emotional state in subsequent segments. By analyzing the emotional progression of both individuals, the model anticipates Person A’s shift to a positive emotional state (“Joy”) in the final segment.

Step 3: Emotion Prediction Based on the temporal dependencies modeled by the LSTM, the system predicts the emotional state for upcoming segments. After processing Segments 1 to 10, the LSTM will predict what Person A will express in Segment 11 (“Joy”). This prediction is informed by the observed progression in positive emotions, particularly in Segments 5 and 7, where Person A exhibits increasing signs of happiness.

Step 4: Model Output and Evaluation The final output of the model consists of predicted emotional labels for the last segment in the video sequence. These predictions are compared to the ground truth labels provided in the IMEmo dataset for evaluation. The next section presents the evaluation of these predictions, showcasing the model’s robustness in accurately capturing complex emotional transitions.

4.3 Experiments and Results

In this section, we detail the experimental setup used to evaluate the performance and robustness of the proposed model using our dataset IMEmo.

4.3.1 Prediction Performance

In Table [4.1](#), we provide a comprehensive overview of the performance of the proposed model through two experiments:

1. **All segments:** We used all segments of a video clip except the last one to predict the emotion of the last segment. Following the notation of [4.2](#), this corresponds to set $k = n$ and $m = n - 1$;
2. **Sequential prediction:** In this experiment, we modified the memory set by predicting the emotion of segment k looking at the emotion of the two previous shots ($m = 2$). This has been done for each set of three consecutive segments within the video, i.e., for $k = 3 : n$, ensuring that we capture the sequential nature of the data. We used $m = 2$ because this represents the minimal memory to include one shot from each of the two subjects.

From the experiments, it turns out emotion forecasting is a complex task. While predicting the general sentiment is quite feasible both in the long- and short-term, predicting the correct emotion from observing previous interactions resulted in a quite low accuracy even for the 6 basic emotions. These results suggest the problem is challenging, and this new protocol can open the way to new investigations in the domain.

4.3.2 Dependency Analysis

In Table [4.2](#), we provide a detailed analysis of the factors and variables that influence the proposed model through three types of experiments using all segments of the dataset in each of the experiments:

1. **Cross-validation:** We employ cross-validation to evaluate the performance of our model.

Table 4.1: Prediction performance: Prediction accuracy (in %) obtained after training and testing the proposed model on the **IMEmo** dataset.

	Sentiments	6 emotions	16 emotions
All segments ($k = n, m = n - 1$)			
Segment n	64.71	35.29	29.41
Sequential prediction ($k = 3, m = 2$)			
3 segments	82.35	47.06	35.29

- **Cross-Validation 1:** We use a unique partitioning method for the data: segments originally designated for the training set are also used in the validation set, and segments from the validation set are included in the training set. This ensures that the model is evaluated on different segments than those it was trained on, maximizing the use of available data while maintaining separation between training and validation data.
- **Cross-Validation 2:** We adjust the partitioning so that some segments initially in the validation set are incorporated into the training set.

The videos are randomly selected for the validation set, ensuring that the videos in Cross-Validation 1 are different from those in Cross-Validation 2. In Cross-Validation 1, the validation set includes videos that are not used in the default validation set provided by the dataset. However, in Cross-Validation 2, some videos originally part of the dataset’s validation set are included.

2. **Temporal Shifts:** We also examined the performance under temporal shifts.

- **Shift by one-segment:** We sequentially shift the segments of the training set by one segment, removing the first segment of each video, while keeping the validation set constant to simulate different temporal contexts.
- **Shift by two-segments:** We sequentially shift the segments of the training set by two segments, removing the first two segments of each video, to simulate different temporal contexts.

3. **Balanced vs. Unbalanced Data:** We balance the training set by duplicating the segments of underrepresented emotion classes. This involves identifying classes with fewer instances and increasing their representation by duplicating their data points. For example, if we have 16 emotions and 6 emotions with 200 segments each, and for sentiment analysis, we have 650 segments for each sentiment.

Table 4.2 presents the performance results across three different evaluation strategies: Cross-validation, Temporal shifts, and Balanced data. Based on the experiment results, it is evident that emotion classification remains a challenging task. While predicting general sentiments shows reasonable consistency, accurately classifying specific emotions, whether it be the 6 basic emotions or 16 emotions, proves to be significantly more difficult.

Table 4.2: Dependency analysis: Prediction accuracy (in %) obtained after training and testing the proposed model on the **IMEmo** dataset.

	Sentiments	6 emotions	16 emotions
Cross-validation			
Cross-validation 1	54.55	36.36	36.36
Cross-validation 2	61.54	30.77	46.15
Temporal shifts			
By 1 segment	64.71	23.53	17.65
By 2 segments	70.59	29.41	29.41
Balanced data			
Balance data	82.35	35.29	11.76

The results show considerable variability, with accuracies for 6 emotions ranging from 36.36% to 30.77% and for 16 emotions from 36.36% to 46.15%. These findings highlight the inherent challenges in emotion prediction and suggest that further refinement and development of the model are necessary. The complexity of this task underscores the intricate nature of human emotions. The Temporal Shifts section evaluates the impact of shifting segments by one and two segments. Shifting by one segment results in values of 64.71 for Sentiments, 23.53 for 6 emotions, and 17.65 for 16 emotions. Shifting by two segments results in an improvement, with values of 70.59 for Sentiments and 29.41 for both 6 and 16 emotions. These findings suggest that shifting segments temporally can affect the performance of the model. The observed improvements when shifting by two segments are likely due to the fact that Person A initiates the emotion, making the consistency of the emotions clearer and more predictable. This underscores the importance of considering temporal dynamics and the initiating factors in emotion prediction models.

From the Balanced data section, it becomes clear that achieving balanced class distributions with the IMEmo dataset does not significantly enhance predictive accuracy. Sentiments can be forecasted with an accuracy of 82.35, but the accuracy for predicting the 6 basic emotions is only 35.29, and it further drops to 11.76 for 16 emotions. These results suggest that while balancing the dataset is a methodological consideration, it may not be the key factor in improving emotion prediction accuracy. From the obtained

results, it turns out that the methods of partitioning and balancing the data can significantly impact model performance. While balanced datasets show particular promise for improving sentiment prediction, the challenge remains in accurately predicting specific emotions. These findings indicate that further exploration in this area could lead to new advancements.

4.3.3 Discussion

In this section, we discuss the results of the experiments of our proposed model by showcasing a confusion matrix and an example of prediction to illustrate its performance on different emotional and sentiment prediction tasks. In fact, Figure 4.3 presents three confusion matrices for the sequential prediction ($k = 3$) scenario: (a) 6 emotions, (b) 16 emotions, and (c) sentiment analysis. A visual representation of the performance of the classification model is provided by illustrating the agreement between the true labels and the predicted labels, highlighting the model’s ability to understand and predict emotional and sentiment transitions over sequences. Specifically, we are predicting the emotions of 100 segments. In the validation set, for example, we use segment 1 and segment 2 to predict segment 3, then segment 2 and segment 3 to predict segment 4, and continue this pattern throughout the dataset. From the confusion matrices, we can see that for the 6 basic emotions: Happy and Sad emotions are predicted with higher accuracy compared to others, with true positive counts of 10 each. Fear is predicted correctly 5 times, while Surprise is the least accurately predicted emotion, often being misclassified as Neutral or Happy. Furthermore, for the 16 emotions: Neutral, Interest, and Sadness show relatively better prediction accuracy. Emotions like Rage, Contempt, and Shame have lower accuracy and are frequently misclassified. Happy and Sadness emotions also show a reasonable level of prediction accuracy, though less than Neutral and Interest. For the 3 sentiments: Positive and Negative sentiments have the highest prediction accuracy, with true positive counts of 27 and 43, respectively. Neutral sentiment has the lowest prediction accuracy, often being misclassified as Negative. Overall, while the model performs well in predicting some emotions and sentiments, there is room for improvement in accurately distinguishing between certain similar emotions, especially those with subtle differences.

Figure 4.4 illustrates the prediction process and the representation of the ground truth and predicted labels in our study. The left panel shows three segments: segment 1 with

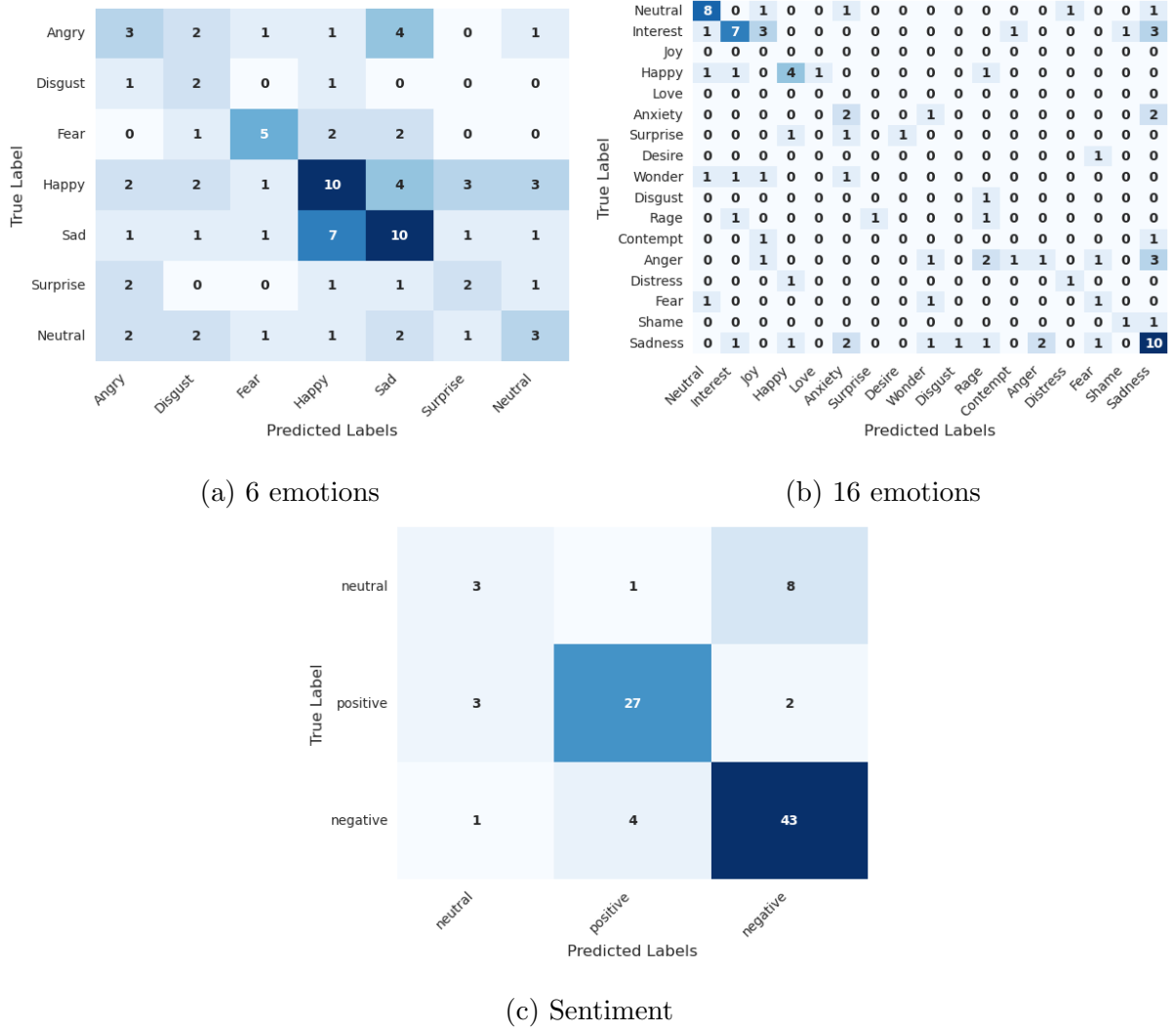


Figure 4.3: Confusion matrix for: (a) 6 emotions, (b) 16 emotions, and (c) sentiment.

Person *A* labeled as "Anger", segment 2 with Person *B* also labeled as "Anger", and segment 3 returning to Person *A*. The right panel compares the ground truth and the predicted labels for Segment 3 across different categorization schemes. For the 16 emotions, the ground truth label is "Sadness", while our proposed model predicted "Anxiety". For the 6 emotions, both the ground truth and predicted labels are "Sadness". For sentiment, the ground truth and predicted labels are "Negative". This comparison reveals several key insights into our model's performance. Firstly, the model demonstrates strong alignment with the ground truth in broader categorization schemes, as evidenced by its accurate predictions in the 6-emotion and sentiment classifications. This suggests the model's robustness in identifying general emotional and sentiment trends. However, the discrepancy observed in the 16-emotion categorization, where "Sadness" was misclassified

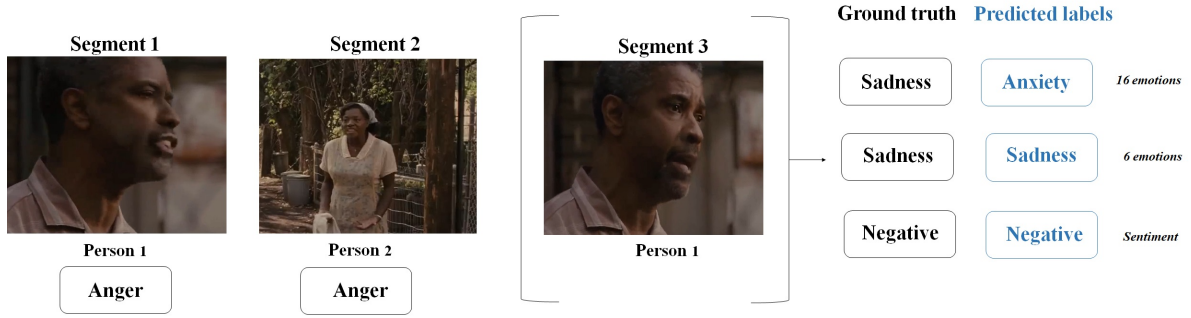


Figure 4.4: Prediction process of the ground-truth and predicted labels.

as "Anxiety," indicates that the model may struggle with distinguishing between more nuanced emotional states. This highlights an area where the model could benefit from further refinement to improve its sensitivity to subtle emotional variations. Overall, while the model shows promising accuracy in broader classifications, enhancing its precision in detailed emotional categorizations remains a crucial objective.

4.3.4 Ablation Study

As an ablation study, aiming to show the relevance of including the emotion of both the subjects for an improved prediction, we have performed the same experiments of Table 3.4, excluding the sentiment one, under the two following settings:

1. **Same subject:** Supposing the shot to predict includes the subject B , we performed the estimation based only on the shots in the clip that includes subject B . So, we predict the emotion of subject B by just looking at her/his emotions in previous shots (either for the case of *All segments* and *Sequential prediction*) completely disregarding the emotion of the subject A ;
2. **Different subject:** Supposing the shot to predict includes the subject B , we performed the estimation based only on the shots in the clip that include the subject A . So, we predict the emotion of subject B just looking at the emotions of the other subject in previous shots (either for the case of *All segments* and *Sequential prediction*).

Table 4.3 shows the results of the above experiments in comparison with the case where the shots of both the subjects are used.

Table 4.3: Prediction performance in percentage: Evaluation of the contribution given by looking in the past to the segments of one subjects at a time (*Same subject* and *Different subject* cases) or to the segments of both the interacting subjects (*Both subjects* case).

	Basic emotions	Fine grained emotions
Same subject		
Segment n	23.53	11.76
3 segments	29.41	23.53
Different subject		
Segment n	23.53	11.76
3 segments	35.29	17.65
Both subjects		
Segment n	35.29	29.41
3 segments	47.06	35.29

Results in Table 4.3 provide a demonstration of the impact of incorporating the emotions of both subjects in the prediction process. When using only the same subject’s previous emotions for prediction (the *Same subject* scenario), the prediction performance is notably lower than in the case the segments of both subject are observed, particularly for fine-grained emotions. This suggests that relying solely on the emotions of the individual being predicted does not capture the full context, which is often influenced by the interaction dynamics between both subjects. When predictions are based solely on the other subject’s emotions (the *Different subject* scenario), the performance is similar to the *Same subject* case when all the *segments* are used. When the prediction is based on just three segments, an increase with respect to the *Same subject* case is noted for the basic emotions, while an opposite behavior appears when the fine grained emotions are used.

Overall, results for the *Both subjects* case clearly outperform the other two scenarios,

emphasizing the importance of considering the interplay between the emotions of both subjects. The significant increase in prediction accuracy when using the emotions of both subjects suggests that the emotional states of individuals in a shared context are interdependent. Therefore, to achieve the best prediction performance, it is crucial to model the emotions of all involved parties, capturing the nuanced dynamics that arise in interpersonal interactions.

4.4 Conclusion

In this chapter, we have introduced an accurate method for predicting the emotion of the next shot in video clips, leveraging the temporal evolution of emotions during social interactions. The proposed emotion prediction method utilizes a neural network architecture that integrates visual features extracted using a "distract your attention" network and emotion embeddings within a temporal framework. By incorporating long short-term memory networks, the model captures temporal dependencies and emotional context, making it well-suited to predict the next emotional state in an ongoing interaction. The design of the prediction pipeline ensures that the model accounts for the sequential nature of emotions and the interactions between subjects, which is crucial for enhancing accuracy in emotion recognition tasks. In addition, the ability to predict emotions not only based on individual shots but also by considering the interplay between two subjects offers a novel contribution to the field. By predicting emotions in this manner, we can better understand the nuances of human interaction, which can have broad applications in areas such as human-computer interaction, affective computing, and social signal processing.

General Conclusion

The accurate prediction of emotional transitions, especially in the context of interpersonal interactions, is a crucial challenge for several fields, including multimedia indexing, affective computing, and human-computer interaction. As digital environments evolve, the ability to understand and predict human emotions has become increasingly vital for creating more engaging and responsive systems. However, the complex and dynamic nature of emotions, particularly during human interactions, presents significant challenges that require advanced models capable of capturing emotional shifts over time. This thesis addresses these challenges by focusing on the development of a robust framework for predicting emotional transitions within dyadic interactions. It began by identifying a significant gap in existing datasets, which tend to capture isolated emotions but lack the depth to model the complex interplay of emotions between individuals. To fill this gap, one of the major contributions of this work was the creation of the IMEmo dataset, a novel and comprehensive dataset designed specifically to capture dynamic emotional exchanges. Annotated with 16 distinct emotions and enriched with demographic information, this dataset offers a rich context for analyzing interpersonal emotional dynamics in natural settings, extracted from commercial movie clips.

Building upon this foundational dataset, the next key contribution was the development of a novel emotion prediction model. Leveraging Long Short-Term Memory networks, the model was designed to capture the temporal dependencies and progression of emotions over time. By incorporating the emotional states of both interacting individuals, the model forecasts future emotional transitions, allowing for a more nuanced understanding of how emotions evolve in response to social interactions. This shift

from static emotion recognition to dynamic prediction marks a significant advancement in the field. Indeed, through extensive experimentation and validation using the IMEmo dataset, the effectiveness of the proposed model was demonstrated, particularly in its ability to predict short-term emotional shifts with high accuracy. Moreover, the ablation study has reinforced the importance of analyzing the emotional interplay between both individuals, showing that incorporating data from both participants in an interaction significantly improves prediction performance. These findings advance the understanding of emotional dynamics in dyadic exchanges and offer new opportunities for applications where emotion prediction plays a critical role, such as human-computer interaction and social robotics.

While this research makes significant contributions to the field of emotion prediction, it also opens up exciting avenues for future exploration. A particularly promising direction is the integration of additional modalities, such as audio, textual data, and physiological signals, to complement the visual cues used in this study. In fact, emotions are inherently multimodal, and capturing them through various channels could provide a richer and more accurate emotional context. By leveraging these diverse modalities, future work could enhance the model’s ability to generalize across different environments, cultures, and scenarios. Moreover, exploring the fusion of multimodal data could also help in addressing challenges such as ambiguity in visual expressions or the subtlety of certain emotional states. For instance, adding speech tone or contextual language information could resolve uncertainties in facial expressions alone, leading to more robust emotion prediction models. Additionally, experimenting with different multimodal architectures and data fusion techniques could unlock new possibilities in improving prediction accuracy and understanding complex emotional dynamics. For instance, the Dempster-Shafer theory can offer a flexible framework to handle uncertainty and incomplete information. By allowing the combination of evidence from diverse sources (e.g., facial expressions, voice, physiological signals), it enhances the reliability and robustness of emotion recognition systems. This approach helps manage conflicting data, improving the accuracy of predictions in complex multimodal real-world environments.

In conclusion, this thesis provides a cohesive journey from the identification of limitations in current emotion recognition methodologies to the development of a new dataset and predictive framework. By tackling the complex nature of emotional transitions in interpersonal interactions, the contributions of this research lay a strong foundation for future advancements in affective computing, offering both theoretical insights and practical applications for creating more emotionally intelligent systems.

Appendix

Table 4.4: The Definition of Terms Denoting Emotions in Terms of Pleasure, Arousal, and Dominance as proposed by Russel [140]

Term	N	Pleasure		Arousal		Dominance	
		Mean	SD	Mean	SD	Mean	SD
1. Bold	27	.44*	.32	.61*	.24	.66*	.30
2. Useful	27	.70*	.20	.44*	.28	.47*	.40
3. Mighty	27	.48*	.37	.51*	.28	.69*	.31
4. Kind	27	.73*	.22	.19*	.32	.57*	.34
5. Self-satisfied	27	.86*	.10	.20	.40	.62*	.31
6. Admired	29	.81*	.21	.44*	.30	.51*	.34
7. Proud	29	.77*	.21	.38*	.34	.65*	.33
8. Interested	29	.64*	.20	.51*	.21	.17	.40
9. Arrogant	29	.00	.51	.34*	.44	.48*	.34
10. Inspired	29	.71*	.30	.63*	.21	.34*	.55
11. Excited	29	.62*	.25	.75*	.20	.38*	.29
12. Influential	28	.68*	.23	.40*	.33	.75*	.18
13. Aggressive	28	.41*	.30	.63*	.25	.62*	.24
14. Strong	28	.58*	.24	.48*	.30	.62*	.30
15. Dignified	28	.55*	.27	.22*	.40	.61*	.30
16. Powerful	28	.54*	.26	.45*	.36	.73*	.25
17. Elated	28	.50*	.47	.42*	.14	.23*	.36
18. Hopeful	29	.51*	.30	.23*	.33	.14	.41
19. Triumphant	29	.69*	.23	.57*	.19	.63*	.26
20. Joyful	29	.76*	.22	.48*	.26	.35*	.31

Table 4.4: *Cont.*

Term	N	Pleasure		Arousal		Dominance	
		Mean	SD	Mean	SD	Mean	SD
21. Capable	29	.70*	.24	.28*	.27	.61*	.31
22. Lucky	30	.71*	.19	.48*	.30	.37*	.29
23. Responsible	30	.35*	.29	.38*	.26	.49*	.37
24. Friendly	30	.69*	.23	.35*	.28	.30*	.27
25. Masterful	30	.58*	.25	.44*	.27	.69*	.25
26. Free	30	.81*	.14	.24*	.38	.46*	.33
27. Devoted	30	.49*	.25	.17*	.27	.10	.37
28. Domineering	27	.23*	.37	.40*	.21	.58*	.26
29. Aroused	30	.24*	.28	.57*	.26	.22*	.33
30. Concentrating	28	.42*	.25	.28*	.27	.39*	.31
31. Happy	29	.81*	.21	.51*	.26	.46*	.38
32. Egotistical	29	.24*	.34	.32*	.25	.50*	.31
33. Carefree	29	.78*	.21	.35*	.39	.41*	.31
34. Affectionate	29	.64*	.26	.35*	.34	.24*	.40
35. Vigorous	30	.58*	.22	.61*	.23	.49*	.24
36. Activated	30	.42*	.26	.58*	.21	.38*	.28
37. Alert	30	.49*	.25	.57*	.20	.45*	.26
38. Alone with responsibility	30	.33*	.34	.34*	.37	.48*	.36
39. Controlling	30	.47*	.26	.34*	.23	.66*	.25
40. Proud and lonely	27	.01	.43	.02	.32	.26*	.40
41. Enjoyment	30	.77*	.17	.44*	.26	.42*	.29
42. Serious	30	.27*	.22	.24*	.22	.42	.27
43. Cooperative	31	.39*	.32	.13*	.27	.03	.34
44. Thankful	27	.61	.25	.10	.34	-.13	.35
45. Respectful	27	.38*	.39	.13	.29	-.08	.49
46. Appreciative	27	.55*	.26	.07	.30	-.14	.35
47. Loved	28	.87*	.17	.54*	.26	-.18*	.30

Table 4.4: *Cont.*

Term	N	Pleasure		Arousal		Dominance	
		Mean	SD	Mean	SD	Mean	SD
48. Grateful	28	.64*	.23	.16	.22	-.21*	.34
49. In love	28	.82*	.17	.65*	.24	-.05	.37
50. Anxious	28	.01	.45	.59*	.31	-.15	.32
51. Impressed	28	.41*	.26	.30*	.25	-.32*	.34
52. Surprised	29	.40*	.30	.67*	.27	-.13	.38
53. Sexually excited	29	.58*	.26	.62*	.20	-.01	.43
54. Wonder	30	.27*	.37	.24*	.35	-.17*	.26
55. Fascinated	29	.55*	.22	.51*	.23	-.07	.35
56. Awed	30	.18*	.34	.40*	.30	-.38*	.21
57. Overwhelmed	30	.14	.39	.45*	.34	-.24*	.32
58. Curious	28	.22*	.30	.62*	.20	-.01	.34
59. Relaxed	27	.68*	.30	-.46*	.38	.06	.49
60. Untroubled	27	.79*	.25	-.01	.46	.33*	.39
61. Modest	29	.27*	.34	-.06	.30	.12	.39
62. Secure	29	.74*	.22	-.13	.32	.03	.37
63. Nonchalant	30	.07	.35	-.25*	.31	.11	.33
64. Aloof	30	.16	.35	-.01	.30	.25*	.32
65. Leisurely	30	.58*	.35	-.32*	.33	.11	.33
66. Reserved	30	.01	.37	-.19*	.28	.02	.37
67. Protected	27	.60*	.35	-.22*	.37	-.42*	.40
68. Consoled	27	.29*	.50	-.19*	.33	-.28*	.38
69. Quiet	27	.19	.57	-.40*	.21	-.04	.39
70. Sheltered	29	.14	.53	-.36*	.32	-.44*	.24
71. Humble	28	.23*	.39	-.28*	.24	-.27*	.35
72. Solemn	29	.03	.39	-.32*	.26	-.11	.33
73. Reverent	28	.31*	.35	-.08	.26	-.29*	.34
74. Astonished	30	.16*	.26	.88*	.19	-.15*	.26

Table 4.4: *Cont.*

Term	N	Pleasure		Arousal		Dominance	
		Mean	SD	Mean	SD	Mean	SD
75. Disgusted	29	-.60*	.20	.35*	.41	.11	.34
76. Insolent	28	-.26*	.43	.21	.44	.20*	.31
77. Cruel	29	-.45*	.35	.48*	.32	.42*	.32
78. Irritated	29	-.58*	.16	.40*	.37	.01	.40
79. Defiant	28	-.16*	.30	.54*	.37	.32*	.42
80. Hate	28	-.56*	.16	.59*	.29	.13	.34
81. Hostile	29	-.42*	.31	.53*	.36	.30*	.32
82. Angry	29	-.51*	.20	.59*	.33	.25*	.39
83. Mildly annoyed	29	-.28*	.16	.17*	.28	.04	.31
84. Enraged	29	-.44*	.25	.72*	.29	.32*	.44
85. Contempt	29	-.23*	.39	.31*	.33	.18*	.29
86. Selfish	29	-.34*	.31	.09	.33	.31*	.43
87. Reprehensible	24	-.09	.36	.11	.24	.06	.32
88. Contemptuous	25	-.24*	.26	.31*	.31	.21*	.32
89. Scornful	28	-.35*	.21	.35*	.27	.29*	.32
90. Suspicious	29	-.25*	.23	.42*	.21	.11	.32
91. Skeptical	29	-.22*	.28	.21*	.25	.03	.33
92. Burdened with responsibility	30	-.08	.41	.28*	.31	.19	.47
93. Cold anger	30	-.43*	.29	.67*	.27	.34*	.44
94. Hostile but controlled	30	-.24*	.31	.42*	.35	.09	.38
95. Crushed	27	-.69*	.50	.03	.47	-.50*	.23
96. Frustrated	27	-.64*	.18	.52*	.37	-.35*	.30
97. Distressed	27	-.61*	.17	.28*	.46	-.36*	.21
98. Insecure	27	-.57*	.34	.14	.42	-.42*	.29
99. Humiliated	27	-.63*	.18	.43*	.34	-.38*	.30
100. Hungry	29	-.44*	.26	.14	.33	-.21*	.27

Table 4.4: *Cont.*

Term	N	Pleasure		Arousal		Dominance	
		Mean	SD	Mean	SD	Mean	SD
101. Fearful	29	-.64*	.20	.60*	.32	-.43*	.30
102. Terrified	29	-.62*	.20	.82*	.25	-.43*	.34
103. Embattled	27	-.37*	.24	.40*	.42	-.02	.39
104. Helpless	29	-.71*	.18	.42*	.45	-.51*	.32
105. Troubled	28	-.63*	.16	.16	.40	-.40*	.24
106. Startled	28	-.09	.34	.65*	.29	-.33*	.32
107. Anguished	29	-.50*	.30	.08	.46	-.20*	.34
108. Shamed	29	-.57*	.23	.01	.46	-.34*	.30
109. Displeased	29	-.55*	.21	.16	.34	-.05	.41
110. Embarrassed	29	-.46*	.30	.54*	.26	-.24*	.40
111. Upset	30	-.63*	.20	.30*	.39	-.24*	.24
112. Defeated	30	-.61*	.24	.06	.39	-.32*	.23
113. Pain	29	-.58*	.21	.41*	.31	-.34*	.28
114. Quietly indignant	26	-.28*	.35	.04	.36	-.16	.40
115. Repentant	25	-.06	.55	.06	.32	-.12	.43
116. Sinful	28	-.30*	.36	.22*	.33	-.01	.41
117. Shy	29	-.15	.33	.06	.30	-.34*	.28
118. Guilty	29	-.57*	.19	.28*	.38	-.34*	.28
119. Weary with responsibility	30	-.27*	.34	.02	.28	-.01	.35
120. Angry but detached	30	-.42*	.22	.28*	.41	-.03	.33
121. Confused	30	-.53*	.20	.27*	.29	-.32*	.28
122. Dissatisfied	30	-.50*	.22	.05	.28	.13	.32
123. Regretful	30	-.52*	.24	.02	.32	-.21*	.28
124. Tense	30	-.33*	.35	.58*	.32	-.11	.39
125. Disdainful	25	.32*	.32	.11	.27	.05	.29
126. Depressed	27	-.72*	.21	-.29*	.44	-.41*	.28
127. Despairing	27	-.72*	.21	-.16	.34	-.38*	.25
128. Lonely	29	-.66*	.35	-.43*	.36	-.32*	.30

Table 4.4: *Cont.*

Term	N	Pleasure		Arousal		Dominance	
		Mean	SD	Mean	SD	Mean	SD
129. Meek	29	-.19	.58	-.25*	.32	-.41*	.42
130. Burdened	29	-.66*	.21	-.03	.41	-.26*	.36
131. Timid	28	-.15	.41	-.12	.37	-.47*	.31
132. Bored	28	-.65*	.19	-.62*	.24	-.33*	.21
133. Feeble	29	-.42*	.35	-.20	.42	-.46*	.31
134. Nauseated	29	-.61*	.25	-.01	.28	-.36*	.33
135. Inhibited	30	-.54*	.26	-.04	.40	-.41*	.23
136. Fatigued	30	-.18	.48	-.57*	.29	-.29*	.29
137. Rejected	29	-.62*	.24	-.01	.38	-.33*	.27
138. Subdued	29	-.17	.35	-.26*	.25	-.18*	.34
139. Impotent	30	-.53*	.35	-.13	.40	-.29*	.32
140. Ennui	16	-.45*	.39	-.43*	.36	-.17	.42
141. Blasé	26	-.29*	.33	-.51*	.25	-.16	.30
142. Haughty and lonely	27	-.47*	.28	-.24*	.34	-.13	.26
143. Listless	27	-.45*	.41	-.59*	.28	-.24*	.28
144. Deactivated	28	-.46*	.32	-.43*	.41	-.46*	.27
145. Weary	29	-.18	.38	-.33*	.42	-.24*	.35
146. Snobbish and lonely	30	-.62*	.26	-.19*	.34	-.14*	.28
147. Uninterested	30	-.47*	.26	-.50*	.22	-.08	.24
148. Detached	30	-.37*	.34	-.26*	.28	-.14*	.29
149. Discontented	30	-.53*	.19	-.16	.41	-.26*	.30
150. Discouraged	30	-.61*	.25	-.15	.32	-.29*	.32
151. Sad	30	-.63*	.23	-.27*	.34	-.33*	.22

Publications

The research work presented in this thesis has produced publications in international conferences and journals.¹

Journal papers

1. **Guerdelli, H.**, Ferrari, C., Barhoumi, W., Ghazouani, H., & Berretti, S. (2022). "Macro-and micro-expressions facial datasets: A survey". *Sensors*, 22(4), 1524.
2. **Guerdelli, H.**, Ferrari, C., & Berretti, S. (2023). "Interpersonal relation recognition: a survey". *Multimedia Tools and Applications*, 82(8), 11417-11439.
3. **Guerdelli, H.**, Ferrari, Berretti, S., Barhoumi, W., & Del Bimbo, A. (2024). "A Video Dataset for Multi-Emotion and Interpersonal Relation Analysis". *Computer Vision and Image Understanding Journal*. A revision is in preparation for resubmission.

International Conferences and Workshops

1. **Guerdelli, H.**, Ferrari, C., Cardia Neto, J. B., Berretti, S., Barhoumi, W., & Del Bimbo, A. (2023, September). "Towards a Better Understanding of Human Emotions: Challenges of Dataset Labeling". *International Conference on Image Analysis and Processing (ICIAP)*, pp.242-254. Cham: Springer Nature Switzerland.
2. **Guerdelli, H.**, Ferrari, C., Berretti, S., & Del Bimbo, A. (2024). "IMEmo: An Interpersonal Relation Multi-Emotion Dataset". *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024. p. 1-10.
3. **Guerdelli, H.**, Ferrari, C., Berretti, S., Barhoumi, W., & Del Bimbo, A. (2024). "Predicting Emotions in Interpersonal Video Interactions: I Know What You Feel". *The European Conference on Computer Vision (ECCV) Workshops*.

¹The author's bibliometric indices are the following: H-index = 2, total number of citations = 38 (source: Google Scholar on October 18, 2024)

Bibliography

- [1] (2013). Psychological image collection at stirring (pics). Available online: <http://pics.stir.ac.uk> (accessed on 29 May 2020).
- [2] Abrilian, S., Devillers, L., Buisine, S., and Martin, J.-C. (2005). Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International Conference*, volume 401, pages 407–408.
- [3] Alvino, C., Kohler, C., Barrett, F., Gur, E., Gur, R., and Verma, R. (2007). Computerized measurement of facial expression of emotions in schizophrenia. *Journal of Neuroscience Methods*, 163(2):350–361.
- [4] Arnold, M. B. (1960). *Emotion and personality*. Columbia University Press.
- [5] Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.
- [6] Arushi, G., Teck, M. K., and Cheston, T. (2019). An end-to-end network for generating social relationship graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [7] Aubrey, A., Marshall, D., Rosin, P., Vandeventer, J., Cunningham, D., and Wallraven, C. (2013). Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282.
- [8] Barrett, L., Adolphs, R., Marsella, S., Martinez, A., and Pollak, S. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

- [9] Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35.
- [10] Becattini, F., Song, X., Baecchi, C., Fang, S.-T., Ferrari, C., Nie, L., and Del Bimbo, A. (2021). Plm-ipe: A pixel-landmark mutual enhanced framework for implicit preference estimation. In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, pages 1–5.
- [11] Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR16)*, pages 5562–5570, Las Vegas, NV, USA.
- [12] Bilakhia, S., Petridis, S., Nijholt, A., and Pantic, M. (2015). The mahnob mimicry database: A database of naturalistic human interactions. *Pattern Recognition Letters*, 66:52–61.
- [13] Bottino, A. G., Simone, M. D., Laurentini, A., and Vieira, T. (2012). A new problem in face image analysis: Finding kinship clues for siblings pairs. In *Conference on Pattern Recognition Application & Methods*.
- [14] Bugental, D. B. (2000). Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126(2):187–219.
- [15] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.
- [16] Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2016). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. on Affective Computing*, 8(1):67–80.
- [17] Cai, H., Bai, C., Tai, Y.-W., and Tang, C.-K. (2018). Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European conference on computer vision (ECCV)*, pages 366–382.

- [18] Campbell, N. (2008). Tools and resources for visualising conversational-speech interaction. In *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*, pages 231–234.
- [19] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2006). The ami meeting corpus: A pre-announcement. In *Proceedings of the Machine Learning for Multimodal Interaction*, pages 28–39.
- [20] Cheng, S., Kotsia, I., Pantic, M., and Zafeiriou, S. (2018). 4dfab: A large scale 4d facial expression database for biometric applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA.
- [21] Clark, M. S. and Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology*, 37(1):12.
- [22] Cohn, J., Kruez, T., Matthews, I., Yang, Y., Nguyen, M., Padilla, M., Zhou, F., and la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. In *Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, Amsterdam, The Netherlands.
- [23] Dai, Q., Carr, P., Sigal, L., and Hoiem, D. (2015). Family member identification from photo collections. In *Applications of Computer Vision*, pages 982–989.
- [24] Davis, J., Senghas, A., and Ochsner, K. (2009). How does facial feedback modulate emotional experience? *Journal of Research in Personality*, 43(5):822–829.
- [25] Davison, A., Lansley, C., Costen, N., Tan, K., and Yap, M. (2018). Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, 9(2):116–129.
- [26] Dehshibi, M. M. and Bastanfard, A. (2010). Unsupervised feature based facial family similarity recognition. In *Proc. International Conference on Image and Video Processing and Computer Vision (IVPCV-10)*, pages 132–138. ISRST.
- [27] Dhall, A., Goecke, R., and Gedeon, T. (2015). Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26.

- [28] Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, Barcelona, Spain.
- [29] Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 3:34–41.
- [30] Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and Heylen, D. (2008). The sensitive artificial listener: An induction technique for generating emotionally coloured conversation. In *Proceedings of the LREC Workshop on Corpora for Research on Emotion and Affect*, pages 1–4.
- [31] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., et al. (2007). The humane database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of the Affective Computing and Intelligent Interaction, Second International Conference*, pages 488–500, Lisbon, Portugal.
- [32] Du, S., Tao, Y., and Martinez, A. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462.
- [33] Ekman, P. (1982). What emotion categories or dimensions can observers judge from facial behavior? *Emotions in the human face*, pages 39–55.
- [34] Ekman, P. (1992a). Are there basic emotions?
- [35] Ekman, P. (1992b). An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200.
- [36] Ekman, P. and Keltner, D. (1997). Universal facial expressions of emotion. In Segerstrale, U. and Molnar, P., editors, *Nonverbal Communication: Where Nature Meets Culture*, volume 27, page 46.
- [37] Erdem, C., Turan, C., and Aydin, Z. (2014). Baum-2: A multilingual audio-visual affective face database. *Multimedia Tools and Applications*, 74:7429–7459.

Bibliography

- [38] Eroglu Erdem, C., Turan, C., and Aydin, Z. (2015). Baum-2: A multilingual audio-visual affective face database. *Multimedia tools and applications*, 74(18):7429–7459.
- [39] Ertugrul, I. O., Cohn, J. F., Jeni, L. A., Zhang, Z., Yin, L., and Ji, Q. (2020). Cross-domain domains for au coding: Perspectives, approaches, and measures. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):158–171.
- [40] Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., and Gool, L. (2010). A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12:591–598.
- [41] Fang, R., Tang, K. D., Snavely, N., and Chen, T. (2010). Towards computational models of kinship verification. In *International Conference on Image Processing (ICIP)*. IEEE.
- [42] Ferrari, C., Lisanti, G., Berretti, S., and Del Bimbo, A. (2018). Investigating nuisances in dcnn-based face recognition. *IEEE Transactions on Image Processing*, 27(11):5638–5651.
- [43] Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99(4):689.
- [44] Foa, E. and Foa, U. (1980). *Resource Theory*. Springer US.
- [45] Frijda, N. H. (1986). *The emotions*. Cambridge University Press.
- [46] Gajsek, R., Struc, V., Mihelic, F., Podlesek, A., Komidar, L., Socan, G., and Bajec, B. (2009). Multi-modal emotional database: Avid. *Informatica*, 33:101–106.
- [47] Gao, J., Qing, L., Li, L., Cheng, Y., and Peng, Y. (2021). Multi-scale features based interpersonal relation recognition using higher-order graph neural network. *Neurocomputing*, 456:243–252.
- [48] Gerlowska, J., Dmitruk, K., and Rejdak, K. (2021). Facial emotion mimicry in older adults with and without cognitive impairments due to alzheimer’s disease. *AIMS Neuroscience*, 8(2):226–238.

- [49] Girard, J., Chu, W., Jeni, L., and Cohn, J. (2017). Sayette group formation task (gft) spontaneous facial expression database. In *Proceedings of the IEEE International Conference on Automated Face and Gesture Recognition*, Washington, DC, USA.
- [50] Goodfellow, I., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2015). Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63.
- [51] Goodfellow, I. J. and others. (2013). Challenges in representation learning: A report on three machine learning contests. In *Int. Conf. on Neural Information Processing (ICONIP)*, pages 117–124. Springer Berlin Heidelberg.
- [52] Gray, J. A. (1982). The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. *Behavioral and brain sciences*, 5(3):469–484.
- [53] Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo*, pages 865–868, Hannover, Germany.
- [54] Guerdelli, H., Ferrari, C., and Berretti, S. (2021). Interpersonal relation recognition: A survey. *Springer Nature*.
- [55] Guerdelli, H., Ferrari, C., and Berretti, S. (2023). Interpersonal relation recognition: a survey. *Multimedia Tools and Applications*, 82(8):11417–11439.
- [56] Guerdelli, H., Ferrari, C., Berretti, S., and Del Bimbo, A. (2024). Imemo: An interpersonal relation multi-emotion dataset. In *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG) Workshops*. IEEE.
- [57] Guo, X., Polanía, L. F., Garcia-Frias, J., and Barner, K. E. (2019). Social relationship recognition based on a hybrid deep neural network. In *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–5, Lille, France.
- [58] Gupta, R., Sharma, V., Jain, Y., Liang, Y., Broeck, G. V. d., and Singla, P. (2021). Towards an interpretable latent space in structured models for video prediction. *arXiv preprint arXiv:2107.07713*.

- [59] Hajer, G., Claudio, F., Walid, B., Haythem, G., and Stefano, B. (2022). Macro-and micro-expressions facial datasets: A survey. *Sensors*, 22(4):1524.
- [60] Happy, S., Patnaik, P., Routray, A., and Guha, R. (2016). The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*, 8:131–142.
- [61] Healey, J. and Picard, R. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166.
- [62] Heaven, D. (2020). Why faces don’t always tell the truth about feelings. *Nature*, 578(7794):502–504.
- [63] Ho, D. (1998). Interpersonal relationships and relationship dominance: An analysis based on methodological relationism. *Asian J. Soc. Psychol.*, 1(1):1–16.
- [64] Hou, R., Chang, H., Ma, B., and Chen, X. (2019). Video prediction with bidirectional constraint network. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE.
- [65] Hou, R., Chang, H., Ma, B., Shan, S., and Chen, X. (2020). Temporal complementary learning for video person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 388–405. Springer.
- [66] Huang, C., Li, Y., Loy, C., and Tang, X. (2019). Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 42(11):2781–2794.
- [67] Hurley, C., Anker, A., Frank, M., Matsumoto, D., and Hwang, H. (2014). Background factors predicting accuracy and improvement in micro expression recognition. *Motivation and Emotion*, 38(5):700–714.
- [68] Husak, P., Cech, J., and Matas, J. (2017). Spotting facial micro-expressions ”in the wild”. In *Proceedings of the 22nd Computer Vision Winter Workshop, Pattern Recognition and Image Processing Group (PRIP) and PRIP Club*.

- [69] Izard, C. (1971). *The Face of Emotion*. Appleton-Century-Crofts, New York, NY, USA.
- [70] James, W. (1948). *What is emotion?* Appleton-Century-Crofts.
- [71] Kiesler, D. J. (1983). The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90(3):185.
- [72] Kim, E. and Vangala, S. (2016). Vinereactor: Crowdsourced spontaneous facial expression data. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*.
- [73] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.
- [74] Kollias, D. and Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv*.
- [75] Kong, Y. and Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401.
- [76] Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). A few-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36.
- [77] Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(4):1022–1040.
- [78] Lan, M., Ning, S., Li, Y., Chen, Q., Chen, X., Han, X., and Cui, S. (2021). From single to multiple: Leveraging multi-level prediction spaces for video forecasting. *arXiv preprint arXiv:2107.10068*.
- [79] Lee, J., Kim, S., Kim, S., and Sohn, K. (2019). Tri-modal recurrent attention networks for emotion recognition. *IEEE Transactions on Image Processing*.

- [80] Lee, J., Kim, S., Kim, S., and Sohn, K. (2020). Multi-modal recurrent attention networks for facial expression recognition. *IEEE Trans. on Image Processing*, 29:6977–6991.
- [81] Li, S. and Deng, W. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370.
- [82] Li, W., Duan, Y., Lu, J., Feng, J., and Zhou, J. (2020). Graph-based social relation reasoning. In *European Conference on Computer Vision (ECCV)*. Springer, Cham.
- [83] Li, X., Pfister, T., Huang, X., Zhao, G., and Pietikainen, M. (2013). A spontaneous micro-expression database: Inducement, collection and baseline. In *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, Shanghai, China.
- [84] Li, Y., Tao, J., Chao, L., Bao, W., and Liu, Y. (2016). Cheavd: A chinese natural emotional audio-visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8:913–924.
- [85] Liu, X. et al. (2019). Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3561–3569, Long Beach, CA, USA.
- [86] Livingstone, S. and Russo, F. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13:e0196391.
- [87] Lotter, W., Kreiman, G., and Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.
- [88] Lu, J., Hu, J., Liong, V. E., Zhou, X., Bottino, A., Islam, I. U., Vieira, T. F., Qin, X., Tan, X., Chen, S., et al. (2015). The fg 2015 kinship verification in the wild evaluation. In *Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7. IEEE.
- [89] Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and

- emotion-specified expression. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101, San Francisco, CA, USA.
- [90] Lucey, P., Cohn, J., Prkachin, K., Solomon, P., and Matthews, I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 57–64, Santa Barbara, CA, USA.
- [91] Lv, J., Liu, W., Zhou, L., Wu, B., and Ma, H. (2018). Multi-stream fusion model for social relation recognition from videos. In *MMM*.
- [92] MacCrimmon, K. R. and Messick, D. M. (1976). A framework for social motives. *Behavioral Science*, 21(2):86–100.
- [93] Mahmoud, M., Baltrusaitis, T., Robinson, P., and Riek, L. (2011). 3d corpus of spontaneous complex mental states. In D’Mello, S., Graesser, A., Schuller, B., and Martin, J.-C., editors, *ACII 2011. LNCS*, volume 6974, pages 205–214.
- [94] Marchetti, F., Becattini, F., Seidenari, L., and Bimbo, A. D. (2020). Mantra: Memory augmented networks for multiple trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7143–7152.
- [95] Matuszewski, B., Quan, W., Shark, L., McLoughlin, A., Lightbody, C., Emsley, H., and Watkins, C. (2012). Hi4d-adsip 3-d dynamic facial articulation database. *Image and Vision Computing*, 30(10):713–727.
- [96] Mavadati, S., Mahoor, M., Bartlett, K., Trinh, P., and Cohn, J. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160.
- [97] McDougall, W. (1926). *An introduction to social psychology (Rev. ed.)*.
- [98] McDuff, D., Amr, M., and Kaliouby, R. (2018). Am-fed+: An extended dataset of naturalistic facial expressions collected in everyday settings. *IEEE Transactions on Affective Computing*, 10(1):7–17.

- [99] McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., and Picard, R. (2013). Affectiva mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected "in the wild". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, Portland, OR, USA.
- [100] McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.
- [101] Merghani, W., Davison, A., and Yap, M. (2018). A review on facial micro-expressions analysis: Datasets, features and metrics. *arXiv*.
- [102] Mesquita, B. and Frijda, N. (1992). Cultural variations in emotions: A review. *Psychological Bulletin*, 112(2):179–204.
- [103] Metallinou, A., Yang, Z., Lee, C.-c., Busso, C., Carnicke, S., and Narayanan, S. (2016). The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language resources and evaluation*, 50:497–521.
- [104] Mollahosseini, A., Hasani, B., and Mahoor, M. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10:18–31.
- [105] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. on Affective Computing*, 10(01):18–31.
- [106] Mollahosseini, A., Hasani, B., Salvador, M., Abdollahi, H., Chan, D., and Mahoor, M. (2016). Facial expression recognition from world wild web. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 58–65, Las Vegas, NV, USA.
- [107] Morshed, M. G., Sultana, T., Alam, A., and Lee, Y.-K. (2023). Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors*, 23(4):2182.

Bibliography

- [108] Mowrer, O. (1960). *Learning theory and behavior*. John Wiley & Sons Inc.
- [109] Myers, D. (2004). *Theories of Emotion*. Worth Publishers, New York, NY, USA, 7th ed. edition.
- [110] Narang, J. (2020). Face detect. <https://github.com/JeeveshN/Face-Detect.git>.
- [111] Nocentini, F., Ferrari, C., and Berretti, S. (2023). Learning landmarks motion from speech for speaker-agnostic 3d talking heads generation. In *International Conference on Image Analysis and Processing*, pages 340–351. Springer.
- [112] Oatley, K. and Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and emotion*, 1(1):29–50.
- [113] Oksuz, K., Cam, B., Kalkan, S., and Akbas, E. (2020). Imbalance problems in object detection: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 43(10):3388–3415.
- [114] Ortony, A. and Turner, T. J. (1990). What’s basic about basic emotions? *Psychological review*, 97(3):315.
- [115] Otberdout, N., Ferrari, C., Daoudi, M., Berretti, S., and Del Bimbo, A. (2023). Generating multiple 4d expression transitions by learning face landmark trajectories. *IEEE Transactions on Affective Computing*, 15(2):566–578.
- [116] O’Toole, A., Harms, J., Snow, S., Hurst, D., Pappas, M., Ayyad, J., and Abdi, H. (2005). A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816.
- [117] Palmero, C., Selva, J., Smeureanu, S., Junior, J., Jacques, C., Clapés, A., Moseguí, A., Zhang, Z., Gallardo, D., Guilera, G., et al. (2021). Context-aware personality inference in dyadic scenarios: Introducing the UDIVA dataset. In *IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–12.
- [118] Palmero, C., Selva, J., Smeureanu, S., Junior, J. C. S. J., Clapés, A., Mosegui, A., Zhang, Z., et al. (2020). Context-aware personality inference in dyadic scenarios: In-

- roducing the udiva dataset. In *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*.
- [119] Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain sciences*, 5(3):407–422.
- [120] Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.
- [121] Pantic, M., Sebe, N., Cohn, J. F., and Huang, T. (2005a). Affective multimodal human-computer interaction. In *ACM Int. Conf. on Multimedia*, pages 669–676.
- [122] Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005b). Web-based database for facial expression analysis. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*, page 5, Amsterdam, Netherlands.
- [123] Parrott, W. (2001). *Emotions in social psychology: Essential readings*. Psychology Press.
- [124] Parsons, T., Shils, E. A., and Smelser, N. J. (1965). *Toward a General Theory of Action: Theoretical Foundations for the Social Sciences*. Transaction Publishers.
- [125] Petridis, S., Martinez, B., and Pantic, M. (2013). The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202.
- [126] Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. In *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1456, Barcelona, Spain.
- [127] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.
- [128] Polikovsky, S., Kameda, Y., and Ohta, Y. (2009). Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP)*, London, UK.
- [129] Poria, S. et al. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

- [130] Radlak, K., Bozek, M., and Smolka, B. (2015). Silesian deception database: Presentation and analysis. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 29–35.
- [131] Ringeval, F., Sonderegger, A., Sauer, J., and Lalande, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Shanghai, China.
- [132] Robinson, J. P. et al. (2020). Families in wild multimedia (fiw-mm): A multi-modal database for recognizing kinship. *ArXiv abs/2007.14509*.
- [133] Robinson, J. P., Shao, M., Wu, Y., and Fu, Y. (2016). Families in the wild (fiw): Large-scale kinship image database and benchmarks. In *MM '16: Proceedings of the 24th ACM international conference on Multimedia*, pages 242–246.
- [134] Roisman, G., Tsai, J., and Chiang, K. (2004). The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Developmental Psychology*, 40(5):776–789.
- [135] Rubin, D. and Talarico, J. (2009). Comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8):802–808.
- [136] Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrilu, D. M., and Arras, K. O. (2020). Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935.
- [137] Rukavina, S., Gruss, S., Walter, S., Hoffmann, H., and Traue, H. (2015a). Open-emorec-ii-a multimodal corpus of human-computer interaction. *International Journal of Computer Electrical Automation Control and Information Engineering*, 9(5):977–983.
- [138] Rukavina, S., Gruss, S., Walter, S., Hoffmann, H., and Traue, H. C. (2015b). Open-emorec-ii-a multimodal corpus of human-computer interaction. *International Journal of Computer and Information Engineering*, 9(5):1216–1222.

- [139] Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- [140] Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.
- [141] Samad, M., Diawara, N., Bobzien, J., Taylor, C., Harrington, J., and Iftekharuddin, K. (2019). A pilot study to identify autism related traits in spontaneous facial actions using computer vision. *Research in Autism Spectrum Disorders*, 65:14–24.
- [142] Savran, A., Ciftci, K., Chanel, G., Mota, J., Hong, V. L., Sankur, B., and Rombaut, M. (2006). Emotion detection in the loop from brain signals and facial images. In *Proceedings of the eNTERFACE 2006 Workshop*, Dubrovnik, Croatia.
- [143] Schmidt, K. and Cohn, J. (2001). Dynamics of facial expression: Normative characteristics and individual differences. In *Proceedings of the IEEE International Conference on Multimedia and Expo, 2001. ICME 2001*, pages 547–550, Tokyo, Japan.
- [144] Schueller, B., Mueller, R., Eyben, F., Gast, J., Hoernler, B., Woellmer, M., Rigoll, G., Hoethker, A., and Konosu, H. (2009). Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774.
- [145] Sebe, N., Lew, M., Cohen, I., Sun, Y., Gevers, T., and Huang, T. (2004). Authentic facial expression analysis. In *Proceedings of the 6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, Seoul, Korea.
- [146] Shivendra, S., Shajulin, B., Thampi, S. M., Hegde, R. M., Sri, K., Jayanta, M., Vipin, C., Oge, M., Selwyn, P., and Corchado, J. M. (2020). Indian semi-acted facial expression (isafe) dataset for human emotions recognition. In *Advances in Signal Processing and Intelligent Recognition Systems. Communications in Computer and Information Science*, pages 150–162. Springer, Singapore.
- [147] Simons, G., Ellgring, H., and Pasqualini, M. (2003). Disturbance of spontaneous and posed facial expressions in parkinson’s disease. *Cognition and Emotion*, 17(5):759–778.

- [148] Singh, S. and Benedict, S. (2020). Indian semi-acted facial expression (isafe) dataset for human emotions recognition. In *Int. Symposium on Advances in Signal Processing and Intelligent Recognition Systems (SIRS)*, pages 150–162. Springer.
- [149] Sneddon, I., McRorie, M., McKeown, G., and Hanratty, J. (2012). The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41.
- [150] Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.
- [151] Sun, Q., Schiele, B., and Fritz, M. (2017). A domain based approach to social relation recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 435–444.
- [152] Svetieva, E. and Frank, M. (2016). Empathy, emotion dysregulation, and enhanced microexpression recognition ability. *Motivation and Emotion*, 40(2):309–320.
- [153] Tanisik, G., Zalluhoglu, C., and Ikizler-Cinbis, N. (2016). Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*.
- [154] Tcherkassof, A., Dupré, D., Meillon, B., et al. (2013). Dynemo: A video database of natural facial expressions of emotions. *International Journal of Multimedia Its Applications*, 5(5):61–80.
- [155] Tomkins, S. (1984). Affect theory”.
- [156] Urbain, J., Bevacqua, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., Picart, B., Tilmanne, J., and Wagner, J. (2010). The avlaughtercycle database. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta.
- [157] Valstar, M. and Pantic, M. (2010). Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In *Proceedings of the Int’l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70.

- [158] Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014—the three dimensional affect and depression challenge. In *Proceedings of the 4th ACM international workshop on Audio/Visual Emotion Challenge*, Orlando, FL, USA.
- [159] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013-the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, Spain.
- [160] Vandeventer, J., Aubrey, A., Rosin, P., and Marshall, D. (2015). 4d cardiff conversation database (4d cddb): A 4d database of natural, dyadic conversations. In *Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, pages 157–162.
- [161] Vicol, P., Tapaswi, M., Castrejon, L., and Fidler, S. (2018). Moviegraphs: Towards understanding human-centric situations from videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [162] Vijayan, V., Bowyer, K. W., Flynn, P. J., Huang, D., Chen, L., Hansen, M., Ocegueda, O., Shah, S. K., and Kakadiaris, I. A. (2011). Twins 3d face recognition challenge. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE.
- [163] Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Proceedings of the Affective Computing and Intelligent Interaction and Workshops*, pages 1–4.
- [164] Wang, M., Du, X., Shu, X., Wang, X., and Tang, J. (2020). Deep supervised feature selection for social relationship recognition. *Pattern Recognition Letters*, 138:410–416.
- [165] Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., and Wang, X. (2010). A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691.
- [166] Wang, Z., Chen, T., Ren, J., Yu, W., Cheng, H., and Lin, L. (2018). Deep rea-

- soning with knowledge graph for social relationship understanding. In *27th Int. Joint Conference on Artificial Intelligence (IJCAI'18)*, pages 1021–1028. AAAI Press.
- [167] Warren, G., Schertler, E., and Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69.
- [168] Watson, J. B. (1930). Behaviorism, rev.
- [169] Weiner, B. and Graham, S. (1984). An attributional approach to emotional development. *Emotions, cognition, and behavior*, pages 167–191.
- [170] Wen, Z., Lin, W., Wang, T., and Xu, G. (2021). Distract your attention: multi-head cross attention network for facial expression recognition. *arXiv preprint:2109.07270*.
- [171] Xu, L., Song, Z., Wang, D., Su, J., Fang, Z., Ding, C., Gan, W., Yan, Y., Jin, X., Yang, X., et al. (2023). Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238.
- [172] Xu, Z., Qing, L., and Miao, J. (2015). Activity auto-completion: Predicting human activities from partial videos. In *Proceedings of the IEEE international conference on computer vision*, pages 3191–3199.
- [173] Yan, H. and Hu, J. (2018). Video-based kinship verification using distance metric learning. *Pattern Recognition*.
- [174] Yan, H. and Song, C. (2019). Semantic three-stream network for social relation recognition. *Pattern Recognition Letters*, 128:78–84.
- [175] Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., and Fu, X. (2014). Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS One*, 9(1):e86041.
- [176] Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J., and Fu, X. (2013). Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, Shanghai, China.

- [177] Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M., and Zhao, G. (2016). Facial affect “in-the-wild”: A survey and a new database. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Affect “in-the-Wild” Workshop*.
- [178] Zara, A., Maffiolo, V., Martin, J., and Devillers, L. (2007). Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 464–475.
- [179] Zhalehpour, S., Onder, O., Akhtar, Z., and Erdem, C. (2016). Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313.
- [180] Zhang, L., Walter, S., Ma, X., Werner, P., Al-Hamadi, A., Traue, H., and Gruss, S. (2016a). Biovid emo db: A multimodal database for emotion analyses validated by subjective ratings. In *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6, Athens, Greece.
- [181] Zhang, M., Liu, X., Liu, W., Zhou, A., Ma, H., and Mei, T. (2019). Multi-granularity reasoning for social relation recognition from images. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1618–1623.
- [182] Zhang, N., Paluri, M., Taigman, Y., Fergus, R., and Bourdev, L. (2015). Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813.
- [183] Zhang, S. et al. (2023). A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 12(17):3595.
- [184] Zhang, X., Yin, L., Cohn, J., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. (2014). Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706.
- [185] Zhang, Z., Girard, J., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al. (2016b). Multimodal spontaneous emotion corpus

Bibliography

- for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, Las Vegas, NV, USA.
- [186] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2016c). From facial expression recognition to interpersonal relation prediction. *arXiv:1609.06426v2*.
- [187] Zhou, L., Wu, B., and Lv, J. (2018). Sre-net model for automatic social relation extraction from video. In *Proceedings of the 2018 Conference on Computer Vision*.