Bimodal emotion recognition through audio-visual cues

BIMODAL EMOTION RECOGNITION THROUGH AUDIO-VISUAL CUES

BIMODAL EMOTION RECOGNITION THROUGH AUDIO-VISUAL CUES

Dissertation

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on
Thursday, July 8, 2021, at 12:00 hours

By

Esam A. H. GHALEB

Supervisors:

Dr. S. Asteriadis
Prof. dr. G. Weiss

Assessment Committee:

Prof. dr. M.H.M. Winands (Chair)
Prof. dr. A. Wilbik
Prof. dr. D.K.J. Heylen (University of Twente)
Dr. E. Gavves (University of Amsterdam)
Dr. K. Driessens


This research was supported by the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS) (http://www.mathisis-project.eu/), under Grant Agreement nr. 687772.


Dissertation Series No. 2021-16
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Affective Computing, Machine Learning, Audio-Visual Emotion Recognition, Shallow and Deep Metric Learning, Attention Mechanisms
Printed by: ProefschriftMaken, de Bilt
Front & Back: Covers designed by Gizem Ustuner, Amsterdam, The Netherlands

ISBN 978-94-6423-307-0

An electronic version of this dissertation is available at
http://repository.maastrichtuniversity.nl/.

To my beloved family.

PREFACE

Pursuing a Ph.D. is an emotional roller coaster with moments of excitement, disappointments, joy, concerns, etc. I appreciate every bit of this experience. I want to take this opportunity to thank whoever was part of my journey by inspiring, motivating, and helping me during my pursuit of knowledge throughout my studies and work.

First and foremost, I am greatly indebted to Stelios. Without his untiring help and dedicated guidance, the research in this thesis and its writing would not have been possible. I was lucky to have Stelios as a supervisor whose vast knowledge and supervision skills provided me with the right balance as he cheered me up when I felt down and frustrated and curbed my enthusiasm when I was too excited. Stelios gave me supportive and critical feedback whenever it was needed. Throughout my Ph.D. journey, Stelios patiently offered his advice and feedback to improve my writing and its readability. Stelios, thank you for helping me become the researcher I am and teaching me academic skills that will help me in the future. I would also like to thank him for the job offer to work in ProCare4Life and be part of his team. I look forward to the continuation of our collaboration on various exciting research problems. I would also like to thank Gerhard, my secondary supervisor, for his kindness and support. As the head of the department of Data Science and Knowledge Engineering (DKE), Gerhard, you were always approachable and willing to help.

I am thankful to Hazim Ekenel and Makarand Tapaswi. During my Master's thesis, their supervision greatly shaped my research interest in computer vision and intensified my work ethic. Makarand, when I visited the CVHCI Lab at Karlsruhe Institute of Technology, your untiring guidance, invaluable academic skills, and knowledge helped me, beyond my Master's thesis, in the course of my Ph.D. research.

I want to thank the "senior" members of the Affective & Visual Computing Lab (AVCL) at DKE. My special thanks go to Mirela, to whom I turned whenever I needed advice and motivation. Mirela, thank you for your guidance in my Ph.D. research. I thank Enrique for his tremendous help in MaTHiSiS. Enrique, thank you for your supervision in the challenging EU project, MaTHiSiS; I learned a lot from you about patience, diplomacy, and communication skills. I thank Jan for many interesting conversations, insightful comments, and valuable feedback on our work in the seventh chapter. I thank Jerry for recommending Stelios to interview me for the Ph.D. position, his sense of humor and the enjoyable side conversations we had.

I also want to thank the "junior" members of AVCL, with whom we had a great working environment. I enjoyed our lunch breaks and gatherings outside work. With the AVCL members, we have had deep philosophical conversations about research, politics, ethics, food, music, movies, etc. I am thankful for my close friendship with Christos, which widened my perspective on many aspects. Christos, thank you for being a friend I could count on, personally and professionally. I thank Dario, whom I learned a lot from during our research collaboration. Dario, thank you and Nofar for the many BBQs and

dinners. I still need to cook Yemenite food :). I am also happy to be a friend of Bulat. I enjoy working with Bulat and look forward to future research and collaborations with him. I thank Danni, with whom I like to talk and discuss various topics. Danni, thank you for your interesting insights. I thank Mado for being a wonderful office and coffee mate. Mado, we need to publish, how about a paper on coffee? I thank Yusuf, who recently joined DKE, for his enthusiasm and work.

I want to thank my colleagues at DKE: Daniel, Katharina, Lucas, Arjun, Yiyong, Amir, Monica, Nasser, Kiril, Chiara, and Seetheu. Hopefully, we come together again once the pandemic is over! I would also like to thank my friends in the UM Ph.D. academy: Alex, Marta, Gaida, Danielle, and many others, with whom I had a great social life in Maastricht when it was still possible.

Life in Maastricht, far away from the family, would have been more difficult without great friends in Maastricht. I thank Athanasios for being a close friend. During the COVID-19 pandemic, my friendship with Athanasios helped me to overcome difficult times. I enjoy our outdoor activities, the random conversations, plans, and hard work towards fitness goals. I am grateful for the sport team members: Andrew, and Francesco, and Niklas, Deni, Sergio, and Jeron. I thank Gizem for her support and for designing the cover of my thesis. I thank Janneke and Ahmed Jawad for giving me lessons about politics when I tried to become an activist.... I want to thank Marie Labussière and Matteo (M&M). I enjoyed our cycling trips and attending your wedding. M&M, I am looking forward to more cycling and Pineau! Through them, I also met amazing friends from FASoS. Imogen and Li Ming, I enjoyed our cycling, dinners, and day trips. I am looking forward to future trips. I thank Giuseppe for being an awesome and positive person. I also thank him for his pleasant social personality and for bringing people together. I thank Ahmad Hosein for his sincere friendship and his support to me when needed. Maria, thank you for your positivity and enthusiasm. I am looking forward to more parties :). I want to thank the Quizards, with whom I enjoyed our pub quizzes (before the COVID-19 pandemic): Martina, Patrick, Lina, Donique, Katherine, Michelle, Matt and others.

Back in Turkey, I have had great friendships that inspired me and helped me on many occasions. I am grateful for having Gencer, Mehmet, Fulya, Hamdi, Ibrahim, Yusuf, Akram, Ugur, Sacida, Safak, and Akin, as close friends in Istanbul. Thank you for believing in me and supporting me during my education at Istanbul Technical University. My great inspiration will always be my friends in high school. I am proud to be a graduate of Jamal Abdulnasser Secondary highschool. I will always remember the intelligence, motivation, passion, and dedication of my friends there. I would like to thank Mohammed Alsakkaf, Jehad, Tariq and Mohammed Fadl, Sameer, Timor, Khaled, Ammar, and many others. I have always enjoyed our conversation. Thank you for believing and motivating me. You are a source of inspiration. Seeing your success around the world is a source of encouragement. I wish one day we can help Yemen and pay our debts back to our struggling country.

My special thanks go to Marie for her love and support. Marie, thank you for being with me, especially during the stressful period when writing this thesis. Thank you for motivating and inspiring me; I am learning a lot from you. You are helping me to become a better person.

Last, therefore the most important, my deepest thanks and gratitude go to my family.

I was lucky to have amazing parents who complemented each other to raise my siblings and me under tough conditions in the best way they could. Without their support and love, I would not have been where I am today. My mother's discipline and sense of responsibility greatly helped our family. Mum, you pushed our limits to get the best of us. Dad, your composure and hard work helped the whole Ghaleb family and us secure life throughout difficult times. Thank you, and I promise that I will always do my best not to disappoint you. I want to thank my sisters and brothers. I love you all, and you inspire me every day, and you are a source of my pride and motivation. Only you know what we went through to achieve our dreams. You are exemplary and extraordinary; keep up!

Esam A. H. Ghaleb
Maastricht, May 2021

ACKNOWLEDGEMENT

This research was supported by the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS) (http://www.mathisis-project.eu/), under Grant Agreement nr. 687772. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

CONTENTS

List of Figures

LIST OF TABLES

GLOSSARY

AC	Affective Computing
AFEW	Acted Facial Expressions in the Wild
AI	Artificial Intelligence
ANN	Artificial Neural Network
AVER	Audio-Video Emotion Recognition
BoW	Bag of visual Words
CNN	Convolutional Neural Network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
DL	Deep Learning
DLF	Decision-Level Fusion
DML	Deep Metric Learning
DNN	Deep Neural Network
DSIFT	Dense Scale-Invariant Feature Transformation
FCL	Fully Connected Layer
FER	Facial Expression Recognition
FV	Fisher Vector
GMM	Gaussian Mixture Model
GMML	Geometric Mean Metric Learning
HCI	Human-Computer Interaction
ITML	Information Theoretical Metric Learning
KNN	K-Nearest Neighbor
LBP	Local Binary Patterns
LMNN	Large Margin Nearest Neighbor
LSTM	Long-Short Term Memory

MATER	Multimodal Attention mechanism for Temporal Emotion Recognition
MER	Multimodal Emotion Recognition
MERML	Multimodal Emotion Recognition Metric Learning
MHSA	Multi-Head Self Attention
MLP	Multilayer Perceptron
MWTSM	Multi Window Triplet Sets Mining
PCA	Principal Component Analysis
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
RBF	Radial Basis Function
RNN	Recurrent Neural Network
ROIs	Regions of Interest
SER	Speech-based Emotion Recognition
SGD	Stochastic Gradient Descent
SIFT	Scale-Invariant Feature Transformation
SR	Speech Recognition
SVM	Support Vector Machine
SVR	Support Vector Regression
TISML	Temporal Identification Similarity Metric Learning
ToF	Theory of Flow
TTM	Temporal Triplet Mining
UER	Uni-modal Emotion Recognition

1

INTRODUCTION

Even insects express anger, terror, jealousy, and love by their stridulation.

Charles Darwin [1]

Emotions are expressions of the human internal states of mind, thinking, and feelings. They are linked to decision making, mood, motivation, and many aspects of cognition and intelligence. As a result, they play a central role in our social interactions and individual well-being. In human-human interaction, emotions enable people to express themselves beyond verbal domains. Emotions could be expressed through means of facial expressions, body posture, speech, and other gestures. In addition, blood pressure, skin, heart-rate, and other body reactions could be a good indicator of the emotion dynamic state. However, representing and computing emotions are difficult tasks. This is due to the fact that emotions as phenomena are ill-defined and have been a controversial topic in the scientific community [2]. In the $19th$ *century, Charles Darwin [1] proposed that like other traits in humans and animals, emotions also evolved and adapted through millions of years. Emotion expression is complex and embedded within several signals in personal, environmental, and cultural contexts. These facts make emotion perception difficult even for humans.*

Recently, emotion recognition (or emotion detection or emotion classification) has gained a notable amount of research [3–5]. Practically, it is part of the Affective Computing (AC) area, which is a multidisciplinary field that aims to automate the process of emotion recognition, simulation, and inducement. AC is conceptually grounded in affective science. Kappas et al. in [2, 6] suggested that affective states are enclosed body expressions, neurological and physiological responses, and cognitive and metacognitive states. Figure 1.1 demonstrates the foundations of AC in the fields of psychology, neurology, and sociology, according to authors in [6], where affective states and the corresponding changes are regulated and modulated by social and environmental contexts.

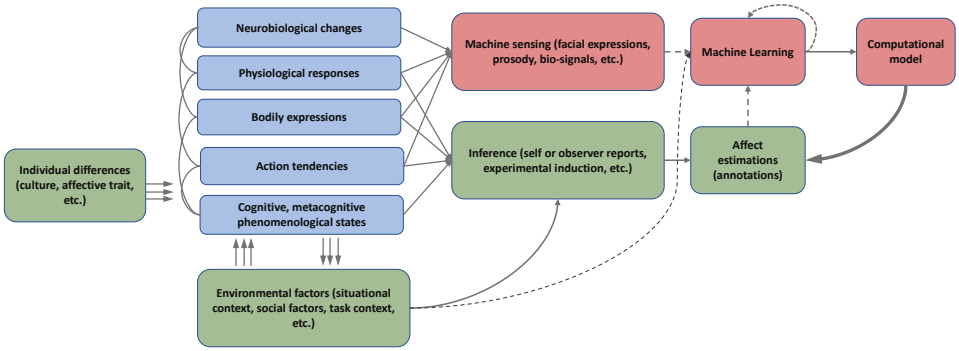


Figure 1.1: Theoretical concepts of Affective Computing (AC) approaches according to D'Mello et al. [2].

However, traditionally, Artificial Intelligence (AI) in human behavior modeling and understanding usually involves cognition processes such as planning, learning and decision making, and perception. Thus, emotion, not being a cognitive process lacked a significant focus of AI research [7], until the beginning of the millennium. However, affect intertwines with cognition and influences the learning process, decision making, and other logical behaviors. Recently, progress has been made in emotion-sensing, detection, and synthesizing. For example, progress in synthesizing emotions is especially achieved for virtual agents [8]. Nonetheless, major advances in AC have been fueled also by recent tremendous improvements in sensing technologies and computational power. In addition, neurological and psychological studies have gained a deeper understanding of emotions which contributed to emotions' modulation and framing.

Therefore, emotion recognition can be addressed through techniques and advances in AC, utilizing pattern recognition approaches. It can be achieved through computational methods by developing classification and regression models to estimate emotions. However, the fact that emotions are multifaceted, complex, socio-psychological and biological concepts, makes automatic emotion recognition a challenging task. This dissertation addresses various problems in multimodal emotion recognition, which is an important task towards achieving AC goals. In particular, it aims to predict emotions through audio-visual cues, which consequently can lead to enhanced interactions between humans and robots, and machines in general. It employs and proposes progressive research towards audio-visual emotion recognition, coming from state-of-the-art techniques in the field of Artificial Intelligence. Earlier research in emotion recognition targeted, either individual modalities (such as facial expressions and acoustic-prosodic cues) or global multimodal emotion recognition [2, 4, 5]. This work focuses on multimodal recognition and exploits temporal interactions between audio-visual channels. It aims to capture modalities' strength for emotion recognition to utilize their complementary information. It adopts recent advances in AC such as Deep Neural Networks (DNNs), Deep Metric Learning (DML), end-to-end learning, and the attention mechanism for Audio-Video Emotion Recognition (AVER).

IN this chapter, Section 1.1 introduces the motivation behind this dissertation based on common theories and concepts in Affective Computing (AC) and presents models for emotions' representation. Section 1.2 elaborates on these concepts behind the motivations and discusses emotion expression and perception from a psychological perspective. Section 1.3 gives an overview of the objectives and the research questions addressed in this dissertation. Finally, Section 1.4 gives an overview of the content of the chapters in the rest of this dissertation.

1.1. MOTIVATION AND CONTEXT

In our daily communication, while we speak, we also use different sensing mechanisms through hearing, feeling, looking, and touching. The multiple means of communication enrich our experiences and provide us with information to process that is used in our interaction with other individuals. In this manner, we get deeper understanding of their needs, behaviors, and intentions. Important factors, while processing multiple sources of data, are emotion expression and perception. Emotion expressivity involves obvious and hidden signals [9]. Obvious signals are vocal utterance, facial expressions, body posture and gestures, gaze, pupil dilation, skin color. Less obvious or hidden cues include, but are not limited to: heart beating, sweating, respiration, temperature, blood pressure, muscular activity. On the other hand, emotion perception involves only obvious signals. All these categories of emotion expression and perception contribute to human-human interaction, as well. People use emotions to synchronize their dialog, signal their pleasure/displeasure, comprehension, and agreement or disagreement [10]. Besides, emotion expressions provide information about people's attitudes, affective state, attractiveness, personality, and cognitive state.

In [11], Minsky suggested that emotions add feelings, values, and other dimensions to rational thinking, just as artists add colors to their black-and-white paintings. In fact, his book "The Emotion Machine" [11] attempts to explain the importance of emotions which are the primary factors to distinguish humans from the rest of animals. In [12], Harari described how emotions played a central role in human evolution during the cognitive revolution, where social interactions and taking care of each other formed the nutshell of human evolution. During this phase, emotions shaped human imaginations and thoughts. For example, research in meta-cognition suggests that emotions could regulate processes such as thinking and acquiring knowledge [13]. There is a scientific consensus with regard to the importance of emotions in human cognition and intelligence.

Similarly, Picard [9] defined Affective Computing (AC) as interdisciplinary research that brings the computational models in computer science together with fields such as psychology, linguistics, and neuroscience. AC aims to add emotional intelligence to Artificial Intelligence (AI). Ultimately, AI should recognize humans' emotions and subsequently adapt its decisions appropriately according to those emotions. AC has gained a notable amount of research in the last two decades. The field itself was established by Picard in 1995 [9]. Despite the impressive progress in AC, there are still many challenges to overcome, and effort is needed to improve technologies, such as affect recognition [3]. These challenges include:

- Intrusive and expensive noisy sensors, such as physiological sensors
- Embedding complex psychosocial experiences in discriminative representations
- Difficulties in collecting and annotating realistic and adequate affective data
- Challenges related to obtaining ground truth labels for supervised learning where inter-rater agreement is usually low
- Lack of universal theory to represent emotions across cultures and individuals due to the difficulty of defining affect [14]

In emotion recognition, initially, many works focused on individual modalities (e.g. emotion recognition using facial expression, speech recognition, or using data from wearable sensors. However, the focus has shifted towards Multimodal Emotion Recognition (MER) through different modalities [2, 4, 5]. MER is more realistic and resembles the way humans detect emotions. Nonetheless, this is a challenging task for several reasons. Firstly, emotions have a highly complex nature which makes it difficult to model and frame them for even humans themselves. Nevertheless, researchers have established a number of theories to simplify these sophisticated experiences (as described in Subsection 1.1.1) [15, 16]. Secondly, multimodal data come from heterogeneous sensors with associated cues obeying different properties and distributions [17]. Therefore, simple fusion or learning algorithms might not be useful to capture the dependencies and complementary information between these modalities, due to the non-linear relationship between them. As a result, there is a need to develop systems to deal with these challenges for accurate and efficient multimodal learning schemes.

In addition, MER has desirable advantages over Uni-modal Emotion Recognition (UER) systems. For instance, it can handle missing data from one modality and overcome noise in the data of other sensors. In particular, e.g. Speech-based Emotion Recognition (SER) can be useless if the user is not speaking, and Facial Expression Recognition (FER) can be misleading if the face is occluded. Nevertheless, a framework with these two modalities can achieve higher performance, due to exploiting the complementarity between them.

Furthermore, recent advancements in sensing technologies and state-of-the-art mathematical methods for data representations and classification procedures have increased research interest in the MER field [3]. In particular, sensing technologies enabled researchers to gather a vast and realistic amount of affective data. Besides, computational methods, such as Deep Neural Networks (DNNs), yielded impressive improvements in emotion recognition due to the increased computational power and the availability of large data corpora [5]. As a result, AC and MER have a tremendous number of applications, which range from education [13, 18, 19], health-care [4, 20, 21], automatic vehicle driving [22–26], to entertainment [27–30]. Emotion recognition through various data channels offers numerous opportunities to improve people's lives while interacting with automatic systems and enhances these systems with emotional intelligence to enable a more human-centered interaction.

Prior to delving into the technical work and the state-of-the-art with regard to MER and AC, the following subsections introduce theories on emotions, visual and vocal emo-

tion perception, and multimodal emotion perception from a psychological perspective. These theories lay down the foundation for the work in this dissertation.

1.1.1. EMOTIONS DESCRIPTION

There is no scientific consensus on the concept of emotions. Nonetheless, psychologists define emotions as subjective affective states that occur in response to stimuli around us. Moreover, theories on emotional states explain emotions as a combination of components such as physiological responses, psychological appraisals, and subjective experiences. For instance, the Arnold-Lazarus appraisal theory of emotions, which is a dominant view on emotions among psychologists [31], suggests that bodily expressions are followed by cognitive appraisals based on subject and context [32, 33]. For example, if someone encounters a bear, his/her emotional response to this stimulus will depend on his/her personal experience. In other words, he/she might feel terrified. However, a well-armed and experienced hunter might be happy. Moreover, emotions are related to concepts such as feelings, mood, personality, temperament, and instincts. Scientists differentiate between emotions, feelings, and mood in terms of intensity, duration, expression, and awareness [34]. Nonetheless, emotions are conscious affective states that tend to be intense with short duration as an instantaneous response to specific stimuli. On the other hand, feelings represent the flow and the experience of emotions. Besides, feelings can be influenced by other factors such as memories and beliefs. Finally, moods are affective states that are long-lived but with low intensity, and they lack a specific contextual stimulus.

It is important to note that, in this dissertation, with regards to the terminology, emotion and affect terms are used interchangeably. They refer to the dynamic state when an individual experiences an emotion. There have been many approaches in the scientific community to frame and model emotions. This section introduces emotion categories that have been widely adopted among psychologists and computer scientists within AC.

DISCRETE MODELS

Emotions can be described in a prototypical way, using discrete categories. This view is common since Darwin's time, which aims to embed emotion description using standard language. This theory implies that emotional states result from an evolutionary process. As a result, they were adapted to help humans in solving specific and recurrent problems [32]. For instance, in this evolutionary process, McDougall [35] referred to the developed information processing mechanisms as instincts [32]. The instincts serve as emotion modules, which include seven basic categories: anger, disgust, elation, fear, subjection, tender emotion, and wonder [35]. Also, James proposed four emotions involving bodily responses: fear, grief, love, and rage [36]. In his theory of emotions, James suggested that emotions are experienced as a result of physiological responses.

Moreover, one of the most prominent works on discrete emotions is the research of Paul Ekman. Paul Ekman [15] suggested that emotions can be categorized into classes, where they can be measured, and they are physiologically discrete, specifically in relation to universal facial expressions. The Ekmanian discrete set of emotions is anger, disgust, happiness, fear, sadness, and surprise. Furthermore, Ekman suggested that facial expressions and emotions are mutually inclusive and usually covary. This is also an

evolutionary view and is influenced by Darwin's work [37]. For instance, widening the eyes, which indicates surprise, can help increasing human vision to navigate through unexpected events. Moreover, in this view, facial expressions, which are associated with discrete emotions, have properties such as short duration. Also, facial muscle actions are symmetric and involuntary. These apparent signals provide a simple yet reliable emotional representation of internal thoughts to coordinate social interactions [32, 38]. Moreover, the Facial Action Coding System (FACS) has been proposed to describe facial muscle movements that are associated with the Ekmanian emotions (FACS is briefly introduced in Subsection 1.2.1).

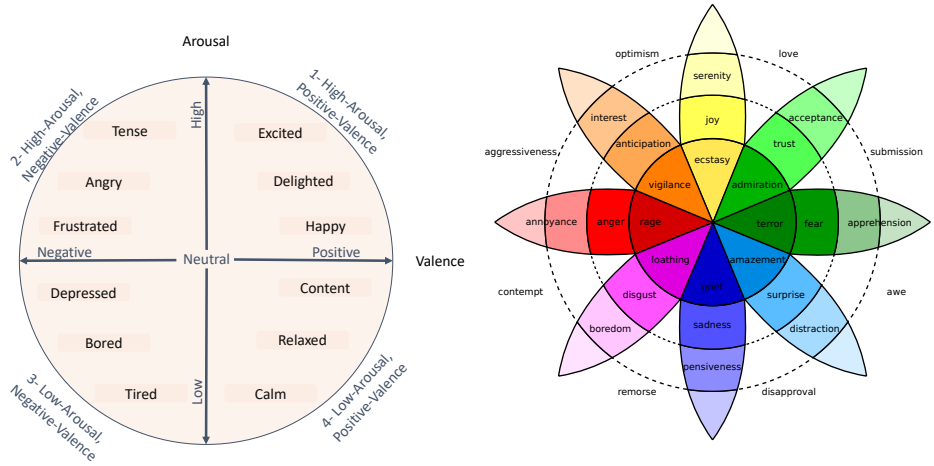
Universality Claims: It is claimed that facial expressions of emotions are universal and can be recognized in different cultures. Studies found high consistency of facial musculature among adults across cultures [14, 37, 38]. For example, empirical research by Ekman [15] showed that people from different cultures were able to recognize the universal facial expressions with high accuracy.

Given the dominant views of basic models of emotion, psychologists have long debated the discrete theories. One of the reasons is that emotions are subjective experiences that are expressed and regulated based on cognitive appraisals. In addition, the universal claim of biologically and physiologically distinct facial responses has been questioned. For instance, the recognition accuracy of universal facial expressions drops among people who are not having Western cultural background [32]. Besides, these models are not descriptive enough to represent all range of emotions in our daily communications [14]. Nonetheless, the idea that emotions are created through natural selection of the evolutionary process is not very controversial since the other cores of the mental subsystem were created by natural selection (e.g., perception, cognition, etc.) [32]. Moreover, proponents of these proposals suggest that emotion activation (the physiological response) is triggered by brain appraisal of perceived stimulus to survive. As a result, these emotions are biologically distinct, as they convey specific purposes. Finally, due to their computability, simplicity, and universality claim, discrete emotions have been extensively studied in psychology and affective computing [38].

DIMENSIONAL MODELS

Dimensional models map emotions on coordinate systems, usually consisting of one or more axes. Within these spaces, distances and similarities between emotions can be captured and represented. Hence, a variety of dimensional models have been proposed in the bibliography. Moreover, dimensional models emphasize concepts such as mood and core affect, where, at any moment, the experiences of affective states are mapped in a continuous space [31]. Most dimensional representations of emotions include arousal and valence. For example, Figure 1.2a shows the scale and regions of arousal and valence in a dimensional representation. Valence indicates how negative or positive the feeling of emotion is, while arousal refers to the energy level of the feeling, from low to high.

In addition, the Pleasure, Arousal, and Dominance (PAD) model uses three scales to represent the nature of emotions [39]. The valence scale measures pleasure (pleasant to unpleasant). For example, fear and anger are on the displeasure side of the scale, while happiness is a pleasant emotion, and it is placed on the pleasure side of the scale. The arousal scale indicates the intensity level of emotion activation. For example, anger



(a) A dimensional representation of emotions (b) Wheel of Emotions by Robert Plutchick [16]

Figure 1.2: Emotion representations

is unpleasant emotion with high intensity, while depression is an emotion with low arousal. Finally, the dominance scale reports how dominant or controlling the experience of an emotion is. For example, anger and fear are unpleasant emotions. However, fear is a submissive emotion, at the same time anger is a dominant emotion. Furthermore, the wheel of emotions, by Robert Plutchick [16], offers a hybrid of both discrete and dimensional representations for emotional classes. According to this wheel, as shown in Figure 1.2b, emotions are grouped on a positive and negative basis, based on their similarities and differences, for example, joy versus sadness and anger versus fear. More importantly, a vertical dimension is used to measure the intensity of the represented emotions.

Finally, although dimensional models are more representative than the discrete models, the projection of these high-dimensional states onto a point on a continuous scale can lead to a loss in information (e.g. some emotions become indistinguishable) [14]. Also, data annotation based on dimensional representations is challenging and requires special training, as these models are not intuitive. Moreover, like the discrete models, dimensional models lack the representation of context and subjective experiences.

APPRAISAL MODELS

Appraisal models resemble the discrete models in the way that they view emotions as a biologically distinct physiological response. The core idea of these models is the claim that emotions are generated following cognitive appraisals (evaluations) of stimuli, contexts, and situations. In addition, the appraisals are influenced by backgrounds, cultures, and personal circumstances. One of the most prominent appraisal models is the OCC model by Ortony, Clore, and Collins [40]. It distinguishes 22 emotion types, which can

change according to the psychological situations they represent. The OCC model studies the structure between appraisal variables and emotions. Specifically, it structures the appraisals and emotions according to three primary reactions: (1) events, which are evaluated by their consequences, (2) actions, which are judged based on agents, and (3) aspects of objects. In particular, the proposed structure by [40] gives the following branches:

- Attraction: Aspects of objects are evaluated based on their intrinsic properties. Affective reactions to objects result in a variation of liking or disliking. In this branch, the appraisals lead to attraction emotions such as *love* or *hate*.
- Attribution: The appraisals of actions of the self-agent or other agents lead to approval or disapproval reactions. These two reactions depend on the agent. For instance, approval and disapproval of self-agent lead to *pride* or *shame*, respectively. On the other hand, when judging other agents' actions, appraisals lead to *admiration* or *reproach*.
- Well-being: When consequences of events are judged, they cause pleasure or displeasure. Appraisals of consequences of positive or negative events related to oneself with irrelevant prospect result in *joy* or *distress* emotions, respectively.
- Prospect-based: when consequences of events are related to oneself with crucial prospects, reactions to the prospect positive or negative events result in *hope* or *fear*, respectively. If the event is already confirmed, *satisfaction* or *fears-confirmed* emotions arise. On the other hand, if events are disconfirmed, reactions lead to *relief* or *disappointment*.
- Fortunes-of-others: If the consequences of the events focus on other people, depending on either the events are desirable or undesirable to the others, appraisals lead to one of the following emotions: *happy-for, resentment, gloating*, or *pity*.
- Well-being/attribution compounds: This last branch is a result of merging the "Well-being" and the "Attribution" branches. It contains the following emotions: *gratification, remorse, gratitude*, and *anger*. This set of emotions arises from the action of agents, as well as the consequences of events.

Computational models derived from the appraisal theory of emotions have been predominant computational emotional models in the design of symbolic AI systems [31]. These systems emphasize explaining the connection between emotions and cognition. For example, computational appraisal models define a set of appraisal variables associated with specific emotion types based on parameters to encode person-environment, and relationship [31]. Subsequently, they apply "if-then" heuristic rules to map emotions based on the defined appraisal variables and context information.

1.2. EMOTION EXPRESSION AND PERCEPTION

In Affective Computing, automatic emotion recognition is based on measuring the physiological responses (expressions) of emotions. These measurements are based on how

the emotions are manifested to and perceived by others. Although these responses and perceptions vary across individuals and cultures, there are universal autonomic patterns [9]. The following subsections discuss theories on emotion expression and perception through obvious signals, such as vocal and visual cues.

1.2.1. FACIAL EXPRESSIONS AND VISUAL CHANNEL

Visual perception of emotions concentrates on the perception of emotions from facial expressions and body gestures and movements [37]. Earlier studies suggested that gestures and body postures are indicators of emotions' intensity [41]. However, recent work, e.g. in [42], showed that bodily movements such as the amount of movement, movement speed, force, fluency, size, and height/vertical position are strong determinants of arousal and potency. Nonetheless, the author of this study noticed the differences between movements in positive and negative emotions are less prevalent. In addition, authors in [41] found that several patterns of body movements occur when portraying emotions, helping in emotion differentiation. These studies showed that body movements and gestures could be an integrated part of a unified nonverbal emotion communication framework. Moreover, visual perception of bodily expressivity has been used in studying personality, which consists in stable patterns of attitudes, mood, and emotions over time [43]. For example, body posture conveys dynamic information about people's expressions (such as excitement and frustration), which are linked to their personality [44].

Facial expressions are one of the two most widely accepted channels for emotion modulation, along with vocal utterance [9]. In the 19th century, Duchenne de Boulogne [45] defined a set of facial muscles for emotions such as attention, lust, disdain, and doubt. Ekman is widely known for his studies in establishing facial expressions that are shared in different age groups and across diverse cultures. Psychologists have also developed the Facial Action Coding System (FACS), which maps measurable facial muscle movements to a discrete emotion space [46]. Action Units (AUs) are used as individual components to break down facial muscle movements. Subsequently, these AUs are employed in a higher-order to identify basic emotions. For example, happiness can be found in two AUs related to raising cheek and pulling lip corners. A comprehensive list of AUs and their mappings to certain emotions can be found in [47], a study by Ekman and Friesen, who proposed the first version of FACS.

In addition, a large body of research suggests that holistic facial configurations could be informative for some affects [48]. Moreover, eyes could be the most important source of affective display. Nonetheless, the combination of facial expression with the voice can shift the attention to the mouth region, which makes the lower part of the face more important during the speech [48]. This proves the fact that combining these sources of information is more relevant for emotion prediction as it gives more confidence in emotion interpretations. Furthermore, there is much uncertainty regarding which information is face perception revealing, in contrast to its valuable information for recognition of personal identities. In other words, the relationship between facial expressions and emotions is ambiguous since there are many social and cultural factors that determine the scope of the emotion spectrum. Since people observe emotions through unified expressive behaviors, the role of multimodal perception of emotions becomes more evident, as

it provides a bigger picture of the expressed behavior, making emotion recognition more reliable. For example, studies show that infants usually pay attention to persons' facial expressions and body gestures, hear persons' voice, and engage in interactions when touched [49], at the same time.

1.2.2. VOCAL-AUDITORY CHANNEL

Voice-based emotion perception and expression have received considerable attention in the last three decades [48, 50]. They are widely acknowledged as the second form of sentic modulation [9, 51]. For instance, emotions are expressed through speech despite inter-speaker variability in the acoustic characteristics of the voice. Researchers have investigated the contribution of acoustic and prosodic features (referred to as emotion encoding or expression) and analyzed how people process these signals as intended by the speaker (a process known as emotion decoding or perception (recognition)) [52]. There is some consensus that prosodic features and changes in speech pitch contribute to the transmission of emotions, whereas loudness seems to be the least important [48]. Besides, acoustic signals contain personal and indexical cues that are nonlinguistic and reveal information about the speaker's gender, identity, age, and emotional state [50]. For example, studies suggest that children recognize emotions in voice before even realizing what is being said [53].

Scherer [54] stated that discrete emotions experienced by the speaker are reflected in specific patterns related to speech. Furthermore, Scherer suggested that the vocalizer's affective expression is accompanied by physiological changes in vocal production [50]. On the contrary, some studies suggested the link between speech acoustics and the activation of some sets of continuous dimensions. For example, Bachorowski et al. [50, 55] mentioned that emotion-related acoustic features are traceable to two orthogonal dimensions of arousal and pleasure. Moreover, Scherer [56] studied human ability to recognize discrete emotions. The research was conducted using 14 professional actors who portrayed 12 emotions. In this study, Scherer found that humans recognize emotions from purely vocal stimuli with an accuracy of 60%. According to Scherer, human perception of emotions from the speech is best for fear, anger, and sadness [50]. However, the perception rate drops for positive emotions (e.g. happiness and interest) and disgust. Finally, some views link vocal signals to inducing emotional responses. For example, Bachorowski and Owren [50] suggested that vocal signals are more about influencing listener emotional responses than communicating emotions. This view is also traced back to Darwin's work on the veridical associations of vocal expressions of emotions and the emotional state of the vocalizer.

1.2.3. MULTIMODAL EXPRESSION AND PERCEPTION OF EMOTIONS

In general, the early work of multimodal research is motivated by what is known as the McGurk effect [57]. The McGurk effect suggests the interaction between vision and speech perception. For example, when people hear the syllable /ba-ba/ and at the same time watching the lips of a speaking person saying /ga-ga/, they perceive a third voice: /da-da/ [57]. This interesting observation indicates the effect of one modality on others, and also how perception could be altered by different sources of information. This phenomenon offers an example of audiovisual integration and contributes to a significant

amount of work that explores the cross-modal interactions [48, 58].

Much research in Multimodal Emotion Recognition (MER) is motivated by the McGurk effect, as well as a large body of work in psychology which proves the importance of multimodal perception for emotion recognition. For example, Auberge et al. [58] show that even in visible emotions (e.g. amusement), audio modality carries important information that is related to the smiling face. Their study suggests that when the McGurk paradigm is applied to discordant amused stimuli, acoustic information clearly interacts with visual decoding. Also, acoustic features of the vocal utterance, such as volume, variability in pitch, rhythm, are associated with certain aspects of facial expressions and body gestures [49].

Moreover, the basic emotion theory suggests that emotional responses are adaptive mechanisms to evolutionary threats and opportunities. These responses are manifested through peripheral physiology, bodily and facial expressions, and other expressive behaviors that govern social interactions [59]. In this framework, emotions are closely related to actions [59]. For example, Darwin claimed that facial expressions are the residual actions of behavioral responses which are accompanied by bodily gestures, postures, vocalization, and other physiological responses [37]. Researchers refer to this process as "emotional packages". For example, the evolutionary view on basic emotions suggests that facial expressions covary with emotional experiences. Research consistently shows that facial expressions convey emotional states across cultures and individuals [52]. Moreover, these facial expressions are coordinated with physiological responses such as somatic activity, skin conductance, and cardiovascular responses. For instance, fear, anger, and disgust are linked to elevated cardiovascular activation [37].

Beyond facial expressions, studying multiple modalities of emotion expression enabled researchers to discover the relative contribution of different modalities in emotion expression and perception [59]. For example, De Gelder [48] postulated that bodily expressions of fear enable us to discern the cause of threats and the subsequent actions. On the other hand, the face can only communicate the existence of a threat. Besides, a large body of research in the field of psychology shows that acoustic features are more reliable in measuring arousal than measuring affectively valence experiences [50]. When it comes to discrete emotions, empirical outcomes suggest the association of acoustic features and expression of anger. As a result, each modality exhibits strength in emotional expression, which indicates the multimodal expression of emotional states.

In addition, emotion perception appears to be multimodal. Humans usually interpret emotions manifested in a variety of behaviors, such as facial and audio expressivities and body gestures. Darwin [1] suggested that emotions are closely related to actions, which means that it does not matter in which way these emotions were displayed. For example, the perception of an angry face or an angry voice leads to the conclusion that the person is angry, not that the face looks angry or the voice sounds angry. De Gelder and Voormen in [48] conducted three experiments on bimodal perception to study the integration mechanism of face and voice. Their study is based on presenting audio-visual stimuli in which they created varying degrees of discordance in the expressed face and in the tone of the accompanying voice. The experiments showed that the identification of emotions in the face is biased in the direction of the way the tone of voice was presented. They also concluded that there exist bidirectional links between affect detection

structures in vision and audition.

Besides, various studies on human ratings for emotion through audio-visual cues revealed interesting observations in terms of which modality is more useful for which kind of emotions. In [52], the authors showed that the recognition of disgust and fear is better with audio-visual cues. On the other hand, anger and happiness are recognized accurately with single modalities. These observations are also consistent with our findings throughout this dissertation. For example, researchers in [52] showed that the human visual perception alone achieved a 69.0% accuracy, while the perception in audio alone was 45.5%. However, the presentation of both visual and auditory signals to human raters increased their perception by at least 5.8%. From a computational perspective, D'Mello and Kory conducted a meta-analysis on multimodal emotion recognition systems. Their study revealed that multimodal emotion detectors are consistently better than their underlying unimodal detectors, with an average improvement of 9.8%.

1.3. RESEARCH OBJECTIVES AND QUESTIONS

Given how pertinent visual and vocal channels are, there is an increased interest in the Affective Computing (AC) and Human-Computer Interaction (HCI) fields towards enhancing digital devices with emotion recognition capabilities through audio-visual cues. This thesis aims at recognizing the displayed affective information through facial expressions and speech signals by contributing to and employing techniques and methods in the fields of AC and Artificial Intelligence (AI). Specifically, the main interest of this dissertation is exploiting the powerful aspects of audio-visual cues for bimodal emotion recognition in video clips. This work represents a step further to enhance automatic systems with basic elements of emotional intelligence to enable a more human-centered interaction. Systems of emotion recognition can be used to achieve richer and human-like communications in settings like HCI. Chapter 3 reviews the state-of-the-art in affective computing in general, focusing on multimodal emotion recognition. It motivates the research problem, technically, by providing information regarding the current state of technologies, mathematical models in AC, and the advantages and disadvantages of various modalities for emotion recognition. This dissertation addresses the problem of automatic recognition of emotions through audio-visual signals. It focuses on achieving the following four main objectives:

- Obtaining robust data modeling and representation for emotion recognition.
- Building an efficient fusion of audio-visual representations.
- Exploiting the temporal dynamics of emotion expression and perception.
- Producing an attentive system to multimodal and temporal expressions of emotions.

These objectives serve as a list of requirements for the developed methods and techniques. In order to achieve the objectives mentioned above, we address them as part of the following formulated four main research questions:

First research question: *How to extract and fuse robust features and which is their contribution to automatic emotion recognition?*

In order to answer this question, two separate studies have been conducted. The first preliminary study focuses on investigating and building temporal features for audio and video representations. It constructs a pipeline to explore and exploit audio-visual data and to efficiently fuse them for emotion recognition. Previous research investigated automatic emotion recognition through individual modalities. For example, the visual modality has been utilized for facial expression recognition using several features. These features include appearance representations based methods (Gabor filters [60], Local Binary Patterns (LBP) [61], Scale-Invariant Feature Transformation (SIFT) [62]), geometric features [63], and unsupervised feature learning methods such as the recently adopted Convolutional Neural Network (CNN) models [64]. In addition, audio-based emotion recognition has shown a promising direction, and features such as prosody, jitter, and fundamental frequencies have been shown to be useful [65]. On top of the extracted audio-visual features, classification algorithms, such as Support Vector Machines (SVMs) with different kernel methods, were employed [66, 67]. Alternatively, weighing approaches, such as a random search for optimal weights on the predictions of each modality obtained from Deep Neural Networks (DNNs) [68], were conducted for multimodal fusion.

In this dissertation, we address this research question by proposing a multimodal framework for emotion recognition in a video-clip by taking advantage of both audio and video features. Moreover, a hierarchical fusion approach is proposed. In this fusion framework, various feature extraction algorithms are employed to obtain representations from audio and video channels. Subsequently, features are encoded via Fisher Vectors (FVs) [69] which project the different types of features onto joint subspaces and also facilitate the analysis of videos of varying durations. Next, the fusion of these representations uses the KullBack-Leibler (KL) divergence minimization in order to obtain optimal configurations for multimodal fusion.

The second study explores the usage of interaction parameters with learning materials as features for affect understanding. These interactions include student's performance, the time needed to attain her/his goal, level of the difficulty, and skill level. The framework uses these features to predict whether a student was engaged, frustrated, or bored during the learning activity. Affective states can be directly linked to a students' performance during learning. Driven by the Theory of Flow (ToF) model, we investigate the correspondence between the prediction of users' self-reported affective states and the interaction features. Cross-subject evaluation on a dataset of 32 users interacting with the platform demonstrated that the proposed framework can achieve a significant precision in affect recognition. Consequently, being able to retrieve the affect of a student can lead to more personalized education, targeting higher degrees of engagement and, thus, optimizing the learning experience and its outcomes.

Second research question: *What is the impact of multimodal learning on automatic emotion recognition?*

As stated in the previous sections, emotion expression and perception intrinsically have a multimodal nature. Studies have shown the impact of multimodal fusion which

can lead to a significant improvement in emotion recognition [3]. Even though each modality exhibits unique characteristics, multimodal learning exploits the advantages of complementary information from diverse modalities. However, modalities' dependencies and relationships are not fully exploited for audio-video emotion recognition. In the literature, most of the recent research studies on multimodal emotion recognition are limited either to feature concatenation of various representations or late fusion of individual modalities in combination with various classification algorithms [66, 68, 70]. Moreover, by employing an efficient metric distance, the accuracy of many classification and retrieval problems can be increased, as it contributes to obtaining an improved performance and robust representation [71, 72]. Metric learning approaches learn distances to bring similar inputs closer and dissimilar ones further, which are more discriminative than the conventional Euclidean distance.

This dissertation proposes a new Multimodal Emotion Recognition Metric Learning (MERML) framework, which leverages the audio-visual information to learn Mahalanobis metrics jointly for each modality. This objective is obtained by learning the discriminative latent subspace contributing to robust emotion classification. Further, this new distance is incorporated efficiently in Radial Basis Function (RBF) based SVMs, benefiting the emotion classification task.

> **Third research question:** *What is the role of temporal dynamics in audio visual cues, in automated emotion recognition?*

Psychological and neurological studies have argued that negative and positive emotions are displayed and recognized at different speeds and rates. In addition, research demonstrated that emotion perception might require a different amount of time for an accurate detection [73]. Thus, these alterations could be exploited efficiently through a temporally-trimmed framework. Literature studies lack in-depth analysis and utilization of emotions variation as a function of time [73–77]. Therefore, building on our findings on the research conducted using metric learning, we address this research question by proposing a novel multimodal temporal deep network framework. The proposed method embeds video clips using their audio-visual content, onto a metric space, where their gap is reduced and their complementary and supplementary information is explored.

> **Fourth research question:** *How can we capture the contributions of the temporal dynamics of affect display using attention mechanisms?*

Emotion display is usually following a common pattern, consisting of on-set, apex, and off-set phases. Since the apex stage is the one usually capturing significant expressivity, it is the phase used most frequently in emotion recognition [78]. Nevertheless, neighboring windows can also benefit the emotion recognition task since the temporal pattern of emotion display can change, depending on the audio-visual modalities. For example, happiness could be initially expressed through facial expressions, while corresponding time windows in the audio channels are not useful yet. However, the following windows could provide valuable information for the auditory cue [77]. We propose a framework which consists of bi-audio-visual time windows that span short video clips labeled

with discrete emotions. Attention is used to weigh these time windows for multimodal learning and fusion. Consequently, we conducted extensive research to evaluate how the framework models the time in both modalities, hence, enhancing identification performance.

1.4. THESIS OVERVIEW

This dissertation is organized into seven chapters, besides this introductory one, which presented the theoretical background of the dissertation topics and formulated the research questions and objectives that guided the research in this dissertation. The structure of the rest of this dissertation is as follows:

Chapter 2 gives an overview of the technical background and tools used in Affective Computing (AC) and audio-visual emotion recognition. In particular, it presents an introduction to the state-of-the-art in the fields of Artificial Intelligence (AI) and machine learning employed in this dissertation.

Chapter 3 introduces a taxonomy of the research problem and a survey of the past work on unimodal and multimodal emotion recognition. Specifically, it concentrates on the recent advances in the field within the scope of the research pertaining to the dissertation.

Chapter 4 presents the two research studies conducted to address the first research question: *how to extract and combine meaningful features and which is their contribution to automatic emotion recognition?* It details feature extraction methods, fusion techniques based on Information Gain (IG) and Genetic Algorithms (GAs) and presents the findings and the obtained results. Next, an additional study about affective state recognition from the interactions with learning materials is detailed. It presents the data collection, tracking, and annotations obtained from students interacting with learning materials. This part of the research demonstrates the potential usage of such features to track students' affective states in terms of engagement, frustration, and boredom.

Chapter 5 addresses the second research question: *what is the impact of multimodal learning on automatic emotion recognition?* A multimodal Mahalanobis distance metric is employed for audio-visual emotion recognition. The proposed metric is incorporated in Radial Basis Function for Support Vector Machine (SVM). The proposed method shows that an efficient fusion of audio-visual representations contributes to enhanced performance.

Chapter 6 demonstrates the impact of using Deep Metric Learning (DML) for audio-visual emotion recognition and answers the third research question: *what is the role of temporal dynamics in audio visual cues, in automated emotion recognition?* It presents the study conducted to evaluate the impact of time on emotion prediction as well as data mining for training a

similarity metric approach for multimodal learning. This study shows that taking advantage of the emotions' temporal display, through incremental perception, is beneficial. In addition, in this proposed framework, the developed method and the associated techniques, such as triplet sets mining and data augmentation, contributed significantly to the stability and the performance of the framework for emotion recognition. The temporal perception of audio-visual cues shows that recognition rates increase faster for positive emotions than for negative ones.

Chapter 7 presents a study addressing the fourth research question: *how can we capture the contributions of the temporal dynamic of affect display using attention mechanisms?* It presents a thorough analysis on using attention, its role in multimodal fusion, and how attention mechanisms are able to utilize embeddings from all time windows and to capture the interactions between video and audio modalities. A comprehensive meta-analysis is presented to explain the multimodal interaction, the benefits of joint modeling of audio-visual cues, and how robust is the framework when exposed to challenging conditions during the training and testing phases. The results show that joint modeling of the audio-visual channel using the attention mechanism brings their entropies closer, hence, improves their performance. Also, the evaluations show that exposing the framework to similar conditions during the training and testing processes results in robust performance.

Chapter 8 concludes the work on audio-visual emotion recognition and highlights some directions for future research in the field.

2

MACHINE LEARNING

I have always been convinced that the only way to get artificial intelligence to work is to do the computation in a way similar to the human brain

Geoffrey Hinton[1]

This chapter presents the state-of-the-art methods in machine learning which are employed in the field of Affective Computing (AC). In particular, it focuses on the approaches which are used throughout the dissertation. In addition, as this work aims to contribute to the automation of emotion recognition, the proposed frameworks in the content chapters utilize techniques from computer vision and machine learning. This chapter introduces the fundamental topics that will be useful in the following chapters. The chapter is structured as follows. Section 2.1 gives an overview of essential concepts in machine learning, which formed a core part of this work. Section 2.2 explains methods that benefit from data similarity for learning, such as metric learning, while Section 2.3 gives a brief introduction to Support Vector Machines (SVMs). Section 2.4 details the fundamentals of Deep Neural Networks (DNNs) and describes a series of related concepts and key architectures. Finally, Section 2.5 points out the context of this chapter within the dissertation and how the content chapters benefited from and contributed to the presented approaches.

[1] https://www.utoronto.ca/news/u-t-computer-scientist-takes-international-prize-groundbreaking-work-ai

The following notions will be used throughout this dissertation:

x	A scalar
$\boldsymbol{x}$	A vector
X	A matrix
$\boldsymbol{X}$	Tensor
$\mathbb{R}$	A set of real numbers
$\{0, 1, ..., n\}$	A set containing n integers between 0 and n
$[a, b]$	An interval of numbers including a, and b
	Indexing
x_i	element i of a vector $\boldsymbol{x}$, and indexing starts at 1
$X_{i,j}$	Element i, j of a matrix $\boldsymbol{X}$
$\boldsymbol{X_i}$	Vector $\boldsymbol{x}$ which corresponds to a row i of a matrix $\boldsymbol{X}$

$f(\boldsymbol{x};\boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parameterized by $\boldsymbol{\theta}$.
	Dataset
$\mathbb{X}$	A set of training examples
$\boldsymbol{x}$	A row (sample) in the training or the testing dataset
d	dimensionality of the data input (sample)
n	Number of samples in the training set
y	A label of a data input
$\boldsymbol{y}$	Vector of the labels
t	Temporal index in a sequential data
c	The number of classes

2.1. MACHINE LEARNING CONCEPTS

MACHINE learning is a family of approaches for Artificial Intelligence (AI), where the foundation of the field is to design computational procedures and mathematical methods that can learn to perform a certain task given data samples [79–81]. A method from the field of machine learning must be scalable and adaptable to different domains, where capturing data patterns and making decisions is important. For example, a task such as sorting spam emails from legitimate ones cannot be easily done with a sequence of instructions and hand-crafted rules. Instead, an approach borrowed from the area of machine learning can acquire the necessary knowledge from a set of spam and non-spam emails. In other words, a proper approach should make use of or even discover related features and build a model that, with high levels of accuracy, can decide on whether an email is a spam or not [80]. In machine learning, there are two main learning categories: supervised and unsupervised learning.

SUPERVISED LEARNING

One large family of techniques in machine learning is making use of example (training) data that are annotated with regards to their category or value. For example, a training set could contain still images of dogs/cats, or in the case of this dissertation, the training set consists of short video-clips labeled with emotions. Supervised learning uses this

labeled data, where each sample is represented by a feature vector and a target class or value. Also, supervised learning can be applied on classification or regression tasks. In a classification task, inputs' labels belong to a discrete class. For example, in facial expression recognition, Ekmanian discrete emotions are widely used as labels for a given face. A learning algorithm should learn to associate a facial image with one of the discrete emotions, such as happiness or surprise. In a regression task, inputs' labels have continuous values. For example, predicting the values of arousal and valence in the dimensional models of emotions (as described in Subsection 1.1.1). The goal of supervised learning goal is to learn the parameters ($\boldsymbol{\theta}$) so as to map the input features to the given output. For example, one of the simplest formulation of the supervised learning could be a linear regression that can be defined as

$$\hat{y} = f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \tag{2.1}$$

where $\boldsymbol{f(.)}$ is the supervised model, $\boldsymbol{x}$ is an input data, and $\boldsymbol{w_0}$ and $\boldsymbol{w}$ are learnable parameters ($\boldsymbol{\theta}$) that allow for an accurate mapping from $\boldsymbol{x}$ to y. The machine learning model is designed to optimize the parameters, $\boldsymbol{\theta}$, such that the approximation error to the target class (value) is minimized [80].

In this dissertation, the discrete emotions are adopted as a model to represent emotions. Therefore, this chapter focuses on reviewing machine learning's classification approaches. In addition, the approaches employed in this dissertation have made use of supervised learning techniques which are presented in the following subsections.

UNSUPERVISED LEARNING

Unsupervised learning is based on discovering patterns, solely relying on data without any supervised output. Unsupervised learning assumes an underlying structure in the input data. Hence, a density estimation function is used to capture this structure [80, 81]. Some patterns occur more often than others and unsupervised learning aims to focus on discovering their probability densities. For example, clustering uses estimation functions that extrapolate algorithmic relationships on data attributes. It groups data onto clusters based on specific patterns and similarities between data samples.

Unsupervised learning is essential in various applications, such as image compression, document clustering, and human genome sequence modeling [80]. More importantly, unsupervised learning can play a role when data annotation is challenging, and the patterns within data could be useful. For example, many Deep Neural Network (DNN) models rely on unsupervised learning, such as autoencoders and Generative Adversarial Networks (GANs). In these models, the parameters could be learned via unlabeled data, yielding impressive results for image generation or feature representation [79].

2.2. SIMILARITY LEARNING

Similarity learning consists in supervised learning methods. In these approaches of machine learning, the main assumption is that similar data produce similar outputs. This similarity is based on distance measurement between objects/classes of a training set ($\mathbb{X}$).

2.2.1. K-NEAREST NEIGHBOR

The K-Nearest Neighbor (KNN) technique assumes that inputs of similar classes are drawn from the same distribution, hence a neighborhood could be defined to infer their similarity [81]. This can be measured with the help of distance functions such as the Euclidean distance. In KNN, the training phase consists in storing all the feature vectors and their labels of training data. Then, when asked to classify a test point $\boldsymbol{x}$, KNN looks up the nearest k entries in the training set ($\mathbb{X}$) and returns their major label as a prediction for the test sample, based on a distance metric.

This procedure is referred to as "majority voting". In other words, a neighborhood ($N^k(\boldsymbol{x}^{(i)})$) is defined by the k closest data samples of the entry i ($\boldsymbol{x^{(i)}}$) in the training data set according to a distance metric [81]. Therefore, $\hat{y}$ will have the label of the dominant class in the defined neighborhood. For example, if $k = 1$, then each test point is labeled with the class of the assigned closest training point.

2.2.2. METRIC LEARNING

Nearest neighbor algorithms could be generalized onto a distance metric other than the L^2 norm, such as learned ones. This family of specialized algorithms can improve the performance of KNN techniques, significantly. Distance metric learning is a family of parametric approaches with a customized distance. Customized distances can be learned using the similarities and dissimilarities in the training examples [82]. Standard distance metric assumes a higher similarity between features of similar inputs and a bigger difference between non-similar inputs, by applying the standard Euclidean distance

$$d(\boldsymbol{x^{(i)}}, \boldsymbol{x^{(j)}}) = \sqrt{(x_1^{(i)} - x_1^{(j)})^2 + (x_2^{(i)} - x_2^{(j)})^2 + ... + (x_d^{(i)} - x_d^{(j)})^2},$$

which can be also written in a vectorized form as follows:

$$d(\boldsymbol{x^{(i)}}, \boldsymbol{x^{(j)}}) = \sqrt{(\boldsymbol{x^{(i)}} - \boldsymbol{x^{(j)}})^T (\boldsymbol{x^{(i)}} - \boldsymbol{x^{(j)}})}$$

In addition, if we consider an identity $\boldsymbol{I}$ matrix, the previous formula can be rewritten as follows:

$$d(\boldsymbol{x^{(i)}}, \boldsymbol{x^{(j)}}) = \sqrt{(\boldsymbol{x^{(i)}} - \boldsymbol{x^{(j)}})^T \boldsymbol{I} (\boldsymbol{x^{(i)}} - \boldsymbol{x^{(j)}})}.$$

As a result, this equation is represented as:

$$d_I^2(\boldsymbol{x^{(i)}}, \boldsymbol{x^{(j)}}) = \begin{pmatrix} (x_1^{(i)} - x_1^{(j)}) & (x_2^{(i)} - x_2^{(j)}) & \cdots & (x_d^{(i)} - x_d^{(j)}) \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} (x_1^{(i)} - x_1^{(j)}) \\ (x_2^{(i)} - x_2^{(j)}) \\ \vdots \\ (x_d^{(i)} - x_d^{(j)}) \end{pmatrix} \tag{2.2}$$

In standard distance metric learning, the goal is to find an optimal metric M according to similarity and dissimilarity constraints. For instance, in the **low-rank Mahalanobis metric learning** approach, a Mahalanobis matrix $M = W^T W$ is learned using convex optimization, where M is symmetric and positive. The learned matrix $W \in \mathbb{R}^{p \times d}, p \ll d$ projects the high dimensional feature vectors (in the original space)

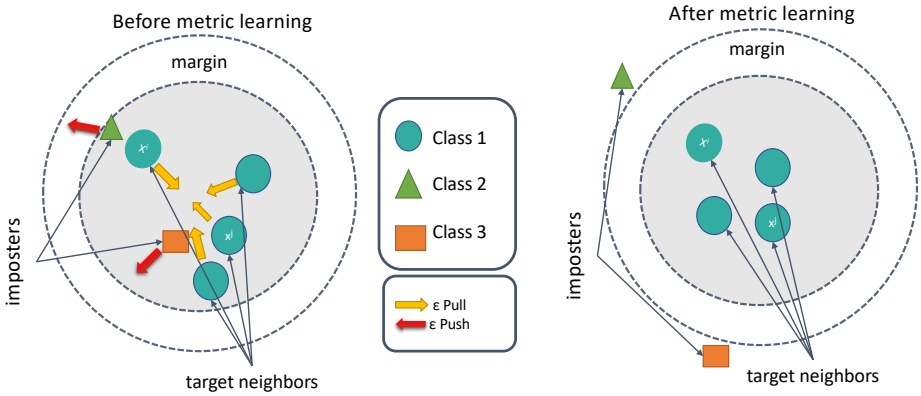


Figure 2.1: Figure from [71] showing the procedure of the distance metric learning, Large Margin Nearest Neighbor (LMNN). In this figure, the target neighbors are set to 3, where the goal is to keep data points with the same label as close as possible within the defined sphere while pushing away the imposters (the data points with different labels than the target class within the sphere).

$\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ to a latent space with lower dimensions $W\boldsymbol{x}^{(i)} \in \mathbb{R}^p$, where p is the dimensionality of the new learned subspace and d is the dimensionality of the original space. As a result, the new formulation of equation (2.2) can be written as follows:

$$\begin{aligned} d^2_M(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) &= (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T M(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) \\ &= \|W\boldsymbol{x}^{(i)} - W\boldsymbol{x}^{(j)}\|^2_2 = \|W(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})\|^2_2 \\ &= (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T W^T W(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) = d^2_W(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \end{aligned} \tag{2.3}$$

In other words, metric learning modifies the standard Euclidean distance to improve its discriminative ability, such that the distance between similar classes would be as small as possible, while enlarging it otherwise. Another benefit of metric learning includes dimensionality reduction of the feature vectors, by the linear transformation matrix (projection): $W \in \mathbb{R}^{p \times d}$, where $p \ll d$, and $p \geq rank(W)$. Note that, if $M = I^{d \times d}$, where $I^{d \times d}$ denotes the identity matrix, then the metric is reduced to a Euclidean distance (as shown in equation (2.2)). Furthermore, a proper metric learning must obey the following properties:

- Non-negativity: $d^2_M(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \geq 0$
- Symmetry: $d^2_M(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = d^2_M(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(i)})$
- Triangle inequality: $d^2_M(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + d^2_M(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(k)}) \geq d^2_M(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(k)})$
- Distinguishability: $d^2_M(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = 0 \iff \boldsymbol{x}^{(i)} = \boldsymbol{x}^{(j)}$

Over the last twenty years, researchers have developed many approaches for distance metric learning. The following list introduces examples, which are related to the methods proposed within this dissertation:

- **Large Margin Nearest Neighbor (LMNN)**: In [71], authors proposed LMNN such that the k-nearest neighbors belong to the same label, while the examples of other classes (labels) are pushed away by a *large margin*. As shown in Figure 2.1, for each data point $\boldsymbol{x}^{(i)}$, LMNN defines target neighbors where they are pulled towards this data point to have minimum distance, while the imposters (data points with different labels) are pushed away. This process is optimized with the following cost function:

$$\sum_{ij} \eta_{ij} d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + c \sum_{ijl} \eta_{ij}(1 - y_{ij})[1 + d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) - d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(l)})]_+$$

where $\eta_{ij} \in \{0,1\}$ indicates whether input $\boldsymbol{x}^{(j)}$ is a target neighbor to $\boldsymbol{x}^{(i)}$ or not, y_{ij} is the target label, and $[]_+$ indicates the standard hinge loss and c is a positive constant.

- **Information Theoretical Metric Learning (ITML)**: In ITML [83], an information theoretic approach is proposed to learn a Mahalanobis distance function. Authors in [83] formulated the problem by minimizing the differential relative entropy between two multivariate Gaussians (M_0 and M),

$$KL(p(\boldsymbol{x}; M_0) || p(\boldsymbol{x}; M)) = \int p(\boldsymbol{x}; M_0) \log \frac{p(\boldsymbol{x}; M_0)}{p(\boldsymbol{x}; M)} d\boldsymbol{x},$$

with the following constraints:

$$d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \leq \alpha_{similar} \ , \ (\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \in \S$$
$$d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \geq \alpha_{dissimilar} \ , \ (\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \in \mathbb{D}$$

where § and $\mathbb{D}$ are sets of similar and dissimilar data points, respectively. As a result, the distance between similar data samples are kept below $\alpha_{similar}$, and the distance between dissimilar samples is pushed above $\alpha_{dissimilar}$.

- **Geometric Mean Metric Learning (GMML)**: In GMML [84], authors take into consideration the impact of the dissimilar data points. As a result, M is proposed to decrease the distance over all similar points, while M^{-1} measures the interpoints distances of dissimilar points using the following objective:

$$\sum_{(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \in \S} d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + \sum_{(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \in \mathbb{D}} d_{M^{-1}}^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$$

- **Diagonal Metric Learning:** This is probably the most straightforward distance metric, where the aim is to learn a weight for each dimension in the feature space. The learning of these weights can be carried by conventional linear Support Vector Machines (SVMs), since they have a similar formulation (as explained in Section 2.3). The relationship of this method to standard Euclidean distance can be observed with the following diagonal metric

$$d_M^2(\boldsymbol{x^{(i)}}, \boldsymbol{x^{(j)}}) = \begin{pmatrix} (x_1^{(i)} - x_1^{(j)}) & (x_2^{(i)} - x_2^{(j)}) & \cdots & (x_d^{(i)} - x_d^{(j)}) \end{pmatrix} \begin{pmatrix} M_{1,1} & 0 & \cdots & 0 \\ 0 & M_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_{d,d} \end{pmatrix} \begin{pmatrix} (x_1^{(i)} - x_1^{(j)}) \\ (x_2^{(i)} - x_2^{(j)}) \\ \vdots \\ (x_d^{(i)} - x_d^{(j)}) \end{pmatrix}$$

- **Low-rank Joint Metric Learning:** This metric learning measures the similarity and difference between feature vectors [85, 86]. It corresponds to the difference between the low-rank Mahalanobis distance, $(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T W^T W(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})$, and the low-rank kernel (inner product), $\boldsymbol{x}^{(i)^T} V^T V \boldsymbol{x}^{(j)}$, of two input feature vectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$. In other words, the resulting distance from the low-rank Mahalanobis metric measures the difference between two data inputs, while the score obtained from the inner product gives the similarity between these two data inputs.

DEFINITION AND OPTIMIZATION OF DISTANCE METRIC LEARNING

Figure 2.1 shows a set of data inputs clustered by a learned distance metric. The clusters are formed according to their qualitative categories, using optimization procedures. There is not a universal approach for the optimization of distance metric learning. However, optimization methods inspired by Expectation Maximization and Gradient Descent are the most widely used techniques. In this dissertation, we mainly use Stochastic Gradient Descent (SGD) in optimizing the adopted metric learning [71, 87]. Chapter 5 concerns with developing a multimodal metric learning for audio-visual recognition. Section 5.4 details the proposed algorithm and defines the training and optimization procedure (using SGD), which can be generalized to other metric learning approaches.

2.3. KERNEL METHODS: SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs) is a discriminant-based method that is written as a sum of the contributions of a subset of the training data samples [80]. In other words, in SVM, it is not necessary to learn classes' densities, rather, it estimates where the class boundaries are. The subset of the training set used to estimate the parameters of the SVM model are called the support vectors. Figure 2.2 shows an illustration of SVM using a hyperplane to separate two classes of data with a margin. SVM is one of the most popular and influential statistical learning methods [88].

The basic implementation of SVM can be illustrated using two classes with $\{+1, -1\}$ labels. SVM uses a linear function $\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b$ to find the boundaries between classes. It imposes a constraint on the output $(y^{(i)})$ to be positive if the class is positive, and negative otherwise. This can be mathematically described through equation (2.4), which covers both the positive, as well as the negative case as follows:

$$y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b) \geq +1, \tag{2.4}$$

where $y^{(i)}$ is the input label. SVM aims to maximize the distance (also referred as the margin ρ) from the closest instance of a class to the optimal separating hyperplane (obtained with the learned weights: $\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b$). Since the distance of $x^{(i)}$ to the closest point in a hyperplane is: $\dfrac{y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b)}{\|\boldsymbol{w}\|}$, thus, for positive and negative samples, the distance should be at least within a margin ρ:

$$\frac{y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b)}{\|\boldsymbol{w}\|} \geq \rho \ , \ \forall i \tag{2.5}$$

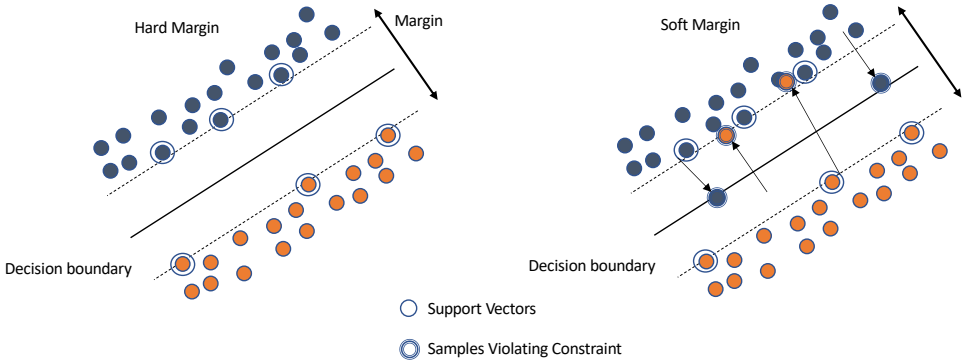


Figure 2.2: An illustration of SVM.

Therefore, ρ should be maximized, and in order to limit the number of the solutions for $\boldsymbol{w}$ to a unique solution, $\rho\|\boldsymbol{w}\|$ is set to 1 [80]. As a result, the learning task can be constrained as follows:

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 \text{ subject to } y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq +1\ , \forall i \tag{2.6}$$

which is a constrained optimization quadratic problem that can be solved by the Lagrangian multiplier method [80]. Therefore, the primal form of the optimization problem can be formulated as follows:

$$\min L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha^{(i)} y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - \sum_{i}^{n} \alpha^{(i)} \text{ wrt } \boldsymbol{w}, b \tag{2.7}$$

where α is a vector of coefficients (Lagrange multipliers) subject to $\alpha^{(i)} \geq 0$ and n is the number of training samples. The derivatives of equation (2.7) with respect to $\boldsymbol{w}$ and b are:

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0 \implies \boldsymbol{w} = \sum_{i=1}^{n} \alpha^{(i)} y^{(i)} \boldsymbol{w}^{(i)} \tag{2.8}$$

$$\frac{\partial L_p}{\partial b} = 0 \implies \sum_{i=1}^{n} \alpha^{(i)} y^{(i)} = 0 \tag{2.9}$$

Notice that $\boldsymbol{w}$ is a linear combination of the training data samples (X) and their outputs $\boldsymbol{y}$, and the Lagrangian values $\boldsymbol{\alpha}$. Furthermore, by substituting these into equation (2.7), the dual formulation is

$$\max L_d = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha^{(i)}\alpha^{(j)} y^{(i)} y^{(j)} \boldsymbol{x^{(i)}}^T \boldsymbol{x^{(j)}} + \sum_{i=1}^{n} \alpha^{(i)} \tag{2.10}$$

where we maximize over $\alpha^{(i)}$, subject to constraints in equations (2.9) and (2.8) and $\alpha^{(i)} \geq 0$. In addition, the dependence on $\boldsymbol{w}$ and b is removed. Notice that this formulation of SVM is written in terms of the dot product between samples.

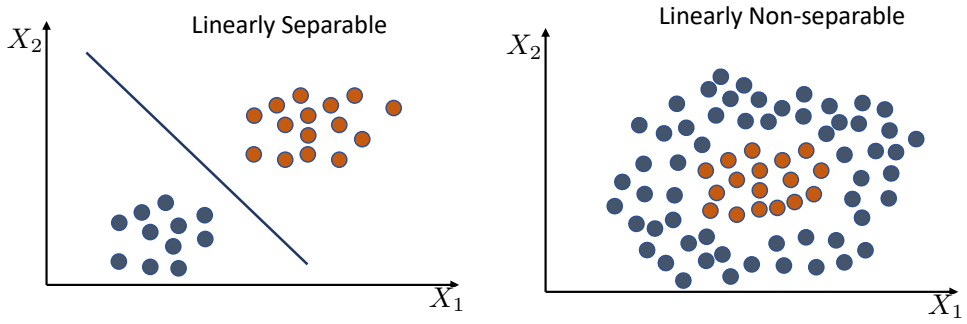


Figure 2.3: Kernel trick in SVM projects data samples onto a subspace where linear model could be applied.

However, in case the problem is nonlinear, nonlinear transformations can be employed to map the problem to a new space instead of fitting nonlinear functions. Hence, the linear model in this new space corresponds to the nonlinear model in the original space [80]. A key innovation in SVM is the kernel trick [79, 80]. The kernel trick is illustrated in Figure 2.3, where it applies a nonlinear transformation (ϕ) using a suitable basis function and then uses a linear model in the new space for the optimization. For example, the dot-product in equation (2.10) can be replaced by a function as follows: $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \phi(\boldsymbol{x}^{(i)})^T \phi(\boldsymbol{x}^{(j)})$, which is called the kernel function. As a result, in this trick, the formulation of SVM in equation (2.10) can be re-written as follows:

$$\max L_d = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + \sum_{i=1}^{n}\alpha^{(i)} \tag{2.11}$$

Besides its computational efficiency, SVM dual formulation and the kernel trick make it possible to learn a model that is nonlinear as a function of $\boldsymbol{x}$ using convex optimization which guarantees an efficient convergence [79]. In other words, it converts the optimization in equation (2.6) to a form in which the complexity depends on the number of samples n rather than the inputs dimensionality d. As a result, the optimization is done mainly on $\boldsymbol{\alpha}$, in which we maximize with respect to $\boldsymbol{\alpha}^i$, subject to the constraints $\sum_i \boldsymbol{\alpha}^{(i)} y^{(i)} = 0$ and $\boldsymbol{\alpha}^{(i)} \geq 0, \forall i$. Finally, there are different types of kernels and the most popular ones are:

- Radial Basis Functions (RBFs): $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = exp\big[-\frac{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|^2}{2s^2}\big]$. In fact, a learned metric (as described in Subsection 2.2.2), such as learned distance metric with a Mahalanobis kernel, can replace the Euclidean distance: $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = exp\big[-\frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T M(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})\big]$

- Polynomial of degree q: $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = (\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(i)} + 1)^q$, where q can be selected by the developer.

- Sigmoidal functions: $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = tanh(2\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} + 1)$

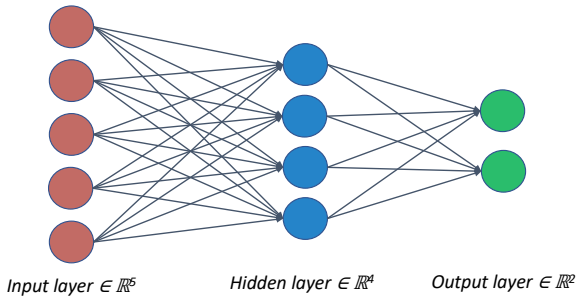


Figure 2.4: An input layer could be pixels of an image, where the Multilayer Perceptron (MLP) tries to learn the mapping functions' parameters to predict the label of this image (e.g. vehicle, cat, or dog).

2.4. DEEP NEURAL NETWORKS

Deep Neural Networks (DNNs) constitute a family of approaches within Machine Learning that are typically more complicated than techniques like kernel methods introduced before. Among the biggest advantages of DNNs is the fact that they can derive feature representations as opposed to hand-crafted features [89]. DNNs learn these representations through multiple processing layers with various levels of abstractions. Based on concepts such as back-propagation and gradient descent, the DNNs' layers change their internal parameters to capture multiple levels of representations. DNNs' architecture is loosely inspired by biological neural networks [79]. The idea is that the brain is an intelligence processor which provides interesting principles that should guide the computational models. However, DNNs are not designed to be realistic models of the biological neural networks [79]. During the last decade, progress in DNNs has been achieved due to the following factors [79, 89, 90]:

- Increased computational power: computation resources were crucial to run very deep DNNs' architectures with multiple layers. One theory in DNNs is that depth and the size of the models contribute to its performance, similarly to animals becoming more intelligent when many neurons work together. Faster Central Processing Units (CPUs) and the advent of general-purpose Graphics Processing Units (GPUs) make it feasible to test these theories and to prove that a larger network can achieve higher accuracy on complex tasks [79, 89].
- The availability of large-scale datasets: sizes of benchmark datasets increased remarkably, which is largely due to extensive data availability on the world wide web, digitization of society, and the vast amount of new sensors to acquire data.

As a result, the availability of efficient and powerful computers allowed researchers to apply machine learning approaches, including DNN methods, to learn sufficiently from the training data [89].

MULTILAYER PERCEPTRONS

The basic building blocks for early implementations of Artificial Neural Networks (ANNs)/DNNs are the feedforward Multilayer Perceptrons (MLPs). Figure 2.4 shows a

MLP, which is a mathematical abstraction to map some input to a certain output. In fact, these computational graphs use functions to map an input $\boldsymbol{x}$ to a category y. For instance, MLPs define the mapping function as $\boldsymbol{y} = f(\boldsymbol{x};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are parameters that MLPs learn in order to achieve the best possible performance conditioned above the imposed hyperparameters (e.g. number of neurons and number of layers), as well as the amount and quality of available training datasets. As depicted in Figure 2.4, a MLP can consist of multiple layers, and the functions could be connected in a chain, $f(\boldsymbol{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\boldsymbol{x})))$, where this chain of functions represents the first layer, the second layer, and the third layer, respectively. The number of these layers represents the depth of a given model. The final layer is usually referred to as the **output (prediction) layer** . The layers between the input and the output layers are referred to as **hidden layers**. In addition, each layer consists of a number of **units** which represents its width. In other words, the design of a DNN refers to the overall structure, which is based on the chain structure of hidden layers and activation units on their outputs. For example, a layer l can be expressed as follows:

$$\boldsymbol{h}^{(l)} = g^{(l)}(\boldsymbol{W}^{(l)^T}\boldsymbol{x} + \boldsymbol{b}^{(l)}) \tag{2.12}$$

where $\boldsymbol{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$ are weight matrices and bias vectors that are learned for each layer (l). $g^{(l)}$ is an activation function and $\boldsymbol{h}^{(l)}$ contains a hidden layer values and their dimensionality can be customized as demonstrated in Figure 2.4. Therefore, the depth of the architecture, as well as the width of its layers (number of hidden units) are important hyperparameters when training DNNs. For instance, studies show that deeper DNNs are often able to generalize to test sets, while shallow architectures are prone to over-fit training sets [79]. The following subsections elaborate on the essential components of a MLP, by explaining the activation functions applied to a hidden unit ($g^{(l)}$), the optimization procedure based on the output unit, and the training algorithms of DNNs' models. These elements are common in existing DNN architectures, which differ in structures (for example operational units and layers' connectivities) according to specific tasks. For example, Convolutional Neural Networks (CNNs) are specialized architectures in computer vision (explained in Subsection 2.4.1), and Recurrent Neural Networks (RNNs) which are widely used in temporal and, sequential, in general, data analysis [89, 90] (see Subsection 2.4.2).

HIDDEN UNITS AND ACTIVATION FUNCTIONS

A perceptron with a single layer can only approximate linear functions. To overcome this limitation, the hidden layers concept was introduced in feed forward networks and nonlinear activation functions are used to compute the values of the hidden layers [79, 80]. Usually, a hidden layer on an input $\boldsymbol{x}$ can be described by computing the linear operation of: $\boldsymbol{z} = \boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b}$, where $\boldsymbol{W} \in \mathbb{R}^{d\times h}$, $\boldsymbol{b} \in \mathbb{R}^h$, and $\boldsymbol{x} \in \mathbb{R}^d$. Subsequently, the non-linear activation functions (g) are applied on these linear operations (as shown in equation (2.12)): $(g(\boldsymbol{z}) = g(\boldsymbol{W}^T\boldsymbol{x} + b))$. For instance, Rectified Linear Units (ReLU) is the most common activation function used in CNNs [91]: $g(z_i) = max(0, z_i)$. Also, step function, with output of either 1 or 0:

$$g(z_i) = \begin{cases} 1 & \text{if } z_i \geq 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}$$

Initially, most neural network models adopted sigmoid activation function, $g(z) = \sigma(z) = \frac{1}{1+e^{-z}}$, or hyperbolic tangent function, $g(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. Sigmoidal functions suffer from saturation as they result in high value if z is very positive, while they give low value if z is very negative [79]. This saturation makes training a neural network, with gradient based learning, difficult. However, hyperbolic tangent units perform better. Besides these units, special operations are applied to the output of DNN (e.g. softmax) as cost (loss) functions, which are explained in the next subsection.

LOSS FUNCTION AND OPTIMIZATION

Like many other machine learning algorithms, the training of DNNs is based on gradient descent. This procedure requires designing a model, a loss function, and specifying the optimization method. However, the nonlinearity of DNNs, caused by the hidden layers and the activation functions, makes loss functions non-convex. Therefore, DNN models are usually optimized through iterative gradient-based approaches.

For a given model of $\boldsymbol{y} = f(\boldsymbol{x};\boldsymbol{\theta})$, its parameters ($\boldsymbol{\theta}$) are learned via an optimization algorithm. These parameters are optimized such that the mapping from $\boldsymbol{x}$ to$\boldsymbol{y}$ leads to robust representations and, hence, performance. The performance of a model can be measured through a loss function (sometimes this is referred to as a score function). It indicates the quality of the learned parameters based on how well the predicted scores agreed with the true labels of the input data. Usually, a Deep Learning (DL) model employs a distribution $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ and utilizes the principle of maximum likelihood for the optimization procedures [79]. For example, as a loss function, the cross-entropy, which is equivalent to the negative log-likelihood, measures the distribution given the model parameters and the input data with respect to the model predictions. The loss function can be defined as:

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\hat{\mathbb{X}}} \log p_{model}(\boldsymbol{y}|\boldsymbol{x}) \tag{2.13}$$

where $\hat{\mathbb{X}}$ is the empirical distribution defined by the training set $\mathbb{X}$. In a classification task, e.g. emotion recognition using audio-visual data, a given model produces c-dimensional probabilities (where c is the number of emotions). Therefore, a softmax function is used on the output to represent the probability distribution over the c different classes, i.e. it produces a vector $\hat{\boldsymbol{y}}$, with $\hat{y}_i = P(y=i|\boldsymbol{x})$. This is a generalization of the sigmoid function which is employed to represent binary variables. Softmax operation guarantees that the prediction vector ($\hat{\boldsymbol{y}}$) sums to 1. More importantly, this operation makes the logarithm operation on the cost function well-behaved for gradient-based optimization [79]. For example, as an output layer, a linear prediction layer produces the non-normalized log probabilities, $\boldsymbol{z}$, as follows:

$$\boldsymbol{z} = \boldsymbol{W}^T \boldsymbol{h} + b \tag{2.14}$$

where $\boldsymbol{z}$ values are defined as follows: $z_i = \log \tilde{P}(y=i|\boldsymbol{x})$. Subsequently, softmax is applied on $\boldsymbol{z}$ to obtain the desired prediction vector ($\hat{\boldsymbol{y}}$), as follows:

$$\text{softmax}(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_j^c e^{z_j}} \tag{2.15}$$

Using an exponential function works well when softmax is employed in the training and optimization processes. Since, as seen in equation (2.13), we aim to maximize the log-likelihood. The log function undoes the exponential, which also helps to prevent the saturation of the gradient. Consequently, this is beneficial for gradient-based learning. Therefore a log is applied to equation (2.15) as following:

$$\log \text{ softmax } (\boldsymbol{z})_i = z_i - \log \sum_j^c e^{z_j} \tag{2.16}$$

CROSS-ENTROPY LOSS

From information theory, the cross-entropy [92] between the actual distribution ($\boldsymbol{y}$) and the predicted distribution ($\hat{\boldsymbol{y}}$) can be defined as:

$$H(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i^c y_i log(\hat{y}_i) \tag{2.17}$$

where y_i means the i^{th} element of the ground truth probabilities. Hence, the network is optimized to minimize the cross-entropy between the predicted class probabilities ($\frac{e^{z_i}}{\sum_j^c e^{z_j}}$) and the actual (true) distribution (e.g. $\boldsymbol{y} = [0, 0, 1, \cdots, 0]$ is a vector that contains a single 1 at the i-th position, for the corresponding class). Therefore, the loss value ($L(\hat{\boldsymbol{y}}, \boldsymbol{y})$), using this score function, guides the training of the model as an optimization cost. This loss value is associated with single training sample $\boldsymbol{x}$, its true label $\boldsymbol{y}$, and the network output prediction $\hat{\boldsymbol{y}}$. Consequently, the cost function as defined in equation (2.13), can be re-written using the cross-entropy for n samples as follows:

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\hat{\mathbb{X}}} \log p_{model}(\boldsymbol{y}|\boldsymbol{x}) = \sum_j^n H(\boldsymbol{y}^{(j)}, \hat{\boldsymbol{y}}^{(j)}) \tag{2.18}$$

Moreover, the cross-entropy could be written in terms of entropy ($H(\boldsymbol{y}) = -\sum_i^c y_i \log(y_i)$) and Kullback-Leibler ($D_{KL}(\boldsymbol{y}||\hat{\boldsymbol{y}}) = \sum_i^c y_i \log(\frac{y_i}{\hat{y}_i})$) divergence [93] as:

$$H(\boldsymbol{y}, \hat{\boldsymbol{y}}) = H(\boldsymbol{y}) + D_{KL}(\boldsymbol{y}||\hat{\boldsymbol{y}}) \tag{2.19}$$

Subsequently, this is equivalent to minimizing the KL-divergence between the two distributions, since the entropy $H(\boldsymbol{y})$ is zero, since $\boldsymbol{y} = [0, 0, 1, \cdots, 0]$.

Softmax classifiers (which give the normalized class (e.g. emotion) probabilities) are intuitive probabilistic methods to design objective functions and to train DNNs. More importantly, information theory concepts have been utilized in this dissertation for multimodal fusion. For example, minimizing the KL-divergence between two modalities predictions with respect to the true distributions leads to a more efficient fusion. These outcomes and decisions are elaborated in Chapters 4 and 7.

TRAINING - BACKPROPAGATION

The information flows through a network when an input data $\boldsymbol{x}$ is feedforwarded to obtain an output $\hat{\boldsymbol{y}}$. The initial output is then back-propagated through the hidden units of each layer in a process called *backpropagation* [79, 94]. *backpropagation* utilizes the cost

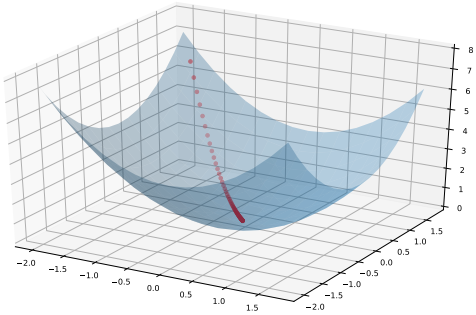


Figure 2.5: An example of how SGD finds the local minimum of a loss function.

value $J(\boldsymbol{\theta})$ in order to compute the gradient of the network. It is important to distinguish between the role of the *backpropagation* and the gradient descent process. The *Back-propagation* algorithm is a method for computing the gradient of the network cost ($J(\boldsymbol{\theta})$) with respect to its parameters. On the other hand, the learning process is performed via the Stochastic Gradient Descent (SGD) using the computed gradient in order to update the network parameters $\boldsymbol{\theta}$ (weights) and minimize the loss function.

Informally, computations in DNNs can be expressed in terms of computational graphs, where each node indicates an operation on a variable [79]. Variables could be a scalar, a vector, or a tensor. The operations, as discussed in the aforementioned sub-sections, could be a cost function on the output layer, activation function on the hidden unit, or a linear transformation on an input layer. Back-propagation is an algorithm that computes the derivatives using chain rules on functions formed by composing other functions. For example, if we have $y = g(x)$ and $z = f(g(x))$, then the chain rule computes the derivatives as follows:

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

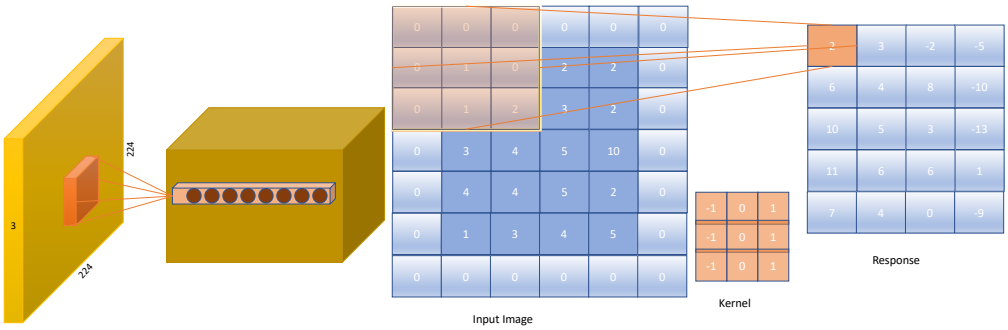
. This can be generalized on vector cases, where $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{y} \in \mathbb{R}^c$, i.e. g maps $\mathbb{R}^d$ to $\mathbb{R}^c$ and f maps $\mathbb{R}^c$ to $\mathbb{R}$. Therefore, the chain rule, if $\boldsymbol{y} = g(\boldsymbol{x})$ and $z = f(\boldsymbol{y})$, is

$$\nabla_{\boldsymbol{x}} z = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^T \nabla_{\boldsymbol{y}} z \tag{2.20}$$

where $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ is the $c \times d$ Jacobian matrix of g. Subsequently, backpropagation performs the Jacobian-gradient product for each computation in the graph [79]. Specifically, we are interested in computing the gradient of the cost function with respect to a network's parameters $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, in a mini-batch of m samples, randomly selected from a training set $\mathbb{X}$

$$\hat{\boldsymbol{g}} = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m} L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)}) \tag{2.21}$$

where L is the loss per sample. Using a learning rate (lr) ϵ, the SGD performs learning



(a) A convolutional layer (b) A convolution operation

Figure 2.6: Convolutional layer illustration.

iteratively (as demonstrated in Figure 2.5) until a stopping criterion is met through updating the network parameters as follows:

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \epsilon \hat{\boldsymbol{g}} \tag{2.22}$$

2.4.1. CONVOLUTIONAL NEURAL NETWORKS

CNNs are specialized architectures of DNNs and are the main network type in computer vision, which are usually applied on data with grid-like topology, e.g. RGB images [89]. The backbone of these networks is a mathematical operation called **convolution**. This operation leverages three principles: *sparse connectivity*, *weight sharing*, and *equivariant representations* [79]. Applying a standard MLP on images can lead to an explosive increase in the number of weights per hidden units. For example, for an image of 32×32 resolution, the first layer has to learn $32 \times 32 = 1024$ weights per hidden unit (neuron). However, this number increases to $196,608$ for an RGB image of $256 \times 256 \times 3$ resolution. In addition, instead of fully-connected units (neurons), units are connected to specific local regions in the previous layers, leading to *sparse connectivity*. This is motivated by the fact that meaningful features like edges can be computed on small numbers of pixels. The main components of CNN models are convolutional layers, pooling layers, and fully connected layers; these are explained in the following subsections.

CONVOLUTIONAL LAYER

The convolution operation between an image and a kernel can be defined as:

$$S(i,j) = (I * K)(i,j) = \sum_n \sum_m I(n,m) K(i-n, j-m) \tag{2.23}$$

where I is an image and K is a given kernel. On the other hand, neural network libraries usually opt for applying the cross-correlation (which similar to convolution but without flipping the kernel), as follows:

$$S(i,j) = (I * K)(i,j) = \sum_n \sum_m I(i+n, j+m) K(n,m) \tag{2.24}$$

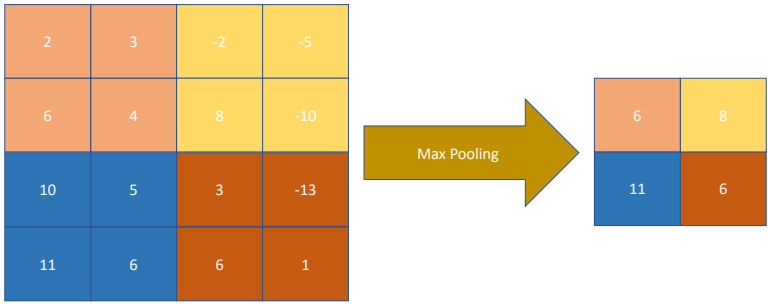


Figure 2.7: Example of max-pooling in CNNs.

The kernel response is computed over image regions, in which the response is higher when applied on similar local regions. In a CNN model, a convolutional layer consists of a learnable set of filters (kernels). Also, a convolutional layer contains a number of channels, where each channel learns one kernel (e.g. with a size of $3 * 3$ as shown in Figure 2.6b). As the size of the kernel is smaller than the size of the image, this satisfies the principle of *sparse connectivity*. Moreover, the kernel is slid over the input image with a stride. If the stride is set to 1, then the kernel is moved over the image by one pixel. In addition, if the kernel size is 3×3, with stride one, the input image then decreases by one pixel at every boundary (as shown in Figure 2.6a). Thus, to prevent this, a padding technique can be used, e.g. zero padding. More importantly, when sliding the kernels, they are activated and generate higher responses when detecting visual features such as edges or blobs with certain colors.

The principle of *weight sharing* in CNNs refers to applying a learned kernel to the whole image. This operation reduces the number of the parameters that a model needs to learn and store. Therefore, a large number of hidden units use the same weights by applying the same convolutional operation. Finally, *weight sharing* leads to a further property which is the equivariance to translation [79]. In particular, equivariance refers to the fact that if the input of a function changes, the output changes in a similar fashion. Mathematically, this can be expressed as $f(g(x)) = g(f(x))$. As a result, convolution is invariant to translation, however, it is not invariant to other transformations such as rotation and scale.

POOLING LAYER

Following the convolutional operation, a non-linear layer such as rectified linear activation is used. After these two processes, a standard practice in the CNN architectures is the usage of a pooling function. At each block, a pooling function is applied on the output of the network at a certain local neighborhood (as shown in Figure 2.7). This function summarizes the content of this neighborhood and reduces its size, which is why it is called a downsampling operation. For instance, the max-pooling operation results in the largest value within a local neighborhood (e.g. 2×2). Other examples of pooling functions include average-pooling, weighted pooling based on the distance from the central pixel, and L^2 norm of a rectangle neighborhood [79].

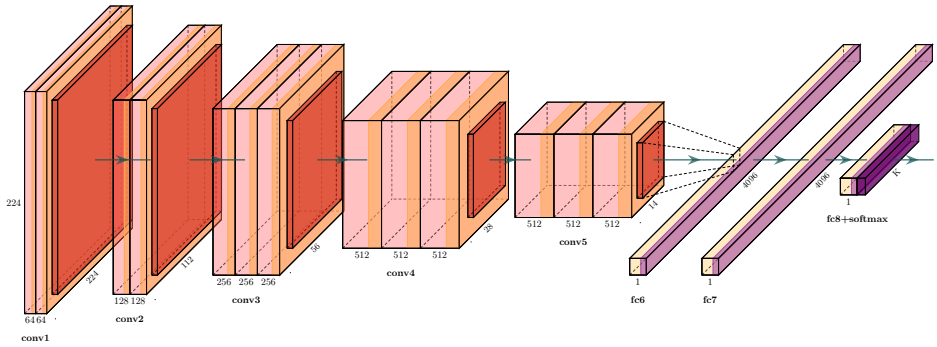


Figure 2.8: A variant of CNN architecture called VGG [95]. This model is called VGG-16 since it has 13 convolutional layers and 3 fully connected layers. The figure was generated using PlotNeuralNetworks Software [2].

FULLY-CONNECTED LAYER (FCL)

The successive blocks of pooling and convolutional layers in CNN architectures (as illustrated in Figure 2.8), constitute the core mechanism through which these topologies extract meaningful features for further classification or regression tasks. A CNN model transforms an input image layer by layer from the pixel-level values to a score-level which represents the prediction of the model for the content of the image. In particular, Fully Connected Layers (FCLs) are usually employed as representations for classification purposes. For example, in Figure 2.8, the two FCLs convert the features' maps (of the last convolution block) onto a feature vector to represent the input image. The last FCL's dimension depends on the number of classes.

Usually, in classification, a softmax layer is applied (as explained in Section 2.4 and in equation (2.15)). The softmax layer's values are between 0 and 1 and their sum is equal to 1. Consequently, the values of the output layer represent the a-posterior probability of the classes in the training data.

VGG MODELS

One of the most common architectures of CNN models is VGG-16 [95], which is illustrated in Figure 2.8. As an input, the figure shows an RGB image with a fixed size of 224×224 pixels. The input image is followed by 5 successive block operations of convolution and pooling. The convolutional layers have small receptive fields, i.e. 3×3. The stride to slide the convolution operation is set to 1 pixel. A max-pooling operation is performed over 2×2 windows with a stride of 2. Following the convolutional layers, two-fully connected layers (each with 4096 dimensions) are applied after the output of the last convolutional layer. Finally, a prediction layer is employed to predict the input class. The dimensions of the prediction layer depend on the number of classes in a dataset. For example, VGG was employed and adapted for the purpose of face recognition in the wild in [96], using a dataset with $2.6K$ people. As a result, the prediction layer of the developed model, which is called VGG-face, has $2.6K$ dimensions. Apart from this, VGG-face

[2] PlotNeuralNetworks Software

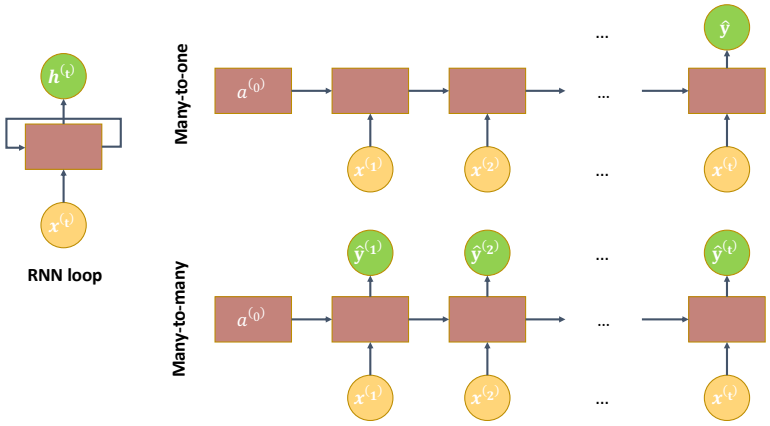


Figure 2.9: Examples of RNNs' loop and structures.

has a similar architecture to the one displayed in Figure 2.8. In this dissertation, VGG-face is used for visual feature extraction from facial expressions' images in Chapters 4 and Chapter 5. In addition, VGG-M [97] is another example of VGG architectures. It is a relatively smaller model, e.g. it has 5 convolutional layers. The convolutional operation is applied with a larger stride, e.g. stride = 2. VGG-M was designed to make computation time efficient and reasonable [97]. In the study of Chapter 7, VGG-M is used for feature extraction from the images of facial expressions, in every of time segments of a video clip.

2.4.2. LONG-SHORT-TERM-MEMORY

RNNs are a special architecture of neural networks, specifically designed to handle sequential data. For example, treating sequential data ($\boldsymbol{x}^{(1)}$, ...,$\boldsymbol{x}^{(t)}$), such as video streams or text, as one large input, does not take into consideration variables like time and data length. Besides, in video processing, it is important to target the whole sequence of the data rather than selecting a small portion of the temporal data. RNNs address these issues and offer good solutions to deal with time-context in the data by adding loop connections in the network graph. Figure 2.9 illustrates how a loop within an RNN cell works. A loop forwards the information from one step to the next one. This idea contributed to a huge success in many areas, e.g. machine translation, language modelling, and speech recognition [98–102] [3].

LSTMs are a specific type of RNNs [103]. They were introduced to handle the problem of vanishing gradient as well as long-term dependencies in long sequences [98, 103]. Hence, they are used widely in modeling sequential data and are adopted in this dissertation as well (in Chapter 6). In addition, LSTMs are preferred types of RNNs due to their ability of propagating salient pieces of information and forgetting less informative ones. Figure 2.10 shows an LSTM cell which contains the following components:

- *Forget gate layer*: it checks the hidden state of the previous layer ($\boldsymbol{h}^{(t-1)} \in \mathbb{R}^h$) and

[3] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

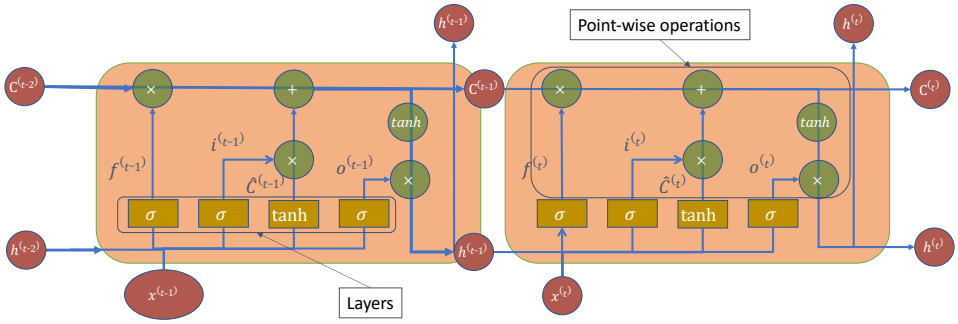


Figure 2.10: Long-Short Term Memory (LSTM) cell. Orange rectangles indicate layers with learned parameters. Green circles refer to the point-wise operations such as addition and multiplication. Arrows indicate the output and the inputs of the layers and the operations.

the current input ($\boldsymbol{x}^{(t)} \in \mathbb{R}^d$) and results in outputs between 0 and 1 for each value in the cell state ($\boldsymbol{C}^{(t-1)} \in \mathbb{R}^k$)

$$\boldsymbol{f}^{(t)} = \sigma(\boldsymbol{W}^f.[\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}] + \boldsymbol{b}^f) \tag{2.25}$$

where [,] indicates a concatenation operation.

- *Input gate layer*: it decides what values to update in the current cell state. It performs the operations necessary to maintain the cell state taken into consideration the previous state

$$\boldsymbol{i}^{(t)} = \sigma(\boldsymbol{W}_i.[\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}] + \boldsymbol{b}^i) \tag{2.26}$$

$$\hat{\boldsymbol{C}^{(t)}} = tanh(\boldsymbol{W}^c.[\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}] + \boldsymbol{b}^C) \tag{2.27}$$

$$\boldsymbol{C}^{(t)} = \boldsymbol{f}^{(t)} * \boldsymbol{C}^{(t-1)} + i^{(t)} * \hat{\boldsymbol{C}^{(t)}} \tag{2.28}$$

- *Output gate layer*: this final layer decides the output using the current state based on sigmoid and tanh functions

$$\boldsymbol{o}^{(t)} = \sigma(\boldsymbol{W}_o.[\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}] + \boldsymbol{b}^o) \tag{2.29}$$

$$\boldsymbol{h}^{(t)} = \boldsymbol{o}^{(t)} * tanh(\boldsymbol{C}^{(t)}) \tag{2.30}$$

It is important to note that $\boldsymbol{W} \in \mathbb{R}^{k\times(d+h)}$ and $\boldsymbol{b} \in \mathbb{R}^k$ are learnable parameters in all the existing gates.

COMMON STRUCTURE OF RNNS

RNNs, in general, can have various structures, depending on the tasks' input and output. Some of these architectures are shown in Figure 2.9. For example, **one-to-many** RNNs are used for tasks such as image-captioning [99]. In this task, the input is an image and the output is a sentence. Another type of RNNs are the **many-to-one** RNNs, which are useful in video-sequence classification [100] and sentiment analysis [101]. For instance,

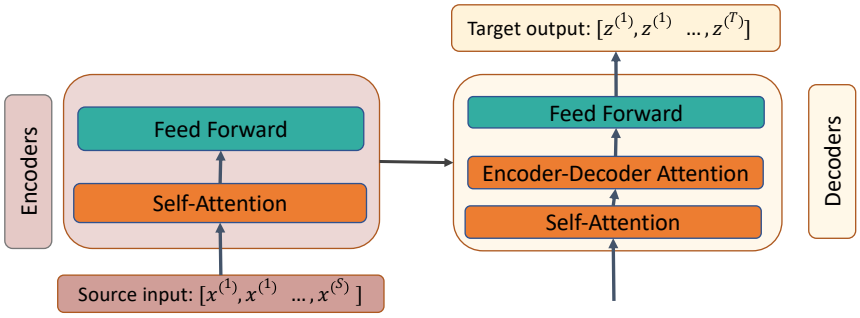


Figure 2.11: A general overview of the Transformer's architecture which consists of stacked encoder and decoder layers.

in sentiment analysis, the goal is to classify, e.g. positive and negative, sentiments of a sentence. Finally, the most popular architectures of RNNs are the **many-to-many** RNNs. They are used for neural machine translation [102], where the task is to translate a text from a language to another.

2.4.3. THE TRANSFORMER

The Transformer is a Sequence to Sequence (seq2seq) neural network architecture that has an encoder-decoder structure. Seq2seq architectures are similar to many-to-many RNNs used for tasks such as machine translation and question answering. For instance, in machine translation, seq2seq models transform input sequence (e.g., a sentence for English) to a target sequence (e.g., a sentence in Arabic). The encoder part of seq2seq models processes the input sequence. Simultaneously, the decoder part processes the encoder's output to transform it onto the target sequence (e.g., using source-target attention).

In 2017, Vaswani et al. [104] proposed the Transformer. The Transformer employs attention to handle sequential data, as well. Attention mechanisms are powerful concepts where the idea is to focus on informative data inputs when processing large amounts of data, such as sequential data. These mechanisms are inspired by cognitive attention. In practice, attention mechanisms allow DNNs to compute weight vectors in a time and input-dependent manner, dynamically [105]. More details about attention mechanisms are explained in the following subsections. As a result, attention mechanisms and the structure of the Transformer eliminate the need for RNNs. It is currently the state-of-the-art model for music generation, protein sequence prediction, machine translation, and many Natural Language Processing (NLP) tasks [104, 105]. It has been shown to have superior performance when compared to other models in terms of performance and efficiency. It has the advantage of being parallelizable and requires much less time compared to models such as LSTMs when working on large scale sequential data. The Transformer's basic building blocks are the stacked encoder and the decoder layers, as shown in Figure 2.11.

Encoders are identical in their structure, in which they consist of stacked self-attention and point-wise feed-forward layers. Similarly, the decoders are identical. At

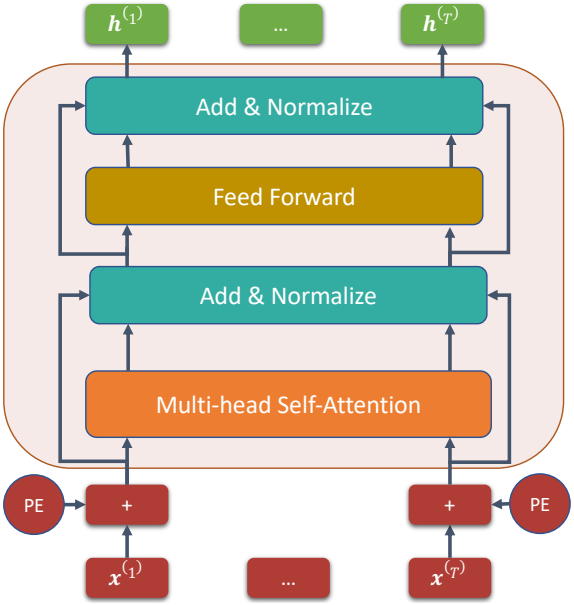


Figure 2.12: The encoder part of the Transformer.

the same time, they have self-attention, encoder-decoder attention, and feed-forward layers. The encoder-decoder architecture is essential for tasks such as machine translation. Nonetheless, the encoders can be separately utilized for many other tasks in Natural Language Processing (NLP) [106], such as sentiment analysis. Similarly, the encoder of the Transformer can be applied to a task such as spatio-temporal audio-visual emotion recognition. For this reason, the overview here focuses on explaining the encoder's components.

ENCODER

Figure 2.12 shows the architecture of an encoder in the Transformer. As shown in the detailed Figure 2.12, an encoder consists of a Multi-Head Self Attention (MHSA) sublayer and is followed by an element-wise fully connected feed-forward sublayer. The number of stacked encoders could vary, and in the original paper [104] it was set to 6. In addition, a residual connection between the two sublayers is employed and followed by layer normalization. As a result, according to the authors of [104]'s terminology, the output of each layer is "$LayerNorm(\boldsymbol{x} + Sublayer(\boldsymbol{x}))$", where $Sublayer(\boldsymbol{x})$ refers to the function of the MHSA or the element-wise feed-forward sublayers, and x is time dependent embeddings (feature vector). Finally, all sublayers in the encoder produce outputs with same dimensions, e.g. $d = 512$, in order to facilitate the residual connections. Nonetheless, prior to feeding the encoder blocks with the sequential data, a positional encoding operation is applied by adding time information to the input embeddings. The following subsections detail the encoder's components (sublayers) and then explain the positional encoding operation. We also detail the underlying concepts of the multi-head

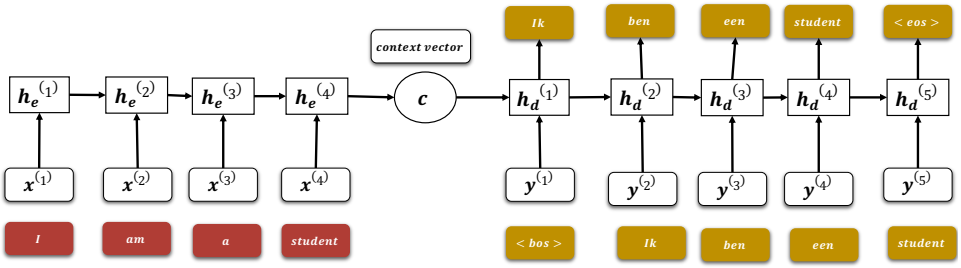


Figure 2.13: An example of a seq2seq model for machine translation without attention.

self-attention sublayer, namely attention and self-attention.

ATTENTION

Attention mechanisms allow DNNs to compute weight vectors in a time and input dependent manner, dynamically [105]. For instance, using attention mechanisms, DNN models adjust weight vectors adaptively in order to pay attention to different parts of the data, e.g. in sequential inputs [105]. In this way, models learn context related to time and proximity of input data.

Initially, attention mechanisms were introduced in many of the RNN architectures [102, 105, 107]. For example, in machine translation, the decoder part of seq2seq models (e.g. many-to-many RNN) uses the context vector resulting from the encoder part (which represents the encoding of the source sequence). We illustrate a seq2seq model in Figure 2.13, based on the initial seq2seq model proposed by Sutskever et al. [108]. As shown in the figure, the context vector ($\boldsymbol{c}$) usually is obtained from the final hidden state ($\boldsymbol{h}$) of the encoder ($\boldsymbol{h}_e$): $\boldsymbol{c} = \boldsymbol{h}_e^{(T)}$. As a result, the decoder has the following form: $\boldsymbol{h}_d^{(t)} = f_d(\boldsymbol{h}_d^{(t-1)}, \boldsymbol{c})$, where $\boldsymbol{h}_d^{(t)}$ is a hidden state at the t position (time step) of the decoder, and f_d can be an LSTM (explained in Subsection 2.4.2). In machine translation, especially for long sentences, this formulation can result in poor performance, since it does give an alignment between two sentences (source and target sentences) from different languages. For instance, word order can be different across languages (e.g. in the Turkish language, the verb is always at the end of the sentence). The alignment can be provided using source-target attention mechanisms. For example, at time step t, the decoder function ($\boldsymbol{h}_d^t = f_d(\boldsymbol{h}_d^{(t-1)}, \boldsymbol{c}^{(t)})$) can operate on context vector, which is computed as a weighted sum of the encoder input vectors as follows:

$$\boldsymbol{c}^{(t)} = \sum_s \alpha^{ts} \boldsymbol{h}_e^{(s)} \tag{2.31}$$

where s indicates an index over the length of the input source (S) of the encoder and $\alpha^{(ts)}$ is a *normalized attention weight*. $\alpha^{(ts)}$ associates the encoder hidden state $\boldsymbol{h}_e^{(s)}$ and the decoder hidden stated $\boldsymbol{h}_d^{(t)}$. $\alpha^{(ts)}$ is a normalized weight through softmax with respect to the other scores computed from the encoder hidden states:

$$\alpha^{(ts)} = \frac{exp(score(\boldsymbol{h}_d^{(t-1)}, \boldsymbol{h}_e^{(s)}))}{\sum_s^S exp(score(\boldsymbol{h}_d^{(t-1)}, \boldsymbol{h}_e^{(s)}))} \tag{2.32}$$

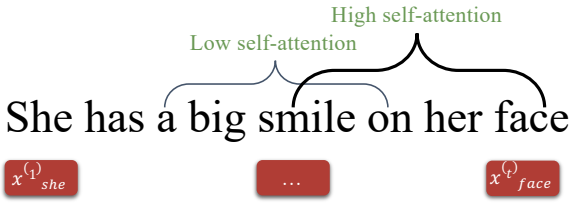


Figure 2.14: Example of self-attention on a sentence, e.g., for sentiment classification.

There have been several ways to define the score function in equation (2.32), which measures similarities between two hidden states (representations). The score can be obtained using learnable parameters through a neural network. For example, Luong et al. [107] proposed a multiplicative scoring mechanism that operates on two hidden states as following:

$$score(\boldsymbol{h}^{(t)}, \boldsymbol{h}^{(s)}) = \boldsymbol{h}^{(t)^T} W^{\alpha} \boldsymbol{h}^{(s)} \tag{2.33}$$

where W^{α} is a trainable parameter. If $\boldsymbol{h}^{(t)}$ and $\boldsymbol{h}^{(s)}$ have the same dimension, i.e. d, W^{α} can be omitted, which results in a dot-product scoring mechanism:

$$score(\boldsymbol{h}^{(t)}, \boldsymbol{h}^{(s)}) = \boldsymbol{h}^{(t)^T} \boldsymbol{h}^{(s)} \tag{2.34}$$

The formulation of attention mechanisms in terms of queries, keys, and values: The concepts of *queries, keys, and values* can be found in application areas such as information retrieval. For instance, when we *query* for an image on Google, the search engine maps this query against a set of *keys* in the database (e.g. captions and tags). Usually, the keys are associated with the query. Eventually, this search results in the best matched *values*. It turned out that the attention mechanisms can be considered as a retrieval process, where we can apply the concepts of *queries, keys, and values*. In particular, the computation in equation (2.34) can be considered as comparing a set of target vectors (i.e. $\boldsymbol{q}^{(t)} \in \mathbb{R}^d$ *queries*), with a set of candidate vectors (i.e. $\boldsymbol{k}^{(s)}) \in \mathbb{R}^d$ *keys*). In a matrix format, this can be written as follows: $A = score(Q, K)$, where $Q \in \mathbb{R}^{T \times d}$, i.e. T is the number of the queries, and $K \in \mathbb{R}^{S \times d}$, i.e. S is the number of the keys (from the source sequence in the encoder). After obtaining these weights (A), we can compute the weighted combination of the *values* $\boldsymbol{v}^{(j)}$ ($\boldsymbol{h}^{(s)}_e$ in equation (2.31)), whose *keys* $\boldsymbol{k}^{(s)}$ are the most compatible (associated) with the t^{th} *query*:

$$\boldsymbol{r}^{(t)} = \sum_s \alpha^{(ts)} \boldsymbol{v}^{(s)} \tag{2.35}$$

The overall matrix formulation of these mappings can be written as follows:

$$R = attention(Q, K, V) = AV \tag{2.36}$$

where R is the resulting matrix of the retrieved (weighted) values.

THE TRANSFORMER: SELF-ATTENTION

In the previous subsection, we demonstrated how source-target attention mechanisms work in an encoder-decoder structure. Instead of the decoder attending to the encoder, the encoder can attend to itself. This is referred to as self-attention. Self-attention is a crucial component of the Transformer architecture. It allows the encoder to capture contextual information within its input sequence.

To illustrate how useful the self-attention mechanism is, consider the following sentence: "*the animal did not cross the street because it was too tired*". In this sentence, "*it*" refers to "*the animal*". However, in the sentence: "*the animal did not cross the street because it was too wide*","*it*" refers to "*the street*". When translating such sequences, it is important to give information for what the pronoun "*it*" refers to. Self-attention offers this possibility by calculating similarities of the input sequence with respect to each other. Specifically, it computes weighted vectors within the input sequence itself. In this way, self-attention improves many tasks such as machine translation. Moreover, let's consider the following: "*she has a big smile on her face*". In tasks such as sentiment classification (where the aim is to classify opinions in text, e.g. as positive or negative, based on emotion expression), self-attention can associate words like "*smile*" and "*face*". Specifically, "*smile*" should be associated with words like "*face*" and "*big*" more than the rest of the words, due to the given context of the sentence and the task of sentiment classification. This process is illustrated in Figure 2.14.

Additionally, the authors of [104] introduced the notion of "Scaled Dot-Product Attention". As shown in Figure 2.15a, the dot product is applied on the queries with the keys, scaled by $\frac{1}{\sqrt{d_k}}$, and then a softmax is applied to obtain the weights on the values:

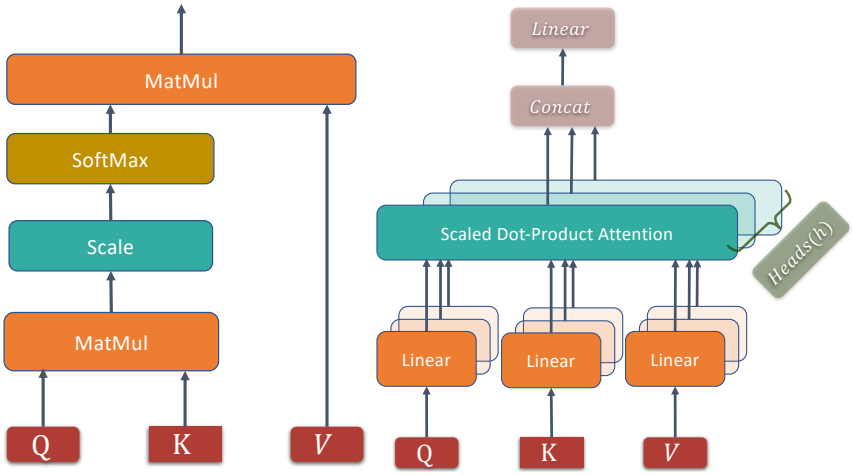
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.37}$$

Note that, in equation (2.37), queries (Q), keys (K), and values (V) matrices are created from the same input in a sequence. This is due to the fact that the encoder part of the Transformer employs a self-attention mechanism, by attending to its input sequence, X.

MULTI-HEAD SELF-ATTENTION

The authors of the Transformer [104] found that applying the self-attention h_{times} on the queries, keys, and values is beneficial. This process is called "Multi-Head Self-Attention (MHSA)". Specifically, MHSA splits the learning loads to learn context information over several heads. In particular, for the queries, keys, and values, we learn linear projections with d_q, d_k d_v dimensions, respectively. For example, if the given input sequence $\boldsymbol{x}^{(t)}$ is a 512 dimensional vector (i.e. $d = 512$) and 8 attention heads are used, then $d_q = d_k = d_v = 64$. Therefore, in a head (i), the linear projection matrices are as follows: W_i^q, W_i^k, and $W_i^v \in \mathbb{R}^{d_k \times d}$. Using these learnable linear transformations helps the self-attention mechanism to get stronger representations and to exploit the context in a given sequence, efficiently.

Subsequently, the resulting outputs from different attention heads, with d_k dimensions each, are concatenated and projected again (with W^o linear projection) to obtain the final weighted vectors. The computations in the MHSA mechanism can be written as follows:



(a) Scaled Dot-Product (b) Multi-Head Self Attention

Figure 2.15: The operations within the MHSA [104].

$$\begin{aligned} MHSA(Q,K,V) &= W^o(concatenate(head_1,...,head_h)) \\ \text{where } head_i &= Attention(W_i^q Q, W_i^k K, W_i^v V) \end{aligned} \tag{2.38}$$

As shown in Figure 2.15b, for each head, the attention is performed on the keys, values, and queries in parallel. Another way to consider the equation in (2.38) is that X is multiplied with each of the three projection matrices to produce the Query, Key, and Value vectors. This notion is equivalent to ours. However, as described in the previous section, X can be used to create the three vectors, where semantically, in the self-attention, Query, Key, and Value vectors are coming from the same source of information. The employed notion is shown in Figure 2.15b.

ELEMENT-WISE FEEDFORWARD NETWORKS

Following the MHSA sublayer, fully connected feed-forward networks (FNNs) are employed. In particular, as shown in Figure 2.12, FNNs are employed on the following output: $\hat{X} = LayerNorm(\text{MHSA}(X) + X)$. The FNN networks are applied to each position (of the time series), separately. The original paper in [104] proposed computing two linear layers, where a ReLU activation is employed between them as follows:

$$FNN(\hat{X}) = max(0, W_1 \hat{X} + W_2 \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{2.39}$$

These linear layers are identical among all positions, however, they vary from layer to layer, in stacked encoders.

[4] https://github.com/jalammar/jalammar.github.io/blob/master/notebookes/transformer/transformer_positional_encoding_graph.ipynb

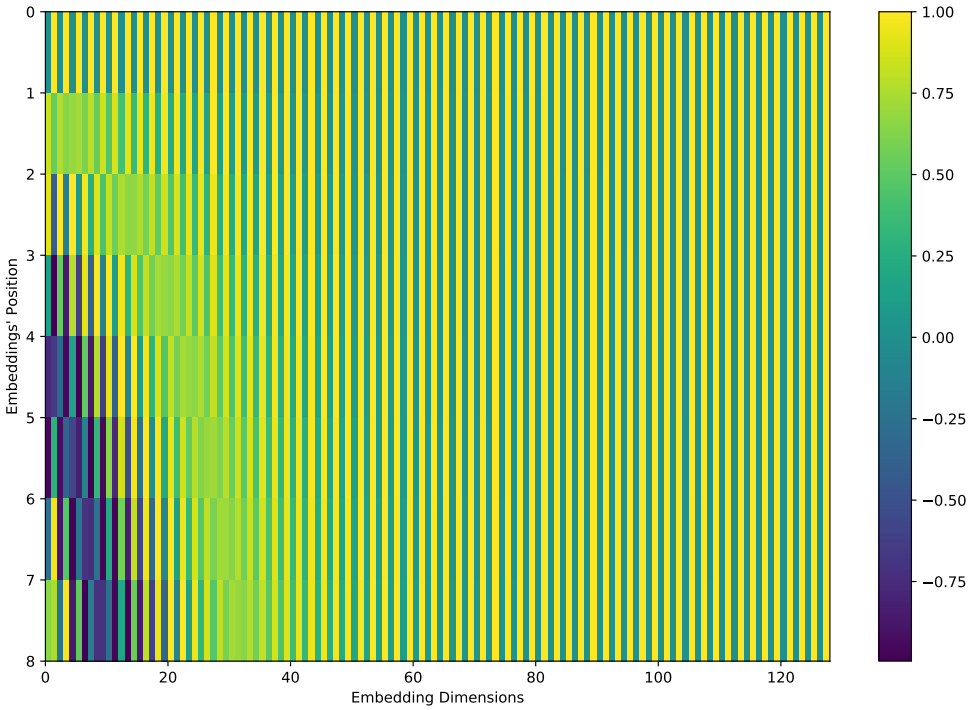


Figure 2.16: An illustration of Positional Encoding for 8 sequential embeddings with 128-dimensions. The figure was generated using transformer_positional_encoding_graph.ipynb [4].

POSITIONAL ENCODING

It is important to note that the Transformer does not have a recurrence operation. It adopts positional encoding to make use of the order in a sequence and its time information, instead of recurrence operations. In other words, Positional Encoding (PE) helps the encoder to embed the position of each time window. In particular, "positional encodings" are added to the sequential input of the encoder, e.g. time windows' audio-visual embeddings in a video clip. The addition (sum) of PEs to the embeddings is applied once, before the flow of the inputs to the encoder. Besides, PEs have the same dimensions (d) as the input embeddings to facilitate their sum. The sine and cosine functions were chosen so that they would help the model to attend by relative positions [104], since they give order information to the embeddings sequence. The authors of the Transformer also experimented with learned positional embeddings but found no significant difference in comparison with fixed ones. As a result, PE employs sine and cosine operations; hence fixed functions, with variant frequencies as follows:

$$\begin{aligned} PE_{(t,2i)} &= sin(pos/10000^{2i/d}) \\ PE_{(t,2i+1)} &= cos(pos/10000^{2i/d}) \end{aligned} \tag{2.40}$$

where t indicates the time position and i refers to a specific dimension, which means that each position corresponds to a sinusoid signal [104]. Note that these Position En-

codings do not have learnable parameters. This operation is illustrated in Figure 2.16.

2.4.4. DEEP METRIC LEARNING

Section 2.2 introduced metric learning approaches, where their objective consists in learning a distance metric based on the similarities and dissimilarities of the data samples. Conventional approaches in distance metric learning seek linear transformations which might not capture the non-linearity where the data lies on [109]. Metric learning has been positively influenced by the boom in DNNs to overcome this challenge. In particular, the learning process has been shifted from learning distance metric onto learning deep feature embeddings which fits the metric learning paradigm where a Euclidean distance between similar samples is small, while it's large otherwise. This paradigm benefits from the hierarchical and nonlinear mappings of DNNs and provides scalable and nonlinear solutions [109].

This process is called Deep Metric Learning (DML). In particular, DNNs' architectures, such as CNNs, have been utilized to learn the mapping functions to obtain feature embeddings. The core idea in DML is based on employing deep architectures ($f(\boldsymbol{x};\boldsymbol{\theta})$) and replacing the categorical loss functions that measure the performance of a model on a single sample by similarity based losses where the optimization of the model is measured by its efficiency to capture the similarities and dissimilarities underlying a given training data set ($\mathbb{X}$). Examples of these loss functions include contrastive loss [82], triplet loss [110], and n-pair loss [111]. In this dissertation, triplet loss has been employed to exploit the complementary information in audio-visual cues for emotion recognition. More details are provided in Chapter 6.

TRIPLET LOSS DEFINITION

Let $\boldsymbol{x} \in \mathbb{X}$ be a data sample and $y \in \{1,...,n\}$ its given label. A contrastive loss [82] employs pairs of examples as input to train a network to separate inputs from different classes and bring samples of the same classes together. The contrastive loss can be written as follows:

$$L_{contrastive} = f(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}; f(.;\boldsymbol{\theta})) = \mathbf{1}\{y^{(i)} = y^{(j)}\} \| f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^{(j)}) \|_2^2 + \mathbf{1}\{y^{(i)} \neq y^{(j)}\} max(0, margin - \| f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^{(j)}) \|_2^2) \quad (2.41)$$

where the $margin$ imposes a distance between data samples from different classes. On the other hand, triplet loss [110] extends the contrastive loss to include an anchor, a positive sample, and a negative sample as follows:

$$L_{triplet} = f(\boldsymbol{x}, \boldsymbol{x}^{(+)}, \boldsymbol{x}^{(-)}; f(.;\boldsymbol{\theta})) = max(0, \| f(\boldsymbol{x}) - f(\boldsymbol{x}^{(+)}) \|_2^2) - \| f(\boldsymbol{x}) - f(\boldsymbol{x}^{(-)}) \|_2^2) + margin) \quad (2.42)$$

In other words, for each data anchor $\boldsymbol{x}$, there exists $\boldsymbol{x}^{(+)}$, which denotes a positive example of the anchor with the same label, and $\boldsymbol{x}^{(-)}$, which refers to a negative example of the anchor with a different label. As demonstrated in Figure 2.17, the mapping function ($f(.;\boldsymbol{\theta})$ of a DNN model, takes $\boldsymbol{x}$ and generates an embedding vector $f(\boldsymbol{x})$. Subsequently, as illustrated in Figure 2.18, the learning process via triplet networks bring the positive

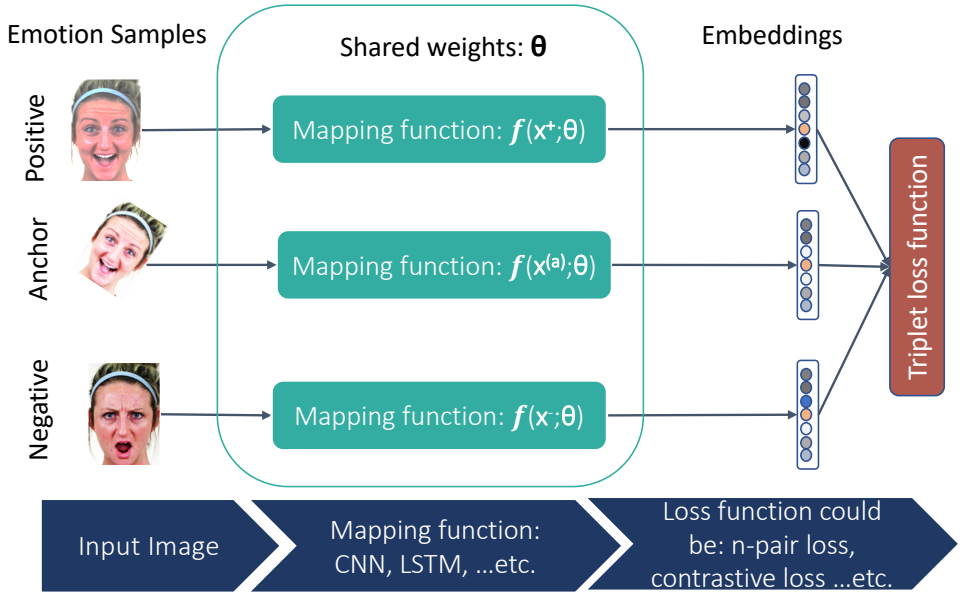


Figure 2.17: An example of DML pipeline using triplet loss to cluster facial expressions based on their similarities-dissimilarities. The lower part of the figure shows the basic steps of DML approaches.

sample closer to the anchor, while pushing away the negative sample. To put it another way, the triplet loss requires the similarity between the anchor and its positive sample to be larger than the similarity between the anchor and its negative sample. These loss functions lead to more discriminative features and project them to a manifold where samples of similar classes are clustered together. Therefore, the contribution of many DML algorithms went to design specialized loss functions that can meet these objectives.

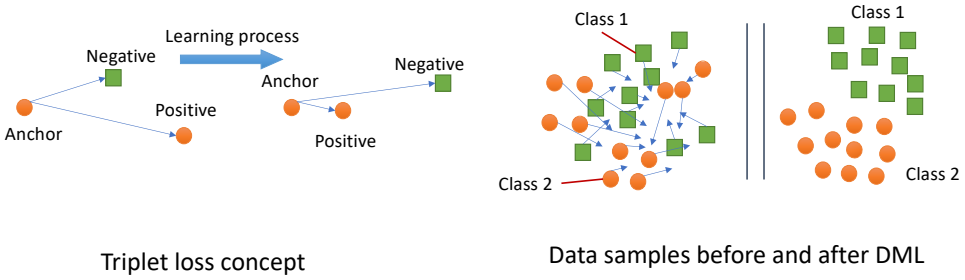


Figure 2.18: Triplet loss concept and its learning goals.

ONLINE HARD NEGATIVE MINING (HNM)

An essential part of training DNNs using metric learning via, e.g. triplet loss, is hard negative mining. However, mining these triplets can be an expensive process since it grows cubically with respect to the number of samples ($O(N^3)$). This is a crucial step in the training process, since a network can fit quickly easy triplets which could harm the discriminative power of the model. As a result, it is common to select hard or moderate negative/positive examples with respect to the anchor. Specifically, it is desirable to mine those triplets that violate the constraint/definition of the triplet loss in equation (2.42). Finally, the HNM step takes a place within batches during the training of DNNs. This online selection of triplets guides the optimization process of the network by providing useful examples that help its improvement. Novel triplet sets mining algorithms are proposed in this dissertation and explained, in detail, in Chapter 6.

2.5. DISCUSSION

This chapter introduced fundamental methods in machine learning. It focused on approaches such as similarity learning, kernel methods, and Deep Neural Networks (DNNs) which are relevant to the research conducted in this dissertation. For instance, this chapter presented important concepts from DNNs such as Convolutional Neural Networks (CNNs), Transforms, Deep Metric Learning (DML), and Long-Short Term Memory (LSTM) that are crucial for the field of Affective Computing (AC). This dissertation proposes novel frameworks utilizing the elaborated methods, to address research questions concerning spatio-temporal audio-visual emotion recognition. For example, Chapter 6 contributes to audio-visual emotion recognition by employing a multimodal triplet loss. While Chapter 7 uses the transforms to re-weight the importance of time windows for temporal emotion recognition. This is an important research area in AC, where the aim is to automatize understanding, inducing, and synthesizing emotions. The next chapter, Chapter 3, complements this overview by demonstrating the usage of machine learning and computer vision for AC.

3

AFFECTIVE COMPUTING: STATE OF THE ART IN EMOTION RECOGNITION

If I have seen further it is by standing on the shoulders of Giants

Isaac Newton[112]

Advances in the field of Artificial Intelligence (AI), sensing technologies, and diverse developments of real-world consumer applications contributed to achieving significant advances in the field of Affective Computing (AC) [2]. These improvements include reading and interpreting body signals, speech, and physiological cues to infer affective states. Moreover, automatic emotion recognition relies on representative data along with accurate and discriminative descriptors for the set of considered emotions. This type of information contributes to obtaining enhanced recognition and classification accuracy. Consequently, various areas, such as education, health-care, and entertainment, can benefit from the advances in AC to improve people's life. This chapter, specifically, focuses on the recent trends of data representation, sensing technologies, datasets, and the applications of AC. On the other hand, Chapters 4 to 7 each provide an overview of the related work on the conducted research in the respective chapter. In other words, this chapter introduces the general state-of-the-art in automatic human emotion recognition.

THis chapter is organized as follows. Section 3.1 presents the benchmarks that are adopted in recent multimodal emotion recognition research works. Section 3.2 reports the work on unimodal emotion recognition using physiological sensors, speech signals, and facial expressions. Then, section 3.3 introduces major directions adopted in multimodal emotion recognition and focuses on Deep Neural Networks (DNNs) methods. Section 3.4 presents examples for applications of AC in different domains. Finally, Section 3.5 concludes the chapter and discusses the context of this dissertation with respect to the literature.

Table 3.1: Public datasets for multimodal emotion recognition using audio (A), video (V), and physiological data (P). Datasets are listed to illustrate the current benchmarks in AC for Audio-Video Emotion Recognition (AVER), as well as other ambitious approaches which include more data channels from bio-signals.

Name	Year	Modalities	Subjects	Samples	Source	Annotation	Labeling
CREMAD [52]	2015	AV	91	7442	Acted (laboratory)	Manual	Discrete
RAVDESS [115]	2018	AV	24	1440	Acted (laboratory)	Manual	Discrete
AFEW6 [117]	2016	AV	–	1749	Movies	Semi-auto.	Discrete
eNTERFACE [116]	2005	AV	42	1166	Acted (laboratory)	Manual	Discrete
SEWA DB [113]	2019	AVP	398	2000 minutes	Spontaneous	Manual	Dimensional (Continuous)
IEMOCAP [119]	2008	AV	10	800	Posed (laboratory)	Manual	Discrete/Dimensional
DEAP [118]	2012	VP	32	40	Induced (laboratory)	Semi-aut.	Dimensional (Continuous)
RECOlA [114]	2013	AVP	23	46	Spontaneous (laboratory)	Manual	Dimensional (Continuous)

3.1. DATASETS

One of the challenges in Affective Computing (AC) is the limited availability of labeled data. This is even more obvious in a multimodal context. The reasons behind this challenge are: (1) data collection and annotation are time consuming procedures, (2) the complexity of generating a multimodal corpus with various sensors, which involves calibration, synchronization of the measurements, and validation, (3) emotions have an ambiguous nature which makes instance labeling even harder, (4) sequential data labeling is difficult as only one label is given to an entire video-clip, even though the indicated emotion is present for a limited amount of time.

Nonetheless, a tremendous effort has been made in terms of producing multimodal datasets. In this section, few datasets, especially the ones adopted in this dissertation, are listed and their properties are discussed. Table 3.1 presents a summary of the most commonly used multimodal datasets in recent studies. It illustrates the differences among them, which lie mainly in the following aspects: environmental setup for data acquisition (e.g. acted or induced sources), data collection and annotation processes, number of subjects and samples involved, and their release date.

Audio (A) and Video (V) modalities are dominant in the existing benchmarks. Some of these datasets are gathered in laboratory environment with spontaneous expressions (e.g. SEWA DB [113] and RECOLA [114]) or posed emotions (e.g. RAVDESS [115] and eNTERFACE [116]). In addition, there has been an attempt to collect audio-visual datasets in the wild from Hollywood movies and TV shows (e.g. AFEW [117]). On the other hand, physiological (P) measurements with various peripheral and EEG signals are included in some datasets (e.g. DEAP [118] and RECOLA [114]).

Another important factor considered in the presented datasets refers to the adopted emotional model. As discussed in Subsection 1.1.1, emotions can roughly be divided into two groups: dimensional with continuous values or categorical with discrete classes. Both of these categories are employed to annotate the data samples. For example, RECOLA, SEWA DB, and DEAP are labeled using the arousal-valence dimensions. However, the Ekmanian discrete emotions' model is the most widely used in the research community and, consequently, in the public datasets.

In this dissertation, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), Acted Facial Expressions in the Wild (AFEW), and eNTERFACE were utilized throughout the chapters. These datasets meet our objectives due to their relatively acceptable size and

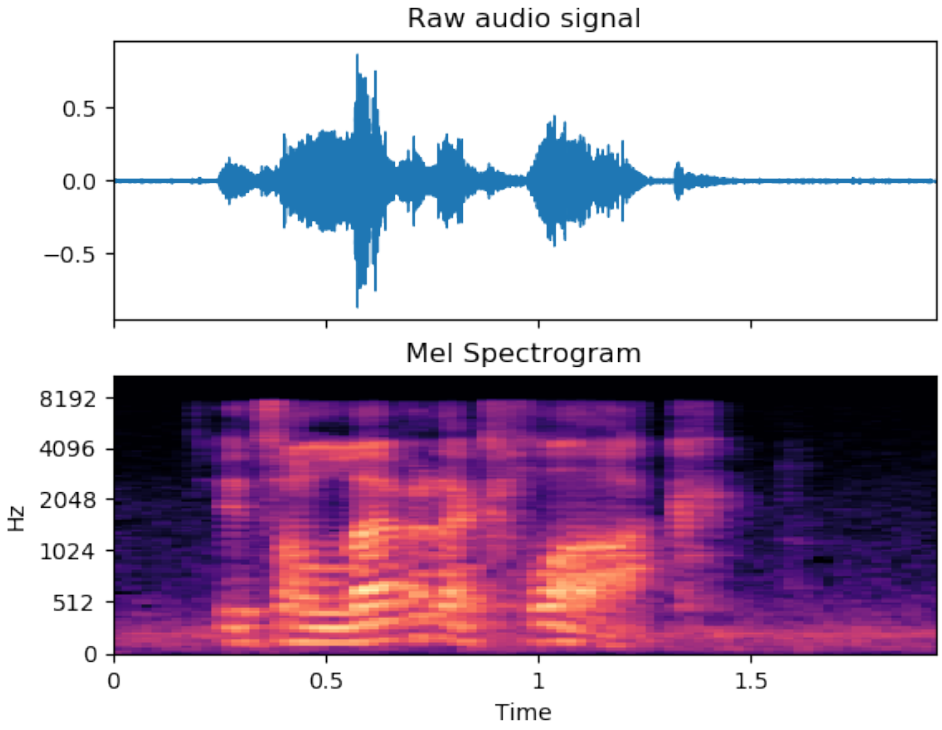


Figure 3.1: Audio raw signal and its spectrogram from the CREMA-D dataset. The facial expressions corresponding to this audio signal are displayed in Figure 3.2.

the affect model used to annotate the data samples. In addition, the recent ones are well studied and they have diverse subjects in terms of cultural backgrounds, ethnicities, and well-balanced biological genders. Ambitious approaches in the literature, such as Deep Neural Networks (DNNs), require a good amount of labeled data. This is a challenging factor in AC. To tackle this problem, advances in regularization methods, transfer learning from other domains such as object recognition, semi-supervised and unsupervised learning are used to initialize the deep models. In this dissertation, transfer learning and regularization methods are adopted and will be explained in the content chapters.

Datasets released recently, namely: CREMA-D and RAVDESS, aim to not only provide sufficient and diverse emotional data but also to enable the training of deep learning methods, which obtained state-of-the-art performance in various application domains. They contain more data samples and a larger number of subjects when compared to older corpora. The following subsections present the adopted datasets in this dissertation, namely: CREMA-D, RAVDESS, AFEW, and eNTERFACE and detail their properties and features.

Figure 3.2: Video clips from the CREMA-D dataset.

CROWD-SOURCED EMOTIONAL MULTIMODAL ACTORS DATASET (CREMA-D)

CREMA-D [52] is an audio-video emotion expression dataset, made public, on GitHub [1]. It contains 7442 clips from 91 actors (43 females and 48 males). Participants' age ranges between 20 and 74, and they come from a variety of races and ethnicities, i.e. Asian, African American, Caucasian, and Hispanic. Actors were asked to speak 12 sentences in five different emotions, namely, anger, disgust, fear, happiness, and sadness, or in neutral. The sentences were spoken with four different levels of intensities: low, medium, high, or unspecified. It is important to note that CREMA-D video clips have an average length of 2.63 ± 0.53 seconds. Furthermore, authors of the CREMA-D dataset asked 2443 participants to rate the emotions and their intensities on three settings: video alone, audio alone, and full audio-video clips. Each participant rated 90 clips (i.e. 30 audio, 30 visual, and 30 audio-visual). 95% of the video-clips have at least 8 ratings. The effort of gathering this dataset was also towards generating standard emotional stimuli for neuroimaging studies. For these studies, it is essential to provide a wide range of expression intensities in visual and auditory modalities, in order to study human-ratings and the activations of brain Regions of Interest (ROIs).

Authors of [52] studied the multimodal expression and perception of the basic acted emotions through raters' responses. An extensive evaluation was then provided on their responses. We report, here, the recognition rates that are based on the relative majority. The relative majority (i.e. a plurality) is measured when an emotion gets the largest share of the votes (ratings) in comparison with the rest of the other emotions. Therefore, this emotion is labeled as the perceived emotion. In this case, the recognition rates for audio-only, video-only, and bimodal audio-video perception are 45.5%, 69.0%, and 74.8%, respectively. Figures 3.1 and 3.2 show an example video-clip from CREMA-D with its audio signal and the corresponding facial expressions, respectively. Throughout this dissertation, we refer to these recognition rates as human benchmarks.

RYERSON AUDIO-VISUAL DATABASE OF EMOTIONAL SPEECH AND SONG (RAVDESS)

RAVDESS [115] is an audio-visual dataset of dynamic expressions for basic emotions. Figure 3.4 shows a face track of a participant's facial expressions, while Figure 3.3 illustrates the corresponding audio raw signal and its spectrogram representation. It contains a large number of songs and speech recordings, each available in audio-only, video-only, and audio-visual formats. Moreover, 247 individuals from North America provided

[1] https://github.com/CheyneyComputerScience/CREMA-D

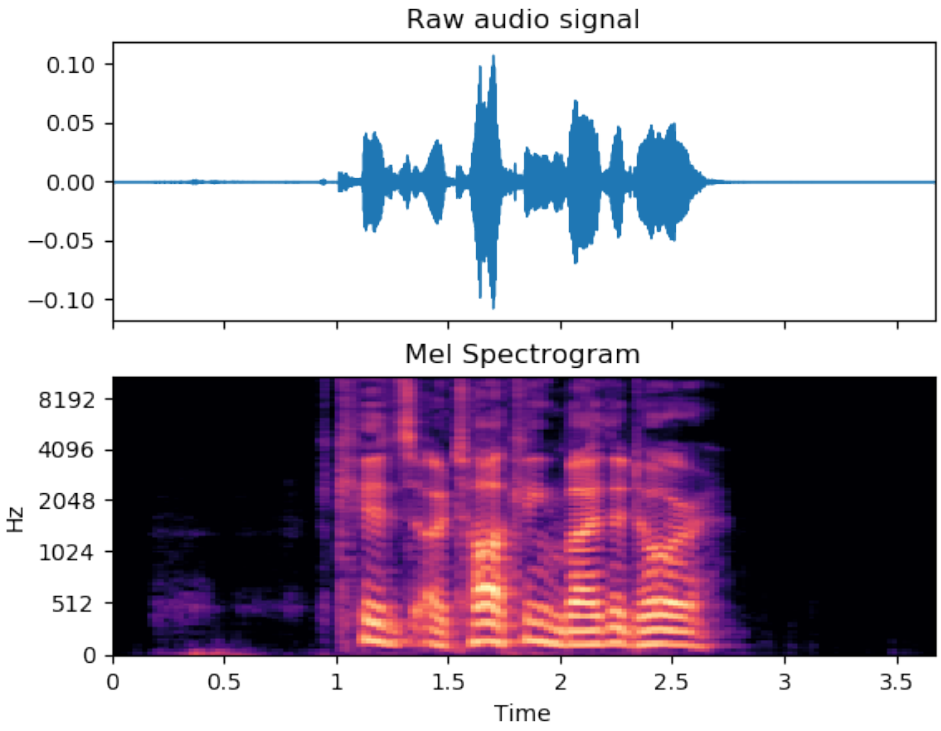


Figure 3.3: RAVDESS: a raw audio signal and its spectrogram representation of a video-clip, where the facial expressions are given in Figure 3.4.

Figure 3.4: A sequence of facial expressions of a video-clip in RAVDESS dataset.

Figure 3.5: An example of still images of affective states and a face track from the AFEW dataset.

ratings. Also, 72 participants provided test-retest data. The dataset's emotions were rigorously validated by participants with high reliability. Each clip was rated 10 times on emotional intensity, validity, and genuineness. High levels of emotional validity, inter-rater reliability, and test-retest intra-rater reliability were reported.

RAVDESS is a gender-balanced dataset of 24 actors who performed vocalizations with emotions that include: anger, calmness, disgust, fear, happiness, sadness, and surprise. Actors also perfromed neutral vocalizations. These emotions were expressed at two levels of emotional intensity, normal and strong. The intensity of emotions is a salient aspect that plays a crucial role in emotion perception. In this dissertation, we chose to use the speech part of the dataset as it is labeled with eight archetypal emotions. This subset contains a total of 2880 recordings. The recordings in RAVDESS have an average duration of 3.82 ± 0.34 seconds. Raters' perception was reported to be: 62.0%, 72.0%, and 80.0%, for audio-only, video-only, and audio-video modalities, respectively. Throughout this dissertation, we refer to these recognition rates as human benchmarks.

ACTED FACES EMOTION IN THE WILD (AFEW)

AFEW [117] is divided into three subsets: training (773 samples), validation (383 samples), and test (593 samples), while only the training and validation sets are publicly available. It has both audio and video modalities. In this dataset, each video clip is labeled with one of the following discrete emotions: anger, disgust, fear, happiness, sadness, and surprise, or labeled as neutral. Developers of AFEW provided baseline results based on automatic emotion recognition. The baseline results of this dataset used handcrafted audio features and Local Binary Patterns on Three orthogonal Planes (LBP-TOP) for visual representations. SVM was applied for classification, achieving 38.8% accuracy for the validation set and 40.47% for the test set. This dataset has in the wild settings, containing wide pose, expression and illumination variation, which reflect real-world challenging conditions. Figure 3.5 illustrates examples of static images and a face track, where the various challenging illumination and pose conditions can be noticed. AFEW is a challenging dataset with occlusions, varying illumination and head poses, which meets real-world conditions.

ENTERFACE

eNTERFACE [116] is an audio-visual dataset which contains six archetypal emotions: anger, happiness, disgust, fear, surprise, and sadness. It includes 42 subjects, who were

asked to simulate the emotions in 5 different reactions, resulting in 1260 video recordings. Among these recordings, 23% were obtained from women and 77% were from men, who have diverse cultural backgrounds. Each subject listened to six short stories. Then, they were asked to react to each situation and two experts judged and validated their reactions. Based on these judgments, clips which have not a clear emotion expression were discarded.

3.2. UNIMODAL EMOTION RECOGNITION

This section introduces literature studies on emotion recognition using individual modalities, namely, visual, audio, and physiological sensors, these are the most informative channels in terms of emotion expressions and predictions [5, 7].

3.2.1. EMOTION RECOGNITION USING PHYSIOLOGICAL SENSORS

Physiological measurements through electroencephalography (EEG), electrocardiography (ECG), electro-dermal activity (EDA), brain signals via EEG, and skin conductance have received significant attention within the machine learning community, for the purpose of emotion recognition [2, 22, 118, 120]. These sensors relate mostly to non-obvious cues, as opposed to facial expressivity or voice prosodics, which are more obvious signals conveying affective content. Figure 3.6 shows a number of sensors to measure biosignals that can be affective indicators.

Healey et al. in [22] collected and analyzed data from the following sensors: electrocardiogram (EKG), electrocardiogram (EMG), and skin conductors. They were used to determine a driver's relative stress level. The proposed system continuously monitored drivers' state and reported every few minutes. The study found out that stress level could be predicted with high accuracy across multiple drivers, a fact which indicates the correlation of physiological features with stress. Stress was also studied at the workplace using EDA signals in [120]. Data were collected from 33 subjects who underwent a laboratory innervation with mild cognitive load and stress factors. Experimental results showed that monitoring EDA could help in distinguishing between cognitive load, which is related to regular office tasks, and psychosocial stress levels with accuracy over 80%. In workplace settings, it is important to discriminate between these two classes for stress monitoring, since office workers can also experience a high cognitive load as part of conducted tasks. Moreover, EDA's peak height and rate carry information about people's stress levels.

Authors in [118] presented a multimodal dataset in which EEG and peripheral signals were collected from 32 participants (DEAP). Participants rated their experiences in terms of the level of arousal, valence, dislike and like, dominance, and familiarity. The authors reported negative correlations between the arousal and the low spectral powers bands of EEG signals. However, valence and liking showed strong correlations with EEG signals in all analyzed frequencies. In addition, a set of features were extracted from the EEG and the physiological signals. Consequentially, classification results were significantly higher than random. This indicates the validity of neurophysiological signals for emotional states.

Moreover, Deep Neural Networks (DNNs) have been utilized to learn spatio-

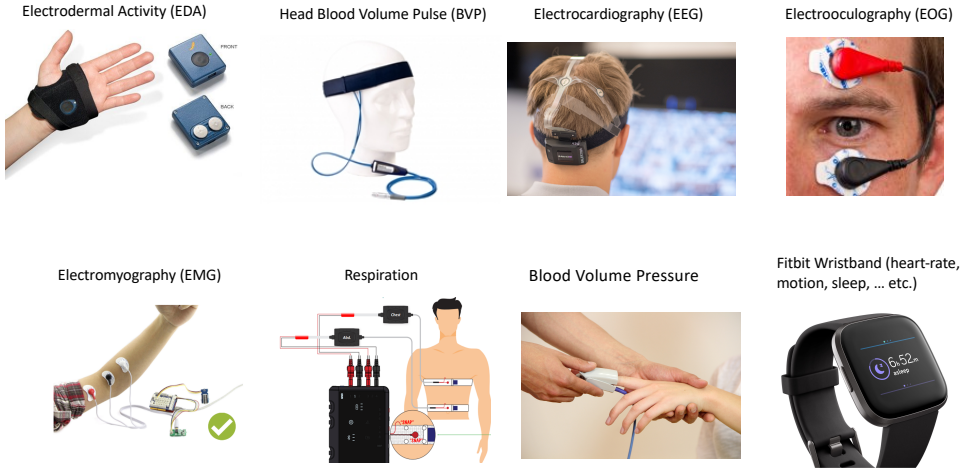


Figure 3.6: Widely used physiological sensors to measure body reactions for emotion recognition.

temporal features from physiological responses, instead of hand-crafted features [5]. $1-D$ Convolutional Neural Networks (CNNs) have been used to learn representations from EEG data. For example, Yanagimoto et al. [121] used 16-channels EEG data to distinguish between positive and negative emotions. A CNN model, with 7 layers and kernels of 10 ms, was implemented on EEG-segments of 1 seconds. The authors reported an improvement of 20% over shallow models. Alternatively, spectrograms representations can be obtained from EEG signals. Li et al. [122] applied spectrograms on segments of 1 second which leads to a grid representation of the signals being convenient for CNN models.

Despite their accurate performance and their characteristics of revealing inner emotions, physiological sensors are considered as intrusive technology [123]. Using body sensors is cumbersome and may interfere with a person's daily activities and emotions. This makes body sensors unsuitable for a significant amount of practical and regular scenarios.

The next two subsections introduce alternative approaches to emotion recognition using audio-visual cues. Audio-visual cues leverage outward expression of emotions and can constitute informative channels for recognizing and understanding people's emotions. However, these approaches do not guarantee to measure inner feelings, as emotion expression can be misleading. For example, a smiling face does not necessarily indicate happy feelings, since this kind of facial expression could have different interpretations based on a certain context or even cultural implications.

3.2.2. SPEECH EMOTION RECOGNITION

Human speech contains rich prosodic, acoustic, and other voice-related features. It is important to note that the usage of Speech-based Emotion Recognition (SER) is beyond the analysis of the spoken word. Speech Recognition (SR) is concerned with the identification of "who said what", while SER can answer a question like "how it is said" for

affect recognition and subsequent natural Human-Computer Interaction (HCI) [124]. This subsection focuses on the studies related to voice features rather than sentiment analysis of the speech itself.

As introduced in Subsection 1.2.2, the audio channel can reveal information such as gender, age, and affect. Psychological perceptual studies of human vocal expressions and processing formed the theoretical bases and scientific ground for SER in the field of Affective Computing (AC) [125]. In these studies, commonly used acoustic low-level descriptors (LLDs) are extracted from the raw audio waveform. There are three main categories of these features [125]. The first category includes prosody related features such as Fundamental Frequency (F_0), and speech energy and rate. These features are related to characteristics of speech, such as rhythm and intonation. The second category measures spectral characteristics of speech, which are related to the harmonic structure of the voice. The most prominent methods of this category are the Mel-frequency Cepstral coefficient and Mel-filter bank energy coefficients. The third category computes quality-related acoustic features of voice, such as shimmer and jitter. These features measure characteristics of voice related to vocal vibrations.

There are open-source toolboxes to extract these sets of features. For example, a well-known set of features can be extracted using the Open Speech & Music Interpretation by Large-space Extraction (OpenSMILE) tool [126]. Usually, these features are followed by data processing approaches to capture the dynamic nature of the voice [125]. For example, statistical functionals are applied which include the following: arithmetic means, standard deviation, skewness, kurtosis, quartiles, quartile ranges, percentile 1%, 99%, percentile range, position max./min, up-level time 75/90, linear regression coefficient, and linear regression error (quadratic/absolute) [127]. For instance, authors in [127] extracted the following set of features: energy and spectral related LLDs, voicing related LLDs, delta coefficients of the voicing related LLDs, and voiced/unvoiced durational features. The functionals, as mentioned earlier, were then applied on the features over a period of time to capture voice-related dynamic representations. Subsequently, these sets of hand-crafted features can be used for SER, in which, Support Vector Machines (SVMs) or HMMs can be applied for emotion recognition. More importantly, research has demonstrated that short-term spectral, energy, and prosodic features have affective information [5].

In a broader aspect, traditionally, SER approaches are influenced by the trend of machine learning approaches for SR. For example, prior to DNNs' era, Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were applied extensively on hand-crafted features [128]. HMMs had been used for modeling the temporal variability of speech, while GMMs had been used to detect how well each state of the HMM model fits a short time window and automatically models multivariate data distributions. On those approaches, acoustic input was represented by Mel-Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Predictive coefficients (PLPs). Those advances made it possible to develop commercially successful SR systems.

In addition, speech features which are related to pitch, energy, or MFCCs have been used. For example, Lin et al. [129] extracted the following set of features: fundamental frequency (F0), energy, the first four formant frequencies (F1 to F4), MFCC1, MFCC2, and five Mel frequency sub-band energies (MBE1 to MBE5). These features proved to

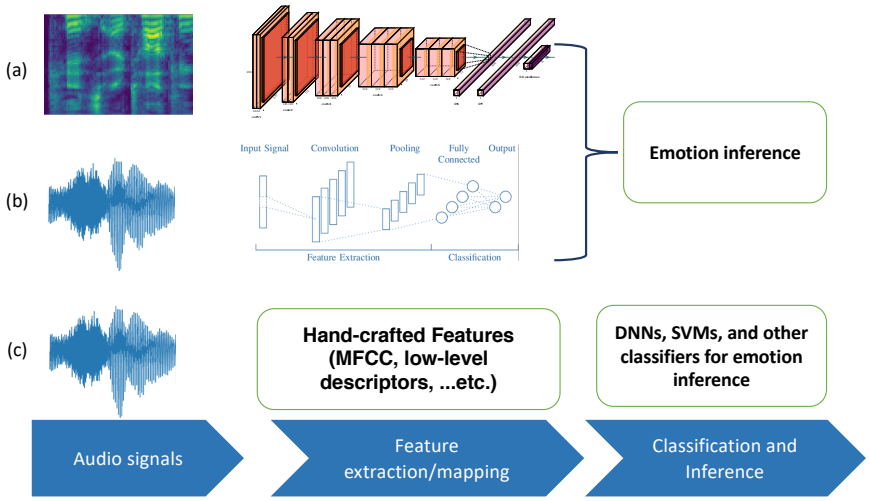


Figure 3.7: Approaches of SER which include: hand-crafted features and end-to-end representation learning on the raw audio signals or spectrograms. (a) starts with obtaining Spectrogram which a visual representation of the spectrum of frequencies of an audio-signal over-time. Then, (a) and (b) apply $2D$ and $1D$ CNNs, respectively, to obtain feature mapping for emotion classification. The bottom-row, (c), shows the traditional pipeline of SER, which starts with extracting hand-crafted features, and then applying machine learning methods, such as HMMs, GMMs for temporal representations and classifiers, like, SVM, DNNs for emotion inference.

be useful for SER. They were used as input for SVMs and HMM classifiers according to whether the data is spatial or spatio-temporal, respectively. Both classifiers achieved high accuracy in detecting five emotional states, namely anger, happiness, sadness, surprise, and neutral.

Furthermore, deep learning has contributed significantly to both SR and SER. Specifically, CNNs are useful in modeling spatio-temporal speech features (based on spectrogram representations) [130, 131], while Recurrent Neural Networks (RNNs) are powerful models to capture the temporal variability and dependencies between acoustic features [130, 132, 133]. To summarize, DNNs have been applied for SER in the following ways:

- CNNs are used to extract and model spatial audio features on either audio-raw signals or spectrograms.
- RNNs are applied to the hand-crafted features or spatial features produced by the CNNs.

For instance, authors in [130] conducted one of the earliest studies to apply one-dimensional CNN models to the audio raw signals. They used two convolutional layers, to replace the need for the hand-engineered features, and they combined the CNN output with Long-Short Term Memory (LSTM) networks to automatically learn best representations from the raw data. Their findings showed that the proposed end-to-end feature learning and regression pipeline was able to double the correlation between the ground-truth and the model prediction for arousal and improve the valence prediction when compared to hand-crafted low-level descriptors.

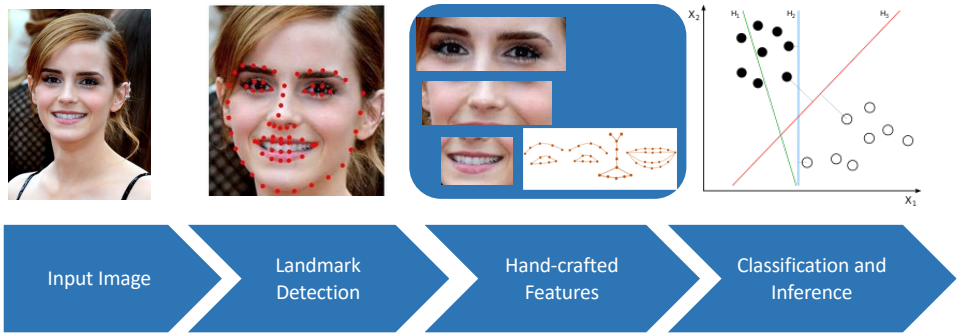


Figure 3.8: Traditional pipeline for Facial Expression Recognition (FER) using crafted features.

Likewise, *spectrograms* can be used as an input to CNN models instead of the raw audio signals. Spectrograms are a visual representation of the spectrum of frequencies of an audio-signal over-time. They can be generated using the Fourier Transform. An example of a spectrogram is illustrated in the top-left (a) of Figure 3.7, which was computed using the magnitude of the Short-Time Fourier Transform (STFT). As spectrograms are computed on multiple frames (times), they can also be interpreted as 2D images. Spectrograms can span a time window that ranges between 250 milliseconds to 1 second. They are usually attributed as heat maps, where the intensity of the frequency is depicted by varying color and brightness.

For example, VGGish[2], which is a variant of VGG models [95], is used on spectrograms of 96×64 resolution. The spectrograms were generated from audio samples with a 16 kHz mono rate. STFT has a window size of $25ms$, a window-hop of $10ms$, and with periodic Hann windows. A Mel spectrogram was computed by mapping the spectrograms to 64 Mel bins with a range between $125 - 7500Hz$. The spectrograms were computed for non-overlapping time windows of duration 0.96 seconds, where they cover 64 Mel bands and 96 frames of 10 milliseconds each. VGGish was trained using Youtube-8M [134], which is a large-scale multi-label audio-video classification dataset.

In this dissertation, we employed VGGish features in Chapter 7, while SoundNet [135] was applied on raw audio signals for feature extraction in Chapter 6. In addition, hand-crafted features were used in Chapters 4 and 5. The reason behind this decision is represented by the limited availability of data in AC. Using a pre-trained model's features can be adopted as an input for a sequential model, such as the Transformers or RNNs, without pre-training. For example, studies in [136, 137] used the VGGish features in a multimodal context for the multi-cultural dimensional emotion recognition task. The adopted features outperformed baseline hand-crafted audio features, on both arousal and valence.

[2] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

3.2.3. FACIAL EXPRESSION RECOGNITION

The face remains the most researched topic in both psychology and AC, due to its properties, including expressivity, and its importance in our daily communications. In this subsection, an overview of FER systems is presented. In particular, it initially introduces studies based on the conventional pipeline consisting of face detection, hand-crafted feature extraction, and classification algorithms. Then, it presents the studies that benefited from DNN models to build spatial and temporal features to encode facial expressions.

Prior to the resurgence of DNNs approaches, hand-crafted features dominated the computer vision field. As a result, conventional approaches to represent facial features by their shape or appearance were proposed [5, 14, 138]. Figure 3.8 illustrates the pipeline of these approaches, which starts by performing face detection and is followed by extracting facial features, based on common detectors and descriptors like Local Binary Patterns (LBP) [61], Gabor wavelets [60], and Scale-Invariant Feature Transformation (SIFT)[62]. In addition, these features were extended to capture the spatio-temporal space, such as Bag of visual Words (BoW) on SIFT features in [139] and LBP- Three Orthogonal Planes (TOP) [140]. On top of these features, classification, or regression methods such as SVMs, Support Vector Regressions (SVRs), and Random-Forest classifiers were employed for emotion inference. Hand-crafted features are robust against illumination, scale, and orientation. However, their generalization and discriminative abilities to capture facial physiognomy are not enough, especially in real-world situations, characterized by a great degree of variability and challenges.

Besides, approaches that rely on shape features have been proposed. These methods use facial landmarks to encode explicitly face geometry [141]. Geometric features are useful since different facial expressions correspond to different facial landmarks' shape deformations (e.g. eyes, mouth, eyebrows, chin, and nose). For example, Active Appearance Models (AMMs) have been used to register deformable visual objects such as faces [63]. The AAMs are useful to capture compact representations of shape and texture. Learned AAMs models are fitted to test facial images by changing the parameters of shape and texture using bounds obtained from a learning set. Besides, Active Shape Models (ASMs) have been utilized for facial expression recognition [142]. ASMs are also shape models that iteratively deform to fit shape features in new images based on distributions learned from training images. They consist of points controlled by the shape model. Moreover, the coordinates of the facial landmarks can be used as features in the classification process. However, this representation results in poor performance since it does not capture the dynamic variations between various individuals [141]. For example, authors in [143, 144] computed geometric features, which can be represented by segments, perimeters, or areas of the figures formed by the coordinates of the facial landmarks. Their features include Euclidean distances, angles, and curvatures between fitted facial landmarks.

Recently, CNN models have been dominating the computer vision area, due to their ability in learning a hierarchy of features that builds from low-level to high-level representations. For example, first layers learn low-level representations like edges, while subsequent layers learn more specific features that can be semantically interpretable [89, 95]. Moreover, through smart design decisions, like parameter sharing or sparsity

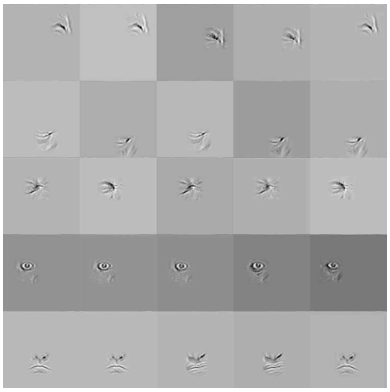


Figure 3.9: Activated facial regions (patches) in five selected filters of the third convolutional layer of a CNN model [145]. Each row corresponds to one filter in the third convolutional layer. The spatial patterns are displayed from the top 5 images. The figure is taken from the study by Khorrami et al. in [145].

modeling, CNNs are able to capture the underlying data distribution. Such representations are currently dominant in the area of face analysis, due to their discriminative and generative power. There are two main categories for processing facial images in FER:

- Learning spatial features from single images
- Encoding sequences of images to learn spatio-temporal features

In the first category, CNNs are the main architecture in DNNs for spatial feature description, the reason why they dominate the first category of facial expression representation from still images. CNNs are powerful methods to extract spatial features from grid-like data, such as facial images [89, 89, 90, 95, 96, 146–149]. Khorrami et al. [145] showed that CNNs could learn features that correspond to specific Facial Action Units (FAUs) without being explicitly trained to do so. Figure 3.9 from [145] shows filters learned in a CNN model that captured automatically the spatial patterns of the facial expressions. The CNN model was able to capture small local changes in facial expressions as well as the global facial behavior. This behavior was obtained by stacking two convolutional layers with small 3×3 filters, which cover the same receptive field, as a bigger 5×5 filter, while requiring fewer parameters. In the literature, it has been shown that this technique is superior to AU's based feature extraction methods [68, 145].

In this dissertation, facial features were extracted using a CNN model called VGG [95] (explained in Subsection 2.4.1). CNN models' success resides also in their ability to leverage a pre-trained model on millions of images and fine-tune it on the problem at hand. Given the high achieved performance, we took the same approach, by fine-tuning the VGG model for emotion recognition. Figure 3.10 displays heat-maps of facial expressions' images. We obtained these heat-maps from a CNN model, i.e. VGG-face. Figure 3.10 shows that different facial regions are more important (displayed in red) than others for each emotional type. At the same time, the entire face carries information, employed in the classification step. From a semantic point of view, our findings are in

Figure 3.10: Heat maps of CNN features on facial expressions.

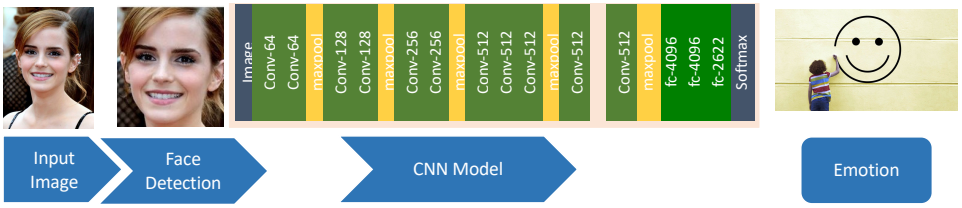


Figure 3.11: End-to-end learning approach in FER systems based on learning representation via DNNs models such as CNN.

line with the work presented by [150], where it is mentioned that sadness and fear rely more on the eyes, while disgust and happiness are depending more on the mouth region.

The second category is inspired by the first one and builds spatio-temporal features from a sequence of images. In fact, facial expressions can span multiple images, the reason why extracting descriptors from spatio-temporal volumes is more effective in FER [5]. Standard CNN models can be extended to represent $3D$ volumes by building $3D$ kernels and extending the pooling and regularization layers of these models to accept $3D$ data. In this way, $3D$CNNs [151] are capable of capturing the motions of facial action units for FER. However, these models have drawbacks in terms of their capacities. In the literature, $3D$CNNs are limited to short segments of videos up to 1 second. Moreover, due to their large number of parameters, they require more data for training [5]. In addition, sequential models such as LSTM can be used to capture the temporal features of a sequence of images. in [152], a hybrid neural network is presented, that combines LSTM with CNN to encode facial motion throughout video frames.

3.3. MULTIMODAL EMOTION RECOGNITION

In general, multimodal learning is an essential step to make substantial progress in understanding human emotions, since it interprets diverse signals [153]. Multimodal Emotion Recognition (MER) is in particular interesting, since it is able to obtain higher reliability and fidelity to model human emotion expressivity due to its multimodal nature. For example, multimodal systems yield better emotion detection over unimodal ones, since they are more suited to model emotion experience [3]. Besides, MER provides the automatic systems with an opportunity to replace missing data in some modalities in order to continuously predict emotions.

For instance, in 2015, a meta-analysis by D'mello et al. [3] on Multimodal Emo-

tion Recognition (MER) revealed that over 85% of then recently published studies indicated an improvement over unimodal emotion recognition. The average improvement is 9.83% with a median of 6.60%, where the improvement was more significant on acted datasets. These improvements show the importance of MER, which outperformed the best Uni-modal Emotion Recognition (UER) counterparts.

In emotion recognition, modalities indicate different sources of information, such as visual, auditory, and physiological sensors' signals. Research in MER faces many challenges as modalities' data is heterogeneous. Multimodal data could be dense or sparse and recording does not always take place in good synchronization. In addition, each modality exhibits unique characteristics, having different data distributions. Figure 3.12 illustrates a multimodal framework that uses a set of sensors observing an environment to obtain multimodal data. MER systems gained notable attention in the 2010*s* with the advent of Deep Neural Networks (DNNs)' architectures. Consequently, this overview focuses primarily on recent studies that fuse various signals based on DNNs. DNNs have largely been utilized to learn data representations in affective computing, mainly according to the contexts below:

- Modeling spatial data: DNNs have been shown to be superior in obtaining automatic representations in comparison to classic computer vision's hand-crafted features. One of the advantages lies on extracting descriptors and building end-to-end models from images [66, 68], video sequences or audio-segments [78].

- DNNs can be also adapted to learn the temporal dynamics of sequential data for affect recognition. Architectures such as Long-Short Term Memorys (LSTMs) and Transformers (see sections 2.4.2 and 2.4.3) address the drawbacks of spatial models and are able to capture the temporal dependencies of sequential data even in asynchronous schemes [154].

- DNNs provide modular and scalable frameworks to handle multimodal data. There is not a linear relationship between raw data of different modalities. Therefore, DNNs have shown to be useful in many tasks of multimodal learning, such as joint-learning of multimodal representations [17], alignment of modalities, such as spoken words to moving lips [155] and translation of one modality to another [156].

Based on the above and, also, the emotional models analyzed in Subsection 1.1.1, multimodal fusion can support integrating multiple modalities for predicting emotions (e.g. happiness and sadness) or measuring continuous values (e.g. arousal and valence). Regarding computational models in MER, there are two main strategies. a.) fusing on feature level (early fusion) and b.) fusing on the decision level (late fusion or decision level fusion). It is important to note that there is a large consensus in the literature of Affective Computing (AC) that late fusion is more robust for automatic emotion recognition than early fusion [5].

EARLY FUSION (EF)

Early fusion can be applied by concatenating features from different modalities or by jointly learning features from different sources of raw data. One benefit of joint feature

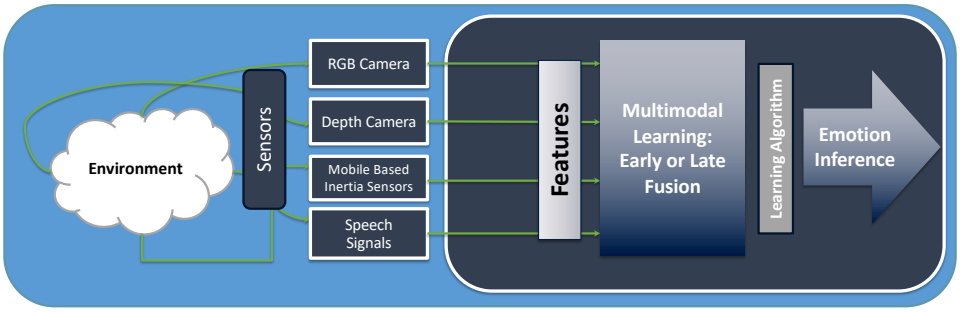


Figure 3.12: A general illustration of a multimodal framework which uses data from various sensors. Learning schemes applied on the features extracted from these sensorial data can be based on early or late fusions.

learning consists in taking into account the temporal context of emotion display. Kim et al. in [157] applied multimodal deep learning to generate joint representations for audio and video modalities. Since the relation between audio and visual information is non-linear, they aimed for cross-modality representations and feature selection. However, their proposed framework resulted in slightly higher classification accuracy than baseline methods, such as Principle Component Analysis (PCA). Alternatively, Wu et al. [158] used attention-based mechanisms to capture temporal information and the Memory Fusion Network with attention across audio, video, and text modalities on the feature level for emotional autobiographical narratives. They adapted the attention mechanism to predict emotional valence over time. Their models led to good results, and in some cases reached a performance comparable with human raters.

More recently, within the Audio/Visual Emotion Challenge (AVEC 2019) [4], Zhao et al. [136] combined features from audio, video, and text modalities in order to investigate knowledge transfer of emotions in a cross-cultural manner. Their multimodal interaction approach gave the best Concordance Cross-Correlation (CCC) results in the challenge for predicting arousal, valence, and likability values. In the same challenge, Haifeng et al. [137] employed an early and late-level fusion based on bi-directional LSTM (BDLSTM). They utilized BDLSTM for audio, video, and geometric features. In addition, they formed new representations, based on the concatenation of audio-geometric, and video-geometric features. On top of these five streams of networks, a late-level fusion based on their prediction is applied. Their approach achieved the second-best results in AVEC2019 challenge [4].

LATE FUSION (LF)

When considering the various methods towards multimodal fusion, Decision-Level Fusion (DLF) or late fusion can be categorized into the following two scenarios: (1) agnostic approaches that use unimodal methods independently and (2) model-based approaches that utilize specific machine learning models in the fusion process [153].

In the agnostic approaches, fusion is applied on the obtained predictions from different modalities. It is based on techniques such as voting schemes, manual weighting, or averaging. For example, in [68], separate methods for each modality were developed,

(e.g. Convolutional Neural Network (CNN) for facial images and Restricted Boltzmann Machines (RBM) for audio information), followed by the late fusion step, consisting in combining the score obtained by each modality using grid search. Similarly in [66], Liu et al. benefited from kernel methods for video feature representation, and the fusion of the different modalities was achieved in a probabilistic manner at a late stage. In [159], different classifiers were trained for each modality, and then combined using late fusion based on a genetic algorithm.

The second set of categories includes model-based approaches where the optimization of the multimodal frameworks and their predictions is conducted simultaneously. These approaches optimize a shared framework, using the same algorithms for the various modalities based on their joint performance. This paradigm uses learned models to combine different modalities during the training and evaluation processes. For example, in [160], Keren et al. applied a series of end-to-end CNNs and Recurrent Neural Network (RNN) models on the raw data of physiological sensors, such as electrocardiogram and electrodermal activity, to predict affect. These end-to-end learning methods were applied separately on sensorial data of ECG, EDA, and skin conductance level (SCL), skin conductance response (SCR), and heart-rate (HR). A multimodal fusion on their predictions was learned to re-weight their relative importance for the arousal and valence prediction. The fusion results outperformed significantly the unimodal performance, by at least .05 in terms of concordance correlation coefficient (CCC). In addition, their study showed that the HR signal dominated the prediction of the arousal and valence. Besides, SCR had more contribution to the valence prediction, while EDA contributed largely in predicting the arousal.

Parts of this dissertation employ model-based approaches. For example, Chapter 6 explains the method which obtains temporal audio and visual features and employs Deep Metric Learning (DML) to cluster emotions, based on a similarity metric. This approach minimizes the distance between the two modalities, as well as benefits from their scores in terms of separating positive and negative pairs of emotions. In addition, Chapter 7 introduces an attention mechanism for emotion recognition, that is employed within a shared-learning framework. The fusion of the two modalities is conducted on the score-level, while the obtained predictions are brought together in order to maximize the performance of audio-visual emotion recognition.

3.4. APPLICATIONS OF AFFECTIVE COMPUTING

Emotions impact various cognitive processes, such as perception, intuition, and decision making [23]. As a result, the applications of affective computing can be tremendous. Inferring human emotions using various data channels offers numerous opportunities to improve people's lives while interacting with automatic systems. The spectrum of these applications can range from education [13, 18, 19], automatic vehicle driving [22–26], health-care [4, 20, 21], to entertainment [27–30].

3.4.1. APPLICATION OF AFFECTIVE COMPUTING IN AUTOMATIC VEHICLE DRIVING

One application where emotion recognition is of vital importance is that of driver emotion recognition, as it can significantly improve the driving experience. Recently, there have been great advances in emotion recognition, focusing on monitoring driving or influencing drivers' performance [23, 24]. Emotions can be detected through several sensorial inputs, such as RGB cameras for facial expression recognition and gaze estimation, and depth cameras for body posture and gesture recognition [24]. In addition, the automotive industry developed new generations of cars, equipped with advanced features to meet the expectation of offering more personalized human-machine interaction [3]. In the literature, emotion recognition can be used to ensure safe driving experience, help the augmented operation of autonomous cars, and to give a personalized driving experience.

For example, in [24], authors utilized facial features, head pose, and gaze estimation to enhance driver's safety by tracking drivers' activity and measuring the driver's focus of attention. Moreover, authors in [161], employed deep learning techniques, such as Convolutional Neural Networks (CNNs), to build a framework for driver's head localization and pose estimation. These inputs can subsequently be used for monitoring driver's emotional states, such as frustration, distraction, and fatigue. A study in [25] suggests that augmenting driving with emotional intelligence could further help the growing industry in real-world scenarios.

Wearable devices are an alternative way of measuring the affective states of drivers. These devices provide sensory information that can be informative about drivers' emotions and can potentially also measure the stress and fatigue which are associated with affect responses. Authors in [22] showed that the stress level of drivers could be inferred with high accuracy from physiological data, such as skin conductors and heart-rate. In addition, the driver's attention and emotional state could be predicted by monitoring the drivers' face and recognizing his/her facial expressions. Authors in [162] performed emotion recognition using facial electromyograms, electrocardiogram, respiration, and electrodermal activity. The identified emotional classes are high stress, low stress, disappointment, and euphoria. Authors in [163] proposed an EEG-Based system for recognizing affective states and mental workload of participants with autism spectrum disorder (ASD) during driving skill training (which occurred in a virtual environment). In [26], physiological signals such as electrocardiogram, galvanic skin response, and respiration were extracted from fourteen drives executed in an instructed route in real driving environments. The research led to the construction of a novel system for driving stress detection based on multimodal feature analysis and kernel-based classifiers. Finally, in [26], the focus of the study was to deduce the emotional state of the drivers based on information derived from electromyography signals of the upper trapezius muscle, photoplethysmography signals of the earlobe, as well as inertial motion sensing of the head movement.

[3] https://blog.affectiva.com/driving-your-emotions-how-emotion-ai-powers-a-safer-and-more-personalized-car

3.4.2. APPLICATION OF AFFECTIVE COMPUTING IN EDUCATION

Of particular interest in Artificial Intelligence (AI) in general, and Affective Computing (AC), in particular, is the construction and the use of e-learning systems able to react and interact with students in a natural manner, similar to that between students and human tutors [13]. According to Picard et al., Technology Enhanced Learning (TEL) systems should incorporate the emotional aspect of the learning process, in addition to the cognitive process [164]. In this manner, students' emotional needs are considered, beyond aspects that address solely productivity and efficiency. However, enhancing learning systems with affective capabilities can be more challenging than adding perception and cognitive functionalities. In an educational context, emotions experienced by a learner directly affect the learning outcome [165, 166]. As a result, a learning system should be able to increase student's motivation and, hence, enhance the students' cognition process [13]. For instance, this could be done through a intelligent system that can detect a student's frustration, boredom, and engagement. In other words, accurate automatic emotion recognition can be useful in enhancing the learning outcomes by providing personalized and adaptive educational processes according to students' emotions, as well as other performance indicators related to productivity and cognitive skills.

Imagine a one-to-one e-tutoring system, in which the developed learning materials, teaching, and evaluation processes are generated based on the needs, emotions, mental abilities, skills, and preferences of the students. This is not possible without having an affective component augmentation within the system [18]. For example, the tutoring system needs to follow the learning curve of the student and must adapt the difficulty level of the learning process according to the student's mood and emotions. In this manner, not only the performance, motivation but also the dedication to complete the study can be increased, due to considering adaptation and personalization aspects. For example, within the course of this dissertation, we contributed to the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS)[4], which is an educational platform that provides a personalized learning experience based on multimodal emotion recognition from diverse cues. This project exploits a wide range of sensors to capture learners' affective states and adapts the learning materials in realtime. For instance, cues from facial expressions, gaze, audio, body posture, and interactions with learning materials are fused to enhance emotion recognition. Subsequently, this paradigm aims to foster students' experience by increasing their engagement and preventing boredom and anxiety [167].

3.4.3. APPLICATION OF AFFECTIVE COMPUTING IN ENTERTAINMENT

In entertainment, user's engagement, frustration, or boredom can be predicted and integrated into games' scenarios, making the gaming industry adaptable and even more entertaining. Shaker et al. [28] showed that players' experience can be predicted with high accuracy. For example, features extracted using behavioral data from gameplay and the player's visual characteristics can model players' experience in terms of frustration, challenge, and engagement. This could help in designing game content, with a final goal to enhance the entertainment sector with systems that can improve gameplay. Indeed, research conducted in [30] found that adapting the game's difficulty, music, characters,

[4] http://mathisis-project.eu/

or mission-based on users' emotional state improved the gaming experience.

3.4.4. APPLICATION OF AFFECTIVE COMPUTING IN HEALTH-CARE

In health-care and medicine, AC is finding practical applications in various areas. For instance, pain detection is used to automatize the health progress monitoring in clinical settings [21]. For example, pain detection can be used in monitoring patients in Intensive Care Units (ICUs) and assessing lower-back pain [21]. For these purposes, studies showed that facial cues are informative for pain detection [21]. Furthermore, depression could be diagnosed through facial activities, head movements, and behavioral signals [4]. Depression is a leading cause of illness and disability. For example, studies showed that a computer agent can detect the level of depression (using the Patient Health Questionnaire PHQ-8 questionnaire). The agent achieved a good Root Mean Square Error (RMSE) in detecting mild depression on the measurement scale. Finally, AC could also help to improve the life of elderly, especially the ones who suffer from dementia. For example, it could help caregivers in detecting signs of apathy as shown in [20]. Furthermore, the study in [20] is part of ICT4Life [5] H2020 European project. ICT4Life is an e-health care platform that aims to provide an integrated monitoring system for people with dementia. Data gathered from patients' trajectories, interactions with a variety of services and cognitive games, the evolution of their symptoms and cognitive abilities, and their health profiles are fused to detect abnormalities in their activities, and moods.

3.5. DISCUSSION

This chapter presented an overview of the multimodal benchmarks for Multimodal Emotion Recognition (MER) systems. Representation learning and fusion techniques have seen tremendous improvements over the last decade. The chapter discussed state-of-the-art methods for extracting representations from raw bio-signals, audio, and video cues for emotion recognition. In addition, it gave an overview of the usability and the benefits of these various sensorial data for emotion recognition and presented a summary of methods for fusion and learning on multimodal data. *Moreover, since late fusion is more robust for automatic emotion recognition than early fusion [5], in this dissertation, the research is based on a joint decision-level fusion.* In particular, each modality produces its predictions for emotion recognition, however, their decisions and learning process are optimized jointly. In addition, over the course of this research, the literature was lacking in-depth analyses regarding automatically extracted, dynamic interactions between audio and video signals in emotionally rich contexts. This dissertation presents studies that investigate the temporal relationships of both modalities and exploit their strength for emotion recognition. Besides, it employs state-of-the-art methods, such as Deep Metric Learning (DML) to perform similarity learning for multimodal emotion recognition.

[5] http://www.ict4life.eu/

4

EMOTION RECOGNITION: THEORY, MULTIMODALITY, AND APPLICATIONS

Recognizing emotions should be interpreted as measuring observations of motor system behavior that correspond with high probability to an underlying emotion or combination of emotions.

R. W. Picard [9]

The pipeline of automatic emotion recognition starts with the selection of the emotional model, and continues with processing certain data structures with varying spatial, temporal, and statistical properties, as well as developing methods to handle the given data for emotion inference. Besides, the aforementioned steps can vary according to the application of emotion recognition. Having this perspective, this chapter consists of two

Parts of this chapter have been published in:

- **E. Ghaleb**, M. Popa, E. Hortal, and S. Asteriadis, "Multimodal fusion based on information gain for emotion recognition in the wild," in 2017 Intelligent Systems Conference (IntelliSys). IEEE, 2017, pp. 814–823.
- J. Schwan, **E. Ghaleb**, E. Hortal, and S. Asteriadis, "High-performance and lightweight real-time deep face emotion recognition," in 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). IEEE, 2017, pp. 76–79.
- **E. Ghaleb**, M. Popa, E. Hortal, S. Asteriadis, and G. Weiss, "Towards affect recognition through interactions with learning materials," in 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2018, pp. 372–379.
- N. Vretos, P. Daras, S. Asteriadis, E. Hortal, **E. Ghaleb**, E. Spyrou, H. C. Leligou, P. Karkazis, P. Trakadas, and K. Assimakopoulos, "Exploiting sensing devices availability in ar/vr deployments to foster engagement," Virtual Reality, vol. 23, no. 4, pp. 399–410, 2019.

studies. The first one proposes a hierarchical framework for multimodal emotion recognition using audio-visual cues in video clips. Part of this study investigates feature engineering approaches, as well as aggregates temporal representations in a video clip using Fisher Vectors (FVs) for emotion recognition. The presented approach for multi-modal emotion recognition is applied on two datasets, namely AFEW'16 and eNTERFACE. Both datasets are composed of video clips labeled with discrete emotions. After a preprocessing stage, we employ different feature extraction techniques, e.g. Convolutional Neural Network (CNN), Dense-Scale-Invariant Feature Transformation (SIFT) on the face and facial Regions of Interest (ROIs), geometric, and audio cues based features. Subsequently, we encode frame-based features using Fisher vector representations. Next, we leverage the properties of each modality using different fusion schemes. Apart from the early-level and the decision level fusion approaches, we propose a hierarchical method based on information gain principles and we optimize its parameters using genetic algorithms. The experimental results prove the suitability of our method, as we obtain 48.9% validation accuracy on the challenging AFEW'16 dataset, surpassing by 10*% the baseline of 38.81%, and good performance of 78.5% on eNTERFACE.*

The second part of this study focuses on understanding affective states through interactions with learning materials. Affective state can be directly linked to a student's performance during learning. Consequently, being able to retrieve the affect of a student can lead to more personalized education, targeting higher degrees of engagement and, thus, optimizing the learning experience and its outcomes. In this study, we apply Machine Learning (ML) and present a novel approach for affect recognition in Technology-Enhanced Learning (TEL) by understanding learners' experience through tracking their interactions with a serious game as a learning platform. We utilize a variety of interaction parameters to examine their potential to be used as an indicator of the learner's affective state. Driven by the Theory of Flow model, we investigate the correspondence between the prediction of users' self-reported affective states and features related to their interactions with the learning material. Cross-subject evaluation using Support Vector Machines (SVMs) on a dataset of 32 participants interacting with the platform demonstrated that the proposed framework could achieve a significant precision in affect recognition. The subject-based evaluation highlighted the benefits of an adaptive personalized learning experience, contributing to achieving optimized levels of engagement.

This chapter is organized as follows: Section 4.1 presents the first study titled "Multimodal Fusion Based on Information Gain for Emotion Recognition". Its subsections elaborate on the related work, the method, and the conducted evaluations. Section 4.2 introduces the second study titled "Towards Affect Recognition through Interactions with Learning Materials". Its subsections detail the collected data, as well as the applied evaluations.

4.1. MULTIMODAL FUSION BASED ON INFORMATION GAIN FOR EMOTION RECOGNITION

FROM a psychological and neurological perspective, research has gained preliminary evidence about how the brain tends to bind information received from different modalities[77]. The binding interaction of different modalities has been demonstrated in the so-called McGurk effect [57], which shows that the perception of audio signals can be altered by the display of incongruent visual information. This multimodal integration is essential for multimodal perception in many cases [58], since it enables accurate perception in a noisy environment or in a state of confusion. For example, people are able to detect a smile from speech signals [58]. In addition, studies show that speech and facial expressions can substitute and complement each other in many tasks such as emotion recognition or speech identifications [77]. For instance, a surprised facial expression might be classified as anger, however, access to auditory signals can resolve this confusion [168]. An extensive overview of multimodal learning's advances and trends in Affective Computing (AC) is introduced in Chapter 3.

As discussed in Chapter 3, computer vision and machine learning algorithms have been employed for recognizing emotions using multiple modalities and various datasets. For example, some of the datasets were gathered in controlled environments such as the Cohn-Kanade [169], the JAFFE [60], the CMU Pose Illumination and Expression (PIE) [170], eNTERFACE [116], or the MMI database [171]. Additionally, efforts were devoted to more challenging datasets, captured in uncontrolled spontaneous conditions such as the Acted Facial Expressions in the Wild (AFEW) dataset [172], containing video clips of unconstrained facial expressions, with varied head poses, occlusions, and challenging illumination conditions. The palette of feature extraction techniques and classification methods employed for Speech-based Emotion Recognition (SER) and Facial Expression Recognition (FER) have been broadly explained in Subsections 3.2.2 and 3.2.3, respectively. For both modalities, these two sections elaborate on approaches based on traditional approaches (e.g. hand-crafted features with Support Vector Machines (SVMs)), as well as recent methods based on Deep Neural Networks (DNNs).

Furthermore, studies in bimodal emotion recognition showed the benefits of fusing visual and acoustic information [173], due to the complementarity of the two modalities. Therefore, in this research, we propose a multi-modal framework for emotion recognition from video sequences, by taking advantage of both visual and audio features. Moreover, one of the main contributions of this study consists in proposing a hierarchical fusion approach, which combines feature level and decision level fusion in an efficient manner, using information gain principles, depicted in Figure 4.1. The proposed fusion framework is general enough to be useful for other tasks such as behavior or object recognition, as long as there are available different types of complementary modalities (with their different feature representations).

In our approach, we take advantage of different feature extraction algorithms. The features are extracted from the audio cues and from the entire face or salient facial Regions of Interest (ROIs) (e.g. eyes, nose, mouth, forehead, and chin). The extracted features (f) include: audio acoustic Low-Level Descriptors (LLDs), Dense Scale-Invariant Feature Transformation (DSIFT), geometric features, and Convolutional Neural Net-

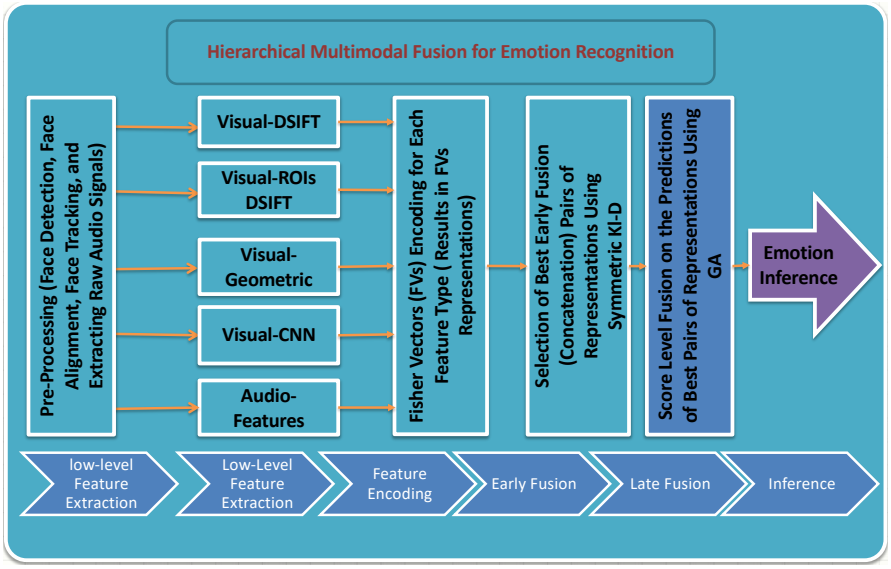


Figure 4.1: Hierarchical multimodal fusion framework based on feature level and score level fusion. The proposed scheme starts with a pre-processing layer for face and facial landmark detection, face alignment, and extracting raw-audio signals. This layer is followed by extracting a set of f low-level features (e.g. DSIFT, CNN, and geometric features). The third and fourth layers include high-level representations of features by Fisher Vector encoding (FV) and selecting pairs of encoded features (representations) based on information gain principles (IG). In the fourth layer are included examples of the selected pairs feature representations. The last layer of the framework depicts score level fusion optimized using a Genetic Algorithm (GA).

work (CNN) based visual features (using VGG-face model [96]). Thus, the $f = 5$ feature channels are coming from audio and video modalities and are displayed on the 2^{nd} layer in Figure 4.1. Another contribution of this study is the representation of the different features using Fisher Vectors (FVs) encoding [69]. FVs are useful at projecting all types of features to the same space and also at facilitating the analysis of videos with different lengths, while efficiently capturing the facial dynamics (3^{rd} layer in Figure 4.1).

Moreover, each of the employed features is useful, while one constraint in their early fusion relates to their different underlying distributions and characteristics. Fisher vectors encoding enables a higher-level representation, in which the representations of the low-level descriptors share similar statistical properties. Also, we use an efficient algorithm for feature-level fusion, which finds the best types of feature representations to be fused in a hierarchical manner. The fusion in the fourth layer (4^{th} layer in Figure 4.1) is based on minimizing the Kullback–Leibler (KL) divergence [174] between the probability distribution function (PDF) of true labels and the PDF of the correctly predicted labels, obtained after employing a classification algorithm (using SVMs) on concatenated pairs of feature representations. For example, at the first stage, the feature representations of both the mouth region and audio cues are concatenated to form a new feature vector. Similarly, DSIFT's and geometric's feature representations can be concatenated in another pair. Then, at the next stage (the 5^{th} layer in Figure 4.1), predictions that are resulted from a linear classification of SVM are fused through a decision-level fusion al-

gorithm which optimizes the weights of each feature representation pair using a Genetic Algorithm (GA).

The proposed framework is useful, as, instead of fusing all feature representations at an early stage as described by [175] or at the end of the pipeline as proposed by [176], it searches for the best combinations at different processing stages for finding complementary modalities. Furthermore, the use of a genetic algorithm facilitates finding the optimum weights for the decision level fusion. We evaluated our proposed approach on the challenging AEFW dataset [117] as well as the eNTERFACE dataset [116].

The remainder of this study is organized as follows: in Subsection 4.1.1, related work is presented showing the popular trends in multimodal emotion recognition. Subsection 4.1.2 explains the pre-processing stage and the feature extraction methods from both video and audio modalities. Our proposed framework towards emotion recognition, based on multimodal fusion of visual and audio modalities, is introduced in Subsection 4.1.3, highlighting different fusion schemes. Next, the experimental results are presented in details in Subsection 4.1.4, while the study ends with conclusions and directions for future work in Subsection 4.1.5.

4.1.1. RELATED WORK

FEATURE EXTRACTION

Previous work on facial emotion recognition mostly uses handcrafted features [14, 138]. The pipeline of these studies starts by performing face detection and is followed by extracting facial features such as Local Binary Patterns (LBP) [61], Gabor wavelets [60], and Scale-Invariant Feature Transformation (SIFT) [62]. In addition, these features can be extended to capture the spatio-temporal space such as Bag of visual Words (BoW) on SIFT features in [139] and LBP-TOP [140]. With the recent improvements in neural networks, deep architectures have become popular and effective for extracting high level features from data and specifically from facial images [89, 89, 90, 96, 146–149, 177]. As discussed in Subsections 3.2.2 and 3.2.3, deep learning approaches for feature extraction have surpassed traditional ones and emerged in an enormous impact and improvement in many pattern recognition and classification tasks [89]. In computer vision, CNNs are well-known deep learning architectures for feature extraction from images. In our work, we benefit from this model as well, by using the state-of-the-art VGG-Face (explained in section 2.4.1) face representation which proved to be discriminative and efficient in face recognition [96]. In [66, 68], CNN features were extracted by fine-tuning pre-trained models for facial emotion recognition. In [152], a hybrid neural network is presented, that combines Recurrent Neural Networks (RNNs) with CNNs to encode facial motion throughout video frames. On the other hand, audio representations are extracted using short-term spectral, prosodic, and energy features, which have affective information [5]. A well-known set of features were extracted using the Open Speech & Music Interpretation by Large-space Extraction (OpenSMILE) tool [178], which are adopted in this work. Audio and video representations for emotion recognition are thoroughly presented in Subsections 3.2.2 and 3.2.3.

MULTIMODAL LEARNING

Data fusion can be achieved with different and complementary modalities such as audio, video, and skeleton joints. The joint analysis of the sensory inputs leads to an improved recognition of the environment, since it enhances the understanding of an event through different channels. However, each modality has its own feature distribution and statistical properties, and different sensory data have high non-linear relationships. There have been many studies that try to optimize a framework and benefit from data of various modalities in order to obtain modality-free description (in other words, joint representations) to capture the correlation between different modalities. For example, authors in [179] developed a probabilistic model to correlate images and their associated captions. In this task, each modality carries different information. For instance, captions can further explain the images' content. The proposed framework aimed to produce joint representations that capture the complementary information. The representations also reflect the similarity of visual and text modalities in the real world. These representations are useful for tasks such as classification and information retrieval [179].

Furthermore, multimodal learning has been applied for several tasks which involve various data sensors, such as person identification, emotion recognition [157], multimedia retrieval [179], and gesture and action recognition [177]. In [17] a deep learning method for audiovisual speech recognition was proposed, where the authors used different settings and scenarios in order to find a framework that would obtain a shared representation for both modalities. One of the main constraints of the scheme presented in that paper is the complexity of the applied architecture. In addition, when analyzing the performance of the late and early fusion, the results show the inefficiency of the deep learning based multimodal learning. This can be traced to the fact that deep learning requires much more data to learn a shared representation than other models.

In our study, we employ a Fisher Vector (FV) representation for encoding low-level features of audio-visual modalities. It functions as a higher layer representation of those features, as it projects them onto a common space, where they share common statistical and distribution properties. Compared with deep learning, FV representation has advantages such as its compactness and efficiency, while it can be computed using a small number of Gaussian Mixture Model (GMM) parameters [180].

MULTIMODAL EMOTION RECOGNITION

There have been various studies that cover multimodal learning for emotion recognition. For example, in [157], multimodal deep learning was applied to learn a shared representation for audiovisual emotion recognition. Other studies exploited late level fusion. In [68], separate methods for each modality were developed, (e.g. CNN for facial images and Restricted Boltzmann Machines (RBM) for audio information), followed by a combination of the score of each modality in late fusion by grid search. Similarly, in [66], the authors benefited from kernel methods for video feature representation and the fusion of the different modalities was achieved in a probabilistic manner at a late stage. In [159], different classifiers were trained for each modality, and then combined using a late fusion approach based on a genetic algorithm. Section 3.3 elaborates on data fusion and representation approaches for multimodal emotion recognition.

In this chapter, we target the task of emotion recognition using both schemes of multimodal learning: early and late fusion. In the early level fusion phase, we first project

modalities' features into a common space that shares similar properties using Fisher vector encoding, and then decide the best combination of the modalities by employing information gain principles for feature selection. In late fusion, we benefit from the resulted combinations of early fusion, and take into consideration the performance of each modality prior to Fisher vector encoding to spot the complementary modalities for robust emotion prediction.

4.1.2. METHOD

In this section, we first explain the preprocessing phase of facial images and how we obtain a face track from a video. Then we present the low-level feature extraction methods implemented in our framework form audio and video (geometric and appearance features). Finally, we describe feature encoding and representation by means of FVs for video modeling and projecting features into the same space.

PREPROCESSING

Facial Landmark Detection: Succeeding the step of face detection via Haar feature-based cascade classifiers [181], we detect 49 landmarks and track them in each frame of a video, using the Supervised Descent Method (SDM) [182]. SDM is a successful face shape regression technique, which begins with facial features (e.g. SIFT), around facial landmarks (S_0), and progressively predicts the final shape of the face in an iterative way. SDM minimizes a Non-linear Least Squares (NLS) function and then uses the learned descent directions to estimate face shape during runtime. Compared to other techniques, this method provides robust and accurate landmark positions in challenging conditions, such as varying illumination and pose, and low-quality images. In addition, it gives a reliable and robust tracking of facial landmarks in the wild, in real-time.

Face Alignment: Face alignment is an essential step in facial emotion recognition. It is the process of registering faces with respect to facial landmarks (e.g. eyes, nose, mouth, and chin) of the canonical frame. This process fixes the landmark positions in aligned images and it is carried out by similarity transformation. In our work, we use facial landmarks provided by the SDM landmark detector and perform a similarity transformation that aligns faces to the fixed canonical frame based on eye centers positions. In addition, facial images are cropped and re-sized to a fixed resolution: 224×224. Figure 4.2 presents examples of tracked facial images from the AFEW dataset.

LOW-LEVEL FEATURE EXTRACTION

Emotion recognition relies on representative data along with accurate and discriminative descriptors. This type of information contributes to enhanced recognition and classification accuracy. Accordingly, in this chapter, we extract a set of low-level descriptors for the visual and audio modalities. Then, we use Fisher vectors for video modeling and projecting them into the same space. In the reminder of this subsection, we outline the low-level features used in our work: DSIFT, handcrafted geometric, CNN and audio features.

Dense Scale-Invariant Feature Transformation (DSIFT) Features: DSIFT has been widely used for image representation in the last decade, in many computer vision recognition tasks [183, 184]. In DSIFT, unlike the regular SIFT where features are extracted

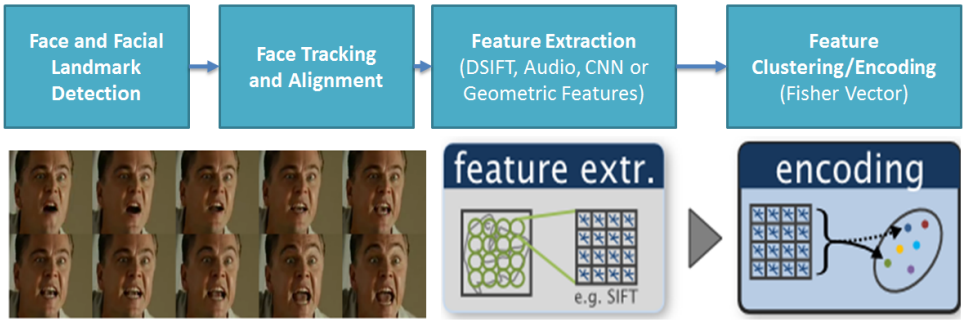


Figure 4.2: Face pre-processing and Feature Extraction and Encoding.

Figure 4.3: An illustration of the six salient facial regions of interest (ROIs): left eye, right eye, forehead, mouth, nose and the region between eyes.

sparsely around facial key-points, we compute the SIFT histograms (of 128 dimensions) densely over a given facial image, using a certain scale factor and step size. This has an advantage since it does not rely on facial landmark detection. We divide the facial images into a grid of overlapping blocks with a step size equal to 1 pixel. Specifically, the block size is 24×24. Later, we compute a DSIFT histogram for each block. This step is repeated in 5 scales, with a scale factor equal to $\sqrt{2}$.

Then, features are aggregated from all patches to a single matrix with size of $128 \times P$, where P is the number of patches. The final descriptor of DSIFT is further reduced by Principal Component Analysis (PCA), from 128 to 64 as suggested by [183]. To help Fisher vector encoding in capturing the spatial information of the face, the normalized spatial coordinates (x, y) of the patches are aggregated to the features. As a result, the PCA-DSIFT feature dimension is 66.

In this work, we compute DSIFT with two approaches: (i) DSIFT on the entire facial image; and (ii) DSIFT on six distinct facial Regions of Interest (ROIs): left eye, right eye, forehead, mouth, nose, and the region between eyes. These six facial ROIs are illustrated in Figure 4.3. We extract and crop the ROIs using the obtained facial landmarks shown in Figure 4.5. Then, DSIFT features are extracted from each region separately. In the

Figure 4.4: A modified VGG-face architecture. The original VGG-face model has the same convolutional/pooling blocks and the Fully Connected Layer (FCL) that follow these blocks. The last two FCLs were added in order to fine-tune the model for the purpose of emotion recognition from facial expressions.

remainder of this chapter, we refer to the DSIFT extracted from the entire facial image as DSIFT, while we call the DSIFT computed on ROIs as ROIs-DSIFT. The dimensionality of ROIs-DSIFT histograms were reduced from 128 to 32 by PCA. As a result, the ROIs-DSIFT feature dimension is 34, since the normalized spatial coordinates (x, y) are added to ROIs-DSIFT.

Convolutional Neural Networks (CNNs) Features: Our CNN face representation is based on the VGG-face model [96] (explained in Subsection 2.4.1), which is a CNN model that consists of five convolutional/pooling blocks and three fully convolutional layers. This model was trained with 2.6M facial images of 2622 people for the purpose of face recognition in the wild [96]. In the literature, it has been shown that CNNs give robust representations that are superior to AU's based feature extraction methods [68, 145]. Nonetheless, as the model is trained for the facial recognition task, and not on emotional data, we fine-tune the model, by keeping the convolutional/pooling blocks and the Fully Connected Layer (FCL) which follows these blocks (denoted as FCL6). The remaining FCLs were discarded, namely, FCL7 and FCL8 which had 4096 and 2622 dimensions, respectively. Instead of these two layers, we appended 1000 and 7 dimensional FCLs. On the top of the final layer, softmax classification with multinomial logistic loss was applied. The modified architecture of the VGG-face model is shown in Figure 4.4. For fine-tuning, we use the training set of the Facial Emotion Recognition (FER) dataset [185], and for validation, we use the public test set of FER. In the feature extraction stage, we employ the FCL6's output as the facial signature. This layer outputs a 4096-dimensional feature vector.

Geometric Features: These features deal with the shape and location of the facial landmarks. Methods to extract geometric features use facial landmarks to encode explicitly face geometry [141]. As explained in Subsection 3.2.3, geometric features are useful since different facial expressions correspond to different facial landmarks' shape deformations (e.g., eyes, mouth, eyebrows, chin, and nose). In this study, the coordinates of the facial landmarks are obtained through the model proposed in [182] and are shown in Figure 4.5. These landmarks are transformed and fitted with the same alignment used for face registration. Moreover, the coordinates of the facial landmarks can be used as features in the classification process. However, this representation results in poor performance since it does not capture the dynamic variations between various individuals [141]. For example, authors in [143, 144] computed geometric features, which can be represented by segments, perimeters, or areas of the figures formed by the coor-

Table 4.1: The extracted handcrafted geometric features. These sets of features are extracted for each face, in a face track. Then, FV encoding aggregate them across a video clip to obtain a single representation. This table is taken from [144].

	What features?	On which landmarks?	Operation
1	Eye aspect ratio (LR)	[20:25], [26:31]	Distance
2	Mouth aspect ratio	32, 35, 38, 41	Distance
3	Upper lip angles (LR)	32, 35, 38	Angle
4	Nose tip - mouth corner angles (LR)	17, 32, 38	Angle
5	Lower lip angles (LR)	[32, 42] , [38, 40]	Angle
6	Eyebrow slope (LR)	[1, 5] , [6, 10]	Angle
7,8	Lower eye angles (LR)	[20, 23, 24, 25], [26, 29, 30, 31]	Angle
9	Mouth corner - mouth bottom angles	32, 38, 41	Angle
10	Upper mouth angles (LR)	[32, 34], [36, 38]	Angle
11	Curvature of lower-outer lips (LR)	[32, 43, 42], [38, 39, 40]	Curvature
12	Curvature of lower-inner lips (LR)	[32, 42, 41], [38, 40, 41]	Curvature
13	Bottom lip curvature	[32, 38, 41]	Curvature
14	Mouth opening / mouth width	45, 48, 32, 38	Distance
15	Mouth up/low	35, 41, 45	Distance
16	Eye - middle eyebrow distance (LR)	[3, 20, 23], [8, 26, 29]	Distance
17	Eye - inner eyebrow distance (LR)	[5, 20, 23], [6, 26, 29]	Distance
18	Inner eye - eyebrow center (LR)	[3, 23], [8, 26]	Distance
19	Inner eye - mouth top distance	23, 26, 35	Distance
20	Mouth width	32, 38	Distance
21	Mouth height	35, 41	Distance
22	Upper mouth height	32, 38, 35	Distance
23	Lower mouth height	32, 38, 41	Distance

dinates of the facial landmarks. Their features include Euclidean distances, angles, and curvatures between fitted facial landmarks.

Following the works in [143] and [144], we obtain a set of features including: Euclidean distances, angles and curvatures between fitted facial landmarks, followed by applying a normalization step. For example, the set of extracted features include but are not limited to: mouth and eyes aspect ratios, lower and upper lips, and mouth corners' angles, nose tip-mouth corner angles, eyebrow slope, mouth corner and mouth bottom angles, and the curvature of lower-outer and lower-inner lips. Table 4.1 gives a comprehensive list of the extracted 23 features, for each frame. Subsequently, these features are pooled and encoded via FVs to obtain a single geometric description of each video clip.

Audio Features: We utilize the speech analysis OpenSMILE toolkit [178] for extracting audio features. This popular and widely used library extracts low-level descriptors (LLDs) that capture both the voice quality and prosodic characteristics of a speaker. We follow the audio feature extraction as explained in [127]. The set of audio features used in this study consists of: 34 energy and spectral related LLDs, 4 voicing related LLDs, 34 delta coefficients of energy and spectral LLDs, 4 delta coefficients of the voicing related LLDs, and 2 voiced/unvoiced durational features. The details for the LLDs are included in Table 4.2. Then, the following set of functionals are computed on the LLDs:

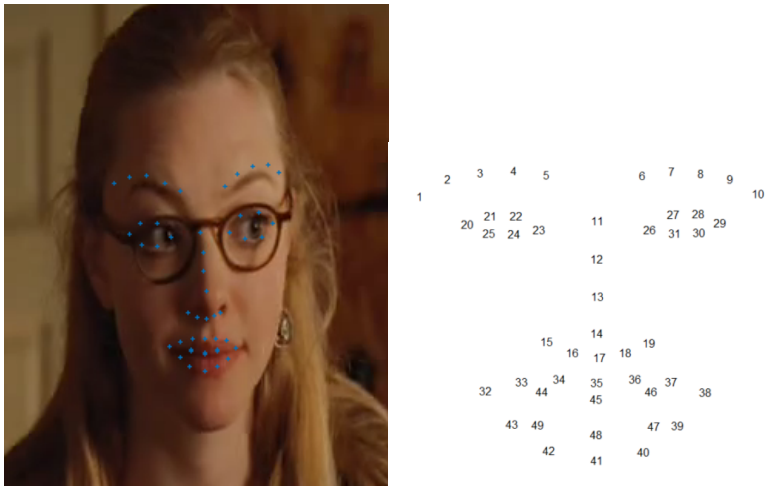


Figure 4.5: Facial landmarks provided by the SDM [182].

Table 4.2: Audio features: Low Level Descriptors (LLDs) [127]

Low Level Descriptors (LLDs)	Audio Features
Energy/Spectral LLDs	PCM Loudness
	MFCC [0-14]
	Log Mel frequency band [0-7]
	Line Spectral Pairs (LSP) frequency [0-7]
	F0 by sub-harmonic sum and F0 Envelope
Voicing related LLDs	Voicing Probability
	Jitter local
	Jitter difference of periods in consecutive frames
	Shimmer local

arithmetic mean, standard deviation, skewness, kurtosis, quartiles, quartile ranges, percentile 1%, 99%, percentile range, position max/min, up-level time 75/90, linear regression coefficient, and linear regression error (quadratic/absolute). The statistical functionals capture the dynamic nature of the voice over time segments. In addition, for each video clip, there is a single feature vector ($N = 1$) describing its acoustic features since the statistical functionals are applied to summarize the LLDs, which are computed over time segments, i.e. 60 ms. Finally, for each video-clip, the LLDs and the applied functionals resulted in a 1582 dimensional feature vector.

FEATURE ENCODING AND VIDEO MODELING

Video Modeling: In this work, we adopt the usage of Fisher vectors for encoding and clustering different low-level features for each feature type of audio and video modalities. The features are not only pooled from one still image. Instead, they are pooled from all the frames across a face track. As suggested in [184], we use video-pooling, where we compute a single fisher vector over the whole face track by pooling together low-level features (e.g. DSIFT, or CNN features) from all facial images in a track. This kind of representation has many advantages compared to still image-based representations for various reasons: (i) it encodes the spatio-temporal information in a face track, (ii) it captures the motion of the face over time, which leads to a robust description of the different low-level features; and (iii) it dramatically reduces the dimensionality of data by producing a single discriminative descriptor for a video.

Fisher Vector Representation: The pipeline for FV encoding typically starts with extracting a set of features (e.g. DSIFT, geometric features, etc.), and then aggregates the large set of feature vectors across all frames in a track into a high dimensional Fisher vector which is well suited for linear classification. This is achieved by fitting a parametric generative model such as GMMs to the features. The GMMs can be referred to as a *probabilistic visual vocabulary*. The next step consists of encoding the gradient of the local descriptors log-likelihood with respect to the GMM parameters, such as GMMs' means, weights, and covariance matrix. The computation of FVs can be summarized as follows:

- Let $I = \{\boldsymbol{x}^{(i)}, \ldots, \boldsymbol{x}^{(N)}\}$ be a large set of local descriptors with d dimensional feature vectors (e.g., the geometric features or DSIFT descriptors). The dimensions (d) of each feature descriptor (type) are listed in Table 4.3. Note that the number of local descriptors (N) depends on the number of frames in a video sequence and the number of patches (for DSIFT and ROIs-DSIFT). However, for each video clip, a single feature vector (N=1) describes its acoustic features, as detailed in the previous subsection. More importantly, it is assumed that the set of feature vectors are independent.

- Authors in [180] defined FVs as a sum of normalized gradient statistics of probability density function which is computed for each feature vector. In Fisher vector encoding, it is also assumed that these vectors are generated by a GMM with K components and, therefore, the probability density function is modeled by a GMM: $p(\boldsymbol{x}) = \sum_{k=1}^{K} w^{(k)} p_k(\boldsymbol{x})$.

- Let $\Theta = \{\boldsymbol{\mu}^{(k)}, \Sigma^{(k)}, w^{(k)}\}$ be the parameters of a GMM (p_k), where $w^{(k)}, \boldsymbol{\mu}^{(k)}, \Sigma^{(k)}$ are the GMMs weights, means, and covariance matrix of the k^{th} GMM, respectively. It is important to note that, in Fisher vector computation, the covariance matrix of the GMMs is assumed to be diagonal and denoted by $\boldsymbol{\sigma}^{2(k)}$, the variance vector (i.e. the diagonal values of $\Sigma^{(k)}$) [180].

- The parameters of GMMs are learned to fit the distribution of given descriptors, e.g. DSIFT or geometric features. These parameters are learned on the local descriptors using the Expectation-Maximization (EM) algorithm. The EM algorithm optimizes a Maximum Likelihood criterion.

- In [180], the authors proposed to use only the gradient with respect to the mean and the variance vectors. They discarded the gradient with respect to the GMMs' weight ($w^{(k)}$) as it brings little information. Therefore, for each component of the GMM, the derivatives with respect to the mean and the diagonal covariance (the variance vector) lead to a vectorial representation, which captures the average first and second order differences between the features and each of the GMM mode, as follows:

$$\Phi_{\mu}^{(k,j)} = \frac{1}{N\sqrt{w^{(k)}}} \sum_{i=1}^{N} \alpha^{(k,i)} \left(\frac{x^{(i,j)} - \mu^{(k,j)}}{\sigma^{(k,j)}}\right) \tag{4.1}$$

$$\Phi_{\sigma}^{(k,j)} = \frac{1}{N\sqrt{2w^{(k)}}} \sum_{i=1}^{N} \alpha^{(k,i)} \left(\left(\frac{(x^{(i,j)} - \mu^{(k,j)}}{\sigma^{(k,j)}}\right)^2 - 1 \right) \tag{4.2}$$

  where $j = 1, 2, \ldots, d$ spans the feature vectors' dimensions (as shown in Table 4.3) and $\alpha^{(k,i)} = p(k|\boldsymbol{x}^{(i)})$ is the soft assignment of a descriptor $\boldsymbol{x}^{(i)}$ to the $(k)^{th}$ Gaussian. In other words, the GMM associates each vector $\boldsymbol{x}^{(i)}$ with each Gaussian component using a strength given by the following posterior probability:

$$\alpha^{(k,i)} = \frac{exp\left(-\frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(k)})^T \Sigma^{-1(k)} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(\boldsymbol{k})})\right)}{\sum_{q=1}^{K} exp\left(-\frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(q)})^T \Sigma^{-1(q)} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(\boldsymbol{q})})\right)} \tag{4.3}$$

- The resulting Fisher vector $\boldsymbol{\phi}$ of a video clip is computed by aggregating the resulted average first and second order differences ($\Phi_{\mu}^{(k,j)}$ and $\Phi_{\sigma}^{(k,j)}$) as follows:

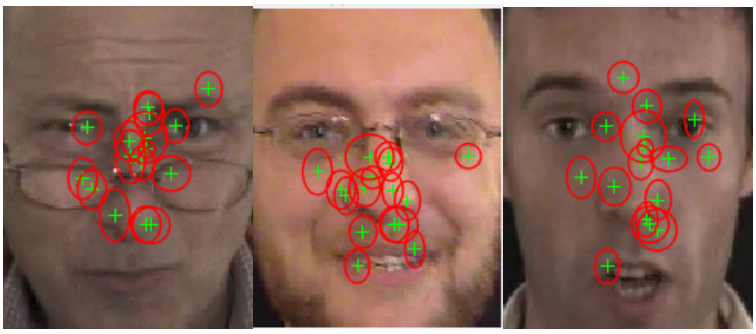
$$\boldsymbol{\phi} = [\Phi_{\mu}^{(1,1)}, \cdots, \Phi_{\mu}^{(K,d)}, \Phi_{\sigma}^{(1,1)}, \cdots, \Phi_{\sigma}^{(K,d)}]^T$$

It is important to note that Fisher vectors' dimensionality is $2 * K * d$ which depends on the number of the GMM components (K), the dimensionality of the employed set of feature types, and the two derivatives with respect to the mean and the variance vectors of the GMMs ($\boldsymbol{\Phi}_{\mu}^{(k)}$ and $\boldsymbol{\Phi}_{\sigma}^{(k)}$). Moreover, in our study, we set different numbers of GMMs for each feature type. These numbers are based on the evaluations and experimentations to what is suitable for each feature type. Table 4.3 summarizes each feature type's dimensionality, the number of used GMMs, and the resulting FVs' dimensionalities. As a result, since we apply the FV on the entire video clip, it is a video representation that is obtained by pooling the spatio-temporal features across the entire video clip. This method is frequently used as a global video and image descriptor in visual classification.

To check the quality of the employed feature extraction and aggregation methods, Figure 4.6 shows the localization of the Gaussians with the highest energy in a linear classification by SVM. The shown Gaussians are obtained from DSIFT descriptors of single images. In DSIFT, each patch descriptor is associated with its spatial information. In particular, as explained in Subsection 4.1.2, to enhance Fisher vector encoding with the spatial information of the face, the (x, y) spatial coordinates of patches are added

Table 4.3: The dimensionalities of feature types and their Fisher vectors representations. Note that even though dimensionality of FVs is large, it is still significantly lower than stacking all the low-level features from all the frames and the patches of a video clip. For example, the dimensionality of $F^{ROIs-DSIFT}$ features can be larger than $5.8M$ for patches with 32×32 in 1 second video clip.

Modalities	Feature Types	Feature Dimensions (d)	GMMs (k)	FV Representations	FVs (ϕ) Dimensions = $2Kd$
Visual	F^{DSIFT}	64 + 2 (x, y) = 66	32	FV^{DSIFT}	2 * 32 * 66 = 4224
	$F^{ROIs-DSIFT}$	32 + 2 (x, y) = 34	32	$FV^{ROIs-DSIFT}$	2 * 32 * 6 (ROIs) * 34 = 13056
	F^{CNNs}	4096	4	FV^{CNNs}	2 * 4 * 4096 = 32768
	$F^{Geometric}$	23	16	$F^{Geometric}$	2 * 16 * 23 = 736
Audio	F^{Audio}	1582	1	FV^{Audio}	2 * 1 * 1582 = 3164



(a) Anger (b) Happiness (c) Fear

Figure 4.6: The most important facial patches localization on facial expressions images, corresponding to anger, happiness and fear. 15 Gaussians are drawn, which correspond to the highest energy in a linear SVM classification.

to their features (descriptors). As a result, these spatial features are encoded along with DSIFT descriptors. In other words, each Gaussian has mean and variance dimensions encoding the spatial information. Subsequently, here, this added information is used to display the important Gaussians in a facial image, given the energy (weights) learned by a linear SVM. In general, the localization of the most influential Gaussians is in the center of the face. Furthermore, as it can be noticed, the most important 15 Gaussian locations corresponding to the highest energy, vary according to the expressed emotion. For example, in the facial expressions of anger and happiness, we observe that the Gaussians are concentrated around the eyes and the mouth, respectively; while for fear, the Gaussians are scattered in the center of the face, in different regions.

4.1.3. MULTIMODAL FUSION

In this section, we present the two fusion approaches employed in the proposed framework, namely, feature level fusion based on information gain and score level fusion using genetic algorithm. We propose a framework for multimodal emotion recognition, which combines the two modalities' feature types in a hierarchical and collaborative fashion, using both early and late fusion schemes. These two techniques aim to maximize the

benefit of employing different features in audio-video emotion recognition. In the rest of this section, first, we introduce our approach by explaining how information gain and Fisher vector representation are involved in early level fusion. Then, we describe our collaborative, late-level fusion method that captures the performance of modalities' paired feature representations per emotion to enhance the final decision-making.

FVs BENEFITS

In our study, we apply various feature extraction techniques. Accordingly, audio-video data comes from heterogeneous input channels. Therefore, each feature has its own distinct distribution properties. However, a multimodal fusion and feature learning methods can be used to capture the correlations between these modalities in real-word data, by employing a feature level representation. As a result, the similarity in the representation space must reflect the similarity in corresponding concepts. For example, speech and facial images are correlated in the real world when people express their emotions. People often tend to speak loudly when they are angry, or they use a certain tone of voice accompanied by a facial expression to indicate their affective states. We use the Fisher vector encoding to map the extracted features onto a common space to achieve a higher layer feature description, which shares similar statistical and discriminative properties. Thus, different feature types for both modalities are projected onto one domain by Fisher vectors, enabling and supporting feature concatenation (explained in the next subsection). The newly obtained Fisher vector-based representations are independent of the input modality, whereas the low-level features are modality-dependent. This is because the encoding of the low-level features is modeled by a generative approach, namely: GMMs. FV encoding assumes that features have a Gaussian distribution, which GMMs can capture, accordingly.

Moreover, the Fisher Vector representation has many advantages over traditional approaches such as Bag of visual Words (BoW). For instance, (i) it is a generic representation that combines the benefits of generative and discriminative approaches, (ii) it can be computed using a small number of parameters (GMM parameters), (iii) more importantly, it is efficient and it shows a significant benefit when used in combination with linear classifiers such as linear-SVM [180]. Furthermore, the employed data representation is useful at capturing the non-linear relationships between the different feature types used in our work. Specifically, since the embedding of the local descriptors is modeled by GMMs which a generative model, the encoding of the feature types shares similar statistical properties.

FEATURE LEVEL FUSION (EARLY FUSION) BASED ON INFORMATION GAIN PRINCIPLES

As described previously, the pipeline of the employed method starts by creating FV representations (explained in Subsection 4.1.2) per feature type (e.g. a different one for VGG and a different one for geometrical features). These feature types are explained in Subsection 4.1.2. Moreover, their dimensionalities and the dimensionalities of the corresponding FVs are summarized in Table 4.3. In the following subsections, we refer to the low-level descriptors as features, and to their Fisher vector encoding as feature representations. The set of the encoded features are as follows: $FV^{DSIFT}, FV^{ROIs-DSIFT}, FV^{audio}, FV^{CNN}$, and $FV^{geometric}\}$.

In the early fusion, we adopt the concatenation of paired feature representations in order to provide the Genetic Algorithm employed for late fusion with the best suitable search space for optimal search. Genetic algorithms constitute a method inspired by the evolutionary process, i.e. natural selection. GAs are used to provide solutions for search problems by means of evolutionary concepts such as mutation, cross-over, and selection. Nonetheless, GAs can work best with a small search space and does not scale well with complexity. As a result, the early fusion results in the best-paired combinations, while the genetic algorithm optimizes the weights for fusing the paired feature representation in the late fusion. In this way, the computational complexity of the search space in the late fusion is reduced, aiming for an efficient fusion.

In order to achieve a hierarchical scheme between this step and the following late fusion, we employ the concatenation on pairs of feature representations. Then, a linear classification using SVMs is applied on these concatenated feature representations. For optimizing the early fusion of audio and video feature representations, and for selecting the best combination among the possible representations, we use measures from information theory, namely the Kullback-Leibler (KL) Divergence [174]. KL-D is useful at measuring the distance between two probability distributions (PDFs). In our study, the PDFs are obtained from the distribution of the labels (emotions) in the overall data samples. In our framework, we aim to choose those pairs with minimum distances between the PDF of their correctly predicted labels (denoted with $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}^{(i)}, i \in \{1,\ldots,f_{conc-feat-rep}\}$) and PDF the true labels (emotions), denoted with $\boldsymbol{y}$. Here, $f_{conc-feat-rep}$ indicates the number of feature representations' pairs. The PDFs of the correctly predicted labels are obtained following an SVM linear classification on top of the i^{th} paired feature representations. Thus, in our case, KL-D is employed as follows:

$$D_{KL}(\hat{\boldsymbol{y}}^{(i)}||\boldsymbol{y}) = \sum_{j=1}^{c} y_j^{(i)} log \frac{\hat{y}_j^{(i)}}{y_j} \tag{4.4}$$

where c is the number of labels. For instance, $\hat{y}_j^{(i)}$ refers to the ratio of samples which were correctly predicted as the j^{th} label (e.g. sadness), when using the i^{th} concatenated feature representations. Specifically, we employ a linear classification (via SVMs) on pairs of concatenated feature representations (e.g. FV^{audio} and $FV^{ROIs-DSIFT}$). Then, the PDF of their correctly predicted emotions ($\hat{\boldsymbol{y}}$), are compared with the actual emotions' distribution ($\boldsymbol{y}$). By having minimum KL divergence, we aim at obtaining PDFs as close as possible to the actual PDFs of the emotions, increasing, in this way, the performance accuracy of our emotion recognition framework. As the KL divergence is not symmetric, we employ in our work the symmetric version [186], for obtaining a general framework, which is not affected by the order of the feature representations in the fusion process:

$$I(\hat{\boldsymbol{y}}^{(i)}, \boldsymbol{y}) = \frac{D_{KL}(\hat{\boldsymbol{y}}^{(i)}||\boldsymbol{y}) + D_{KL}(\boldsymbol{y}||\hat{\boldsymbol{y}}^{(i)})}{2} \tag{4.5}$$

This stage uses a set of fused feature representations' pairs (from both audio and video modalities), achieving in this way a result as close as possible to the expected actual PDFs:

$$\text{argmin}\, I([\hat{\boldsymbol{y}}^{[j,k]}, \boldsymbol{y}), \text{ where } k, j \in \{1, \ldots, f_{feat-rep}\} \tag{4.6}$$

where $[j,k]$ indicates the PDF obtained from i^{th} concatenated (paired) j and k feature representations. The best pairs of feature representations are carried to the next level, based on the symmetric KL-D. In the next stage, only the best pairs of feature representations (based on the symmetric KL-Divergence values) are employed in a late fusion using Genetic Algorithm (GA), a late fusion technique which is explained in the following subsection.

SCORE LEVEL FUSION

In our work, we observed that emotional states are more dominant depending on the existing modalities, e.g. some of them prevail through the visual channel, while others are stronger displayed through the audio modality. As modalities, and their feature representations, can be complementary to each other and display varying performance characteristics across emotions, we take advantage of this aspect for predicting emotional states in a collaborative manner at the decision level. We apply this scheme in two stages: First, we select the best pairs of feature representations based on the symmetric KL-Divergence, resulted from the early fusion. Second, we combine the scores of the resulted combinations at the decision level. In the first stage, each feature representations' classifier is regarded as an expert model due to its distinctive performance in emotion prediction. In this phase, we take advantage of the best fused feature representations obtained using the information gain principles (presented in the previous subsection). Then, we apply a weighing scheme that takes into consideration the performance of each combination with respect to each affective state. The final decision is obtained using a weighted sum of the prediction given by each pair of feature representations. For optimizing our results, we employ a genetic algorithm (GA) to assign weights to the resulting scores for each affective state.

GA is a metaheuristic optimization algorithm and was first proposed by Holland [187] in 1970, inspired by natural selection. It can be applied in search problems to find the fittest candidate solution in the search space (S). This algorithm applies selection, crossover, and mutation using a fitness function. The initial population can first be set randomly. Nonetheless, this population is improved through an iterative selection based on the fitness score. For example, our case's fitness score is based on the candidate solution's performance to re-weight paired feature representations' predictions in the training set of emotion recognition. The iterative selection relies on the two pairs of individuals (parents) from a previous iteration. Children with good fitness are more likely to be selected. The crossover involves the cross point to mate parents, and this point is chosen randomly. Finally, the mutation adds randomness to the candidate offspring with low probability by changing their values.

We applied a re-weighing on emotion predictions coming from the best classifiers on the pairs of concatenated FVs' representations. This is the search space which is a hyperparameter that assigns scores for each emotion and the predictions of feature representations' pairs. Accordingly, GA learns the weights of the final decision for the pairs' combination and their predictions. The search space S of GA depends on the number of feature representations pairs fed into it: $f_{feat-rep-pairs}$, and the number of labels c (i.e. the number of basic emotions). Therefore, the search space matrix S has $[f_{feat-rep-pairs} \times c]$

dimensions. Note that $f_{feat-rep-pairs}$ is the number of best selected paired feature representations which are resulted from the late fusion.

Prior to learning the weighing scheme of the selected pairs of feature representations (that can come from audio and video modalities), we consider lower and upper bounds constraints to avoid over-fitting a specific pair of feature Representations by GA. We use the following constraints to regularize the learning during the weight search:

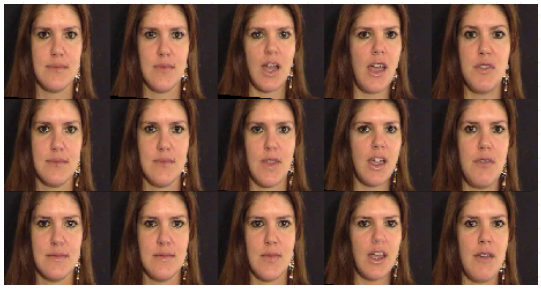
$$0 \leqslant S(k,i) \leqslant 1, \& k \in \{1,\ldots,f_{feat-rep-pairs}\}, i \in \{1,\ldots,c\} \tag{4.7}$$

Note that k, here, has the predictions of two concatenated feature representations (e.g. the linear SVM predictions of the concatenated FV^{audio} and FV^{CNNs} representations).

Algorithm 1 High Level Description of the Hierarchical Multimodal Fusion.

procedure AUDIO-VIDEO EMOTION RECOGNITION
Inputs:
    Raw audio and video signals in video clips
Low-level Feature Extraction:
    Extract set of audio and video features (f)
    ▷ As shown in Figure 4.1 and described in Subsection 4.1.2
Video Modeling and Feature Encoding:
    Employ feature encoding and aggregation via FVs
    ▷ As elaborated in Subsection 4.1.2, so that we obtain a single descriptor per feature type for each video: $FV^{(feature-type)}$
Early Fusion:
    Concatenate feature representations in pairs, $[FV^{Audio}, FV^{Geometric}]$
    ▷ Here, it results in 10 pairs since we have 5 feature types.
    Perform Linear SVM on each concatenated feature representations' pair
    ▷ E.g.: $[FV^{ROIs-DSIFT}, FV^{Geometric}]$
    Calculate the symmetric KL-D between the PDF of the correctly classified emotions coming from the concatenated pairs of features and the PDF of the actual emotions
    Select best pairs of feature representations based on lowest symmetric KL-D values
    ▷ E.g.: two, three, or four pairs
    ▷ In our study, three pairs were selected since they outperformed other permutations (including all feature representations combined)
Late Fusion:
    Employ a re-weighing scheme for a late fusion on the predictions of the best pairs of feature representations using GA
Evaluation:
    Use the obtained weights for evaluating new samples
end procedure

Finally, Algorithm 1 details the proposed procedure. It summarizes the steps of (1) the extraction of feature types, (2) the feature encoding via Fisher vectors, (3) the feature-level fusion based on concatenating pairs of feature representations, and (4) the late level



(a) eNTERFACE: An example of a face track

(b) AFEW: Static images

(c) AFEW: Aligned and cropped facial images

Figure 4.7: Examples of a face track and static images from eNTERFACE and AFEW, respectively

fusion benefiting from a weighing mechanism using GA. In our study, three pairs were fed to the last step in the weighing mechanism via GA. We noticed that three pairs of feature representations significantly outperformed other permutations (including all feature representations combined). Also, there was not a significant gain when more than three pairs of feature representations were employed.

4.1.4. RESULTS

In this section, we first introduce the chosen datasets for our experiments, then we present an extensive study and evaluation of the employed feature types and their encoding. We first evaluate each feature representation separately to assess their discriminative properties and to estimate their efficiency. Then, we apply feature level fusion on all feature representations. Finally, we apply the proposed hierarchical scheme for selecting the combination of feature representations' pairs based on information gain principles and genetic algorithm optimization.

DATASETS

Acted Faces Emotion In The Wild (AFEW) [117] is divided into three subsets: training (773 samples), validation (383 samples), and test (593 samples), while only the training and validation sets are publicly available. It has both audio and video modalities. In this dataset, each video clip is labeled with one of the basic emotions: anger, disgust, fear, happiness, sadness, and surprise or as neutral. Several facial expressions' datasets are gathered in controlled environments, which mainly contain static images or videos of frontal faces. Furthermore, the facial expressions are posed, limiting the capacity of the data to reflect challenging real-world conditions. However, AFEW has the following properties: (i) it is a challenging dataset with occlusions, varying illumination, and head poses, which meets real-world conditions; (ii) it provides baseline results and an evaluation protocol which is useful to evaluate our scheme's efficiency, and (iii) it is currently studied by the research community, as it has been the subject of several competitions over the last few years.

eNTERFACE is a multimodal dataset which contains six archetypal emotions: anger,

Table 4.4: Performance of individual modalities and their representations on the AFEW validation set and eNTERFACE using linear SVM classifier. Statistical bounds are provided for eNTERFACE in terms of standard deviations for the average accuracies across the datasets' folds. AFEW has only a validation set, so it is common in the literature to produce accuracy on its validation set. Comparisons with other methods are shown at the end of this section.

Modalities	Feature Representations	AFEW Validation Set Accuracy (%)	eNTERFACE Average Accuracy (%)$\pm std$
Visual	FV^{DSIFT}	39.4	60.3±3.5
	$FV^{ROIs-DSIFT}$	39.2	60.6±3.3
	FV^{CNNs}	40.0	61.5±5.0
	$FV^{Geometric}$	32.8	43.5±5.9
Audio	FV^{Audio}	36.4	55.7±4.8

happiness, disgust, fear, surprise, and sadness. It includes 42 subjects who were asked to simulate the emotions in 5 different reactions, resulting in 1260 video recordings. 23% of the recordings were obtained from women, and 77% were gathered from men. The respondents have diverse cultural backgrounds. In the evaluation, we follow the protocol described in [188]. In this protocol, for cross-validation, we split the data samples into ten folds.

More details on these two datasets are available in Section 3.1. Furthermore, as we described in Subsection 4.1.2, we use video-pooling, where the low-level features are pooled from all the frames across a face track in each video of the AFEW and eNTERFACE sets. Then, we compute a single Fisher vector over the whole face track by aggregating and encoding low-level features (e.g. DSIFT or CNN features) of all frames. Examples of static facial images and cropped faces from the AFEW dataset are depicted in Figure 4.7b and Figure 4.7c, respectively. Figure 4.7a shows a face track from the eNTERFACE dataset.

EVALUATIONS METRICS

In our experiments, we take into consideration several evaluation criteria: (i) Accuracy, which is the number of correctly classified video samples (ii) Confusion Matrix between the ground truth and the predicted emotion labels and (iii) Symmetric KL-Divergence, where we aim to minimize the symmetric KL-divergence between the predicted labels and the true labels. For the AFEW dataset, we train our proposed approach on the training set and test it on the validation set. For eNTERFACE, in each fold, we trained the whole pipeline with the training folds and evaluated the framework on the validation fold. The reported results are the average results of the validation folds in terms of accuracy and their standard deviations.

UNIMODAL EXPERIMENTS

Firstly, we apply the evaluation metrics for each feature representation separately on the eNTERFACE and AFEW validation sets. These experiments aim to show the performance of different feature representations for both visual and audio modalities. The results are presented in Table 4.4. In addition, the developers of AFEW [117] provided baseline results using handcrafted audio features (i.e. similar to the ones employed in our study) and Local Binary Patterns on Three orthogonal Planes (LBP-TOP) for visual

Table 4.5: Performance of feature level fusion (FLF) on concatenated pair modalities of AFEW validation set. Note that [,] indicates concatenated representations.

Fused Modalities/Feature Representations	Sym-KLDV	Validation Set Accuracy (%)
$[FV^{ROIs-DSIFT}, FV^{Geometric}]$	0.2622	43.6
$[FV^{Audio}, FV^{DSIFT}]$	0.2626	43.3
$[FV^{Geometric}, FV^{DSIFT}]$	0.3244	40.6

Table 4.6: Average performance of feature level fusion (FLF) on concatenated pair modalities of eNTERFACE. Note that [,] indicates concatenated representations.

Fused Modalities/Feature Representations	Sym-KLDV	Average Accuracy (%)±std
$[FV^{Audio}, FV^{CNN}]$	0.015	74.6±3.9
$[FV^{ROIs-DSIFT}, FV^{Audio}]$	0.025	73.7±1.9
$[FV^{Geometric}, FV^{Audio}]$	0.030	64.3±4.4

representations. The baseline results are based on feature level fusion by concatenating audio-video representations. SVM was applied for classification, achieving 38.8% accuracy for the validation set [117]. The presented results in Table 4.4 show that best performance is obtained through the visual modality, where CNN is the leading representation. Moreover, for AFEW, CNN appearance-based feature representations resulted in slightly higher than the baseline (38.8%) results. Another interesting finding is represented by obtaining an improved accuracy of audio features when encoded with Fisher vectors in comparison to the raw audio features. For instance, we noticed that, in the AFEW validation set, the gain in the performance was significant, at around 6%. This improvement can be attributed to the employed robust feature representation and modeling in our study.

MULTIMODAL EMOTION PREDICTION

Feature Level Fusion was introduced in Subections 4.1.2 and 4.1.3. We first encode the low-level features of audio and visual modalities using a Fisher vector representation. To such an extent, we create feature representations per feature type that come from similar distributions. Next, we concatenate the Fisher vectors of pair feature types (which can come from audio or video modalities) and then perform the classification task using linear Support Vector Machines (linear-SVM). In the case of concatenating the Fisher vectors of all feature types, the accuracies on eNTERFACE and the AFEW validation set are 77.0% and 45.6%, respectively. However, fusing all the feature representations into one feature vector is less efficient for the classification task and slower than the following scheme of score level fusion, which is based on the fusion of the best pair feature representations/modalities.

It is important to note that, in the AFEW dataset, there are several videos for which it is very hard to decide their emotion label only from the visual information. For example,

Table 4.7: Score Level Fusion (SLF) of pair modalities in table 4.5 and 4.6 AFEW validation set and eNTERFACE.

Score Level Fusion	AFEW Validation Set Accuracy (%)	eNTERFACE Average Accuracy (%)±std
Genetic Algorithm Based Fusion	48.9	78.5±2.9
Performance Based Weights Fusion	44.4	77.0±2.9

in many videos, we noticed that facial expressions, labeled as surprise, have been classified as an angry emotion by several human annotators, an observation also supported by [189]. Therefore, we need complementary information to enhance accuracy and to classify these ambiguous videos correctly. Thus, the audio modality represents one way to boost the classification task's performance by adding the needed complementary information. Besides, we observed that, in both fusion schemes, employing different features and modalities led to high accuracy. For instance, tables 4.5 and 4.6 illustrate the performance for feature level fusion on AFEW and eNTERFACE, respectively. We notice that the fusion of visual and audio feature representations increases the performance, for example to 43.3% and 74.6%, for AFEW and eNTERFACE, respectively. This improvement is mainly due to the complementarity of the two channels.

Furthermore, as we aim to investigate the advantages of a hierarchical fusion scheme, we apply the information gain theory to minimize the symmetric KL-divergence. This approach aims at deciding the best pairs of feature representations (coming from audio and video modalities) to be combined in feature level fusion. In Tables 4.5 and 4.6, the three best pairs of feature representations are included, for AFEW and eNTERFACE. These tables show the KL-divergence values and the obtained accuracies for both datasets. We notice the increase in the performance over the unimodal results in Table 4.4 in all cases, proving the benefits of both feature-level fusion and the Fisher vector representation. Interestingly, audio is present in all the feature representations' pairs of eNTERFACE, in addition to CNN, ROIs-DSIFT, and geometric features. However, for AFEW, the best pairs are obtained by concatenating the FV of the following feature types: (i) ROIs-DSIFT and geometric features, (ii) audio and DSIFT features, and (iii) geometric and DSIFT features.

Score Level Fusion: Following the feature-level fusion step, we fed the obtained prediction scores from pairs of feature representations into late level fusion. Late fusion searches for the best weights to fuse them. As emotions are more dominant depending on the audio or visual modalities, score level fusion aims to breakdown the fusion into this level, where we assign weights per-feature representation/modality and per-emotion. For achieving this purpose, we employ two approaches: (i) firstly, we use as weights of each feature representation the diagonal elements of the confusion matrix; (ii) the second technique uses GA for searching the best weights to fuse the given paired feature representations. In the first case of using the performance-based weights, the overall accuracy for eNTERFACE and AFEW is 44.4% and 77.0%, respectively. However, in the second case, we apply a genetic algorithm as an optimization search algorithm, using 5-fold cross-validation. Figure 4.8 gives an example of the score level fusion approach together with the weights per-feature representation and per-emotion in the best performing case for the AFEW dataset. The GA-optimized search resulted in enhanced

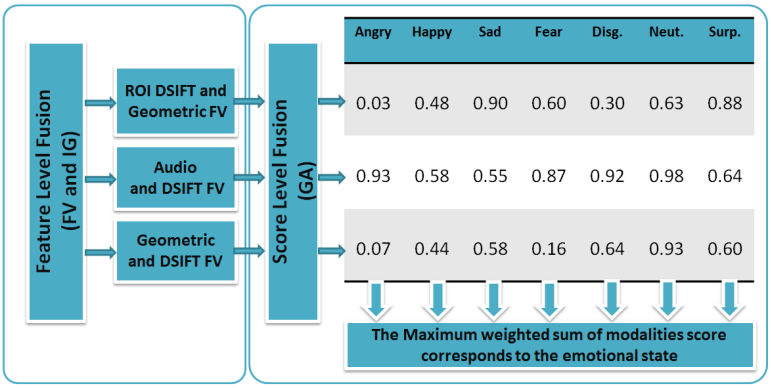


Figure 4.8: The resulting modalities and feature representations from feature level fusion by FV and IG, and the weights per-modality and per emotion obtained by score level fusion using GA.

Table 4.8: Performance of different methods on the AFEW validation set and eNTERFACE

Dataset	Approach	Accuracy %
eNTERFACE	Feature level fusion [190]	71.0
	Score-level bimodal SVM [188]	87.4
	Late fusion on C3D-DBN[191]	**89.4**
	Hierarchical early and late fusion on audio-visual representations (ours)	78.5
AFEW	AFEW Baseline [117]	38.8
	Audio + C3D [152]	52.0
	Late fusion on audio-CNN-DSIFT [192]	51.2
	Audio + C3D + ResNet-LSTM [193]	**53.9**
	Hierarchical early and late fusion on audio-visual representations (ours)	48.9

performance with an average accuracy of 48.9% and 78.5% for AFEW and eNTERFACE, respectively. The results obtained in both cases are shown in Table 4.7. Compared to the feature level fusion and the performance-based weights late fusion, the genetic algorithm outperformed both approaches, leading to a accurate fusion model.

Comparisons with Other Methods: Finally, we report the performance of other methods on both eNTERFACE and the AFEW validation set in Table 4.8. For example, on eNTERFACE, our proposed system achieves better results when compared to the approach in [190]. Nonetheless, it underperforms in comparison to the state-of-the-art results (89.4%) in [191] and the results reported in [188]. For instance, Nguyen *et al.* [191] employed end-to-end DNNs, namely, 3D CNN (C3D) cascaded with deep belief networks (DBN). Similarly, for the AFEW dataset, the work described in [152] achieves a better recognition rate than what we obtained by employing pretraining for fine-tuning 3D-CNNs' features. It is important to mention that, in our approach, we only used the training set available in the AFEW dataset, limited to 773 video samples. The state-of-the-art results on AFEW are usually obtained through approaches that rely, to a signif-

icant extent, on extensive pretraining and ensembling of multiple deep learning methods. These approaches require unmatched computational power. An example is a work reported in [193], where authors employed several visual representations through different deep learning models such as Residual Neural Networks (ResNet), Long-Short-Term-Memory (LSTM) with CNN, and 3-D CNN (C3D), to perform later fusion with audio features. The late fusion weights were assigned manually based on the researchers' observations on the performance of each representation. We think that hierarchical fusion is an intuitive approach; however, it underperforms in comparison with existing deep learning methods, which rely on massively annotated data and techniques such as transfer learning. The following chapters overcome these limitations by employing progressive fusion approaches to enhance the bimodal recognition of emotions using the best pairs of audio-visual representations, namely CNN and audio representations.

4.1.5. DISCUSSION

In this research, we proposed a framework for multimodal hierarchical emotion recognition, tested on the AFEW'16 and the eNTERFACE datasets. We employed a Fisher vector representation for capturing the discriminative and temporal information across the frames in each video sample. This encoding was applied on different types of features (e.g. DSIFT, geometric, CNN, audio). It enables mapping the features into a common space, where early fusion is performed. Next, we also applied a decision-level fusion approach on top of the best feature and modality combinations obtained through feature-level fusion, surpassing the baselines and the unimodal performances. In this study, we show how robust video modeling and feature encoding is crucial for performance. Besides, the fusion schemes, either the early or the late fusions, improved the recognition rates. The early fusion is important as it brings the features scores closer to the true labels, which makes it possible for the late fusion to perform an efficient re-weighing procedure. We optimized the features and modalities weights for each emotional state using a genetic search algorithm, which leads to an overall accuracy of 48.9% and 78.5% for AFEW and eNTERFACE, respectively.

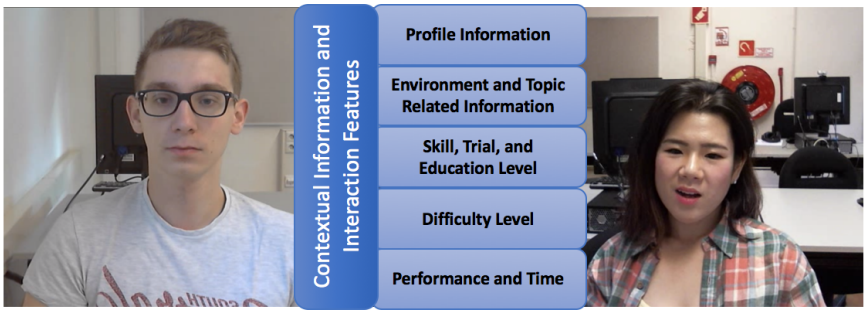


Figure 4.9: Participants interacting with the learning game.

4.2. TOWARDS AFFECT RECOGNITION THROUGH INTERACTIONS WITH LEARNING MATERIALS

EMOTIONS are multilayered subjective affective states that can vary according to many factors, such as personal and environmental context, mood, personality, and culture. Besides, demographic data, personal characteristics, cultural background, and context information could enable data-driven methods to achieve technical breakthroughs with accurate performance in emotion recognition. In this section, we study affective state recognition, where we include context information related to students' interaction with learning materials, besides capturing their facial expressions (as shown in Figure 4.9). Students' affective states are reported through self-assessment. Self-assessment has been proven to be beneficial in reporting emotions when collecting context-related information [6]. In particular, within this research, we address affective state recognition utilizing contextual information and interaction parameters as features, despite the fact that this input is usually viewed solely through the prism of performance. Our motivation behind this research was to incorporate visual information during students' key interaction moments with learning materials, which would enable multimodal affect recognition. Nonetheless, a study by [194], using the video clips of the students while interacting with the adopted serious game, concluded that the correlation between students' facial expressions and their self-reported affective states was low. As a result, in this section, we focus on measuring students' affective states through their interaction features with learning materials.

The automatic recognition of affective states could enable Technology Enhanced Learning (TEL) with personalized learning (as described in Section 3.4.1). Personalized learning is an important topic in education, and, therefore, a significant amount of research is now driven towards personalized interfaces that can motivate and, actually, support the student in interacting with learning materials in a highly engaging manner [195, 196]. While currently there is a huge interest in TEL, such as Massive Open Online Courses (MOOCs), gamification of learning, and blogs, these systems, to a large extent, lack the capability of affect recognition which is a vital step towards achieving personalization. Conversely, in traditional education, teachers are capable of factoring the students' affective states in their presentation and feedback strategies. Similarly, TEL must be able to understand the affective states of the users.

Goleman in [165, 166] stated that the brain operates to either boost or lower our performance in different domains of abilities such as learning and work. More specifically, he analyzed the relationship between one's affective state and memory capacity. His research suggests that in educational contexts, stress, anxiety, and frustration experienced by a learner directly affect the learning outcome negatively. This challenge can be tackled through various Machine Learning (ML) techniques [195]. ML's role relies on providing a timely and an appropriate guidance of students' cognitive states where high-level affective content can be predicted from low-level human-centered signals, using different sources such as depth and RGB cameras, inertial sensors, speech signals or log-data [197–201]. Along with these extensively utilized signals, log-data and contextual information of the learner while interacting with TEL can provide an essential channel for affect recognition due to its wide availability and relevance in digital learning materials [202, 203].

In this study, we demonstrate the feasibility of identifying student's affective state such as boredom, engagement and frustration in a learning environment, inspired by the Theory of Flow [204, 205], which is a well-known and grounded paradigm of how affect is modeled during interacting with learning and gamified applications [204, 206–211]. Towards this aim, the main goal of this work is to model and design a learning activity-based tracker, utilizing digital interaction parameters and learning analytics information related to performance, difficulty level, the complexity of the learning activity, number of attempts to perform a task, and time needed to accomplish this task. Figure 4.9 displays two students playing the learning game we use in our study, while the system is tracking their interaction features and is utilizing machine learning techniques to predict their affective state.

Contributions. The contribution of this work can be summarized as follows: We propose an affect recognition approach, by developing a generic framework that adopts interaction tracking for affect recognition to be used in a variety of scenarios through utilizing a leading standard in learning analytics, the Experience API (xAPI) (Subsection 4.2.2). Through this framework, we are able to describe a series of parameters according to user-platform interaction, related to performance, the time needed to accomplish a goal, and level of difficulty of the task in hand. The benefit of the proposed approach relies upon its suitability to be adopted in various setups and devices, where specific sources of information are not available, such as depth or RGB information. Secondly, we use the model of the Theory of Flow [205] (Subsection 4.2.3) as a model for affect inference in the learning environment. This model has three different states (namely, boredom, engagement, and frustration). They were obtained through self-annotation in a real operating environment of students interacting with an e-learning platform (Subsection 4.2.2). Finally, we present an extensive analysis of our case-study research (Subsection 4.2.4) and showcase the applicability of our model cross-subjects and per-subject for inferring affect during learning. Summarizing, we introduce in this chapter an application of machine learning algorithms for predicting the user's affective state based on interaction parameters contributing to the education domain and specifically supporting TEL.

4.2.1. RELATED WORK

Affect recognition is a challenging problem due to its multimodal nature and the variability of affect expressivity across different individuals. These factors make sensing, modeling and recognizing affects a hard task [212]. This becomes even more evident in educational settings, where the research community has proposed various approaches for affect recognition, employing, most of the times, silo approaches that often make use of log-files [202, 203] and ambient sensing techniques [199–201, 213]. Unfortunately, these studies concluded with results that are not easily applicable on a satisfactorily wide range of educational settings.

The idea of correlating human behaviors with features related to interaction and context has been researched in many problems. For example, these studies aim at interpreting the attention level of a user towards a certain goal using ML algorithms. For example, in educational settings, the authors in [202] studied affect prediction such as confusion, boredom and engagement through log-data of students in a math tutoring system. Their study shows a negative correlation between boredom and performance during problem-solving while concentration and frustration are associated with positive learning. Similarly, authors in [203] used solely students' interaction with a digital agent teaching algebra for affect recognition. Their work showed a performance which is higher than a random classifier at identifying engaged concentration, boredom, confusion and frustration. Moreover, many previous works have dealt with studying affect in gamification [214, 215], where the link between cognition and affect is the focus of attention.

In addition, facial expressions have been studied in TEL as an affect informative channel [199, 200]. However, many of these studies have focused on posed affect and did not take into account the nature of the classroom environment or the type of education provided, such as MOOCs, distance learning and digital learning. For example, in [201], Kapoor et al. studied affect using multimodal sensory information from the face, postural shifts and learners' activity on the computer. Their approach employed Mixture of Gaussian Processes for data fusion. Additionally, several studies have been using speech as an input for affect recognition. In [197], authors proposed a framework for online feedback based on the learner's vocal affect expression. Asteriadis et al. in [213] analyzed the case of using head movements and gaze patterns towards learning material to study the mapping between these cues and learners' affective states. This would eventually lead to adaptation of the learning process towards more personalized learning.

As detailed in Chapter 3 (Subsection 3.2.1), physiological signals are also popular modalities in affect recognition [118]. However, since most of the related approaches rely on specialized hardware, which is often expensive and hard to calibrate, they may result in cumbersome settings and hampered spontaneous interaction [28]. On the other hand, recently, there have been a lot of studies that validated the potential of making use of inertial sensors, such as gyroscopes and accelerometers for recognizing affective states. As described in [167, 216], these sensors, which are embedded in most of the mobile devices nowadays, can be used for computing descriptors that contribute to tracking affects while interacting with a mobile device.

Table 4.9: The xAPI framework's attributes.

Field's type	Attributes
Mandatory fields	Actor, Verb and Object
Optional fields	Result, Context, Timestamp (internal recording of timestamp), Authority, Version and Attachments

4.2.2. DATA COLLECTION

In TEL, there are different approaches proposed for data modeling and analysis with regards to affect recognition. One large family of approaches mainly deals with low-level data channels such as facial expressions, keystrokes, mouse events or gestures [199, 201, 213]. However, as mentioned in the related work discussion, approaches which focus on high-level features, for example, learner's activities, digital interactions and so forth, have not been extensively studied in the field of affect analysis.

Therefore, the focus of this research is placed on the latter family of techniques, which targets the high-level interaction features. In learning analytics, currently, there are different methods of standardization. One of the most recent specifications applied in learning contexts is the Experience API (xAPI) which focuses on high-level activities [217, 218]. We gathered data using the xAPI standard through an interaction with an e-learning platform, developed for the scope of this research. This section details the developed game, the way the xAPI tracker is used, the necessary steps and instructions for the users, the statistics of the collected dataset, and the affective states model used.

INTERACTION TRACKING

xAPI is a leading framework and an emerging standard for tracking and storing educational data [217]. The xAPI specification is two-fold, as it can both focus on the syntax of the data and, at the same time, define the characteristics of a Learning Record Storage (LRS), which serves as the end-point collecting and exchanging learning analytics traces [219], such as learning content, and learner's activities and performance. When an activity is recorded, xAPI sends statements to the LRS. These experience statements are the core of xAPI, detailing the trajectories of learning activities for learning analytics, and providing the digital interactions for affect recognition in this study. The xAPI data format defines the experience statements with the attributes[1] as listed in Table 4.9.

The rationale behind choosing this standard to track the interactions of the learner with learning materials is mainly due to the following reasons: (1) xAPI is event-centered, covering a large range of actions, as the field verb in xAPI statements can define arbitrary actions; (2) it is designed for optimized interoperability in different educational systems; (3) xAPI has many adopters [220] and it is also supported by commercial Learning Record Store services such as Learning Locker [221]; and (4) the xAPI framework can be easily extended and integrated into new software components in a learning platform, allowing the adaptation of the system driven by the digitally tracked interactions in combination with ML techniques.

[1] https://experienceapi.com/statement-design/

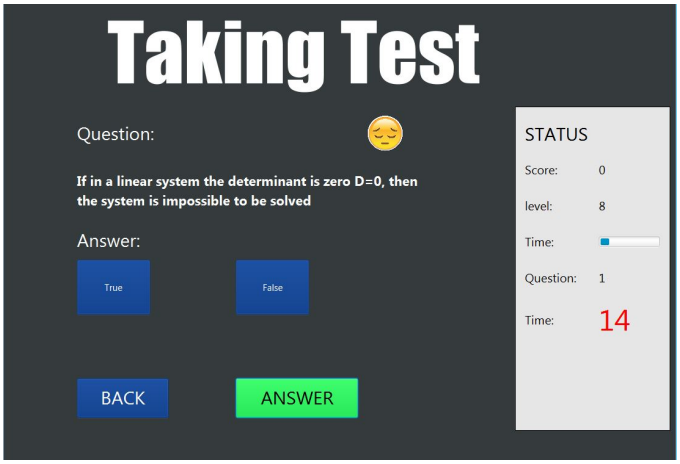


Figure 4.10: A screen-shot of a question example from the learning game.

The standard is used in our framework as follows: "Actor" defines who is the subject that performs an activity, either a learner, a tutor, or the device being used. Field "Verb" describes an action taken by the actor (e.g. passed a question). "Object" represents the subject of the actor's action (e.g. question). In this work, "Timestamp" provides essential information that allows the system to get historical information about all the actions performed and the exact moment when they occurred while it also enables the calculation of meaningful information such as the time required to perform specific activities. The "Result" element gives the outcome of the experience (such as parameters regarding success, completion or score during the activity).

GAME DESCRIPTION

A serious game has been utilized for the scope of this research and was part of study for personalized learning experience in [222, 223] by C. Athanasiadis et al., consisting of a learning platform which, according to its complexity and levels of difficulty of its materials, is expected to generate different affect states to the student, who will then report them at the end of each session. Figure 4.10 shows a screen-shot of this game, where the students can perform a question-based learning interaction with the platform. The game contains three types of questions, namely: "choose an answer", "true or false" and "fill in the blank", where students are required to enter the whole answer. For collecting as much data as possible, a database that includes a total of 800 questions has been developed, covering four major topics: Mathematics, History, Sports and Geography. Moreover, the questions used have a varying difficulty level from 1 to 9 (in ascending order). Further details about the game flow and how the students can use it are listed below:

- A student logs in after creating an account, which includes her/his demographics information, competence (skills) on each topic presented in the game and her/his education level.

- When ready, the student is directed towards an interface where she/he can choose one of the following topics: mathematics, history, sports, and geography. Subsequently, a new learning session starts.

- In each learning session, the student is asked consecutively seven questions related to the chosen topic (belonging to one of the aforementioned three types of questions).

- After answering each question, the student is able to see whether her/his answer was correct or not.

- At the end of each session, the student is asked to provide self-annotations regarding the experienced affective state, as boredom, engagement or frustration. Specifically, students reported their affective states on a Likert scale (from zero to five), the degree of engagement, frustration, and boredom they experienced. In our study, for each session, we select the affective state with the highest rating (obtained from the Likert Scale) as a label for the corresponding session.

- Next, the platform provides the student with her/his score in the topic so far, and a comparison with the results of other students.

- The same procedure can be followed up to four different trials for each topic, while the level of difficulty increases after each session. Finally, the student has the choice of either continuing playing the game or logging out.

Regarding the modification of the level of difficulty, initially, the first level of a topic was selected randomly among the lowest three ones, while the subsequent levels were increased according to the student's performance. It is important to note that users were given oral instructions about the game process they had to follow, the number of sessions they should complete and regarding self-annotation of the affective states. Their self-annotation has been used as the ground truth label of their interaction in a learning session.

xAPI STATEMENT DESCRIPTION

Each of the steps mentioned above is associated with an xAPI statement for recording students' activities and gathering their interaction parameters. In order to achieve this goal, a set of statements has been defined. The set which has been applied in this study is listed in Table 4.10.

Using these pre-defined statements, it is possible to track, for example, when a student starts a new session (triggering the "initialized" verb), when a question is presented ("Game asked a question") or when a learner responds correctly/wrongly (Learner "passed" or "failed" a question'). To track when a learner does not provide an answer or skips a question, the "completion" field of the "Result" element is sent as *False* to indicate this event. This set of statements can be easily expanded to monitor a broader range of interactions.

Table 4.10: The xAPI statements used to track features while interacting with the serious game.

Actor	Verb	Object	Related activities and features
Tutor/game	initialized	an interaction	Timer is triggered and student's profile information is collected such as skill level, study background, and age.
Learner	passed/failed	a question	Performance and question-related data are streamed, for instance, the time needed to answer a question, score attained, the difficulty level of the question, the topic of the question, and the number of attempts in a session (i.e. number of the performed trials in a topic, e.g., math and history).
Game	asked	a question	A learner is asked a question related to the session topic.
Learner	terminated	an interaction	Self-annotation of affective state is asked, and the summary of the interaction details is stored.

4.2.3. AFFECTIVE STATES MODELLING

Previous studies used various models in affect recognition, such as the discrete model, proposed by Ekman [224] or dimensional ones that model affects on the valence and arousal dimensions [225]. In [224], Ekman described a theory that addresses the basic emotions (namely: anger, contempt, disgust, fear, interest, happiness, sadness and surprise). This model is widely used in affect recognition. However, in [226], authors suggested the use of a simplified model that students mostly use to express their state, interest, distress (frustration) or pleasure (enjoyment) rather than the general models of basic affects. Another model based on Yerken-Dodson Law, the so-called "sweet spot for achievement [166]", describes affect through three major states, namely disengagement, frazzle, and flow. These states have an impact on a person's performance whether at work or during learning activities.

One of the models that has been proposed for modelling affect is the Theory of Flow (ToF), developed by Csíkszentmihályi in [204, 205]. Flow is the mental state experienced when a person conducts an activity in a state of deep involvement, in a feeling of energized focus, maximum engagement and enjoyment in the process [205]. ToF has been shown to be a very effective tool for measuring affect in the learning procedure, as it encapsulates the way learners perceive the learning content as a function of their skills and the challenge imposed by the learning materials [206–208]. In the research reported in [208], the authors studied students' engagement using longitudinal data of 526 schools in U.S.. This work investigated the conditions of adolescents under which they reported their engagement. The research stated the increase of engagement when

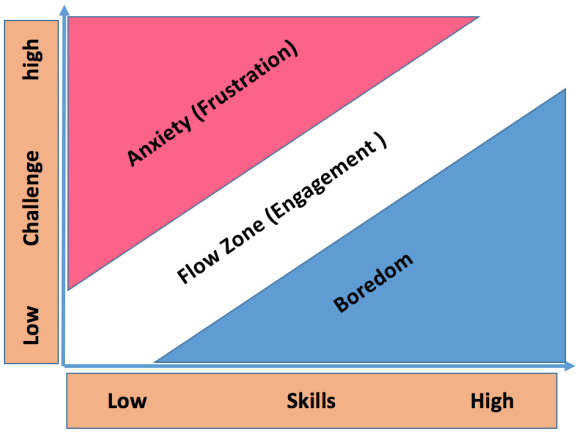


Figure 4.11: The model of the Theory of Flow consisting of the three affective states relevant to learning experience as defined in [204].

there is a balance between the students' skill and the challenge of the given task. Moreover, the work in [207] examined students' affective responses in distance learning by studying the cause and effect of flow experience and the impact of interaction on it. In addition, ToF has been widely applied in serious games [227, 228] and this theory has been extensively used and investigated in the domain of computational intelligence and the video game industry [209–211]. Therefore, ToF is presented as a convenient affect model in serious games, since it includes important requirements, such as: clear goals, clear challenge-skill balance, immediate feedback, concentration, timelessness, and the ability of experiencing temporary loss of the feeling of self-consciousness.

Motivated by ToF and its use in both education and serious games, and due to the gamified character of our learning application, we adopted the basic model of ToF as defined in early work of Csíkszentmihályi in [204] and described in [209]. Figure 4.11 presents the details of this model, which consists of three different zones, namely, boredom, flow (engagement) and anxiety (frustration). These states are the most relevant to the users' skills and the amount of difficulty imposed by the learning application. In this model, for a certain user, keeping the balance between skill level and challenge dimensions leads to a positive effect, while a wrong selection can provoke either learners' frustration or boredom. The ground truth utilized in the collected data is based on the self-annotation of the experienced affective state by the participants.

DATASET STATISTICS

During data collection, diversity in participants' population has been carefully considered in order to collect data from interactions of users with varying profiles in terms of age, education level and gender. Specifically, a dataset of 32 subjects has been collected where 18 females and 14 males voluntarily participated in the process. The users were either bachelor or masters students (12 master and 20 bachelor) with an average age of 22.40 ± 3.13. The participants can be grouped in two different knowledge profiles: 22 students from the Department of Data Science and Knowledge Engineering (DKE), and

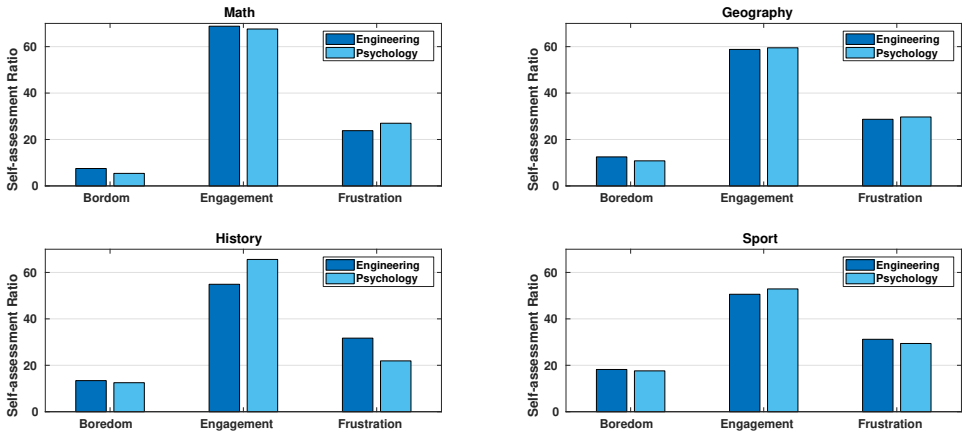


Figure 4.12: Percentage of the self-assessment (its histograms) among the students in two faculties for the game's topics.

10 are from the Faculty of Psychology and Neuroscience (FPN) at Maastricht University. The total duration of each experiment per participant lasted an average of 26 ± 5 minutes. Each student was asked to perform four learning sessions on the available four topics. However, some students performed less than four sessions leading to a total number of 459 recorded sessions, instead of 512. Data gathering was conducted in adjusted environments like classrooms.

4.2.4. RESULTS

This section presents statistical data related to the gathered sessions and the interaction features extracted and modeled according to the xAPI standard stream. The approaches applied to evaluate the data (cross-subjects and subjects-based) and the implications of the results are discussed. Various classification methods had been applied, including Adaboost, Random Forest classifier, while the best results, which are reported next were obtained using radial basis Support Vector Machines (SVM).

SESSIONS STATISTICS

The collected 459 sessions were labeled as follows: 57 sessions as boredom, 272 sessions as engagement and 130 sessions as frustration using the provided self-annotations by the users. As we intentionally collected data from two different faculties (Engineering and Psychology), Figure 4.12 presents the histogram of each affective state for the collected sessions, in each topic of the game among the students of both faculties independently. For example, the second column shows the ratio of the reported self-annotation in mathematics among engineering students. Figure 4.12 shows comparable distributions of the states among both profiles, in terms of psychology and engineering. Extensive statistical analysis on the validity of the collected data, the distributions of the self-reported affective states among demographic and educational levels are reported in the study in [222, 223]. For example, the study in [223] found that students with an engineering background performed and concentrated better in mathematics topics than

Table 4.11: Feature vector per session

Performance features	Sessions' summary features	Profile information
• Score attained and time needed to answer each question	• Ratio of correct, wrong and skipped answers. • Difficulty level of the session • Duration of the session • Number of trial	• Skill level in the session's topic, • Level of education, • Age • Gender

students with a background in psychology. The disparity between students' grades with different backgrounds was validated through Welch's test, where the null hypothesis for distribution equality was rejected with p = 0.05. Moreover, this observation is supported by the fact that the average grades for engineering/ psychology students were 0.62/0.53, respectively. Furthermore, another interesting observation is the difference between the self-assessment (in terms of affective states) of males and females of boredom on topics related to sports. Female students showed a higher tendency to annotate their affective state as boredom than male students. Overall, the diversity of the students contributed to the validity and the richness of the collected data. This research focuses on using the collected data to study the self-reported affective states for students' emotion recognition.

FEATURE VECTORS

Table 4.11 details the constructed feature vector for each session according to each student's performance in it, the session information and the profile information provided by the students. Since a session consists of seven questions, there are 14 features related to the obtained scores and the time needed to answer each question. In addition, each session contains questions related to the same difficulty level and a trial value which represents the student's stage in the game for a given topic. We also include the session duration, and a summary of the results in it through computing the ratio of correct, wrong and skipped answers. In the third column, four features regarding the student's profile are included, where skill level is based on the students' report of their level of knowledge in each topic presented in the game (graded in a scale ranging between 0 and 10), while the level of education can be either bachelor or master. Therefore, the resulting feature vector's length is 24. These features provide overall descriptors, while our aim is to associate them with the learner's affective state. Feature vectors were normalized by

Table 4.12: Precision, recall and f1-score results obtained by SVM using the two validation schemes. Note that the standard deviation (std) on cross-subject validation is higher than the 10-fold cross-validations. The disparity between the two validations' std is expected. Cross-subject evaluation is more challenging than the 10-fold based cross-validation due to the inter subjects variability and the high number of subjects in our study (i.e., 32 subjects). We report the standard error of the mean (sem) to give a direct comparison between the two validation approaches.

Measure	cross-subject validation			10-fold validation		
	Mean value	std	sem	Mean value	std	sem
Precision	66.6 %	±14.6	2.6	67.7%	±3.4	1.1
Recall	59.3 %	±14.0	2.5	59.5%	±4.9	1.6
F1 score	61.4 %	±13.3	2.4	61.2%	±4.1	1.4

subtracting the mean and dividing them by their standard deviation.

CLASSIFICATION AND EVALUATION SCHEMES

We trained radial basis Support Vector Machine (SVM)[2] classifiers making use of a one-vs-all strategy, in order to fit a classifier for each class against all other classes. This gives us the advantage of inspecting each class and its corresponding classifier. For instance, a separate classifier has been trained to infer frustration against both boredom and engagement classes. A similar approach was followed for boredom and engagement. During the testing phase, the three classifiers were fit on the test samples. Consequently, for a test sample, a classifier with maximal value was selected as a predictor of the affective state corresponding to a given session. We introduce next two evaluation strategies, ***cross-subject*** for testing the generality of our approach across students with different profiles, personalities and knowledge base and ***subject-based*** for enabling an adaptive learning according to user's personality, skills, and overall profile, capable of detecting changes in the affective state of the user and proposing a way to improve his/her productivity and learning experience.

CROSS-SUBJECT EVALUATION

The cross-subject evaluation assesses the system performance in predicting the self-reported affective states, using cross-validation. As we aim at testing the system applicability to a new subject, we employ the leave-one-out cross-validation scheme (LOOCV), by dividing the data according to the users where, in each fold, we exclude one users' data for testing while the rest of the users' data is used for training. Furthermore, we also evaluate the system using 10-fold cross-validation, by dividing the samples of each class into ten partitions.

The performance on the collected data is reported in terms of precision, recall, and f1 score. The second and the third column in Table 4.12 display the results of the two evaluation schemes on the collected data. In LOOCV, the classifier achieves good results obtaining a precision of 66%. Although the expectation is that experienced affective

[2]SVM implementation provided by scikit-learn [229] is used. Scikit-learn is a machine learning library for Python programming language. Soft margin cost function parameter C and the regulation parameter gamma are set to 100 and 0.1, respectively.

Table 4.13: Confusion matrix obtained with the subject-based Leave-one-out cross-validation (LOOCV).

States	Boredom	Engagement	Frustration
Boredom	29.8%	31.6%	38.6%
Engagement	16.2%	62.1%	21.7%
Frustration	23.8%	11.5%	64.6%

Table 4.14: The classification performance is reported on the collected data using engagement and non-engagement labels through subject-based cross-validation. Subject-based evaluation is obtained using Leave-one-out cross-validation (LOOCV).

Measure	**Mean Value**	**standard deviation (std)**	**standard error of the mean (sem)**
Precision	75.4%	±11.4	2.0
Recall	74.9%	±13.4	2.3
F1 score	74.3%	±11.5	2.0

states and learning experience are person-specific notions, the achieved performance indicates the potential of the affect recognition system in TEL based on interaction parameters. Likewise, the results of the 10-folds evaluation in the third column are promising. It is important to note that the standard deviations of the performance measures with cross-subject validation are higher than the 10-fold cross-validations. The disparity between the two validations' stds is expected due to the inter subjects variability and the high number of subjects in our study (i.e., 32 subjects). The subjects have different education levels with varied context information, resulting in unique educational and affective experiences. For this reason, we report the standard error of the mean (sem) to give a direct comparison between the two validation approaches, where we notice that the disparity between the two validation approaches is reduced.

The performance, in this case, is similar to the LOOCV which supports the generality of the framework for the prediction of the self-reported affective states in the digital learning environment. It is important to note that the standard deviations of the performance measures on cross-subject validation are higher than the 10-fold cross-validations. The disparity between the two validations' std is expected due to the inter subjects variability and the high number of subjects in our study (i.e., 32 subjects). For this reason, we report the standard error of the mean (sem) to give a direct comparison between the two validation approaches, where we notice that the disparity between the two validation approaches is reduced.

Table 4.13 details the confusion matrix of the three states obtained from the LOOCV. As expected, the accuracy of engagement and frustration is higher compared to boredom. Indeed, most of the participants in the data collection reported engagement or frustration while only 50% of them reported the boredom state in any of their self-annotation. Additionally, this result can be due to the fact that the boredom state was harder to experience due to the dynamic nature of the game and the possibility to leave the session anytime.

Engagement vs Non-engagement Evaluation: A further analysis was applied on the

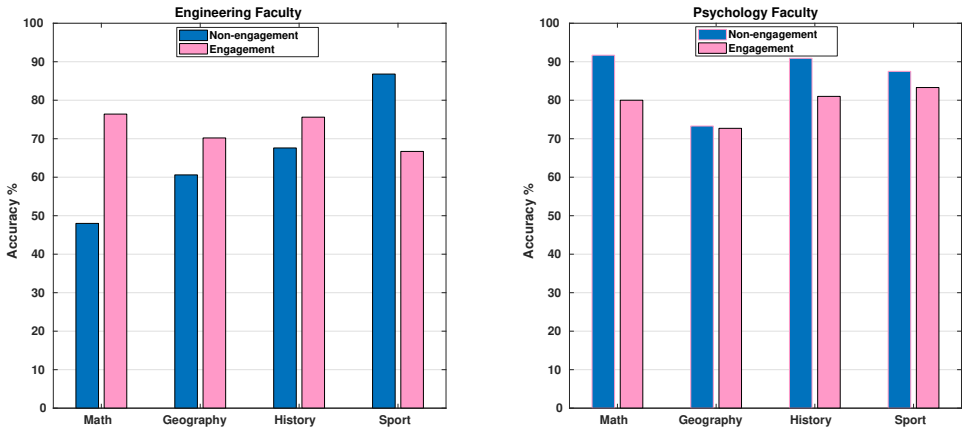


Figure 4.13: The accuracy of the engagement and non-engagement detection among the students in two faculties and different topics.

collected data for investigating the performance of the classification when considering the scenario of engagement versus non-engagement states. In this strategy, boredom and frustration sessions were assumed to represent the non-flow zone in the adopted affect model, as shown in Figure 4.11. This simplification can assist in cases when the focus is on detecting the non-engagement states to adapt the learning experience according to learner's competence and by prompting the right level of difficulty. Table 4.14 illustrates the results of this scenario using LOOCV. Recall and precision have high values, fact which proves the ability of the proposed approach to distinguish between the states of engagement and non-engagement.

Figure 4.13 gives a closer look at the results of this scenario among the two different faculties and the topics used in the game. When observing the obtained results of each faculty independently, we notice a higher detection accuracy of engagement than non-engagement for engineering students with an exception in sports, whereas for psychology students, in all topics, the detection of non-engagement is relatively higher than engagement. Indeed, the overall performance of engineering students was higher than psychology students which might explain their engagement in playing the game, while for psychology students the non-engagement can be due to the perceived challenge and their lower interest towards the game topics. These observations and results are consistent with the Theory of Flow model. The obtained system accuracy shows how the game's topic, challenge level, and students' educational background can contribute to affective learning technologies.

SUBJECT-BASED EVALUATION

Both affect and personality traits contribute to the learning process. In addition, the difficulty level of the presented learning material can be perceived individually according to user's competence, which can also influence her/his affective state. For this reason, a subject-based analysis is performed to study the ability of the system to predict the self-reported affective states, based on interaction features. This evaluation was applied

Table 4.15: Confusion matrix obtained with the subject-based evaluation.

States	Boredom	Engagement	Frustration
Boredom	61.4%	26.3%	12.3%
Engagement	8.5%	79.4%	12.1%
Frustration	7.7%	26.9%	65.4%

for each user independently, by employing each session once for testing and the rest of the user sessions for training. Our evaluations showed a precision of 74%, which is supported by the confusion matrix presented in Table 4.15 where a good accuracy detection is achieved for each state in the adopted affect model. It is important to mention, that the system was able to detect sudden changes of affective states (e.g. frustration after several engagement sessions or vice versa). The achieved results highlight the benefit of employing interaction features in customized and personalized learning activities according to the user's affective state.

4.2.5. DISCUSSION

In this study, we presented a framework for affect recognition in TEL, based on contextual information and the tracking of interaction features during learning activities. For this purpose, we utilized a standard tool, the xAPI framework, for learning analytics and high-level activities tracking. We have developed a serious game, as a learning platform for data collection and evaluation of the proposed framework. We adopted the Theory of Flow as an affect model, used by the students to self-report their experienced affective state during the interaction with the platform. We employed two evaluation strategies in combination with machine learning algorithms. The cross-subjects analysis highlighted the generalization ability of our system across students with different profiles (namely engineering and psychology), obtaining a good precision of 67% for three affective states and 75% when considering engagement vs. non-engagement states. Furthermore, the subject-based evaluation, in particular, is useful in boosting the adaptive nature of the learning process to enhance the outcome of the learning experience, maximizing the learners' knowledge acquisition and enabling personalization. The obtained results are promising in this direction, with a precision of 74% for the recognition of the three affective states.

4.3. CONCLUSIONS

This chapter studied affect recognition in two approaches, the first one focused on the identification of discrete emotions using various representations of audio and visual cues. This procedure introduced an extensive study of audio and video modalities, the importance of aggregating their temporal information through the Fisher Vectors (FVs), and the subsequent fusion of the two modalities' features. We noticed the advantages of our multimodal learning scheme for combining the feature level and the score level fusion in a hierarchical manner based on the Information Gain (IG) principles and the Genetic Algorithm (GA) optimization. Additionally, score level fusion has the advantage

of re-weighing existing modalities to benefit from their individual expertise and performance on specific emotions. The following chapters focus on two directions, by exploiting more advanced fusion algorithms such as metric learning, and developing end-to-end learning paradigms for audio and visual representations, which is the focus of the rest of this dissertation.

The second approach targeted affect recognition within educational settings. In particular, it proposed a framework for tracking students' interaction with learning materials to identify their affects using the Theory of Flow (ToF) emotional model. The results of affect recognition within the proposed framework indicate the potential usage of interaction parameters with learning materials as a useful channel for measuring students' affective states. One possible research direction of this study was to incorporate the visual information during the key interaction moments for enabling multi-modal affect recognition. However, a study by [194], using the video clips of the students while interacting with the adopted serious game, concluded that the correlation between students' facial expressions and their self-reported affective states was quite low. As a result, in this dissertation, the visual modality was not part of this study.

Overall, this chapter demonstrates how emotion recognition depends on emotion annotation, the perceived modalities, modalities' data representations, and computational modeling. As shown in the second study of the chapter, these steps might vary greatly according to emotion recognition applications. We also showed that having robust modeling and representation impacts automatic emotion recognition, where the performance can differ according to the audio-visual representations. Nonetheless, progressive fusion techniques for audio-visual representations are essential to improve their performance. The following chapters overcome the limitations of this chapter by focusing on the audio-visual cues and employing progressive fusion approaches to enhance the bimodal recognition of emotions. Specifically, the following chapter aims at producing more robust representations and efficient solutions for audio-visual representations.

5

METRIC LEARNING BASED MULTIMODAL AUDIO-VISUAL EMOTION RECOGNITION

People express their emotions through multiple channels such as visual and audio cues. Consequently, as demonstrated in Chapter 4, automatic emotion recognition can be significantly benefited by multimodal learning. Even-though each modality exhibits unique characteristics, multimodal learning takes advantage of the complementary information of diverse modalities when measuring the same instance, resulting in an enhanced understanding of emotions. Yet, their dependencies and relations are not fully exploited in audio-video emotion recognition. Furthermore, learning an effective metric through multimodality is a crucial goal for many applications in machine learning. Therefore, in this chapter, we propose Multimodal Emotion Recognition Metric Learning (MERML), learned jointly to obtain a discriminative score and a robust representation in a latent space for both modalities. The learned metric is efficiently used through the Radial Basis Function (RBF) based Support Vector Machine (SVM) kernel. The evaluation of our framework shows a significant performance, improving the recognition rates of emotions on different datasets, compared to the first study of the previous chapter.

Parts of this chapter have been published in:

- **E. Ghaleb**, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," IEEE MultiMedia, vol. 27, no. 1, pp. 37–48, 2019.

5.1. INTRODUCTION

THE previous chapter addressed multimodal emotion recognition using feature engineering to obtain various representations and late fusion of individual modalities, utilizing various algorithms. This chapter targets the challenge of multimodal learning and focuses on obtaining robust representations and an efficient fusion algorithm. This work is inspired by the success of similarity learning (which is elaborated in Section 2.2). In this learning scheme, by employing an efficient metric distance, the accuracy of many classification and retrieval problems can be increased [71], as this can contribute to obtaining improved performance and more robust representations. Metric learning approaches learn distances to bring similar inputs closer and dissimilar ones further apart, which are more discriminative than the conventional Euclidean distance. This transformation is done through convex optimization with pairwise constraints [71]. Typical examples include Large Margin Nearest Neighbor (LMNN) [71], Geometric Mean Metric Learning (GMML) [84], and Information Theoretical Metric Learning (ITML) [83].

Motivated by the success of unimodal metric learning and the multimodal nature of emotion recognition, we propose a novel Multimodal Emotion Recognition Metric Learning (MERML) framework, which leverages the audio-visual information to jointly learn modality-specific Mahalanobis metrics. The approach simultaneously optimizes the distance and similarity metrics for audio and video representations, such that they are more discriminative in the learned latent space, contributing to accurate emotion classification.

A substantial benefit of our proposed learning approach relies on its generalization ability. In other words, it can be applied in various multimodal learning domains beyond emotion recognition, for jointly learning and fusing complementary data channels. The contributions of the chapter are summarized as follows:

- We propose MERML for the challenging audio-video emotion recognition task (Section 5.4). We jointly learn the modality-specific metric, that aims to not only capture and explain the complex relationship between these two modalities, but also to efficiently learn a latent space for improved representations and enhanced classification.

- The proposed distance is incorporated efficiently in RBF-based SVM, benefiting the emotion classification task. Besides, the developed MERML is scalable, since it learns modality-specific metrics without imposing any constraint on them, such as the dimension of their features or the number of data samples. Furthermore, the rationale of the proposed distance is intuitive, contributing to the explainability of the model, in terms of modalities' input importance for each emotion. These details are further elaborated in Subsection 5.4.5.

- The efficacy of the proposed method is evaluated extensively on various public benchmarks for audio-video emotion recognition, Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [52], eNTERFACE [116], Acted Facial Expressions in the Wild (AFEW) [117], and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [115]. We provide extensive experimental results, showing a notable performance of our method in comparison with baseline

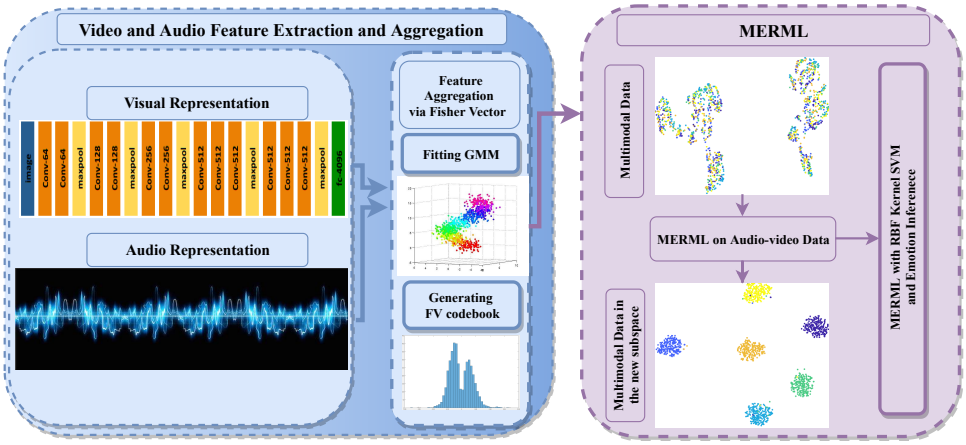


Figure 5.1: The proposed method starts with extracting visual and audio features and aggregates them via Fisher vectors. Then, MERML implementation is applied, with subsequent incorporation through SVM for emotion recognition.

5

metric learning approaches such as LMNN [71], GMML [84], ITML [83] or other methods (Section 5.5).

The proposed method is depicted in Figure 5.1, where, first, we extract audio-visual features and then aggregate them through Fisher vectors to obtain spatio-temporal representations (Section 5.3). Subsequently, the MERML is learned, followed by the audio-video emotional content inference phase, discussed in Section 5.5.

The rest of the chapter is organized as follows. Section 5.2, we briefly discuss related work on audio-video emotion recognition and metric learning. Section 5.3 describes the employed feature extraction and aggregation methods. In Section 5.4, we introduce the formulation and optimization of the proposed MERML framework. In Section 5.5, we present the experimental setup and the obtained results of the proposed algorithm and compare them with the results of other methods. Finally, we conclude our work in Section 5.6.

5.2. RELATED WORK

5.2.1. MULTIMODAL EMOTION RECOGNITION

During the past decades, many researchers investigated human emotional states through different channels and descriptions [66, 173, 201]. In Chapter 3, Section 3.3 provides an overview of the state-of-the-art data representation and fusion. For example, some of these studies utilize public benchmark datasets annotated with either discrete Ekmanian emotions [230] such as AFEW [117] and eNTERFACE [116], or annotated continuously on two dimensions: arousal and valence in the case of RECOLA database [114]. For instance, authors in [157]applied multimodal deep learning to generate joint representations for both modalities. Since the relation between audio and visual information is non-linear, they aimed for cross-modality representations and feature selection. How-

ever, their proposed framework resulted in slightly higher classification accuracy than baseline methods such as Principle Component Analysis (PCA). Moreover, many of these studies focus on concatenating the audio-visual representations or on the late fusion of individual modalities. However, in this work, we propose a joint multimodal emotion recognition metric for audio-visual fusion. We efficiently learn this distance metric by capturing the non-linearity between the high-dimensional modalities, contributing to obtaining an enhanced performance.

5.2.2. METRIC LEARNING

Many studies have implemented metric learning in a unimodal context for a plethora of domains such as multimedia retrieval, computer vision, and machine learning [71, 83, 86, 183, 231]. In [71], Weinberger et al. defined the Large Margin Nearest Neighbor (LMNN) metric under which the *k-nearest* neighbors of a sample belong to the same class. They effectively obtain a Mahalanobis distance metric as the solution to a semi-definite program. In [83], authors proposed the Information Theoretical Metric Learning (ITML) algorithm to minimize the differential relative entropy between two multivariate Gaussians conditioned with a constraint in the distance function. Additionally, many distance metric algorithms have explored other structures such as sparse representations [232, 233] and low-rank constraints [233]. An extensive review of metric learning is given in [87], while in Subsection 2.2.2, the technical background of metric learning techniques, in general, is provided.

Multimodal metric learning has also shown success in miscellaneous domains utilizing various representations of visual data [231, 234, 235], as well as other data channels such as text [236]. A simple way to learn metric for multimodal data can be done by concatenating the features of the modalities and then applying classical metric learning such as LMNN or ITML [237]. For example, in [237], the authors used LMNN metric for facial expression recognition by applying it on the concatenated set of visual representations of facial images. That said, this approach lacks the potential of exploiting metric learning for a multimodal problem.

In [235], the authors applied multi-metric learning (L3LM) for face and kinship verification, aiming for maximum correlation for various representations. Similarly, in [238], Xie et al. developed a metric aiming to embed data of various modalities into a single latent subspace. The parameters of the developed metric are learned by optimizing the data likelihood under the latent subspace model. Authors in [236] use heterogeneous metric learning for cross-media retrieval. Recently, in [234], Zhang et al. applied hierarchical multimodal metric learning, in which, modality-specific metric learning is used with a general one to map the representations to a common space. Even though learning a shared metric could be useful, one limitation of the proposed approach in [234] is the constraint of having modalities with the same dimensions, and the concatenation of the projected features in the classification step. This concatenation might affect the potential discriminative power of learned metric and result in an expensive computation.

In this chapter, we propose a scalable and flexible metric learning approach for emotion recognition, namely, the Multimodal Emotion Recognition Metric Learning (MERML), developed for the challenging audio-video emotion recognition task. We jointly learn a modality-specific metric that aims to not only capture and explain the

complex relationship between these two modalities but also to efficiently learn a latent subspace for improved representations and enhanced classification. An additional advantage of the proposed metric is a heterogeneous emotion inference. In other words, even if only a single modality exists during an emotion query, we can still benefit from the jointly learned metric in the prediction phase. Finally, a non-negative weighing scheme is applied in terms of the contribution of each modality. This strategy is particularly beneficial as it aims to exploit the nature of our task, where visual and audio information is complementary and can have varied contributions depending on the emotional state [66, 68].

More importantly, this chapter's study contributes to the state-of-the-art fusion in multimodal emotion recognition by using similarity learning and by proposing and audio-visual metric learning. To the best of our knowledge, this is the first study that demonstrates how progressive fusion of audio-visual representations using metric learning can improve their contribution to emotion recognition.

5.3. FEATURE EXTRACTION AND AGGREGATION

Face Tracking and Alignment: We use the Ensemble of Regression Trees (ERT) method (described in [239, 240]) in order to detect and track faces in a video sample. ERT provides robust and accurate facial landmarks under challenging conditions, such as varying illumination and poses, with reliable and robust tracking in real-time. In a face track, each face is aligned by registering it with respect to facial landmarks, such as eye's centers, nose, mouth, and chin, to a canonical frame through a similarity transformation. Facial images are cropped and re-sized to a fixed resolution of 224×224 pixels.

Visual Features: Similar to the Convolutional Neural Network (CNN) representations in Chapter 4, we use a CNN representation based on the VGG-face model [96]. VGG-face is a 16-layer CNN model, trained with 2.6M facial images of 2.6K people for the task of face recognition in the wild. In the feature extraction stage, we employ the FCL 6's output as the facial signature. This layer outputs a 4096-d feature vector. Further details about this model are given in Subsection 2.4.1, as well as Chapter 4 (Subsection 4.1.2).

Audio Features: Also, we extract the same set of handcrafted audio features as the ones in Chapter 4. These features are obtained by utilizing the speech analysis toolkit openSMILE [126]. This open-source library is popular and widely used for extracting audio features that capture both the voice quality and prosodic characteristics of the speaker (explained in Subsection 3.2.2). We extract a set of features as explained in [127], including the following descriptors: 34 delta coefficients of energy and spectral Low-Level Descriptors (LLDs), 34 energy and spectral related LLDs, 4 voicing related LLDs, 4 delta coefficients of the voicing related LLDs, and 2 voiced/unvoiced durational features. Moreover, a set of statistical functionals are applied to capture the dynamic nature of the voice and summarize the LLDs over time segments [125]. For each video-clip, the resulting feature vector dimension is 1582. The complete list of the LLDs and the applied functionals are provided in Subsection 4.1.2.

Temporal Feature Aggregation using Fisher Vectors: as the extracted set of features has high dimensions, (e.g. each video frame has a 4096-d feature vector), and each modality description has its unique characteristics and data distribution, we employ a high-level representation, through Fisher Vector (FV) encoding. FV is used for aggre-

gating and clustering low-level features of each frame (e.g. CNN and audio features), obtaining one FV from all the frames in a sequence (e.g. video-clip's face track), by fitting a parametric generative model such as Gaussian Mixture Model (GMM) to the features. GMM can be referred to as a *probabilistic visual vocabulary*, while FV encodes the gradient of the local descriptors log-likelihood with respect to the GMM parameters. Further details about FVs are given in Subsection 4.1.2.

FV encoding is used to obtain a compact and single feature vector for each modality of a given video clip. As explained in Subsection 4.1.2, FV dimensionality is $2Kd$, where K is the number GMM components, which in our case was set to 4 and 2 for video and audio features, respectively. For example, in our work, for a given AVER dataset ($\mathbb{D}$), the audio and visual features are extracted and then aggregated via FV for each sample (i). Therefore, $\boldsymbol{x}^{(i)(v)}$ and $\boldsymbol{x}^{(i)(a)}$ denote video and audio FVs corresponding to the i^{th} sample of video $X^{d^v \times n}$ and audio $X^{d^a \times n}$ data matrices.

5.4. METHOD: MULTIMODAL EMOTION RECOGNITION METRIC LEARNING

In this section, we introduce a set of definitions and notions, prior to our Multimodal Emotion Recognition Metric Learning (MERML) formulation. Then, we explain the optimization of the method and its usage for the classification task.

5.4.1. DEFINITIONS AND NOTIONS

Audio-Visual Emotion Recognition (AVER): Given a dataset $\mathbb{D}$ with audio-visual emotional content, consisting of n samples, each annotated with a discrete emotion c:

$$\mathbb{D} = \{(\boldsymbol{x}^{(1)(v)}, \boldsymbol{x}^{(1)(a)}, c^{(1)}), (\boldsymbol{x}^{(2)(v)}, \boldsymbol{x}^{(2)(a)}, c^2), ..., (\boldsymbol{x}^{(n)(v)}, \boldsymbol{x}^{(n)(a)}, c^n)\}$$

where $\boldsymbol{x}^{(i)(v)} \in X^{d^v}$ and $x^{(i)(a)} \in X^{d^a}$ denotes video and audio feature vectors corresponding to the i^{th} sample of video $X^{d^v \times n}$ and audio $X^{d^a \times n}$ data matrices, while $c^{(i)}, i \in \{1, \ldots, e_n\}$ refers to the given discrete emotion (e) label. The goal, in AVER, is to predict the emotional content of a given sample test. In this chapter, like the rest of the dissertation, we use uppercase letters to denote a matrix, e.g. $X \in \mathbb{R}^{m \times d}$, and lowercase bold letters for vectors, e.g. $\boldsymbol{x} \in \mathbb{R}^d$. Additionally, we define the following operators and terms:

- n: the number of samples.
- d^a, d^v: the dimensionality of audio and video features.
- $M^a = {W^a}^T W^a$ and $M^v = {W^v}^T W^v$: $M^a \in \mathbb{R}^{d^a \times d^a}$ and $M^v \in \mathbb{R}^{d^v \times d^v}$ are distance matrices for audio and video modalities. Besides, $W^a \in \mathbb{R}^{p^a \times d^a}$ and $W^v \in \mathbb{R}^{p^v \times d^v}$ are the linear transformation matrices .
- p^v and p^a: the dimensionality of the new subspace.
- $c^{(ij)}$: is 1 if i and j belong to the same class, or -1 otherwise.
- $\mathbb{S}$ and $\mathbb{DS}$: two sets of similar (positive) and dissimilar (negative) instance pairs

$$\mathbb{S} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}, c^{(ij)}) | c^{(ij)} = 1\}, \text{ and } \mathbb{DS} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}, c^{(ij)}) | c^{(ij)} = -1\}$$

- $d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$: the multimodal distance of two samples x_i and x_j given their audio and video modalities.

Figure 5.1 provides an overview of the proposed framework. We define a Mahalanobis distance for each modality, with these distances being learned jointly. Finally, the learned multimodal distance is used within the RBF kernel based for SVM. The technical background of Metric Learning (ML) is given in Subsection 2.2.2, where this concept is defined and its properties are elaborated on.

5.4.2. A BRIEF REVIEW OF METRIC LEARNING

Standard distance metric assumes a higher similarity between two samples' representations of a similar class and bigger difference otherwise. Given two samples, $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$, the standard Euclidean distance is:

$$d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \sqrt{(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})} \tag{5.1}$$

In standard distance ML, the goal is to find an optimal metric M according to the similarity and dissimilarity constraints. This is done through a convex optimization, to learn the transformation matrix W for the original features, such that $M = W^T W$, where M is symmetric and positive. As a result, the new formulation of equation (5.1) is:

$$\begin{aligned} d_M^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) &= (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T M (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) \\ &= \| W\boldsymbol{x}^{(i)} - W\boldsymbol{x}^{(j)} \|_2^2 = \| W(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) \|_2^2 \\ &= (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T W^T W (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) = d_W^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \end{aligned} \tag{5.2}$$

In more details, ML modifies the standard Euclidean distance to improve its discriminative ability, such that the distance between similar classes would be as small as possible, while enlarging it otherwise. Another benefit of ML includes dimensionality reduction of the feature vectors, by the linear transformation matrix (projection): $W \in \mathbb{R}^{p \times d}$, where $p \ll d$, and $p \geq rank(W)$. Note that, if $M = I^{d \times d}$, where $I^{d \times d}$ denotes the identity matrix, then the metric is reduced to a Euclidean distance.

5.4.3. FORMULATION

In this section, we define Multimodal Emotion Recognition Metric Learning (MERML) for audio-visual emotion recognition or similar tasks of a multimodal nature. Like the conventional unimodal ML, similar samples should be projected, in the new space, as close as possible to each other, while dissimilar ones must be placed further apart from each other. In this work, we aim to jointly learn modality-specific metrics, M^a and M^v, that maximize the prediction accuracy utilizing both audio and video data channels. We formulate this metric as follows:

$$\begin{aligned} d_{W^v, W^a}^2(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)}) &= d_{W^v}^2(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}) + d_{W^a}^2(\boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)}) \\ &= \| W^v \boldsymbol{x}^{(i)(v)} - W^v \boldsymbol{x}^{(j)(v)} \|_2^2 + \| W^a \boldsymbol{x}^{(i)(a)} - W^a \boldsymbol{x}^{(j)(a)} \|_2^2 \\ &= d_{W^v, W^a}^2(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)}) \end{aligned} \tag{5.3}$$

Both the dimensionality of modalities' feature vectors, $d^v$and d^a, and the linear transformation matrices, W^v and W^a, in the new subspace can be different. The aim is to learn Mahalanobis matrices $M^v = W^{v^T} W^v$ and $M^a = W^{a^T} W^a$, using a convex formulation in the low-rank subspace, such that $W^v \in R^{p^v \times d^v}$ and $W^a \in \mathbb{R}^{p^a \times d^a}$. These two matrices can also serve in reducing the high dimensionality of audio-visual modalities $X^v \in R^{d^v}$ and $X^a \in R^{d^a}$, which makes the developed method applicable for high dimensionality datasets.

In addition, a non-negative weighting scheme is applied in MERML. This strategy is specifically useful in audio-visual emotion recognition, as the varied contribution of audio-visual modalities in emotion prediction has been supported in many studies. For example, in [52], visual cues have been reported to have a higher impact than audio cues on the final decision of emotion perception by humans. Furthermore, the assigned weighting scheme, $(\omega, 1-\omega)$ can further help the algorithm to converge faster. As a result, equation (5.3) becomes as follows:

$$d^2_{W^v,W^a}(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)}) = \omega d^2_{W^v}(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}) + (1-\omega) d^2_{W^a}(\boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)}) \tag{5.4}$$

5

In these new subspaces, the weighted distance of both modalities and the linear projections are more efficient, such that using both audio and face, the distance between two samples i and j becomes smaller than a learned threshold $b \in R$, if they belong to the same class or larger otherwise. As explained in [71], this condition can be further imposed with a margin larger than *one* by the following constraint

$$c^{(ij)}\Big(b - d^2_{W^v,W^a}(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)})\Big) > 1 \tag{5.5}$$

5.4.4. OPTIMIZATION

To solve the defined formulation of MERML in equation (5.4) with the weighting scheme and the constraint condition, the following hinge-loss function is applied and optimized through Stochastic Gradient Descent (SGD)

$$\underset{W^v,W^a,\omega,b}{\operatorname{argmin}} L^{(ij)} = \sum_{i,j} max\Big[1 - c^{(ij)}\big(b - d^2_{W^v,W^a}(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)})\big), 0\Big] \tag{5.6}$$

This loss-function is non-differentiable due to the max-operation. However, the sub-gradient is usually used instead, through the following condition

$$\begin{aligned} \mathbb{1} = (c^{(ij)}\Big(b - \omega(\boldsymbol{x}^{(i)(v)} - \boldsymbol{x}^{(j)(v)})^T W^{v^T} W^v (\boldsymbol{x}^{(i)(v)} - \boldsymbol{x}^{(j)(v)}) \\ + (1-\omega)(\boldsymbol{x}^{(i)(a)} - \boldsymbol{x}^{(j)(a)})^T W^{a^T} W^a (\boldsymbol{x}^{(i)(a)} - \boldsymbol{x}^{(j)(a)})\Big) > 1) \end{aligned} \tag{5.7}$$

Where $\mathbb{1}$ is a logical operator, and it is *0* if the condition holds or *1* otherwise. The sub-gradient is applied to solve for W^v, W^a, b, and ω. Therefore, through an on-line SGD, at each iteration t, based on pairs of audio-visual samples (i, j), either negative or positive with the same frequency, we perform the following updates

$$W^{a^{t+1}} = W^{a^t} - lr \frac{\partial L^{(ij)}}{\partial W^{a^t}} = W^{a^t} - \mathbb{1}(c^{(ij)} \psi^a) \tag{5.8}$$

$$W^{v^{t+1}} = W^{v^t} - lr\frac{\partial L^{(ij)}}{\partial W^{v^t}} = W^{v^t} - \mathbb{1}(c^{(ij)}\psi^v) \tag{5.9}$$

$$b^{t+1} = b^t - \gamma\frac{\partial L^{(ij)}}{\partial b^t} = b^t - \mathbb{1}(-c^{(ij)}) \tag{5.10}$$

$$\omega^{t+1} = \omega^t - \beta\frac{L^{(ij)}}{\partial \omega^t} = \omega^t - \mathbb{1}(\mathbb{W}^v - \mathbb{W}^a) \tag{5.11}$$

Where lr, β and γ are the learning rates. ψ^a, ψ^v and $(\mathbb{W}^v - \mathbb{W}^a)$ are defined as follows:

$$\psi^v = \omega W^v(\boldsymbol{x}^{(i)(v)} - \boldsymbol{x}^{(j)(v)})(\boldsymbol{x}^{(i)(v)} - \boldsymbol{x}^{(j)(v)})^T$$

$$\psi^a = (1-\omega)W^a(\boldsymbol{x}^{(i)(a)} - \boldsymbol{x}^{(j)(a)})(\boldsymbol{x}^{(i)(a)} - \boldsymbol{x}^{(j)(a)})^T$$

$$\begin{aligned}(\mathbb{W}^v - \mathbb{W}^a) = \Big((\boldsymbol{x}^{(i)(v)} - \boldsymbol{x}^{(j)(v)})^T W^{v^T} W^v(\boldsymbol{x}^{(i)(v)} - \boldsymbol{x}^{(j)(v)}) \\ -(\boldsymbol{x}^{(i)(a)} - \boldsymbol{x}^{(j)(a)})^T W^{a^T} W^a(\boldsymbol{x}^{(i)(a)} - \boldsymbol{x}^{(j)(a)})\Big)\end{aligned}$$

PROJECTION MATRICES INITIALIZATION

Prior to learning W^v and W^a, the two matrices are initialized either randomly or by Principal Component Analysis (PCA). When initialized by PCA, we followed authors in [183] by initializing W^v and W^a with PCA dimensionality reduction. We learn PCA dimensionality reductions from the video $X^{d^v \times n}$ and audio $X^{d^a \times n}$ data matrices. PCA initialization matrices are only used to initialize the learning process. In other words, we set the values of W_0^v and W_0^a projection matrices only in the first iteration of the SGD, using the obtained PCA initialization. Moreover, it is important to note that the learned metric is applied on the original dimensions ($X^{d^v \times n}$ and audio $X^{d^a \times n}$ data matrices) to learn W^v and W^a. Furthermore, the evaluation section compares the two initialization procedures. It also shows how performance improves substantially when the proposed metric is applied.

5.4.5. CLASSIFICATION

Following the metric learning computation, we apply the learned distance, through the kernel trick in Support Vector Machine (SVM) (as described in Section 2.3). To that end, the standard Euclidean distance function $d^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ in the RBF kernel (equation (5.12)) [241] is replaced by the proposed multimodal metric distance as follows:

$$K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \exp\Big(-\frac{d^2(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})}{2\sigma^2}\Big) \tag{5.12}$$

Consequently, the classification is carried out by SVM with the new distance. Therefore, in our proposal of MERML, we efficiently employ the multimodal distance to obtain the following non-negative kernel function as follows:

$$K(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)}) = \exp\Big(-\frac{d^2_{W^v,W^a}(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)})}{2\sigma^2}\Big) \tag{5.13}$$

The motivations behind replacing the RBF kernel with the proposed metric can be listed as follows:

- Concatenating the projected features of each modality and then using SVM affects the potential discriminative power of the learned metric and results in an expensive computation. In our experiments, the performance was less accurate, than with the proposed metric. In other words, instead of using a classification algorithm on the concatenated features resulting from the linear projections of different modalities, we utilize the weighted score in the kernel trick directly.
- This approach makes the method scalable. Scalability refers to the property of projecting the high dimensional data to a latent space with a compact representation. As a result, the approach is able to deal with big sizes of training data. Plugging the scores, resulting from each modality in equation (5.13), allows an efficient emotion inference. Specifically, in AVER, for some instances, the face might not be detected, or vice-versa the audio may be missing. For example, the audio modality is prone to be missing in instances associated with emotions related to happiness, which is manifested mainly through facial expressions.

To summarize, the proposed MERML approach is detailed in Algorithm 2.

Algorithm 2 MERML

procedure MERML($\mathbb{D}$)▷ $\mathbb{D}$: Multimodal Emotion Recognition Dataset
Inputs:
    Obtain similar ($\mathbb{S}$) and dissimilar ($\mathbb{DS}$) sets from the training dataset
    Define MERML as in equation (5.4) constrained as described in equation (5.5)
    Number of iterations: T, learning rates: lr, β, and γ,
Initialization:
    Initialization of the parameters in MERML, which includes the following ones: W^a and W^v (by PCA or random initialization), $\omega = 0.5$, and b as in equation (5.5)
SGD:
    for `t = 1:T` **do**
        if *the condition in equation* (5.7) *satisfied* **then**
            simultaneously perform the updates of MERML parameters in equations 5.8, (5.9), (5.10), and (5.11)
        end if
    end for
Output:
    W^a, W^v, b, and ω
MERML score in SVM
    Use MERML distance in RBF kernel SVM as defined in equation (5.13)
end procedure

5.4.6. POSITIVE AND NEGATIVE SAMPLES MINING

One of the main challenges in Metric Learning (ML) is that the optimization through SGD often suffers from slow convergence when applied on a large scale dataset. There-

fore, we propose an optimized process of selecting positive-negative pairs, using several criteria. When the emotional classes are not balanced, we ensure that an equal percentage of samples from all emotional classes are always present during training. Then, using the confusion matrix between emotions, we balance the ratio of difficult vs. easy pairs. We consider difficult pairs, the ones belonging to emotions which are difficult to be discriminated, and they are especially relevant for MERML for obtaining an efficient high-level representation. Furthermore, the sampling is done using the multimodal information, such that the MERML distance defined in equation (5.4) selects hard negative samples $(\boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(j)(a)})$ from both modalities, that maximize the $d^2_{W^v,W^a}(\boldsymbol{x}^{(i)(v)}, \boldsymbol{x}^{(j)(v)}, \boldsymbol{x}^{(i)(a)}, \boldsymbol{x}^{(j)(a)})$ distance.

5.5. RESULTS

To illustrate the efficacy of the proposed method, we present an extensive experimental evaluation and report the results on different benchmarks for audio-video emotion recognition datasets: Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [52], eNTERFACE [116], Acted Facial Expressions in the Wild (AFEW) [117], and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [115]. As each database has a different number of emotion categories and due to their varying settings and recording setups, we provide an independent evaluation for each dataset.

5.5.1. EXPERIMENTAL SETUP

In the experiments on each dataset, we present the following results:

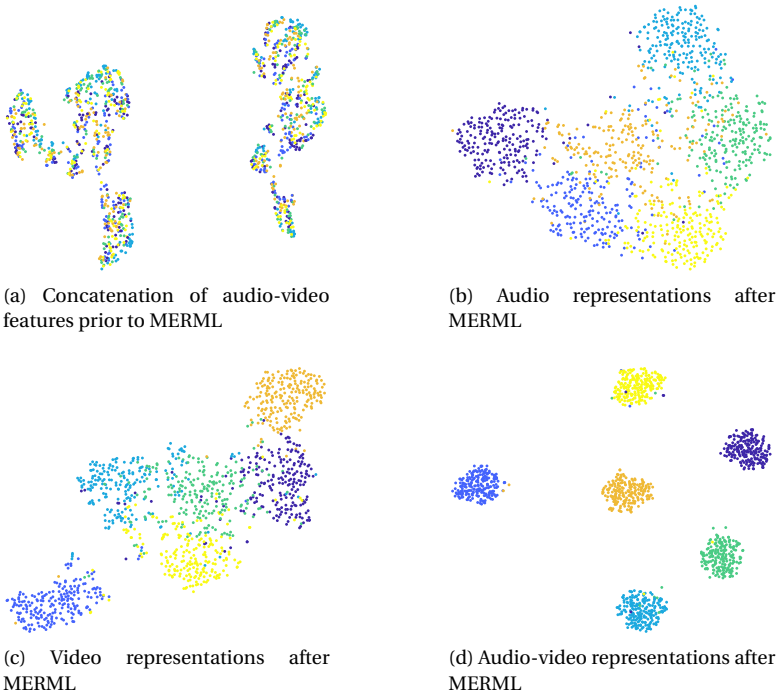
1. Unimodal evaluation of each representation separately using RBF-Support Vector Machine (SVM), showing the initial performances of the employed audio and video features in a unimodal emotion recognition task. In addition, a direct classification based on RBF-SVM is reported on the concatenated audio and video features.

2. The results of the baseline metric learning techniques Large Margin Nearest Neighbor (LMNN) [71], Information Theoretical Metric Learning (ITML) [83] and Geometric Mean Metric Learning (GMML) [84] on the concatenated representations of audio-video features. RBF-SVM is applied on the resulting features from these metric learning techniques for emotion classification. LMNN, ITML, and GMML are explained in Section 2.2 (Chapter 2).

3. The results of Multimodal Emotion Recognition Metric Learning (MERML) on the pairs of audio-video (AV) features according to its formulation in Section 5.4, illustrating the benefits of the proposed method.

Unlike the previous chapter's evaluations (Chapter 4), we provide the results of MERML on pairs of visual and audio representations, since we aim to show the effectiveness of the method in a bi-modal setting, rather than bestowing more importance to the visual descriptors through using more than one of them each time.

First, we qualitatively evaluate the MERML framework by visualizing the projected data in the newly learned subspace. For this purpose, we utilize t-SNE, a popular visu-



(a) Concatenation of audio-video features prior to MERML

(b) Audio representations after MERML

(c) Video representations after MERML

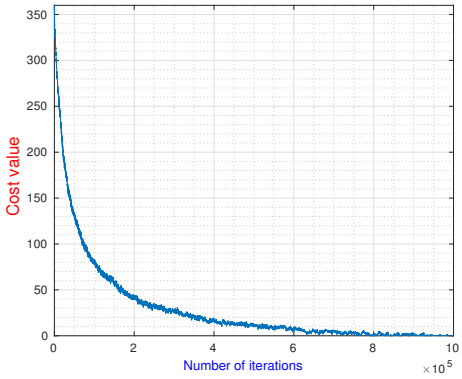
(d) Audio-video representations after MERML

Figure 5.2: t-SNE embedding of the eNTERFACE features, in the original subspace and in the new learned subspace. Features in (a) are concatenated from both modalities in the original space. Note that they are highly correlated and the number of formed clusters is not well defined. In (b), (c) and (d), we visualize the audio, video and audio-video data following MERML, where the clusters are well structured. The cluster colors represent the emotional classes. (The figure is better viewed in color.)

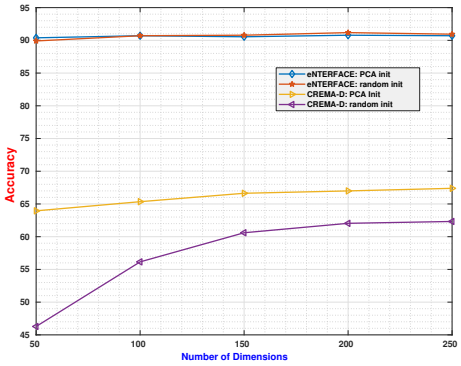
alization, and unsupervised dimensionality reduction tool [242]. Figure 5.2 shows eNTERFACE sample features before and following MERML. We can observe that prior to MERML, the emotional classes are highly correlated, hence form similar clusters. However, applying MERML improved the cluster formed based on the emotional content, leading to well-separated emotions in the new subspace.

5.5.2. SENSITIVITY ANALYSIS

Since the performance of MERML depends on the W^a and W^b projection matrices, as formulated in equation (5.4), we conduct a sensitivity analysis with regards to these variables. In particular, we check how the initialization of W^a and W^b and their dimensions contribute to the overall performance (explained in Subsection 5.4.4). We tested two initialization approaches, random and PCA initialization with various dimensions. Figure 5.3b details the sensitivity analysis on both datasets. We notice that, in the case of CREMA-D, Principal Component Analysis (PCA) initialization of W^a and W^b gives higher performance. However, various dimensions of these projection matrices give very close results, especially, when these dimensions are in a range of 150 to 250. This shows



(a) Cost value of MERML over iterations.



(b) Sensitivity analysis on W_a and W_v with regards to their initialization and dimensions.

Figure 5.3: A cost value illustration and the sensitivity analysis

that MERML is robust under different configurations. In the rest of our experiments using MERML, we use PCA initialization while the dimensions of W^a and W^b are set to 200.

In addition, to illustrate how the proposed algorithm converges on a given task, we included in Figure 5.3a the loss values (as defined in the cost function (5.5)) over a number of iterations. In our training setup, the SGD optimization of MERML has a fixed number of iterations, which was set to 1 million in our case. It is important to note that, with equal frequency, in each iteration, we pick positive or negative pairs according to Algorithm 2. The early stop of the SGD (e.g. at 800K) does not influence the results. In all benchmarks, we use the following parameters for training MERML: learning rate=0.001 and modalities weight learning rate=0.0001. For the incorporation of MERML in the SVM-RBF kernel, we use gamma=0.001. These parameters are fixed for evaluation on the three benchmarks. The following sections detail MERML evaluation on each benchmark.

5.5.3. EVALUATIONS ON CREMA-D

CREMA-D [52] is an audio-video emotion expression dataset. It contains 7442 clips from 91 actors (43 females and 48 males). Participants' age ranges between 20 and 74, and they come from a variety of races and ethnicities, i.e. Asian, African American, Caucasian, and Hispanic. Actors were asked to speak 12 sentences in five different emotions, namely, anger, disgust, fear, happiness, and sadness, or in neutral. The sentences were spoken with four different levels of intensities: low, medium, high, or unspecified. It is important to note that video clips in CREMA-D have an average length of 2.63 ± 0.53 seconds. As explained in Section 3.1, in this dataset, recognition rates of human perception based on the relative majority are reported. The relative majority is obtained when an emotion gets the highest share of the votes compared to the rest of the other emotions. In this case, the recognition rates of human benchmark for audio-only, video-only, and the bimodal audio-video perception, are 45.5%, 69.0%, and 74.8%, respectively.

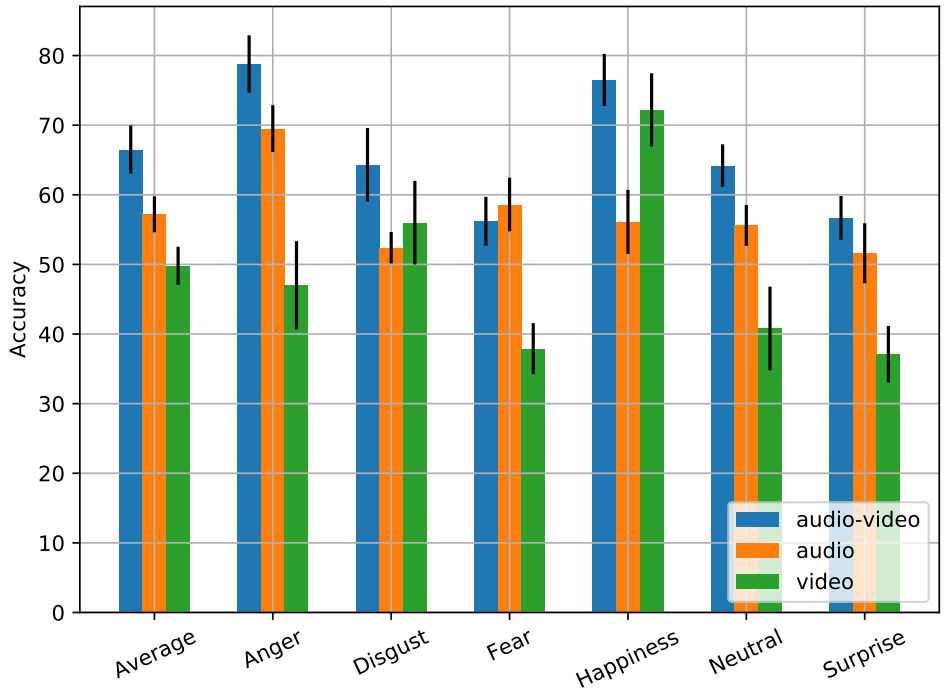


Figure 5.4: Bar diagrams (with error bars in standard deviation) for average performance (accuracy) for all emotions and per-emotion accuracy on CREMA-D. The figure shows the multimodal and unimodal accuracies of RBF-SVM on audio-only, RBF-SVM on video-only, and MERML on audio-video. The first three bars indicate the average accuracies for all emotions.

RESULTS ON CREMA-D

To the best of our knowledge, as there is not a technical evaluation baseline for this dataset, we divided the dataset into 10-folds to perform cross-validation based on subjects (the actors). In other words, for each fold, a subject's clips are either in the training or the testing sets. Subsequently, in each fold, we train the whole pipeline of FV encoding, learning MERML and SVM on the training folds, and then test it on the remaining fold. The reported results are the average of these 10-folds.

Positive and negative mining: Due to the size of CREMA-D, the number of possible positive and negative pairs is huge, approximately 45 million pairs. For this reason, we followed a careful selection of the used pairs during the SGD iterations for MERML optimization. Firstly, we place more emphasis on the positive samples of the same emotion, e.g. anger, to be selected from different subjects. Secondly, we pooled more negative pairs of emotions from the same subject. In this way, we ensured that the MERML training focus is on capturing the intra-class and inter-class variations depending on the emotions, rather than the subjects of the video clips.

Methods and features	Average Accuracy (%)
OpenFace features (V) + LSTM + COVAREP Features (A) + LSTM + Dual Attention [243]	65.0
Human Perception (relative majority recognition)	74.8
RBF-SVM on the concatenated audio-visual representations	65.2±3.0
ITML on the concatenated audio-visual representations	60.5±4.2
GMML on the concatenated audio-visual representations	65.0±4.7
LMNN on the concatenated audio-visual representations	63.5±2.8
MERML (Section 5.4) on audio-visual representations	**66.5**±3.5

Table 5.1: The average recognition accuracy of MERML and other methods on CREMA-D. Note that we provide the standard deviations (stds) for MERML and other metric learning methods. However, other methods did not report the standard deviations as a statistical bound on the average accuracies.

UNI-MODAL AND MULTIMODAL INTERACTION:

To check the contribution of the audio and video modalities, Figure 5.4 provides the unimodal and MERML results of the employed features, in terms of average accuracy and the accuracy on each emotion of CREMA-D. The unimodal results are obtained using RBF-SVM classifier on each modality's representation, separately. The first bar in each group gives the average result of MERML on the audio-visual representations. MERML was able to capture the complementary and supplementary information of both modalities, achieving an accuracy of 66.5%. In addition, MERML outperformed the individual modalities and yielded a performance increase of 9.3% and 16.7% over the audio-only and video-only perception, respectively. In addition, as shown in Figure 5.4, the recognition accuracy for each emotion (with a slight exception for fear) increased with the multimodal learning through MERML. These results show that the best results are achieved when using both audio and video information, the contribution per modality can be different for each emotion. For example, audio appears to be more significant than facial expressions for anger, and video has more impact in recognizing happiness. On the other hand, disgust and neutral emotions require both data channels for accurate recognition, proving the benefits of multimodal learning in any case.

In terms of unimodal results, audio outperforms visual representations on CREMA-D. This result is due to the limited number of sentences in CREMA-D, which led to learning good prosodic features and vocal expressions for each emotion. Moreover, in the case of the audio modality, authors of the CREMA-D dataset [52] reported that the emotion recognition rate increased by 10% when human raters interacted with and responded to more clips. However, in the case of the video modality, the interaction with more video clips did not have an impact on the emotion recognition rates of the human raters. In other words, getting exposed to audio information makes it easier to recognize new samples for both human raters and automatic systems.

MULTIMODAL EVALUATION

MERML's results and a comparison of its performance with LMNN, ITML, GMML, and RBF-SVM baselines are given in Table 5.1, showing that MERML resulted in at least 1.3% performance gain. In addition, the MERML approach outperformed the published results in [243]. In [243], authors employed audio and visual features using COVAREP [244] and OpenFace [245], respectively. OpenFace [245] visual representations provide the fol-

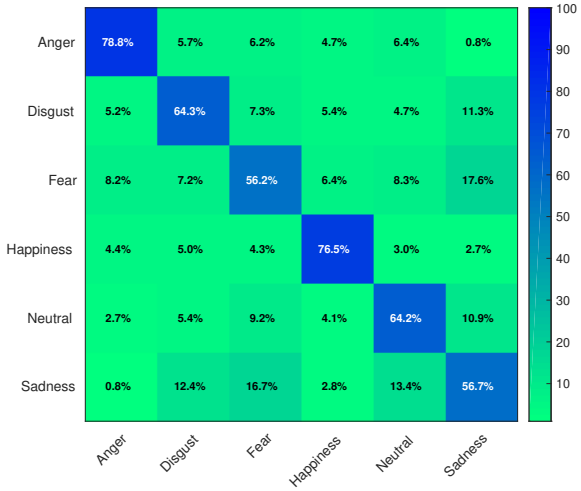


Figure 5.5: The confusion matrix of MERML on CREMA-D dataset, between ground-truth and clip predictions.

lowing features: Histogram of Oriented Gradients (HOGs), gaze direction, and head pose (3D position, and orientation of the head). COVAREP [244] is an open-source speech analysis toolkit which extracts Prosodic, spectral, and voice quality related features. Subsequently, to fuse these features, R. Beard *et al.* used recursive multi-attention Recurrent Neural Networks (RNNs) with dual-attention. In [243], the performance (65.0% accuracy) was obtained by combining facial and audio temporal features with Long-Short Term Memory (LSTM). These results show the efficiency of our approach for an enhanced joint multimodal learning and fusion.

The **Confusion Matrix (CM)** displayed in Figure 5.5 shows the achieved performance of our approach on CREMA-D classes and the degree to which an emotion gets misclassified. Without exception, the diagonal elements have the highest accuracies, a fact which indicates the high classification accuracy of the intended emotions. In particular, anger, neutral, disgust, and happiness have higher accuracy detection, compared to fear and sadness. The figure shows that fear and sadness were confused with each other by approximately 16%. These results are also compatible with the reported human perception and confusion in [52].

5.5.4. EVALUATIONS ON ENTERFACE

eNTERFACE is a multimodal dataset which contains six archetypal emotions: anger, happiness, disgust, fear, surprise, and sadness. It includes 42 subjects, who were asked to simulate the emotions in 5 different reactions, resulting in 1260 video recordings. 23% of the recordings are obtained from women and 77% are from men, including respondents of diverse cultural backgrounds. In the evaluation, we follow the protocol in [188] by splitting the data samples into 10 folds for cross-validation. Consecutively, in each fold, we train the whole pipeline of FV encoding, learning MERML and SVM on the training folds, and then testing it on the remaining fold.

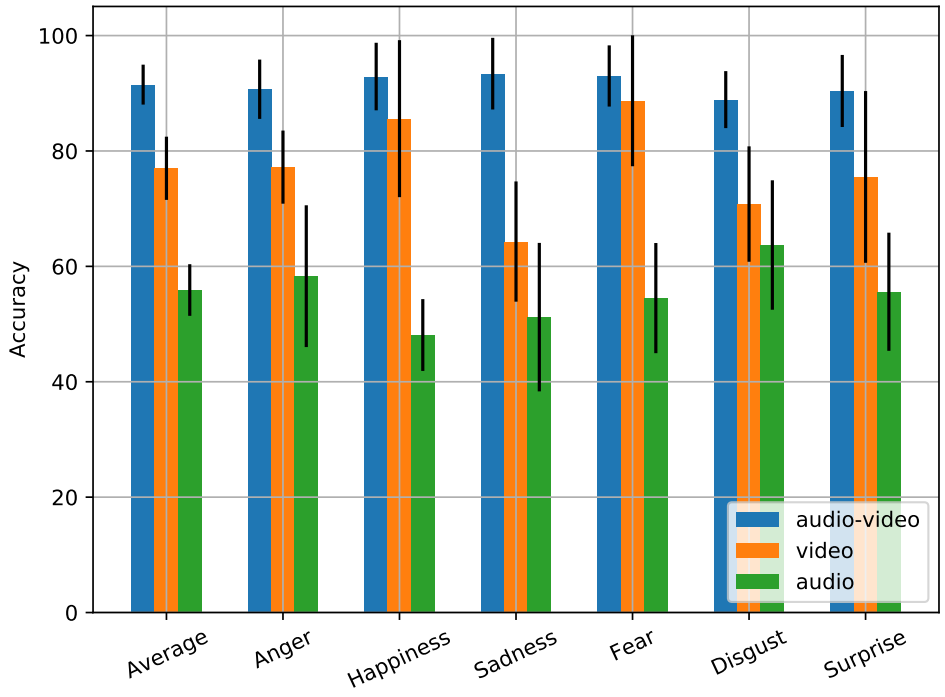


Figure 5.6: Bar diagrams (with error bars in standard deviation) for average performance (accuracy) for all emotions and per-emotion accuracy on eNTERFACE. The figure shows the multimodal and unimodal accuracies of RBF-SVM on audio-only, RBF-SVM on video-only, and MERML on audio-video. For example, the first three bars indicate the average accuracy of the three modalities in each dataset.

UNI-MODAL AND MULTIMODAL INTERACTION

Figure 5.6 provides the detailed performance of the visual, audio and MERML fusion for each emotion in eNTERFACE. On average, MERML obtains an accuracy of 91.5%, which helped to increase the performance of individual modalities by 14.5% and 35.6% for visual-only and audio-only perceptions, respectively. Furthermore, the fusion of audio-visual modalities through MERML boosted the performance by improving the classification accuracy for each emotion. In particular, performance for sadness was largely higher following MERML. This shows how the presence of both modalities is important in multimodal perception and subsequently emotion recognition.

MULTIMODAL EVALUATION

TABLE 5.2 presents the average recognition accuracies for the eNTERFACE dataset. We compare MERML with the current state-of-the-art audio-visual approaches in [188, 191] on eNTERFACE. Our MERML approach on audio-visual representations gives the highest recognition accuracy (91.5%), which is competitive to the state-of-the-art result (89.4%) reported by [191], that was obtained through end-to-end DL methods, namely, 3D CNN (C3D) cascaded with deep-belief networks (DBN). The reason behind this im-

Methods and Features	Average Accuracy (%)
Feature level fusion [190]	71.0
Score-level bimodal SVM [188]	87.4
Late fusion on C3D-DBN[191]	**89.4**
Hierarchical early and late fusion on audio-visual representations [168] (Chapter 4)	78.5±2.9
RBF-SVM on the concatenated audio-visual representations	84.7±3.4
LMNN on the concatenated audio-visual representations	85.1±3.5
GMML on the concatenated audio-vusual representations	81.9±4.9
ITML on the concatenated audio-visual representations	77.5±4.9
MERML according to Algorithm 2	**91.5**±3.7

Table 5.2: The average recognition accuracy of MERML and other methods on eNTERFACE. Note that we provide the standard deviations (stds) for MERML and other metric learning methods. However, other methods did not report the standard deviations as a statistical bound on the average accuracies.

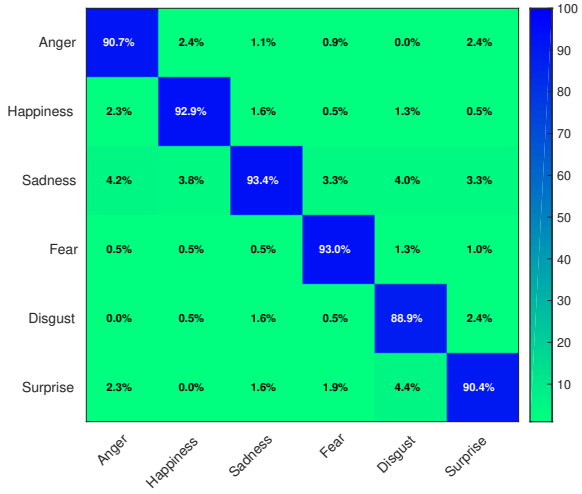


Figure 5.7: CM of MERML on eNTERFACE dataset.

provement is that DL methods have more parameters to train and learn, which needs to be properly initialized, e.g. via transfer learning.

Interestingly, we can observe that MERML increases the performance of the audio-visual representations, compared with the case when using only SVM or LMNN, GMML, ITML on the concatenation of audio and video representations. To illustrate this fact, the recognition rate of audio-visual representations increased by 6.8%, 6.4%, 9.6%, and 14% when using MERML in comparison with SVM, LMNN, GMML, and ITML respectively. MERML also outperformed the results we obtained in our previous study in [168] (Chapter 4). In addition, comparing the baseline of LMNN, GMML, ITML, or SVM and MERML using significance testing, we were able to validate the substantial gains (p-value $\ll 0.05$). These results show that MERML can provide a distance measure that enhances the performance of audio-video emotion classification.

The **Confusion Matrix (CM)** of MERML on audio-visual features is illustrated in Fig-

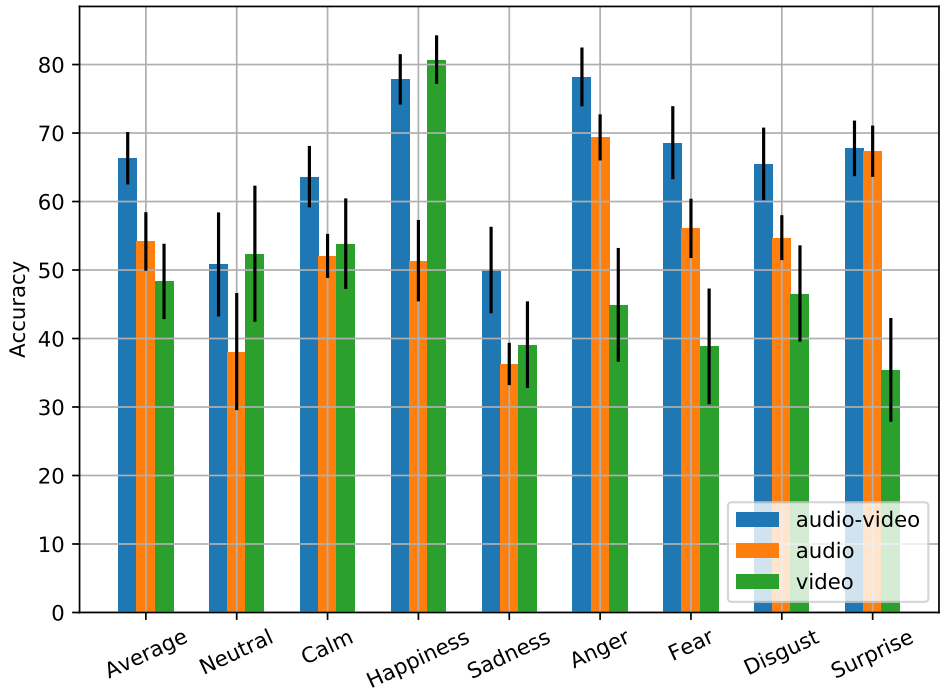


Figure 5.8: Bar diagrams (with error bars in standard deviation) for average performance (accuracy) for all emotions and per-emotion accuracy on RAVDESS. The figure shows the multimodal and unimodal accuracies of RBF-SVM on audio-only, RBF-SVM on video-only, and MERML on audio-video. For example, the first three bars indicate the average accuracy of the three modalities in RAVDESS.

ure 5.7. From the CM, it can be observed that all the emotions were detected with high accuracy. Disgust was the emotion to be confused with the rest of the intended emotions. For example, it was confused with sadness by 4% and with surprise by 4.4%.

5.5.5. EVALUATIONS ON RAVDESS

RAVDESS [115] has two sets: speeches and songs subsets. We use the speech set as it is labeled with eight archetypal emotions: anger, happiness, disgust, fear, surprise, sadness, calmness, and neutral. The dataset contains 24 subjects, 12 males, and 12 females, with an age range of [21,33]. It contains short speech video-clips of an average of 3.82 ± 0.34 seconds. The total number of videos is 1444. On RAVDESS, 247 individuals from North America provided human ratings of emotions. Human perception (benchmark) was reported as follows: 62.0%, 72.0%, and 80.0%, for audio-only, video-only, and audio-video modalities, respectively.

UNI-MODAL AND MULTIMODAL INTERACTION

Figure 5.8 presents the role of the multimodal fusion and its interaction with the audio-only and video only modalities on the AFEW validation set. The multimodal perception

Methods and features	Average Accuracy (%)
OpenFace features (V) + LSTM + COVAREP Features (A) + LSTM + Dual Attention [243]	58.3
Human Perception [115]	80.0
RBF-SVM on the concatenated audio-visual representations	**67.3**±4.2
ITML on the concatenated audio-visual representations	56.3±3.8
GMML on the concatenated audio-vusual representations	52.8±3.8
LMNN on the concatenated audio-visual representations	52.8±4.3
MERML (Section 5.4) on audio-visual representations	66.3±3.7

Table 5.3: The average recognition accuracy of MERML and other methods on RAVDESS. Note that we report the standard deviations (stds) for MERML and other metric learning methods. However, other methods did not report the standard deviations as a statistical bound on the average accuracies.

through MERML achieved an average accuracy of 66.3%, which helped to increase the recognition accuracy by 18.0% and 12.1% for the video-only and audio-only emotion recognition, respectively. Here, in the RAVDESS dataset, MERML improved the detection accuracy of most emotions (with the exceptions in Neutral and Happiness). Neutral was added in the RAVDESS dataset, and human raters confused it mainly with calmness and surprise, while video modality is leading the recognition rates of happiness. Another reason behind this performance is that the unimodal perception of neutral was lower than that of other emotions. Furthermore, Figure 5.8 shows the individual modalities' performance on the RAVDESS dataset. Audio representations showed higher performance than the video modality, with an average accuracy of 54.2%, and 48.3% for video and audio modalities, respectively.

MULTIMODAL EVALUATION

Table 5.3 details the recognition rates of MERML and other state-of-the-art methods. MERML gives good performance when applied to audio-visual features, achieving 66.3% accuracy. This is the only case when MERML slightly underperformed, by 1% when compared to SVM classification directly on their concatenated features. However, MERML improved significantly the results on audio-visual representations, compared to applying a vanilla LMNN as a baseline method. For example, MERML resulted in a gain of 13.5% in comparison with LMNN. As presented in Table 5.3, the MERML approach outperformed the baselines provided by LMNN, ITML, and GMML. For example, MERML outperformed both ITML and GMML by 2.2% and 10.3%, respectively. However, SVM-RBF resulted in a slightly higher recognition rate (67.3%).

Figure 5.9 presents the **confusion matrix** on RAVDESS, where the aim is to show how emotions are detected or misclassified for other emotions. The diagonal values are higher than the rest of the values. For example, the method gives good recognition rates for anger, happiness, fear, calm, and disgust. On the other hand, sadness and neutral emotions have less accuracy and were confused with each other and the rest. MERML also misclassified neutral state, as this expression is challenging even for human raters [115].

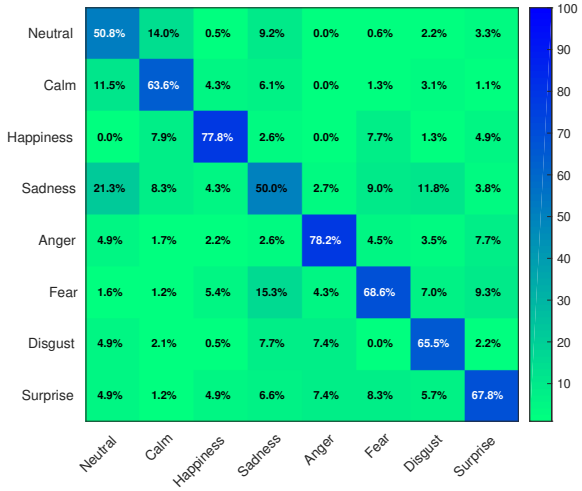


Figure 5.9: CM of MERML on RAVDESS.

Set	Anger	Happiness	Sadness	Fear	Disgust	Neutral	Surprise	Total
Training Set	133	150	117	81	74	144	74	773
Validation set	64	63	61	46	40	63	46	383

Table 5.4: Emotion categories distribution on training and validation sets of AFEW

5.5.6. EVALUATIONS ON AFEW

AFEW [117] is a multimodal dataset and consists of two public sets: training (773 samples) and validation (383 samples) sets. Each video clip is labeled with one of the six discrete Ekmanian emotions and neutral. AFEW is a challenging dataset with occlusions, varying illumination and head poses, harvested from Hollywood movies. It reports the baseline results (38.8% accuracy) on the validation set, which are based on feature level fusion of audio-video representations. Local Binary Patterns-TOP (LBP) features are used for the visual representations. For audio features, the authors of [117] employed similar audio features to the ones employed in our study. SVM is used for the classification. In our evaluation, we list and compare the results on the AFEW's public validation set, while the training set is used during learning and for optimizing FV and MERML parameters in Algorithm 2. In addition, we report the studies that benefit from both audio and video modalities for emotion recognition in AFEW.

POSITIVE AND NEGATIVE PAIRS MINING

The AFEW dataset is not balanced, in terms of the number of samples per emotion. Emotion categories' distribution is given in Table 5.4. For example, disgust and surprise have much fewer samples compared to happiness, anger, and neutral. Therefore, positive and negative samples mining could be biased towards those classes which have a higher number of samples. Eventually, this aspect might affect the performance of MERML on the final results. To avoid this bias, we generate an equal number of positive and neg-

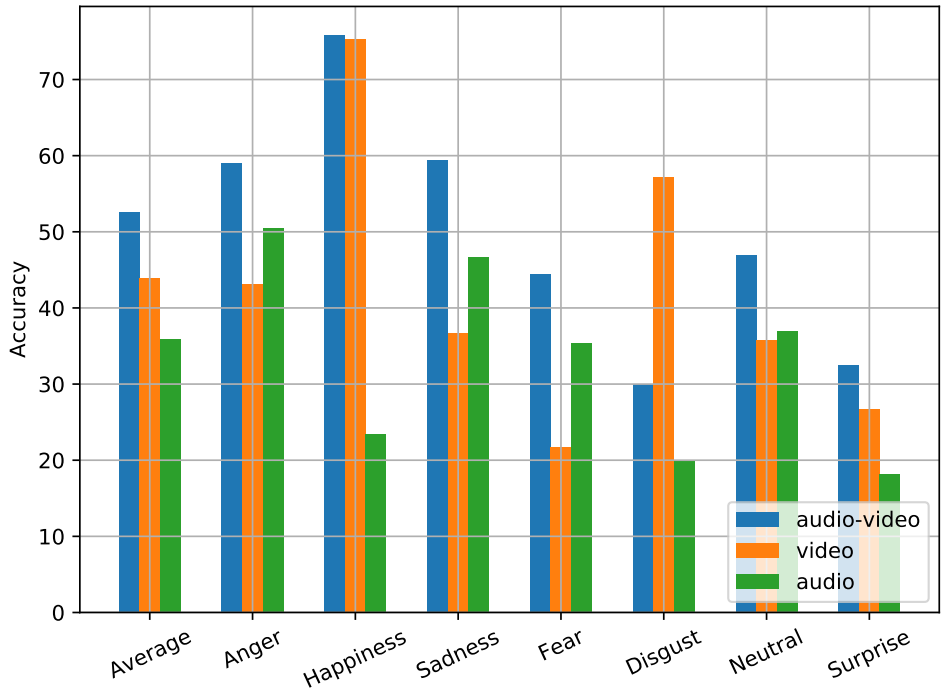


Figure 5.10: Bar diagrams for average and per-emotion performance on AFEW validation set, to show the multimodal and unimodal accuracies of RBF-SVM on audio-only, RBF-SVM on video-only, and MERML on audio-video. For example, the first three bars indicate the average accuracy of the three modalities in the AFEW validation set. The distribution of the emotions' categories is given in Table 5.4

ative pairs for each emotion, to be used during the SGD optimization of MERML. This strategy increased the performance of MERML by at least %1.

UNI-MODAL AND MULTIMODAL INTERACTION

Figure 5.10 presents the role of the multimodal fusion and its interaction with the audio-only and video only modalities on the AFEW validation set. AFEW is a challenging dataset, where the presence of both audio and visual modalities is vital. The multimodal perception through MERML achieved an average accuracy of 52.6%, which helped to increase the recognition accuracy by 8.6% and 16.7% for the video-only and audio-only emotion recognition, respectively. In addition, MERML was able to enhance the detection accuracy of most emotions (with only one exception in Disgust, which has the smallest number of video-clips in both training and validation sets (Table 5.4)), taking advantage of their complementary information that these two modalities provide.

Furthermore, Figure 5.10 shows the individual modalities performance on the AFEW validation set. Visual representations showed higher performance than the audio modality, with an average accuracy of 44%, and 35.9% for video and audio modalities, respectively. Many AFEW clips contain background noise, such as music, or irrelevant

Methods and features	Accuracy on the validation set (%)
AFEW Baseline [117]	38.8
Audio + C3D [152]	52.0
Late fusion on audio-CNN-DSIFT [192]	51.2
Audio + C3D + ResNet-LSTM [193]	**53.9**
Hierarchical early and late fusion on audio-visual representations [168] (Chapter 4)	48.9
RBF-SVM on the concatenated audio-visual representations	48.9
ITML on the concatenated audio-visual representations	47.6
GMML on the concatenated audio-visual representations	45.0
LMNN on the concatenated audio-visual representations	49.5
MERML according to Algorithm 2	**52.6**

Table 5.5: The recognition accuracy of MERML and other methods on AFEW validation set.

conversations, other than those of the main character of the clip. This fact degrades the performance of the audio modality. Multimodal recognition is even more interesting under these circumstances, where both modalities complement each other and make the multimodal perception more important for understanding emotions.

MULTIMODAL EVALUATION

Table 5.5 details the recognition rates of MERML and other state-of-the-art methods. MERML gives good performance when applied to audio-visual features, achieving 52.6% accuracy. This result is due to the discriminative power of MERML, in which features are further optimized by projecting them in a latent sub-space for capturing their complementary information. Subsequently, MERML improves the performance of visual and audio representations, when compared to RBF-SVM classification directly on their concatenated features or after learning LMNN, ITML, and GMML as baseline methods. For example, MERML resulted in a gain of 3.1% in comparison with LMNN. A significant improvement is also seen when comparing with our previous study in [168] (Chapter 4), where a combination of early and late fusions is applied on various audio-visual representations.

As presented in Table 5.5, the MERML approach does not only achieve higher accuracies than the baseline provided by the AFEW, LMNN, ITML, or GMML, but also provides competitive results when compared to other approaches. A key point to mention is that many approaches deal with audio and video modalities separately, and perform a late fusion for the multimodal learning and prediction steps. For example, in [192], authors employ similar representations to the ones in our work (namely: CNN and audio) and Dense Scale-Invariant Feature Transformation (DSIFT), and apply Score Level Fusion (SLF) by logistic regression [192]. When comparing results using MERML with those reported in [192], it becomes obvious that a single visual and audio representation outperforms their combined representations.

Finally, there are approaches that rely, to a significant extent, on feature engineering. An example is the work reported in [193], where authors employed several visual representations through different deep learning models such as Residual Neural Networks (ResNet), Long-Short-Term-Memory (LSTM) with CNN, and 3-D CNN (C3D), in order to perform later fusion with audio features. The late fusion weights were assigned manually based on the researchers' observations on the performance of each represen-

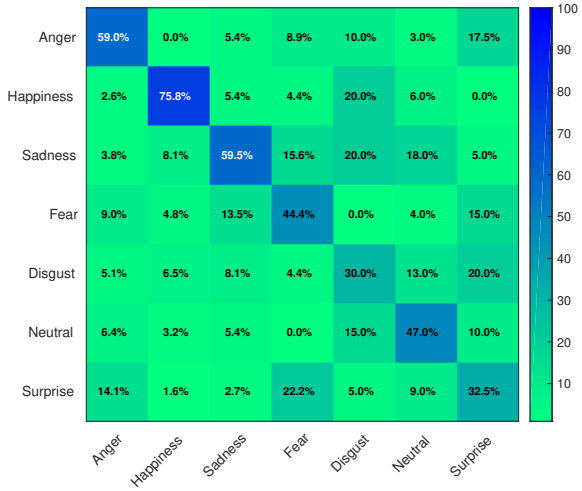


Figure 5.11: CM of MERML on AFEW validation set.

tation. However, in this work, the focus is not on feature engineering but, instead, we are aiming at learning a joint metric for audio and video modalities such that their representations in a latent subspace are robust to maximize the recognition accuracy. As a result, MERML has the advantage of focusing on multimodal learning on any given pair of audio-visual representations.

Figure 5.11 presents the **confusion matrix** on the AFEW validation set, where the aim is to show how emotions are detected or confused with other emotions. The diagonal values are higher than the rest of the values, which shows high detection of the video-clips' ground-truth. For example, the method gives good recognition rates for anger, happiness, and sadness, while other emotions have less accuracy, e.g. surprise, which was misclassified as anger even by human annotators [189]. In AFEW, other studies also reported the confusion between neutral, fear and disgust, which resulted in lower accuracies [189, 192, 192, 246].

5.5.7. MULTIMODAL INTERACTIONS

To provide an explanation of MERML and the contribution of the audio and video modalities, we check the agreement in emotion prediction, based on audio and video, compared to the audio-video representation. Our analysis shows that while audio-visual fusion has the best results, the contribution of its sub-modality varies in some emotions. For example, in CREMA-D, the audio is more significant in anger classification, where the recognition rate is 20% higher than the facial expression's classification. However, the video has more impact on recognizing happiness, which could be at least 10% higher than the audio modality. On the other hand, disgust requires both data channels for higher recognition.

5.6. CONCLUSIONS

We presented a joint multimodal metric learning for audio-video emotion recognition. Multimodal Emotion Recognition Metric Learning (MERML) can be applied in various multimodal contexts in which data complementarity could be exploited for increasing the performance, through an improved latent space representation. Our approach exploited successfully the dependencies and the complementary information of audio and video modalities in the context of emotion recognition, as their representations are well structured in the newly learned subspace, and their mutual emotion recognition is maximized. The quantitative and qualitative evaluation of the method on two datasets, utilizing distinct pairs of visual and audio representations, demonstrated the significant contribution of the method to an increased classification accuracy, achieving more robust performance than baseline results in different datasets. Furthermore, the comparison with the Large Margin Nearest Neighbor (LMNN), Geometric Mean Metric Learning (GMML) and Information Theoretical Metric Learning (ITML) baseline metric learning approaches showed the benefits of our method, which is efficiently learning the two modalities and optimizing their contribution for an enhanced performance.

This chapter's study contributed to the state-of-the-art in emotion recognition by using similarity learning and by proposing audio-visual metric learning. It demonstrates how multimodal emotion recognition can benefit from similarity-based learning methods, a key family of methods in machine learning, due to its applicability to a wide range of tasks, e.g., those tasks where less data is available. The study demonstrated that progressive fusion of audio-visual representations could improve their contribution to emotion recognition. A limitation of the study in the current approach is that it does not consider two factors: learning embeddings (representations) in an end-to-end manner, exploring or modeling the temporal expressivity of emotions. The next chapter focuses on the direction of using Deep Metric Learning (DML) to improve the audio and visual representations, addressing the limitation of shallow metric learning through learning audio-visual representations in an end-to-end manner. Using deep metric learning with the concept of similarity learning can also further enhance the fusion of audio-visual representations. It also adjusts the proposed method for another multimodal domain, namely, personality computing. The study in the next chapter also aims at capturing the temporal dynamics of emotional expressivity and perception. It focuses on exploiting the importance of visual and audio modalities over time in bimodal emotion recognition. These insights are obtained through state-of-the-art methods, namely, DML and Long-Short Term Memorys (LSTMs).

6

MULTIMODAL DEEP METRIC LEARNING FOR EMOTION RECOGNITION

The previous chapters explored the problem of multimodal emotion recognition from different points of view, looking into unimodal or late fusion methods. However, there is a need to bring these two modalities into a unified framework, to effectively learn joint multimodal fusion for Audio-Video Emotion Recognition (AVER). Besides, literature has not yet studied thoroughly the relation between time and emotion-related cues coming from audio and visual information. For instance, although studies from psychology and neuroscience underline the impact of time in recognizing negative and positive emotions, such knowledge has not been sufficiently supported by computational models, yet. In this chapter, a novel multimodal temporal deep network framework is proposed, that embeds video clips using their audio-visual content, onto a metric space, where their gap is reduced and their complementary and supplementary information is explored. This chapter addresses the research question of how audio-visual cues' temporal dynamics impact the recognition rate and speed of emotions. The proposed method is evaluated on two datasets, Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The study findings are promising, achieving significant performance on both datasets, showing a crucial impact

Parts of this chapter have been published in:

- **E. Ghaleb**, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019, pp. 552–558.
- D. Dotti, **E. Ghaleb**, and S. Asteriadis, "Temporal triplet mining for personality recognition," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020, pp. 379–386.

of multimodal and temporal information on emotion perception.

6.1. INTRODUCTION

A FUNDAMENTAL question in Multimodal Emotion Recognition (MER) is the one related to temporal expressivity and incremental perception of emotions. Emotion expressivity varies as a function of time, and this temporal process is also emotion-dependent [75–77]. For example, in [75], Kari Edwards investigated the temporal structure of facial expressions of emotions (FEEs). This study showed the importance of the temporal status of facial expressions in emotion perception. For example, the temporal dynamics of FEEs were detected and decoded accurately by observers. In addition, observers were more sensitive to the initial states of temporal displays of an expression. Moreover, authors in [76] studied how listeners recognize emotions over time and whether this functionality varies by emotion type or not. The study was conducted on five basic emotions, namely, anger, disgust, fear, sadness, and happiness. Their study suggested that anger, sadness, fear, and neutral expressions are detected more accurately at shorter time windows than happiness. For instance, the authors noticed that the recognition rate of happiness improves significantly by the end of vocal utterance, while fear was recognized at the fastest rate compared to the rest of emotions.

Much of the focus in Audio-Video Emotion Recognition (AVER) systems relies on multimodal learning and fusion through the selection of apex moments [78]. The literature lacks a thorough exploration and analysis of multimodal interaction over time [73]. For instance, in the literature of multimodal emotion recognition, much of the effort is focused either on a late fusion of modalities [152] or on building temporal features. Those directions of research are based on the assumption that emotions are expressed in audio-visual cues, simultaneously, in a global manner. As a result, studies have concentrated on obtaining global information to represent the emotional content of a video clip. However, these studies might overlook an essential aspect of how multimodal information binds and evolves over time, which is the aim of this chapter.

In this work, we addressed the following research question: what is the role of temporal dynamics in audiovisual cues, in automated emotion recognition? The study aims to efficiently connect information from different modalities and to deal with incremental emotion display. For this purpose, we designed a data-driven unified multimodal-temporal deep learning method to explore the variation of emotion expression over time through audio-visual modalities. The proposed method aligns the visual and audio representations using multi-stages integration and learning. The integrated multimodal framework is inspired by a gating paradigm introduced in [247] by Grosjean. In this paradigm, a stimulus is presented in successive segments of increasing duration. It is shown that emotion perception improves over time, by providing people with a richer context. Similarly, automatic emotion recognition will be capable of factoring affective states when having an incremental and multimodal presentation.

Furthermore, by employing an efficient metric distance, the accuracy of many classification and retrieval problems [71, 72] can be increased, as it contributes to obtaining an improved performance and robust representation. In metric learning, the task is to learn a distance function that is efficient to measure the similarity and dissimilarity of data samples. The efficiency of metric learning has been discussed in Sections 2.2.2 and

5.1 of Chapters 2 and 5, respectively.

In this chapter, there are two main contributions. First, to address the previously introduced research question, and motivated by the success of deep metric learning, we propose an end-to-end multimodal deep metric learning architecture. Moreover, this strategy addresses the limitation of the proposed shallow metric learning, the Multimodal Emotion Recognition Metric Learning (MERML), which was introduced in Chapter 5. Deep Metric Learning (DML) models overcome the limitations of shallow metric learning by capturing the non-linearity between highly heterogeneous audiovisual cues, as they optimize their representations from the raw data jointly. As a result, the proposed integrated temporal paradigm aims to learn audio-visual embeddings (representations) that are aware of emotional content in both auditory signals and facial expressions. The proposed method is illustrated in Figure 6.1 and explained in details in Section 6.3. Second, we develop an efficient learning algorithm through multimodal and incremental triplet sets' mining [110] and data augmentation, which is crucial to train the proposed method and to achieve significant performance. The proposed solutions are explained in Section 6.4. In addition, to address the research question of how emotion recognition varies as a function of time, the proposed method is designed to capture and exploit temporal information. Finally, we demonstrate these contributions in Section 6.5, by providing an extensive evaluation on two audio-video emotion datasets, namely: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [115] Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [52], showing how the proposed method efficiently fuses the different contributions of each modality over time.

This chapter is organized as follows. Section 6.2 introduces the related work on utilizing temporal information for multimodal emotion recognition and a brief overview of DML. Section 6.3 defines the architecture of the proposed method, while Section 6.4 presents the technical details of the implementations, such as the visual and audio mappings and data augmentations. Section 6.5 details the experimental evaluations of the proposed method. Next, Section 6.6 summarizes a study which extends this work in another subarea of affective computing. Specifically, it discusses the applicability of triplet loss in bodily analysis for personality recognition. Finally, Section 6.7 concludes the research and highlights its findings.

6.2. RELATED WORK

6.2.1. TEMPORAL EMOTION RECOGNITION

Emotion perception might vary over time according to the expressed emotion. Previous studies on Audio-Video Emotion Recognition (AVER) indicate the importance of multimodal and temporal information [5]. However, to a large extent, existing systems for emotion recognition lack the focus of studying how to bind multimodal information over time. For instance, many studies attempt to model the onset, apex, and offset phases of expressions. In addition, some studies select apex expressions and speaking face tracks for multimodal learning and fusion [78]. In [73], Kim and Provost proposed a data-driven framework to explore the impact of timing and expressivity duration on emotion classes. Their work is based on window averaging of audio-visual cues. It aims to spot the influential time windows for emotion inference in order to study how different utterances

impact emotion recognition. For instance, the evaluation found that anger, sadness, fear, and neutral emotions are recognized more accurately with shorter time windows. This finding was also aligned with a speech emotion recognition study in [76]. Moreover, the spotted windows showed consistency across speakers which are aligned with related findings from the point of view of psychology [75–77].

Furthermore, in [74], Zheng et al. adopted a discriminative Graph Regularized Extreme Learning Machines algorithm to identify the stable patterns of electroencephalogram (EEG) over time for emotion recognition. Stable EEG patterns indicate neural activities in different human brain areas under emotional classes. This study suggested that neural patterns exist for three emotion categories: positive, neutral, and negative. For example, their findings reveal that the human brain's lateral temporal areas activate more for positive emotions than negative ones. Furthermore, Deep Learning (DL) has been used for learning spatial and temporal features and in joint feature representations for multimodal data [5]. Further details regarding the use of Deep Learning (DL) in unimodal and multimodal temporal emotion recognition are provided in Sections 3.2 and 3.3, respectively, in Chapter 3.

Nevertheless, our approach, in this chapter, focuses on building multimodal incremental embeddings and checking how they contribute to emotion recognition over time. The proposed paradigm benefits from initial time windows of emotion expressivity and transmits the learned semantics to subsequent time windows.

6.2.2. DEEP METRIC LEARNING

The basic concept of metric learning is to modify a conventional metric, such as Euclidean distance, by including an efficient mapping function: $f : \boldsymbol{x} \rightarrow \mathbb{R}^n$. In this mapping process, the aim is to bring similar samples closer, and the dissimilar ones further, given the distance: $d(f(\boldsymbol{x}^{(i)}), f(\boldsymbol{x}^{(j)}) = \|f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^{(j)})\|_2^2$. Conventional metric learning approaches, such as Large Margin Nearest Neighbours (LMNN), usually learn a linear mapping. This linear mapping might suffer from the non-linear relationships between data samples, especially in multimodal learning tasks. These details are further explained in Subsection 2.4.4.

Moreover, metric learning has been applied in the context of deep learning, using two approaches. The first one applies metric learning on top of a deep learning architecture. This method uses the standard metric learning algorithm of the learned representation, by learning the low rank or full rank M matrix [146, 248]. In this way, it guarantees a robust description by learning a compact and discriminative representation. The second approach involves distance metric embedding within the architecture to measure the effectiveness of the learned representation, such as triplet loss [111, 149]. Deep Learning (DL) has been widely accepted as an effective model in highly non-linear data, and currently is proven to be the state-of-the-art in data representation and perception tasks [89, 153]. Therefore, DL can be used explicitly in learning mapping functions for metric learning through a set of non-linear transformations. This incorporating of metric learning within Deep Neural Networks (DNNs)' architectures is known as Deep Metric Learning (DML). DML is explained in Subsection 2.4.4. Moreover, multimodal deep learning is a way to exploit the dependencies and complementary information in multimodal tasks such as AVER [153, 249].

6.2.3. MULTIMODAL LEARNING

Multimodal learning is one of the challenging frontiers in machine learning [153]. Different approaches have been proposed for multimodal learning and can be categorized as follows [250]: multi-representation alignment (e.g. correlation-based models [17] and distance and similarity-based models [251]) and multi-view representation fusion (e.g. graphical models[250] and neural network models [252]).

This work follows the category of multi-view representation alignment, by employing DML. In this study, DML is adopted, given its efficiency in learning robust representations. Involving distance metric embeddings within the proposed architecture guarantees learning compact and discriminative features [111, 149]. Specifically, we propose learning deep embeddings through the use of triplet loss. Triplet loss involves negative and positive samples for a given anchor, which makes it more suitable for a classification task such as the one in AVER (more details are given in Subection 6.3.2). Moreover, by employing triplet loss, one can efficiently tackle the lack of sufficient data to train deep models, since it exploits similarities between samples, which generates a larger pool of data to train DL efficiently.

6.3. METHOD: TEMPORAL AND JOINT DEEP METRIC LEARNING

In this chapter, we aim to generate temporal audio-visual embeddings for accurate multimodal and temporal (incremental) emotion perception. Specifically, we aim at producing discriminative embeddings, by taking into consideration the binding information between audio-visual modalities across time. When designing the multimodal deep learning framework based on metric loss, we have the following objectives and motivations:

- It should exploit complementary information in the audio-visual representations.
- It is expected to produce discriminative and representative features and to reduce the gap between the audio-visual representations. In AVER, one of the challenges is that usually there is not a perfect alignment between the two data channels in terms of emotion expression. For example, happiness could be initially expressed through facial expressions, while corresponding time segments in the audio channels are not useful yet. However, the following audio time slices could provide valuable information [74].
- The framework should take into consideration the temporal evolution of emotion expressivity in video-clips. Not only temporal windows can have a different impact on each emotion, but also the impact might vary according to the subjects' personalities. Emotion expressivity could have gradual descent or ascent, while it can also peak at certain moments. Studies demonstrated that emotion perception might require a varying amount of time for an accurate detection [73, 75, 76]. Therefore, these alterations could be exploited efficiently through a temporally structured framework.

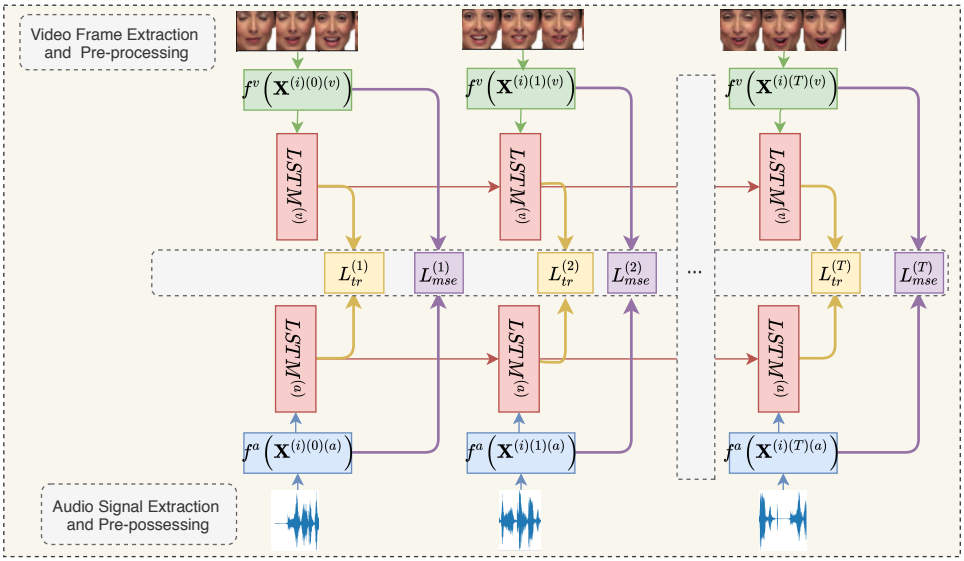


Figure 6.1: The proposed framework of multimodal temporal deep metric learning for AVER. It has two streams of audio ($f^a(\boldsymbol{X}^{(a)})$) and video ($f^v(\boldsymbol{X}^{(v)})$) sub-networks, diachronically connected via LSTM cells as a gating paradigm. In each gate, identification and discriminative signals guide the training of the network. We employed soundNet [135] and 3D-CNN [253] as sub-networks for audio ($f^a(\boldsymbol{X}^{(a)})$) and visual ($f^v(\boldsymbol{X}^{(v)})$) mappings, respectively.

6

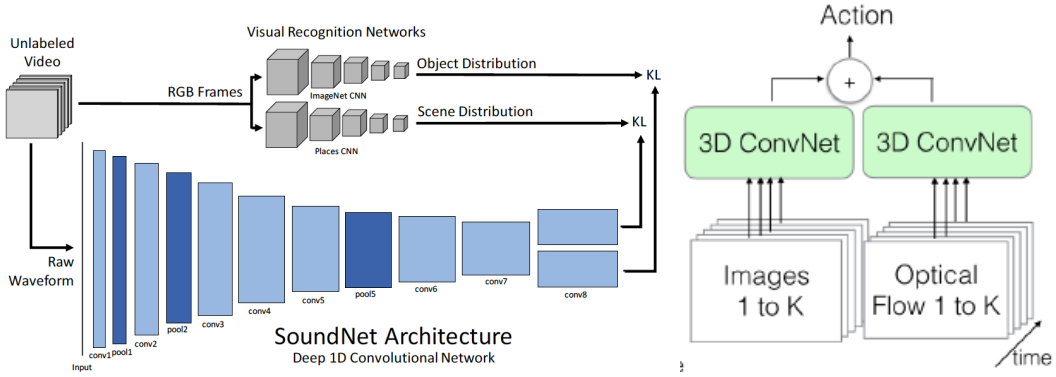
In our work, we apply Deep Metric Learning (DML) based on triplet networks. The loss function of this type of architecture uses triplet sets: $\{\boldsymbol{x}, \boldsymbol{x}^{(+)}, \boldsymbol{x}^{(-)}\}$, where $\boldsymbol{x}$ is an anchor, $\boldsymbol{x}^{(+)}$ and $\boldsymbol{x}^{(-)}$ are similar and dissimilar examples to $\boldsymbol{x}$, respectively:

$$d_f(\boldsymbol{x}, \boldsymbol{x}^{(+)}, \boldsymbol{x}^{(-)}) = \|f(\boldsymbol{x}) - f(\boldsymbol{x}^{(+)})\|_2^2 - \|f(\boldsymbol{x}) - f(\boldsymbol{x}^{(-)})\|_2^2 + margin \tag{6.1}$$

The optimization procedure aims to minimize the distance between the anchor (baseline) input to a positive pair while maximizing the distance from the anchor to the negative pair [149]. More details on the DML and the triplet loss are provided in Subsection 2.4.4.

Figure 6.1 outlines the proposed architecture for achieving incremental shared representations, modeling the relationships between the two modalities over time. This is pursued through similarity and discriminative loss functions in each time window that are averaged to obtain a temporal score. The proposed architecture shows how the sub-networks are connected incrementally via Long-Short Term Memory (LSTM) cells. Moreover, in each sub-network, gates are connected by LSTM cells to explore and utilize temporal dependency between video clip segments (time windows). Prior to feeding LSTM cells with the audiovisual mappings, the gap between the two modalities is reduced, through minimizing the distance between the two representations via Mean Square Error (MSE) loss functions (see Subsection 6.3.3).

The following subsections explain the Audio-Video Emotion Recognition (AVER) dataset in the context of this study, and how audio and visual mappings are employed on



(a) SoundNet [135] (b) I3D [253]

Figure 6.2: SoundNet and I3D were employed for audio-visual mappings.

the raw audio-visual data to build temporal joint embeddings. Then, the developed multimodal triplet loss, its formulation and optimization, and triplet sets' mining methods are elaborated.

6

6.3.1. AUDIO AND VISUAL INPUTS AND MAPPING FUNCTIONS

To target our research objectives, we learn joint embeddings via two-stream networks, one for audio and one for video cues. Moreover, this chapter employs an end-to-end deep learning approach, where we use mapping functions on audio and video cues as follows:

- $\boldsymbol{X}^{(t)(a)}$ is the raw audio data input of the t time window, which is fed forwarded to an audio mapping function ($f^a(\boldsymbol{X}^{(a)(t)})$), namely soundNet [135] (explained in the following subsection). SoundNet is a 1-D Convolutional Neural Network (CNN) that maps high dimensional audio data in a video clip onto a latent subspace as follows: $f^a(\boldsymbol{X}^{(t)(a)}) : \mathbb{R}^{d^{(a)}} \rightarrow \mathbb{R}^p$.

- Similarly, $\boldsymbol{X}^{(t)(v)}$ is the video data input of a t time window (e.g. $\boldsymbol{X}^{(t)(v)} \in \mathbb{R}^{16\times3\times96\times96}$ is the video segments of 16 RGB frames with 96×96 resolution). Subsequently, $\boldsymbol{X}^{(t)(v)}$ is fed forwarded to a video mapping function ($f^v(\boldsymbol{X}^{(t)(v)})$), namely I3D [253] (explained in the following subsection). I3D is a 3D-CNN that maps the high dimensional video content onto a latent subspace as follows: $f^v(\boldsymbol{X}^{(t)(v)}) : \mathbb{R}^{d^{(v)}} \rightarrow \mathbb{R}^p$.

In other words, for the audio and video mapping, we employ SoundNet[135] and the 3D-CNN on RGB image sequences branch of Two-Stream Inflated 3D ConvNet (I3D) [253], respectively. The two architectures are pre-trained using softmax. Moreover, one of the challenges in AVER is the fact that data samples usually constitute clips of short duration [5], therefore, modifications on the adopted architectures of visual and audio

mapping are required (as elaborated in the following two subsections). In our work, we adopted these architectures as audio and visual mapping models for the proposed framework, due to the following reasons: (1) their ability to capture and model temporal data up to several seconds, and (2) they are state-of-the-art models and have been shown to have good discriminative power for many other audio-video recognition tasks.

I3D

Modeling spatio-temporal RGB information is considered as one of the major challenges within the computer vision community, especially when the focus is on processing data acquired in noisy and varying conditions in terms of occlusions, luminance, clutter, etc. For this reason, the extension of the promising 2-D CNN models into the 3-D domain appears to be a promising strategy, which has been proposed through different research works, lately [253]. $3D-$CNNs are similar to the conventional CNN models, but with spatio-temporal filters. I3D is a framework which consists of two $3D-$CNN streams, as shown in Figure 6.2b. One stream utilizes RGB image sequences, while the other one is applied on optical-flow data. In this study, we employed the $3D-$CNN branch on RGB image sequences (the first branch in Figure 6.2b), which was pre-trained jointly with the optical-flow branch. This $3D-$CNN model is developed by inflating a $2D-$CNN into $3D$. For this purpose, authors of [253] adopted the Inception-v1 with batch normalization [254] and performed the required modifications, accordingly. More specifically, these adjustments can be done by inflating the convolutional filters and pooling kernels, by endowing them with an additional temporal dimension. For example, if a filter is square with $N \times N$ dimensions, it becomes cubic with $N \times N \times N$ dimensions.

The parameters of the $2D-$CNNs were bootstrapped for the $3D$ filters and kernels from pre-trained ImageNet models. This is done by converting an image into a video, by copying it, repeatedly, to make a video. Then, the $3D$ models can be pre-trained on the ImageNet dataset. The motivation behind this conversion is to avoid repeating the process of developing spatio-temporal models and to benefit directly from the weights of models which are pre-trained on large scale datasets [253]. More details about the architecture of these models are given in the corresponding paper [253]. The video mappings resulting from I3D have 400 dimensions.

SOUNDNET

SoundNet [135] is a deep convolutional architecture developed to obtain representations from audio signals for sound recognition. Figure 6.2a displays the architecture of SoundNet. As shown in the figure, the network benefited from a student-teacher training procedure to transfer knowledge from discriminative visual recognition models into the sound modality. For this purpose, the authors used a dataset that contains two million videos. SoundNet yielded impressive performance, even though it was trained without ground truth labels. SoundNet (the audio model in Figure 6.2a) consists of a series of one-dimensional convolutions, followed by non-linear operations (e.g. ReLU layers). This architecture benefits from the convolution networks for audio representations, since the convolutions are well suited for the audio waveform. Similar to applying CNN models on RGB images, CNNs can also result in detecting high-level concepts through low-level detectors (the 1-dimensional convolutional layers) from audio signals [135].

Another advantage of the architecture of this model is that it can support inputs of variable lengths since audio samples in a dataset can vary in their temporal length. To do so, a fully convolutional network, with pooling layers was proposed by the authors in [135]. In this manner, the model can apply convolution and pooling, in each layer, resulting in representations for audio signals with varying lengths. In addition, to make the representations adapt to the temporal length of the audio signals, the network was applied over multiple time windows with fixed lengths (e.g. dividing the waveforms into time windows of 5 seconds). In this way, it produces an output for each time window in a video. In this manner, the network is able to handle all the information from a video clip, without discarding any useful information. This strategy also inspired the design of our framework, since we trained the proposed method with visual information (using I3D in our case), where video clips have variable lengths. Finally, we added one convolutional layer with 400-dimensions on the top of the seventh layer of the SoundNet, in order to have similar dimensionalities for audio and video mappings.

6.3.2. INPUT EMBEDDINGS

In AVER, a dataset ($\mathbb{D}$) contains n video clips with audio and visual signals, while each clip is annotated with a discrete emotion c. In this study, the m modality in the $\mathbb{D}$ dataset can be defined as following:

$$\mathbb{D}^m = \{([\boldsymbol{x}^{(1)(1)(m)}; \cdots ; \boldsymbol{x}^{(1)(T)(m)}], c^{(1)}), [\boldsymbol{x}^{(2)(2)(m)}; \cdots ; \boldsymbol{x}^{(2)(T)(m)}], c^{(2)}), \\ \cdots , [\boldsymbol{x}^{(n)(1)(m)}; \cdots ; \boldsymbol{x}^{(n)(T)(m)}], c^{(n)})\}$$

where $\boldsymbol{x}^{(i)(t)(m)} \in \boldsymbol{X}^{d^m}$ is the resulting embeddings of the m^{th} modality from the mapping function $f^m(\boldsymbol{X}^{(i)(t)(m)})$ for the time window t, and corresponding to the i^{th} data sample. As a result, $\boldsymbol{x}^{(i)(t)([v,a])} \in \boldsymbol{X}^{d^{\{v,a\}}}$ denotes the video and the audio feature vectors corresponding to the i^{th} sample of video $\boldsymbol{X}^{d^v \times n}$ and audio $\boldsymbol{X}^{d^a \times n}$ data samples (tensors), respectively. $c^{(i)} \in \boldsymbol{c}$ refers to the given discrete emotion label. Ultimately, the goal, in AVER, is to predict the emotional content of a given sample test.

6.3.3. FORMULATION

To optimize and learn the parameters of each audio-visual mapping, and the temporal connections between the cells, in each time window (gate), we employ a temporal metric that has two terms: (1) multimodal triplet L_{tr} and (2) MSE L_{mse} loss functions. As shown in Figure 6.1 L_{tr} is applied on the LSTMs' outputs. For instance, for a sample i, at time window t, the corresponding LSTM output of audio modality, for an i^{th} sample, is referred to as $\hat{\boldsymbol{x}}^{(i)(t)(a)}$. On the other hand, the MSE L_{mse} is computed on the audio and visual mappings, e.g. the audio mapping of the i^{th} sample at t is indicated as: $f^a(\boldsymbol{X}^{(i)(t)(a)})$. As a result, we optimize the following objective:

$$\underset{f^v, f^a}{\operatorname{argmin}} L_{total} = \frac{1}{2NT} \sum_{i=0}^{N} \sum_{t=0}^{T} L_{total}^{(i)(t)} = \frac{1}{2TN} \sum_{i=0}^{N} \sum_{t=0}^{T} L_{tr}^{(i)(t)} + L_{mse}^{(i)(t)} \tag{6.2}$$

where L_{tr} and L_{mse} are defined as follows:

$$L_{tr} = \sum_{i=0}^{N} \sum_{t=0}^{T} \sum_{m=\{v,a\}} max\big(d_{f^m}(\hat{\boldsymbol{x}}^{(i)(t)(m)}, \hat{\boldsymbol{x}}^{(t)(+)(m)}, \hat{\boldsymbol{x}}^{(t)(-)(m)}), 0\big)$$
$$L_{mse} = \sum_{i=0}^{N} \sum_{t=0}^{T} \| f^{v}(\boldsymbol{X}^{(i)(t)(v)}) - f^{a}(\boldsymbol{X}^{(i)(t)(a)}) \|_2^2 \tag{6.3}$$

where t refers to a time window in the architecture, of a sequence consisting of T time windows; N is the number of samples per mini-batch; $\hat{\boldsymbol{x}}^{(i)(t)(\{a,v\})}$ indicate the corresponding segment of LSTM output for (a) audio or (v) video data in a given video clip. We formulate the multimodal triplet loss L_{tr} that optimizes $f^{v}(\boldsymbol{X}^{(v)})$ and $f^{a}(\boldsymbol{X}^{(a)})$ to minimize the distance between an anchor and a positive sample, while increasing the distance to a negative sample. It is important to note that L_{tr} is a modality specific loss, where the original loss in equation (6.1) is utilized in the overall loss in equation (6.2).

In each time window and mini-batch, for both modalities, we sample two sets $T_{a,v}$ of triplets: $\{\hat{\boldsymbol{x}}^{(i)(t)(\{a,v\})}, \hat{\boldsymbol{x}}^{(t)(+)(\{a,v\})}, \hat{\boldsymbol{x}}^{(t)(-)(\{a,v\})}\}$, where $\hat{\boldsymbol{x}}^{(i)(t)(\{a,v\})}$ is an anchor, and $\hat{\boldsymbol{x}}^{(t)(+)(\{a,v\})}$ and $\hat{\boldsymbol{x}}^{(t)(-)(\{a,v\})}$ are similar and dissimilar examples to $\hat{\boldsymbol{x}}^{(i)(t)(\{a,v\})}$, respectively. Note that we use an anchor i, and we refer to its positive and negative samples with $\hat{\boldsymbol{x}}^{(t)(+)(\{a,v\})}$ and $\hat{\boldsymbol{x}}^{(t)(-)(\{a,v\})}$, respectively, since the three samples are coming from three different video clips. In other words, triplet loss minimizes the intra-class variations and maximizes the inter-class variations, and provides an identification signal, which is conducted on the audio and video embeddings through a multimodal and incremental negative and positive triplet sets mining (explained in Subsection 6.3.4).

The second term of the temporal loss formulation, L_{mse}, is responsible for leveraging similar information in both modalities, by minimizing the distance between the audio and video mappings. Therefore, the main advantage of our framework consists in not only capturing the complementary and supplementary information between the audio-video channels in a global manner, but also in modeling them across time, contributing to an accurate overall emotion understanding, regarding its display pattern.

6.3.4. MULTI WINDOWS TRIPLET SETS MINING

One of the main challenges in triplet networks based DML is that triplet loss often suffers from slow convergence. In each time window, as the possible number of the triplet sets is large, DML learns to map correctly easy samples. However, hard negative mining is essential to improve the performance of the network and to provide it with useful training guidance. Therefore, as suggested by [111, 255], online hard negative mining based on mini-batches is an effective solution. Figure 6.3 illustrates the process of hard and semi hard triplet sets' mining. For instance, a hard triplet set is considered when the negative sample is closer to the anchor than the positive sample: $d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(-)}) < d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(+)})$. Moreover, for each mini-batch, we consider valid triplet sets, which yield positive loss (according to equation (6.1)). In particular, we compute the loss on the triplet sets that satisfy the following constraint:

$$d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(-)}) < d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(+)}) + margin \tag{6.4}$$

As a result, the loss is computed on the hard and semi hard negatives triplets. A crucial step is to not take into account the easy negatives (i.e. $d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(-)}) > d_f(\hat{\boldsymbol{x}}, \boldsymbol{x}^{(+)}) + margin$),

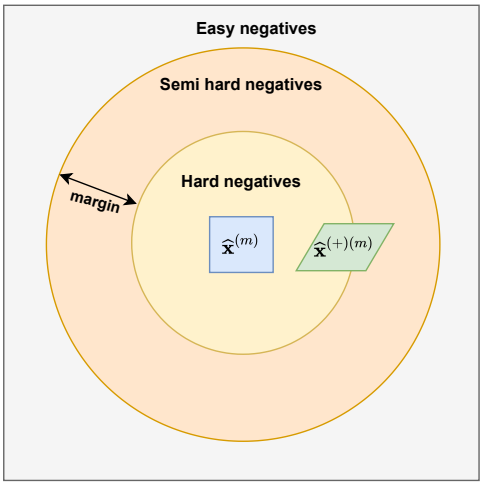


Figure 6.3: Three setups of triplet sets' mining, for a modality m. For example, a hard negative set is considered when a negative sample is closer to an anchor than the positive sample: $d_f(\hat{\boldsymbol{x}}^{(m)}, \hat{\boldsymbol{x}}^{(m)(-)}) < d_f(\hat{\boldsymbol{x}}^{(m)}, \hat{\boldsymbol{x}}^{(m)(+)})$, where m indicate that the embeddings belong to either audio or video modalities. A semi hard triplet set is considered when the positive sample is closer to the anchor, but using the negative sample still gives a positive loss (due to the margin as seen in the orange circle): $d_f(\hat{\boldsymbol{x}}^{(m)}, \hat{\boldsymbol{x}}^{(m)(-)}) < d_f(\hat{\boldsymbol{x}}^{(m)}, \hat{\boldsymbol{x}}^{(m)(+)}) + margin$. The easy negative samples are those that are far from the anchor, and using them yields a negative loss.

a strategy that is cumbersome for the learning process. Since this strategy can easily over-fit similar samples, which is not useful for the learning procedure.

Furthermore, in our work, we propose Multi Window Triplet Sets Mining (MWTSM), an effective technique for triplet sets mining across time windows. MWTSM can be explained as follows:

1. In our study, each mini-batch contains N-samples (video-clips) for each modality, and we use T time windows.

2. When mining the triplet sets, for each anchor, we select the hardest negative (smallest distance in $d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(-)})$) and the hardest positive (biggest distance in $d_f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(+)})$). This selection is referred to as *batch hard* [256]. It is called *batch hard* since it takes into consideration one hardest positive and one hardest negative samples to the anchor (there can be more candidates).

3. Batch hard sampling is applied for each time window, separately. As a result, each time window has N triplet sets. In other words, to obtain hard negative triplet sets, we search for corresponding time windows, independently. For example, the first window in a sequence is always compared to the first time windows of other sequences, and the second time window of another sequence is compared only with the second time windows of other sequences.

4. Considering that in our framework we have 2 modalities (audiovisual), T time windows, the resulting number of triplet sets is $2TN$. As a result, the constraint in 6.4

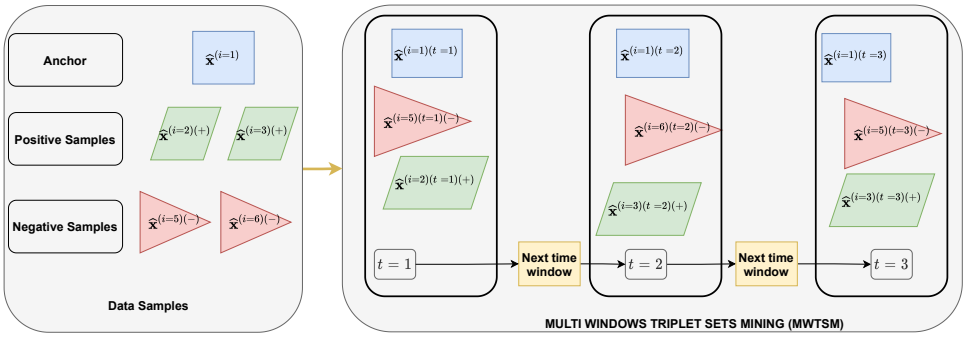


Figure 6.4: A toy example of MWTSM. Here, we show the workflow of MWTSM on five data samples which have 3 time windows. Also, we show how it works for one anchor. We consider the first data sample as an anchor. The third and the fourth data samples are positive samples of the anchor (i.e. with the same label). The fifth and the sixth samples are negative samples for the anchor (i.e. they have different labels). Note that the triplet sets mining (using the batch hard strategy) is done for each time window separately. At the first time window, we have the following triplet set: $\{\hat{\boldsymbol{x}}^{(i=1)(t=1)}, \hat{\boldsymbol{x}}^{(i=2)(t=1)(+)}, \hat{\boldsymbol{x}}^{(i=5)(t=1)(-)}\}$. In the second time window, we opt for choosing a different set, i.e. $\{\hat{\boldsymbol{x}}^{(i=1)(t=2)}, \hat{\boldsymbol{x}}^{(i=3)(t=2)(+)}, \boldsymbol{x}^{(i=6)(t=2)(-)}\}$. Note that the first choice of the anchor ($\hat{\boldsymbol{x}}^{(i=1)(t=2)}$) could be the positive and negative samples from the same data samples in the first time window (but with the embeddings of the second time window), namely, $\{\hat{\boldsymbol{x}}^{(i=2)(t=2)(+)}, \hat{\boldsymbol{x}}^{(i=5)(t=2)(-)}\}$. However, we aim to have a distinct selection of triplet sets to provide the loss function in equation (6.3) with as many as possible useful triplet sets for the optimization of the framework across time windows.

can be re-written as follows:

$$d_f(\hat{\boldsymbol{x}}^{(t)(\{a,v\})}, \hat{\boldsymbol{x}}^{(t)(-)(\{a,v\})}) < d_f(\hat{\boldsymbol{x}}^{(t)(\{a,v\})}, \hat{\boldsymbol{x}}^{(t)(+)(\{a,v\})} + margin \tag{6.5}$$

5. Additionally, the selected triplet sets in a time window, are discarded from the selection pool when mining triplet sets in the next time window. A detailed description of this procedure is provided in Figure 8.2, where we illustrate the way we select distinct triplet sets across time windows. The MWTSM strategy avoids duplicates triplets sets. This is done so that, in a new time window, the triplet sets mining selects new samples for each anchor and discards the ones from the previous time windows. As a result, in a time window, an anchor can target new pairs of negative and positive samples in the mining procedure. This strategy enables the framework to have many useful triplets and avoids repeated sets to optimize the learning process.

Moreover, Figure 6.5 displays the process of pulling positive samples closer to the anchor and pushing away the negative samples when employing audio and video modalities. To summarize, the proposed approach is detailed in Algorithm 3.

6.4. IMPLEMENTATION DETAILS

6.4.1. DATA AUGMENTATION

For data augmentation, we apply frame cropping across video sequences. In each mini-batch, we fix the coordinates for cropping images such that the same cropping is applied on all the video samples of a mini-batch. However, the coordinates for cropping

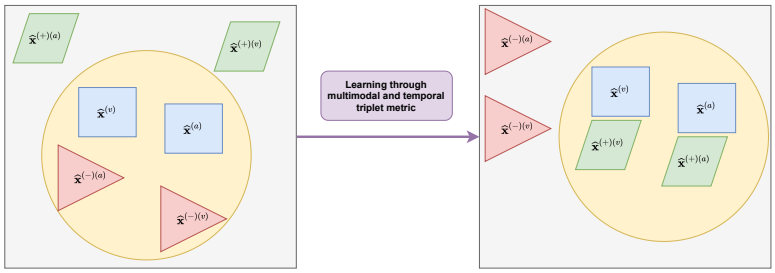


Figure 6.5: An illustration of the learning process using triplet loss based DML, which minimizes the distance between the anchor video-clip and its positive sample, while maximizing the distance to its negative sample, given the information of both modalities.

images might vary across mini-batches. For example, cropping is applied by re-sizing video frames into 112×112 resolutions, then cropping 96×96 patches with the given coordinates for the mini-batch. In addition, if the time windows' length is less than the required length (the default length as defined in Algorithm 3 (e.g. 16 frames), we looped on the time window as many times as necessary to add frames by randomly selecting frames within this time window in order to satisfy each model's input interface. The same temporal cropping is applied for audio signals. In addition, other image processing techniques, such as flipping of video frames consistently, rotations, and Gaussian noise are added to video frames during the training phase. Similarly, data augmentation techniques are applied on audio raw signals such as adding noise, changing the pitch, and the speed of the signal. However, during evaluation, models are applied over given video clips segments, and, for visual mapping, we select 96×96 center crops. However, no pre-processing is applied on audio signals during testing.

6.4.2. TRAINING PROCEDURE

The proposed architecture is trained on two Titan XP GPUs for 100 epochs. The batch size is 5 times the number of emotion classes. For example, if we have 8 classes and 5 time windows, the formulated triplet loss (defined in equation (6.2)) will have 8 (emotion types) $\times$ 5 (time windows) $\times$ 5 (samples) = 200 anchors. We used Adam optimizer [257] with a 0.0001 learning rate, and 0.0005 weight decay. The margin in the triplet loss (equation (6.1)) was set to 1.

6.5. RESULTS

6.5.1. EXPERIMENTAL SETUP

The experiments of incremental multimodal perception are inspired by the Gating Paradigm (*GP*) proposed in [247] and employed in [77]. It is used to test the recognition speed of emotions from a speaker's face. For example, in [77], video clips were presented to human raters in successive intervals (gates), with increasing duration. Authors of [77] noticed that human raters were more successful in recognizing emotions, as they already have seen the first gate (segments of 160 ms). In other words, in the standard GP, audio-visual cues are presented in successive segments (with increasing durations

Algorithm 3 Implementation of multimodal and incremental DML for AVER

procedure IMPLEMENTATION OF DML FOR AVER($\mathbb{D}$)▷ $\mathbb{D}$: AVER dataset

Inputs:
  $\mathbb{D}$: AVER dataset
  Define and formulate the method as shown in Figure 6.1
  Audio ($f^a(\boldsymbol{X}^{(a)})$) and visual ($f^v(\boldsymbol{X}^{(v)})$) mappings
  Mini-Batch Sampler, MWTSM algorithm, and data augmentation techniques
  Number of Epochs: NE, number of samples in a mini-batch: N, number of time windows: T, windows lengths, learning rates, lr scheduler
Initialization:
  Pre-trained audio and visual mappings, parameters initialization as described in Subsection 6.4.2
Training:
  for `epoch=1:NE` **do**
    Apply mini-batch sampler and data augmentation
    for `n = 1:N` **do**
      for *t = 1: T* **do**
        Perform hard positive and negative mining and utilize MWTSM (as elaborated in Subsection 6.3.4)
        Apply L_{tr} to capture the similarities between samples' audio and video representations
        Apply L_{mse} to reduce the gap between audio and video representations
        Keep track of the triplet sets of each time window, and avoid duplicating them in the next time windows: $t+1$
      end for
    end for
    Apply SGD guided by the formulation in equation (6.2)
  end for
Output:
  Optimized Audio ($f^a(\boldsymbol{X}^{(a)})$) and visual ($f^v(\boldsymbol{X}^{(v)})$) mappings, and LSTM cells
Evaluation
  For each sample, i, use the output of audio-visual LSTM cells ($\hat{\boldsymbol{x}}^{(i)(t)(a)}$, $\hat{\boldsymbol{x}}^{(i)(t)(v)}$) to check the performance of embeddings over time
  Apply K-Nearest Neighbor (KNN) for the classification
end procedure

as follows: 160 ms, 320 ms, etc.), in a forward manner. Next, in each gate, participants give confidence by rating emotions of each segment. Similarly, in our approach, during training and evaluation, these segments are presented to the framework, where the temporal layer, based on Long-Short Term Memory (LSTM), is accumulating and learning the contribution of each time segment. Our approach resembles the standard GP, in which participants rate segments, by evaluating the multimodal presentation in each step using the similarity measure provided by the triplet loss.

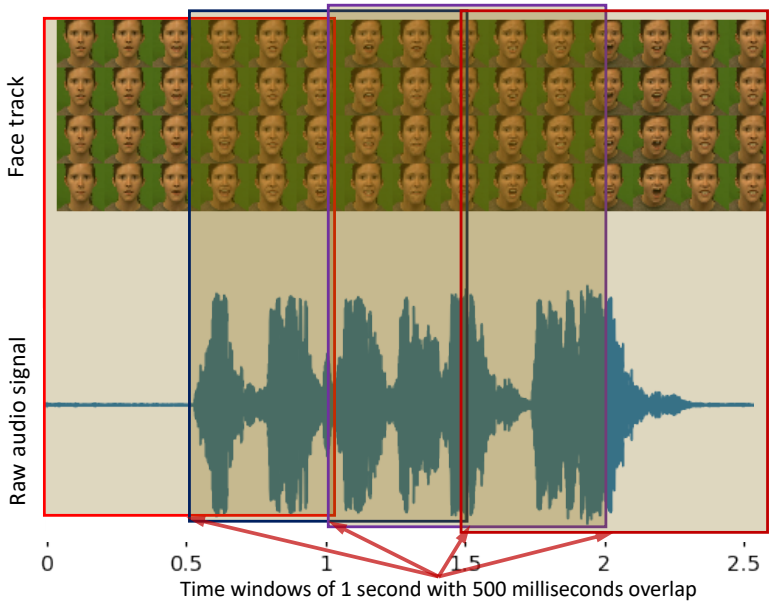


Figure 6.6: Time windows illustration.

Each audio-visual segment's length was set to either 1 or 2 seconds. As a result, at time window t, each LSTM output ($\hat{\boldsymbol{x}}^{(t)(\{a,v\})}$) includes audio and video representations up to that time window. For example, the first LSTM cell has information of the first level (i.e. the embeddings coming from the first time window/second), the second (2 seconds), the third (3 seconds), and so forth. In addition, these windows can have an overlap of 500 ms. Figure 6.6 illustrates the division and the overlap of audio-visual windows. The frame per second (FPS) rate is 30, and 16 of these frames were sub-sampled as an input for the visual mapping ($I3D$). In addition, audio signals were sampled at 48000 samples per second. Next, in each gate (LSTM cell), audio-visual embeddings ($\hat{\boldsymbol{x}}^{(t)(\{a,v\})}$) are assessed through the similarity loss (triplet loss), and the gap between the two modalities is reduced by the MSE loss (as defined in equation (6.2)).

6.5.2. EVALUATION'S SCENARIOS AND DATASETS

VALIDATION PROTOCOL

The efficiency of the proposed method is evaluated on two public Audio-Video Emotion Recognition (AVER) datasets, namely, Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [52] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [115]. To have a balanced number of samples per fold, we divide the two datasets into 10 (for CREMA-D) and 12 (for RAVDESS) folds to perform cross-validation based on subjects. In other words, for each fold, subjects' clips are either in the training or testing sets. Subsequently, for each fold, we train the proposed framework on the training folds and then test it on the remaining fold. The reported results are the average of all the folds. The reason for having different numbers of folds for cross-validation is to

have balanced samples across the folds. For instance, there are 24 subjects in RAVDESS datasets, where it is best to divide the subjects among the 12 folds since ten folds yield an unbalanced number of samples across the folds.

In this Chapter and the following Chapter, we conduct our experiments on RAVDESS and CREMA-D. Unlike eNTERFACE and AFEW, the RAVDESS and CREMA-D datasets provide extensive studies on human perception of emotions. For this reason, we focus our experimental analysis on the two selected datasets where the human perception forms a benchmark and a reference for our evaluations. In addition, this Chapter focuses on exploiting the temporal dynamics of emotions within short video clips. Studies on RAVDESS and CREMA-D provide human benchmarks of humans' incremental and temporal perception of emotions.

CREMA-D [52] is an audio-video emotion expression dataset. It contains 7442 clips from 91 actors (43 females and 48 males). Participants' age ranges between 20 and 74, and they come from a variety of races and ethnicities, i.e. Asian, African American, Caucasian, and Hispanic. Actors were asked to speak 12 sentences in five different emotions, namely, anger, disgust, fear, happiness, and sadness, or neutral. The sentences were spoken with four different levels of intensities: low, medium, high, or unspecified.

RAVDESS[115] is a multimodal emotional speech and songs database. In this work, we chose to use the speech part of the dataset as it is labeled with eight archetypal emotions: anger, happiness, disgust, fear, surprise, sadness, calmness, and neutral. This subset contains a total of 2880 recordings. More details about the CREMA-D and RAVDESS datasets are discussed in Section 3.1.

CLASSIFICATION AND EVALUATION SCENARIOS

In each time window, the LSTM cell's output (hidden representation) is considered as the corresponding embeddings, representing this time window: $\hat{\boldsymbol{x}}^{(t)(\{a,v\})}$. The dimension of each modality's embeddings is 400, as described in Subsection 6.3.1. The embeddings produced by the proposed framework are suitable to be evaluated by a simple classifier such as K-Nearest Neighbor (KNN). The evaluation is applied on the audio, video and the concatenated audio-video embeddings. K was set to 15, while the distance used is the Euclidean distance. The method is tested according to the following scenarios:

- **A baseline on middle time windows:** This baseline is formed based on the unimodal and bimodal perceptions of audio-visual cues coming from the video clips' apex moments. In particular, it uses the middle time windows (i.e. with 1 second duration). It uses SoundNet and I3D embeddings of these time windows and optimizes the mapping functions (the models) using triplet and L_{mse} losses. Hence, the temporal loss on multiple time windows is not employed in this baseline. It is important to note that since we employ single time windows, for each modality, LSTMs are not employed. This baseline is shown in Figure 6.7. The resulting audio and visual mappings are used for the evaluations: $\boldsymbol{x}^{(t)(\{a,v\})} \in \mathbb{R}^{800}$.

- **A baseline using LSTM without GP:** This baseline is based on the concatenation of audio-video embeddings and trained using a Deep Metric Learning (DML) approach. In particular, this baseline uses LSTMs on the concatenation of audio-video mappings at each time window. However, DML, using triplet and L_{mse}

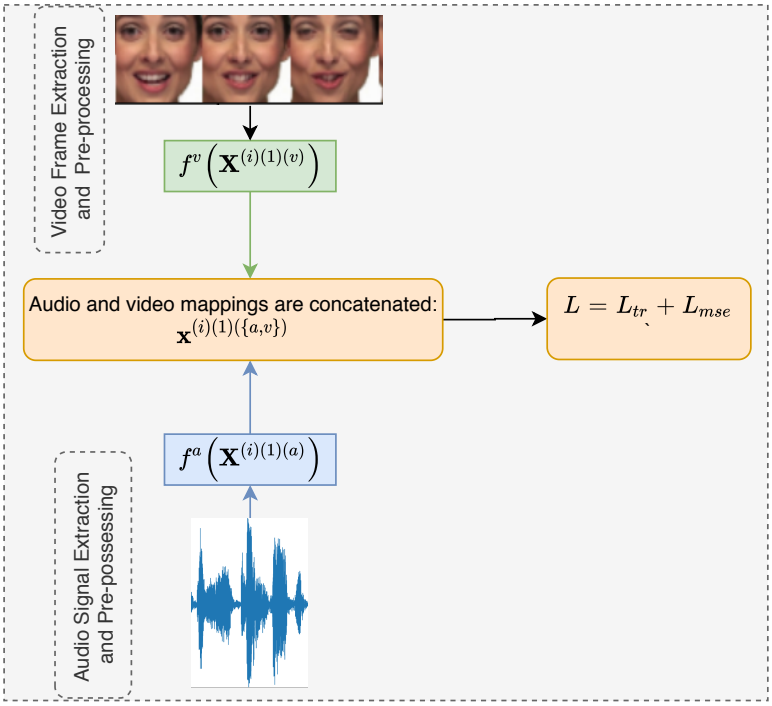


Figure 6.7: A baseline based on the middle time windows which usually considered the most expressive intervals, and hence the apex moments. The concatenated audio-visual mappings (obtained from SoundNet and I3D, respectively) are then fed onto triplet and MSE losses.

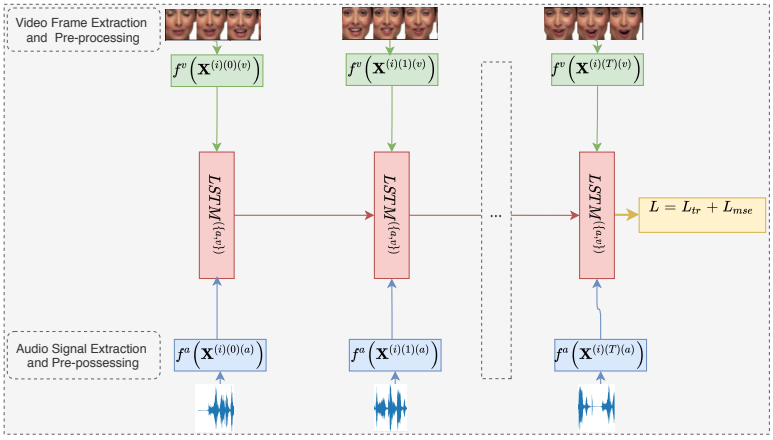
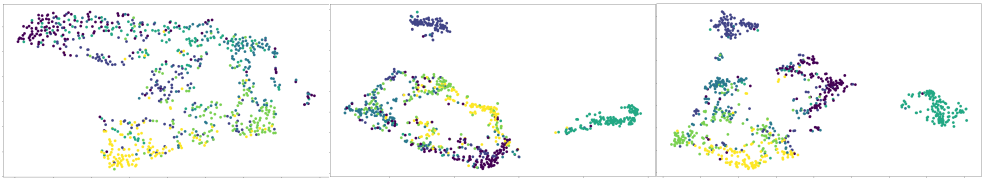


Figure 6.8: LSTM baseline. MSE and triplet losses are applied on the outputs of the last time window. The topology of this baseline is similar to the approach described in Section 6.3. However, the incremental learning, at each time window, is not applied in this case.



(a) audio (b) video (c) audio-video

Figure 6.9: t-SNE plot for a subset of CREMA-D, in the learned subspace. In (a), (b) and (c), we visualize the audio, video and the concatenated audio-video data following the proposed approach, where the clusters are well structured, and best separated when both modalities exist.

losses, is applied on the concatenated embeddings of the last time window. Figure 6.8 shows this baseline, where the audio and video mappings are concatenated. We refer to it as LSTM's feature concatenation. When evaluating this approach, the resulting audio-visual embeddings from the last time window are used: $\hat{\boldsymbol{x}}^{(t)(\{a,v\})} \in \mathbb{R}^{800}$.

- **Incremental perception with the GP:** This is based on the gating paradigm through the proposed method (our approach). It is different from the second baseline, as it applies DML on audio and video representations (without concatenations), as explained in Section 6.3 and algorithm 3. Here, the evaluations can be reported for each time window, due to the nature of the proposed procedure. As a result, to evaluate time window t, we concatenate the resulting audio-visual representations (from the output of the corresponding two LSTMs): $\hat{\boldsymbol{x}}^{(t)(\{a,v\})} \in \mathbb{R}^{800}$.

A **visualization** of embeddings from the final time window of our approach ($\hat{\boldsymbol{x}}^{(t)(\{a,v\})}$) is provided in Figure 6.9. The figure illustrates the clusters formed based on emotion classes. More importantly, we can observe that the clustering is improved when the two modalities are combined, compared to only visual or audio information. In addition, in our experiments, we observe that the contribution of visual information in the multimodal perception is greater than the audio information.

6.5.3. MODEL'S HYPER PARAMETERS EVALUATION

We evaluated the framework parameters, such as the number of cells (windows), length of audio and video inputs, and whether these inputs are overlapping or not. Table 6.1 shows the results on these parameters. Due to different lengths of video-clips in CREMA-D and RAVDESS, the number of windows was set differently. For example, the average length of video clips in CREMA-D is on average lower by 1.2 seconds than the length of RAVDESS video clips. The RAVDESS results show that increasing the number of windows with overlapping segments helps significantly to increase the performance. Overlapping has less impact due to the length of the considered audio-visual signals, which is either 1 or 2 seconds. The best performance was obtained with overlapping one-second segments, and the number of time windows was 8 and 4, for RAVDESS and CREMA-D, respectively.

Table 6.1: Tests for various configurations. RAVDESS and CREMA-D have an average of 3.82 ± 0.34, and 2.63 ± 0.53 seconds length video clips, respectively. The reported accuracies are averaged among the 10 and 12 folds of CREMA-D and RAVDESS datasets, respectively, using cross subjects validations.

Dataset	Num of windows	Window length	Overlap	Average Accuracy (%)±std
RAVDESS	2	1	✓	65.8±5.3
	2	2	✓	67.8±5.0
	4	1	✓	69.1±4.3
	4	2	✓	68.9±4.7
	6	1	✓	68.7±4.2
	6	2	✓	68.0±4.5
	8	1	✓	70.1±4.4
	8	2	✓	68.7±4.8
	2	1	✗	66.1±5.3
	2	2	✗	68.5±4.7
	4	1	✗	68.9±4.7
CREMAD	2	1	✓	73.7±3.0
	2	2	✓	73.9±3.3
	4	1	✓	74.3±3.3
	6	1	✓	73.8±3.5

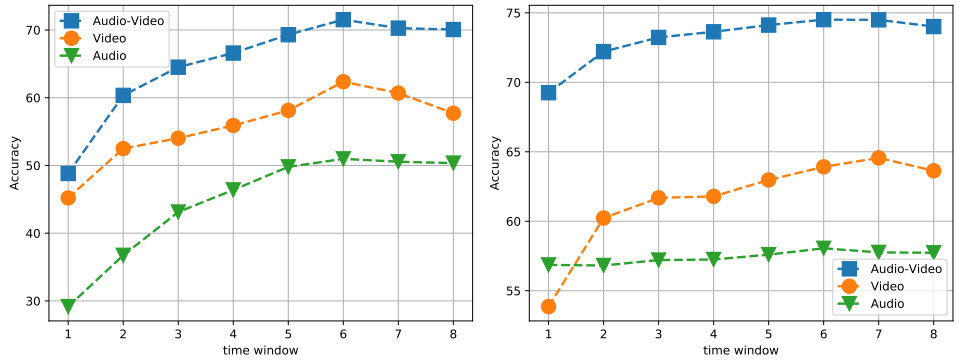
The standard deviations of the average accuracies across RAVDESS folds are higher than those of the CREMA-D dataset. In each fold, subjects' in the train and the test sets are different, giving a high variability in their emotion expressivities. Moreover, RAVDESS has fewer data samples (1440 in total) and a higher number of cross-validation folds (12) than CREMA-D, which has 7442 data samples and 10-folds for cross-validation. These facts contribute to having higher standard deviations in the RAVDESS dataset's results, making the training procedure of the proposed method a more challenging task.

6.5.4. IMPACT OF THE MULTI WINDOW TRIPLET SETS MINING

The strategy of Multi Window Triplet Sets Mining (MWTSM) helps to increase the performance and guides the training procedure. This scheme is specifically important when the model starts to over-fit the training data. The anchors of triplet sets could have different options for positive and negative samples. Based on the batch hard sampling, we selected the ones that are not previously chosen in the previous windows. In any given configuration, we noticed that the accuracy of the system increased by at least 3%. In addition, the proposed data augmentation helped the training in terms of generalization and learning process, unlike a total random augmentation that harmed the performance and prevented the system convergence.

6.5.5. UNI-MODAL AND MULTIMODAL EVALUATIONS

Figure 6.10 gives a closer look into the recognition rates of the embeddings of audio-only (AO), video-only (VO), and audio-video (AV) fusion over time. Their representations are taken from the output of LSTM cells at each gate of the proposed framework. For both



(a) RAVDESS (b) CREMA-D

Figure 6.10: AV, VO, and AO accuracies over-time for RAVDESS and CREMA-D.

datasets, AV fusion outperformed both AO and VO modalities. Most importantly, the results of multimodal perception prove that emotion recognition is a function of time, where rates are gradually ascending. Specifically, we notice that the high impact of time is more evident in the video modality and the audio-video fusion. This shows that our framework can utilize both the multimodal and the time impact for audio-video emotion recognition.

In addition, Table 6.2 presents the recognition rates (accuracies) of the visual and audio embeddings, as well as their concatenation ($\hat{\boldsymbol{x}}^{(t)(\{a,v\})} \in \mathbb{R}^{800}$) as the embeddings of the audio-video fusion. We report the performances of the embeddings for three approaches, namely, (1) the baseline based on the middle time windows, (2) the baseline which uses LSTMs to concatenate audio-visual representations (without the gating paradigm), and (3) our approach which uses LSTMs and the temporal loss on each time window. Finally, we provide the results of MERML, the published results in [243], and the recognition rates of human raters, which are reported in both datasets.

As shown in the table, perception rates increased when the two audio-video modalities embeddings are efficiently employed. For example, compared to the other two baselines, the gating paradigm increased the audio-video accuracy by at least 1.4% and 1.9%, in CREMA-D and RAVDESS, respectively. Moreover, in CREMA-D, our approach achieved good accuracy of 74.3%, which is slightly less than human-raters' recognition rate (based on relative majority). In addition, the proposed approach achieved significant results, with an accuracy of 70.1% on RAVDESS, and improved the recognition rates over the audio and video modalities. Also, on both datasets, our method achieved higher performance than the recently published results in [243]. In [243], authors employed audio and visual features using COVAREP [244] and OpenFace [245], respectively. OpenFace [245] visual representations provide the following features: Histogram of Oriented Gradients (HOGs), gaze direction, and head pose (3D position and orientation of the head). COVAREP [244] is an open-source speech analysis toolkit which extracts Prosodic, spectral, and voice quality related features. Subsequently, to fuse these fea-

Table 6.2: The recognition accuracies (%) of unimodal and multimodal embeddings, with and without the temporal and multimodal DML.

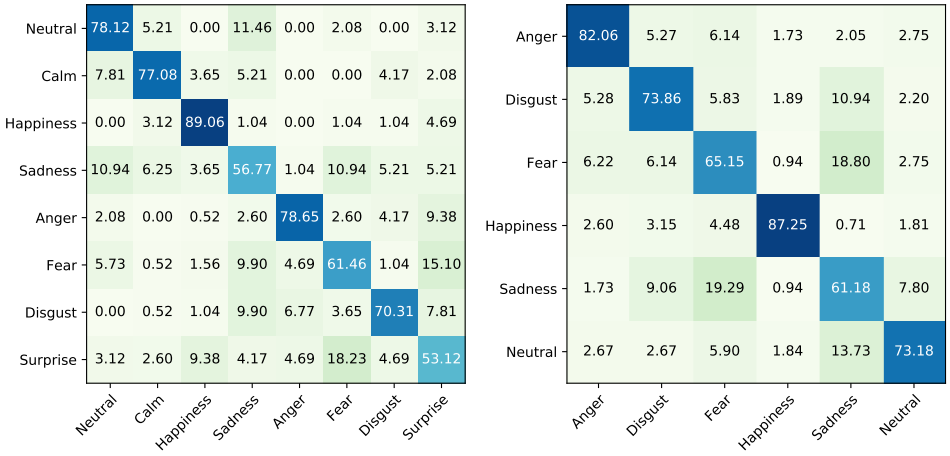
Used Embeddings	Approach	CREMA-D	RAVDESS
Audio	Global: based on middle time windows	56.4	48.0
	LSTMs without gating paradigm	50.2	48.1
	Gating Paradigm (our approach)	57.0	50.3
Video	Global: based on middle time windows	63.1	57.1
	LSTMs without gating paradigm	66.8	58.1
	Gating Paradigm (our approach)	65.0	57.7
Audio-Video	Global: based on middle time windows	69.0	68.2
	LSTMs without gating paradigm	72.9	66.2
	Gating Paradigm (our approach)	**74.3**	**70.1**
Audio-Video	Human Perception	**74.8**	**80.0**
Audio-Video	OpenFace features (V) + LSTM + COVAREP Features (A) + LSTM + Dual Attention [243]	65.0	58.3
	Multimodal Emotion Recognition Metric Learning (MERML) [168] (Chapter 5): AV	66.5	66.3
	RBF-SVM on the concatenated audio-visual representations [168] (Chapter 5): AV	65.2	67.3

tures, R. Beard *et al.* used recursive multi-attention Recurrent Neural Networks (RNNs) with dual-attention. In [243], the performance (65.0% accuracy) was obtained by combining facial and audio temporal features with LSTM.

In addition, the proposed framework in this chapter outperformed the results obtained from MERML, which is shallow metric learning for audio-video fusion, by a large margin. Moreover, the performance difference between MERML's results and the results obtained in this study shows the advantages of employing a Deep Metric Learning approach compared to shallow metric learning. Deep Metric Learning is a powerful procedure, especially when employed with end-to-end mapping functions, namely SoundNet and I3D. Also, MERML used FVs representations on handcrafted audio-features and CNN visual features. Both feature representations are not applicable in the settings we use in this chapter's study, where end-to-end mapping functions are employed for both modalities. These results, on both datasets, show how the proposed framework captures dependencies and complementary information overtime for AVER.

CONFUSION MATRICES (CMS)

Finally, for each dataset, Figure 6.11 displays the CMs between samples' actual labels and the predicted labels in the last segment (the final time window). The figure shows that the actual emotions are correctly recognized since the matrix's diagonal elements have (by far) the highest accuracies. For example, in CREMA-D, the most common misclassification occurs between fear and sadness, with 19%. Most of these observations are aligned with the study presented in [52], where the predictions are based on human raters. In RAVDESS [115], human raters (with average recognition of 80.0%) confused calmness and neutral emotions greatly, which is similar to the case in our system. On the other hand, the proposed framework eliminated most of the confusion between all



(a) RAVDESS (b) CREMA-D

Figure 6.11: CM between true and predicted labels.

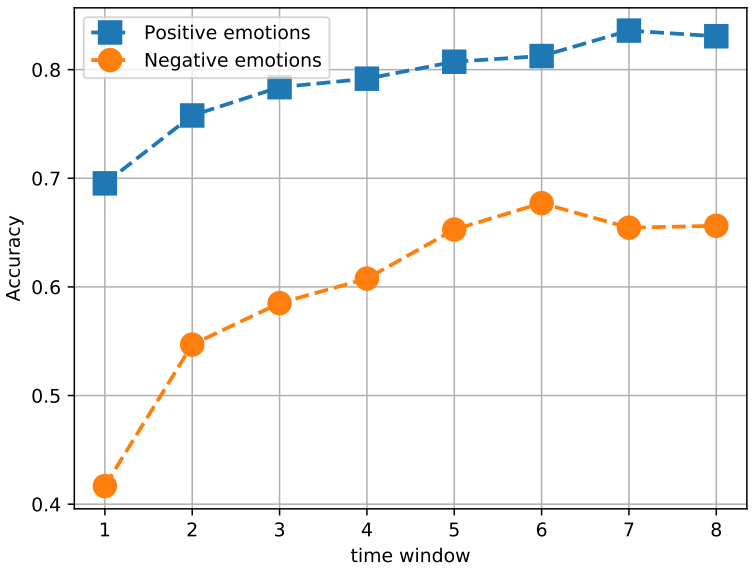


Figure 6.12: Recognition speed of positive and negative emotions over-time.

the emotions and the neutral state. According to the CM in RAVDESS, while the recognition rates at the diagonal elements are the highest, the recognition accuracy of positive emotions is higher than those of negative ones.

6.5.6. RECOGNITION OF POSITIVE AND NEGATIVE EMOTIONS

Studies in the literature suggest that the impact of temporal perception can be different when perceiving negative and positive emotions. Motivated by the study in [77], in this subsection, we examine the recognition speed of negative and positive emotions. As explained at the beginning of this section (Subsection 6.5.1), Barkhuysen et al. employed the Gating Paradigm, where video segments were consecutively presented to human raters but with incremental duration. The study found that positive emotions are recognized faster than negative ones.

In this evaluation, we use RAVDESS since it has a reasonable number of positive and negative emotions. We report the performance on its negative emotions (namely: sadness, anger, and fear) and positive emotions (namely: calm and happiness). Figure 6.12 provides the recognition speed over time for these two categories using our framework. Interestingly, we noticed that the recognition scores increase faster for positive emotions than for the negative ones. These results are in alignment and slightly similar to the findings of the study in [77], where video clips were rated by humans. Indeed, the figure shows that time has an immediate impact on the negative emotions, while the recognition plateau is reached earlier for positive emotions.

Another interesting outcome is that positive emotions are recognized with higher accuracy compared to negative ones. In the employed framework of this chapter, video modality has higher performance than audio modality (as it can be noticed in Figure 6.10). Moreover, visual information is more potent in the detection of positive emotions. For example, the studies of emotion perception by human raters in [52, 115] found that raters' recognition of happiness is high when perceiving only facial expressions. The recognition rate slightly dropped when perceiving both audio and video modalities. For example, in RAVDESS [115], human raters recognized happiness with 89% and 84% accuracies, through video only and the bimodal audio-video perception, respectively. As a result, the gap performance, seen in Figure 6.10, between the recognition accuracy of positive emotions and negative emotions can be attributed to the fact that video modality has a higher performance than audio modality. Besides, video modality has been found to have a higher contribution in bimodal perception than audio modality [52].

6.6. DEEP METRIC LEARNING FOR PERSONALITY RECOGNITION

The previous ideas and the developed models can be adjusted for another sub-field of Affective Computing (AC), namely, personality computing[1]. Studies in psychology showed that attitude, mood, and personality are directly connected to human behavioral patterns [43]. Since these human characteristics are often subtle, the affective computing field still faces several challenges. Personality computing applications achieved

[1] The study in this section was conducted in collaboration with Dario Dotti:

- D. Dotti, **E. Ghaleb**, and S. Asteriadis, "Temporal triplet mining for personality recognition," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020, pp. 379–386.

As a second author, I contributed to the design of the framework and the experimental setup.

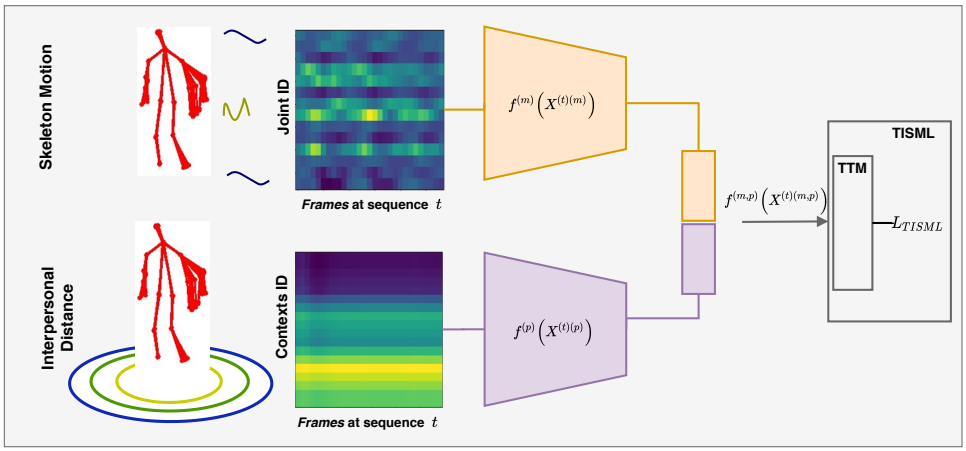


Figure 6.13: Two descriptors representing the skeleton temporal motion as well as the spatial interaction are extracted in every frame sequences t. The descriptor images show the evolution over time (x-axis) of the reference information (y-axis). The reference information is joints motion evolution for the person descriptor and proxemics to the surrounding contexts for the context descriptor. Deep CNN models are then used to obtain a compact representation of each spatio-temporal patch. The outputs of the CNN models are concatenated and fed into the proposed learning framework TISML. Temporal Triplet Mining (TTM) is employed to select temporally related positive samples encouraging the model to learn meaningful behavioral sequences that bear a higher discriminative power. Finally, a double objective loss function L_{TISML} is adopted for personality recognition and retrieval.

reliable results in analyzing faces [258], body postures [259], and multimodal information [260]. Personality recognition aims to identify personality labels given via self-assessment. In this study, we use personality labels provided by [261]. In particular, the Big Five personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) are projected onto three semantically higher categories called personality types [262] (Resilient, Overcontrolled, and Undercontrolled).

In personality computing, motion features can be applied to capture the dynamics of the human body, and proxemic features (e.g. interpersonal distances in a social context) can be employed for representing the dynamics in the scene [263]. In other words, motion and proxemic features can embed the identity and the context that correlate with personality. We employ Deep Metric Learning (DML) via triplet loss to exploit the similarity between temporally related samples and to encode higher semantic movements in order to map them into personality labels.

6.6.1. THE PROPOSED FRAMEWORK

In the previous sections, our experiments showed that DML can obtain meaningful features of short video clips to represent audio and facial cues. In this research [261], we propose a framework to encode local motion dynamics from the human body in combination with global interpersonal distances (proxemics) to encode personality-dependent behavioral patterns. Our work employs DML to map spatio-temporal descriptors to an optimized latent space, where, behaviors with discriminative power are learned and

grouped together, whereas non-informative sequences are positioned far apart. As human behaviors are very dynamic and change according to the situation, it is very difficult to find semantic similarities between them [264]. Therefore, a novel Temporal Triplet Mining (TTM) strategy, tailored for behavioral data, is proposed. We argue that taking advantage of the triplet mining scheme, short-term spatio-temporal descriptors are implicitly matched, allowing the creation of an embedding space that encodes behavioral patterns of varying sizes optimized to retrieve personality-conditioned behaviors.

Figure 6.13 shows the proposed framework's architecture, where skeleton motion, as well as proxemics descriptors, are extracted for every frame in a sequence (t). As the two descriptors capture the motion and the spatial dynamics of a sequence, two separate Convolutional Neural Network (CNN) architectures are leveraged to obtain compact representations of the input features. The obtained representations are concatenated and fed to the Temporal Identification Similarity Metric Learning (TISML) loss component. TISML aims to project the concatenated motion (m) and proxemics (p) embeddings produced by the two CNNs (which serve as mapping functions of the raw features) $f^{(m,p)}(X^{(m,p)}) : \mathbb{R}^{d^{(m,p)}}$ onto a shared feature space $\mathbb{R}^d$. Note that, in this study, $X \in \mathbb{R}^{joints \times frames}$ refers to, e.g. motion raw matrix at a sequence (t), for a predefined number of joints and frames. Similar features are positioned closely and dissimilar ones are placed far apart from each other based on data similarity and personality class. Towards this goal, within TISML, a simple but effective Temporal Triplet Mining (TTM) approach is proposed to facilitate the overall learning effort. The details of the employed motion and proxemics features can be found in [261].

6.6.2. TEMPORAL IDENTIFICATION SIMILARITY METRIC LEARNING

In this research, we introduce the notion of TISML, which is used to train the framework and consists of two major components: The first one is an identification signal based on personality labels, while the second one is a similarity signal based on Deep Metric Learning (DML). The general goal of the DML approach is to construct models that bring samples with similar labels (positive examples) closer together while pushing apart samples with different labels (negative examples). Additionally, in the training stage, our intuition is to add another constraint for the selection of positive/negative samples. We select positive samples in the temporal proximity of the anchors (within a time window) to encourage the model to generate embeddings with temporal relation while maintaining a high discriminative power for personality recognition. We assume that samples in the temporal proximity are more likely to have a semantic relation with the anchor (i.e. belonging to the same behavior) and, therefore, they can carry important information for the personality recognition task.

A given sequence (t) of motion and proxemics features (embeddings) of a subject s ($\boldsymbol{x}^{(s)(t)(m,p)}$) is associated to a discrete label ($y^{(s)}$) to estimate the corresponding subject personality label. Each frame sequence is represented by motion and proxemics embeddings $\boldsymbol{x}^{(s)(t)(m,p)} = f^{(m,p)}(X^{(s)(t)(m,p)})$, where $f^{(m,p)}$ denotes the motion and proxemics mapping function, and $X^{(s)(t)(m,p)}$ refers to the motion and proxemics raw data matrices which are fed into the mapping functions. For simplicity, we refer to $f^{(m,p)}(X^{(s)(t)(m,p)})$ as $f(X^{(s)(t)})$, which includes both motion and proxemics embeddings.

Formulation: TISML optimizes $f(X^{(s)(t)})$ to generate embeddings correlated with a

personality class. In our work, the personality recognition task is carried out using two loss functions

$$\underset{f^{(m,p)}}{\operatorname{argmin}} L_{TISML} = L_{Sim}(\boldsymbol{\Theta}_{sim}) + L_{Ident}(\boldsymbol{\Theta}_{ident}) \tag{6.6}$$

where $\boldsymbol{\Theta}s$ are the parameters of the mapping functions, which are associated with each loss function and optimized jointly. The first function is a similarity measure based on a DML loss (triplet loss) which positions semantically related embeddings closer to each other (decreasing the intra-class variations), and positions the semantically unrelated embeddings far apart (increasing the inter-class variations) [265]. Triplet loss uses triplet sets: $\{f(X^{(s)(t)}), f(X^{(s+)(t+)}), f(X^{(s-)(t-)})\}$, where $f(X^{(s)(n)})$ is an anchor (baseline), $f(X^{(s+)(t+)})$ is a positive (similar) sample to $f(X^{(s)(n)})$, and $f(X^{(s-)(n-)})$ is a negative sample (i.e. different label) to $f(X^{(s)(n)})$. As shown in eq. (6.7), the optimization procedure aims to minimize the distance between the anchor (baseline) input to a positive sample while maximizing the distance from the anchor to the negative sample within a margin [149].

$$L_{Sim}(f(X^{(s)(t)}), f(X^{(s+)(t+)}), f(\boldsymbol{x}^{(s-)(n-)}) = \|f(X^{(s)(t)}) - f(X^{(s+)(t+)})\|_2^2 - \|f(X^{(s)(t)}) - f(X^{(s-)(n-)})\|_2^2 + \text{margin} \tag{6.7}$$

The second loss in our work is an identification signal, which classifies a given embedding into one of the given personality type labels, namely, resilient, overcontrolled, and undercontrolled ($\boldsymbol{y} \in \mathbb{R}^3$). The identification signal is computed through a softmax-layer to predict the probability distribution over the labels [248]. In particular, in our work, the network is optimized via minimizing a categorical loss using softmax activation plus cross-entropy as follows:

$$L_{Ident}(f(X^{(s)(n)}), \boldsymbol{y}) = -\sum_{i=1}^{3} y_i \log \hat{y}_i \tag{6.8}$$

where $f(X^{(s)(n)})$ refers to the mapping functions that produced the motion and proxemics embeddings, $\boldsymbol{y}$ is the target class. In addition, y_i is personalities groundtruths distribution, where $y_i = 0$ for all i except $y_i = 1$ for the target personality i and $\hat{y}_i$ is the predicted probability value for personality i. Further details about Temporal Triplet Mining (TTM), and the employed features, the mapping functions, and the implementation details can be found in [261].

6.6.3. RESULTS AND DISCUSSION

Empirical experiments showed that TISML discovered meaningful behavioral patterns that improve the state-of-the-art results. Moreover, as these sequences contain a higher semantical value, they are easier to compare with respect to short term spatio-temporal descriptors. They also facilitate the discovery of critical behavioral patterns linked to the personality descriptions. The evaluations on two publicly available datasets, namely the Salsa [266] and Nonsocial [267] datasets, showed a significant increase in the performance, compared to state-of-the-art results. For example, the proposed method

achieved 75.6% and 74.9% f1-scores for Salsa [266] and non-social datasets [267], respectively. Specifically, we achieve higher results than the state-of-the-art models that use only motion by 3.8% for the salsa dataset and by 0.5% on the nonsocial dataset. When using skeleton motion and proxemics, we improve the personality recognition state-of-the-art results by 2.6% on the salsa dataset and by 2.3% on the nonsocial dataset. Furthermore, the TISML trained using a double objective loss reaches higher performance results than when trained using individual signals, proving that using a double term is beneficial to create more informative embeddings leading to better recognition performance on both datasets.

6.7. CONCLUSIONS

In this chapter, we proposed an end-to-end multimodal and temporal Deep Metric Learning (DML) for Audio-Video Emotion Recognition (AVER). Our proposed methodology embeds audio-visual cues across time, leveraging temporal information. The proposed incremental perception, based on the acquired representations from the framework, shows its efficiency at modeling the temporal context inherent to emotionally rich video sequences. Within this framework, algorithms for triplet sets mining and data augmentation were developed. The developed method and the associated techniques, such as Multi Window Triplet Sets Mining (MWTSM), contributed significantly to the stability and the performance of the framework. The obtained results are significantly higher than the baseline ones and produced high accuracies for both the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets. Our experiments have demonstrated that the incremental perception of both audio and visual cues enhances the recognition rates overtime. We noticed that increasing the number and the length of the time windows improves the accuracy of emotion recognition. Additionally, the recognition speed for positive and negative emotions differ, where positive emotions are recognized faster than the negative ones. Finally, the developed framework was adjusted for personality recognition, a challenging task due to the nature of its long-term sequential data. We employed DML for this task using motion and proxemics features. Both features contributed to personality recognition. More importantly, the experimental evaluations showed that the proposed Temporal Triplet Mining (TTM) algorithm outperformed a random selection of triplets, for Salsa and non-social datasets.

This chapter's studies demonstrate how the temporal information embedded in audio, visual, and context information is essential for emotion and personality recognition. They show the benefits of similarity-based learning when used in Deep Neural Networks (DNNs) to extract robust features and obtain efficient fusion over time. Modeling the temporal dynamics of audio and video modalities represents an important aspect of emotion recognition. The next chapter (Chapter 7) is building on outcomes of the studies in this chapter, which highlight the essence of time in emotional expressivity through facial expressions and speech prosody. In particular, the next chapter is exploiting the multimodal and temporal interaction between audio-visual channels for automatic AVER. It employs a state-of-the-art approach, namely, the attention mechanisms, to weigh the importance of each time window, for each modality, with a final goal to construct a model that attends to the right 'pieces' of information per cue.

7

JOINT MODELLING OF AUDIO-VISUAL CUES USING ATTENTION MECHANISMS

Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought.

William James [268]

Exploiting the multimodal and temporal interaction between audio-visual channels is essential for automatic audio-video emotion recognition (AVER). The importance of each modality in emotion recognition, as well as informative versus less informative time segments, can be considered through a family of methods in artificial intelligence, namely, attention mechanisms. Chapter 6 exploited the temporal and the multimodal information through Long-Short Term Memorys (LSTMs) and similarity learning, i.e., Deep Metric Learning (DML). The study aimed to capture the holistic temporal and incremental display of audio and video cues. However, in this chapter, our motivation is to spot the informative time segments of audio and video modalities through attention mechanisms. Attention mechanisms are family of a powerful approaches for sequence modeling, which can be employed to fuse audio-video cues over-time. In this manner, we incorporate both

Parts of this chapter have been published in:

- **E. Ghaleb**, J. Niehues, and S. Asteriadis, "Multimodal attention mechanism for temporal emotion recognition," in 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020, pp. 251–255.
- **E. Ghaleb**, J. Niehues, and S. Asteriadis, "Joint modelling of audio-visual cues using attention mechanisms for emotion recognition", under submission.

modalities' temporal display and mainly focus on the informative time windows automatically. We propose a novel framework that consists of bi-modal (audio and video) time windows spanning short video-clips labeled with discrete emotions. Attention is used to weight these time windows for multimodal learning and fusion. Experimental results on two datasets show that the proposed framework can achieve improved accuracies in emotion recognition, compared to state-of-the-art techniques and baseline techniques not making use of the notion of attention. The research in this chapter also introduces detailed studies and meta-analysis findings, linking the outputs of our proposition to research from psychology. Specifically, it presents a framework to understand underlying principles of emotion recognition as functions of three separate setups in terms of modalities: audio only, video only, and the fusion of audio and video. It also analyses the joint modeling of audio-visual cues and how attention helps to model their fusion to enhance multimodal recognition. Our experiments show that attention mechanisms reduce the gap between the entropies of unimodal predictions, which increases the bimodal predictions' certainty and, therefore, improves the bimodal recognition rates. Furthermore, evaluations on emotion recognition as a function of time are extensively discussed. The study shows that the middle time intervals of a video clip are essential in the case of using audio modality. However, in the case of video modality, the importance of time windows is distributed equally. Besides, we introduce visualizations to demonstrate the interactions of audio-video performances in terms of their complementarity, redundancy, or agreement in the bimodal emotion recognition. Finally, to check the framework's consistency and the attention mechanism's behavior, evaluations with noisy data in different scenarios are presented during the training and testing processes. The results show that the framework is robust when exposed to similar conditions during the training and the testing phases.

7.1. INTRODUCTION

EMOTIONS play a central role in human-human interaction [5]. As described in Chapter 1, they are highly sophisticated sub-conscious reactions, that are expressed through multiple cues, among which, the most prominent ones are visual and audio signals. Emotion-related cues are usually complementary with each other and observing both visual data (e.g. facial expression), along with voice characteristics in audio (e.g. prosodics, voice frequential components, or deep features) can help in an overall improvement of emotion perception and recognition [153]. As elaborated in Chapter 6, audio-video modalities' importance varies over time according to the expressed and perceived emotions [73]. For example, a number of works coming from the field of psychology have demonstrated that positive and negative emotions can be recognized at an early or late stage during the expression, depending on the available modalities [77]. This chapter focuses further on attending to the informative time segments in audio and visual cues for Multimodal Emotion Recognition (MER). It addresses the following research question: *how can we capture the contributions of the temporal dynamics of affect display using attention mechanism?*

A large body of research has recently shown that attention mechanisms result in great success when modeling and interpreting sequential data, with applications in machine translation [104] and natural human-machine communications [106] (e.g. chatbots). In human perception, studies show that sound could boost the awareness of visual events

through attention [269]. In many cases, people pause their activities when hearing auditory warnings. This is probably an inherent attention mechanism, expressed in the form of a natural reaction, that commonly summons human consciousness to act in a certain way. This process is applied to emotion perception as well. However, in automatic emotion recognition, there is a need for cross-modal integration of audio and visual cues to capture their multimodal interaction, which is the aim of the research in this chapter.

This research aims to model and exploit the temporal strength of audio-video cues by spotting their informative segments. For this purpose, we utilize the Transformer's self-attention mechanism [104]. As explained in Subsection 2.4.3, the Transformer is currently the state-of-the-art approach for many tasks with sequential data. We, thus, propose a novel Multimodal Attention mechanism for Temporal Emotion Recognition (MATER) framework, adapted to the needs of multimodal fusion across time windows of audiovisual cues. We address the research question of how to efficiently utilize these signals over time according to each modality's strength on emotions to maximize the automatic AVER performance.

MATER is a modality-specific framework, where learning is based on decision-level fusion. This design allows the specialization of the framework to leverage modality-specific properties in their data-stream. In this study, we investigate the benefit of attention mechanism for AVER. Besides, the model is extensively evaluated against several baselines and approaches such as Long-Short Term Memory (LSTM). We conclude, based on our experimental findings, that employing attention mechanisms can benefit computational models, as was our initial intuition.

This chapter is organized as follows. Section 7.2 introduces the related work of attention for Audio-Video Emotion Recognition (AVER). Section 7.3 presents the technical details of MATER. Section 7.4 summarizes the general experimental evaluation of the proposed approach. Section 7.5 gives an extended evaluation of the framework in terms of the performance of the multimodal fusion, modalities' interactions, the role of time in emotion perception, and the robustness of the framework when noisy data is used. Section 7.6 presents a number of training protocols for MATER and investigates the role of sequence length in recognizing emotions. In addition, it explains the re-training strategies of MATER using noisy data and investigates its robustness under challenging conditions in the two modalities and their fusion. Finally, Section 7.7 concludes the research and highlights its findings.

7.2. RELATED WORK

7.2.1. ATTENTION MECHANISMS FOR MULTIMODAL LEARNING

In a multimodal context, attention mechanisms have been applied for tasks such as Audio-Visual Speech Recognition AVSR [270], video captioning [271], and dialog systems [272]. For example, authors in [270] used transformer architectures with Connectionist Temporal Classification (CTC) loss for recognizing phrases and sentences from audio and video signals. In [271], self multimodal attention was used with LSTMs to boost video captioning by learning from audio-video streams jointly. This approach exploited the multimodal input to generate coherent sentences.

In addition, attention mechanisms have been applied for emotion recognition. For

example, authors in [273] utilized a self-attention mechanism to learn the alignment between text and audio for emotion recognition in speech. A self-attention layer was used to learn the alignment weights between speech frames and text words from different time-stamps. In addition, Wu et.al., in [158], employed transformer-based self-attention to attend the emotional autobiographical narratives. In their study, attention mechanisms were found to be powerful in a combination of Memory Fusion Network for multimodal fusion of audio, video, and text modalities. Authors in [243] proposed a recursive multi-attention with shared external memory based on Memory Networks. Their cross-modal approach showed that gated memory can achieve robust results in multimodal emotion recognition. In our work, we address emotion expression as a function of time windows and model the joint learning of audio-visual cues using the attention mechanisms. Our method not only succeeds in enhancing the multimodal recognition but, also, offers a framework to understand the behavior of temporal audio-video emotion recognition and the benefit of their joint modeling.

7.2.2. EMOTION PERCEPTION

Audio-Video Emotion Recognition (AVER) has been studied in terms of human perception using different modalities. For example, authors in [77] studied how visual cues from the speaker's face relate to emotions. The study found out that positive emotions can be detected accurately with visual information, while negative emotions are perceived more accurately using the audio modality. However, audio-visual modalities usually increase perception accuracy. Similarly, automatic emotion recognition should be capable of factoring affective states when having a multimodal presentation. Further details on the recent technologies and trends of the multimodal fusion are introduced in Chapter 3 (Section 3.3), while the technical background of the attention mechanisms is explained in Subsection 2.4.3. An extensive survey and taxonomy about the recent technologies and approaches on general multimodal fusion can be found in [153].

7.3. METHOD: ATTENTION MECHANISMS FOR EMOTION RECOGNITION

In this section, we describe the main components of the proposed method: namely, we provide details regarding the extraction of audio-visual embeddings, application of the Transformer attention mechanisms on the embeddings of time windows in a video clip, and their joint multimodal fusion.

MATER, shown in Figure 7.1, consists of two networks, with each one being dedicated to one modality. As input to a modality-specific network, we consider a time-dependent signal deriving from the embeddings of the video ($X^{(v)}$) and audio ($X^{(a)}$) modalities. We employ the encoder part of the Transformer [104] on the visual embeddings: $X^{(v)}$ and another one on the audio embeddings: $X^{(a)}$. The embeddings are obtained from VGG models, and the applied audio and video encoders have exactly the same architecture. The novel bi-modal framework aims to study the temporal presentation of audio-visual cues for emotion recognition. Using these sub-networks on the visual and audio embeddings, the model is optimized to learn the proper class of the given video clip.

As discussed in [274, 275], multimodal deep learning is a challenging task, due to

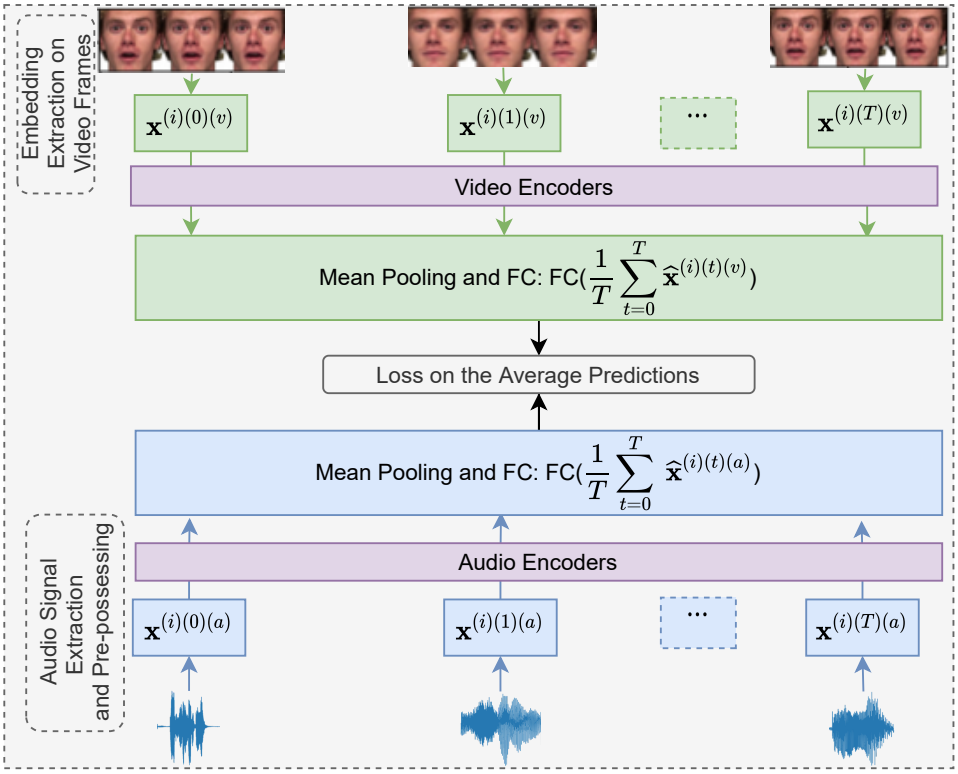


Figure 7.1: An illustration of the proposed framework, Multimodal Attention mechanism for Temporal Emotion Recognition (MATER), for Audio-Video Emotion Recognition (AVER). It has two data streams, composed of two sub-networks, applied on raw audio ($f^{(a)(i)}(x^{(a)(i)})$) and video ($f^{(v)}(x^{(v)})$) data coming from a video clip, i.

the increased capacities of Deep Neural Networks (DNNs), in the case of more than one modality exists. By the increased capacities, we refer to the usage of different DNN models for each modality. For instance, having dedicated models for each modality increases the number of learnable parameters in a given network. For this reason, in this study, the architecture is based on a late joint fusion to avoid overfitting one modality and to allow the two sub-networks to generalize at different rates. In addition, from an analytical point of view, the design of MATER is based on the following motivations:

- Emotion display consists of on-set, apex, and off-set phases, while the apex captures the maximum expressivity, thus, it is the segment considered in most research works [73]. Nevertheless, it is better not to pre-define these phases, since they depend on the emotions and the presented modalities. MATER is specialized in exploring and utilizing modalities' correlation with emotions on these phases for robust performance.
- Research has demonstrated that emotion perception might require a different amount of time for an accurate detection [73], depending on the expressed emo-

tion and involved modalities. Thus, these alterations could be exploited efficiently through a temporal framework.

7.3.1. INPUT MODALITIES' EMBEDDINGS

In AVER, a dataset ($\mathbb{D}$) contains n short video clips with **a**udio and **v**isual (video) modalities, and each clip is annotated with a discrete emotion c

$$\mathbb{D} = \{(\boldsymbol{x}^{(v)(1)}, \boldsymbol{x}^{(a)(1)}, c^{(1)}), (\boldsymbol{x}^{(v)(2)}, \boldsymbol{x}^{(a)(2)}, c^{(2)}), ..., (\boldsymbol{x}^{(v)(n)}, \boldsymbol{x}^{(a)(n)}, c^{(n)})\}$$

where $\boldsymbol{x}^{(a,v)}$ are the embeddings extracted from the audio or video raw-data. In this work, we consider non-overlapping time windows of 0.25 and 0.5 seconds, as inputs to the audio and visual models for embeddings extraction. These embeddings are then normalized with l_2*-normalization* to have zero mean and unit length.

VIDEO EMBEDDINGS

In each time window of a video clip, faces are detected and tracked using the Dlib library [239, 240]. Subsequently, faces are cropped to 96×96 resolution. A pre-trained VGG-M model [97, 276] on the Facial Emotion Recognition (FER) dataset [185] is used to extract representations of a given facial image. VGG-M is explained in Subsection 2.4.1, as part of the VGG models. We used the output from the final convolutional layer, which corresponds to a 512-dimensional vector. As these representations are for each frame, we found out that mean-pooling through time window frames' features has resulted in a good representation. Another alternative pooling scheme can be max-pooling, however, we observed that this scheme was inferior to the adopted one.

AUDIO EMBEDDINGS

We extract audio embeddings for a time window using VGGish [277]. VGGish is a variant of VGG models, which was trained to generate high level and semantically useful embeddings for audio recordings. It was pre-trained with the YouTube-8M dataset [134], and we use the output of the last convolutional layer, which corresponds to a 512-dimensional vector. These embeddings can be fed onto a downstream classification model. VGGish was trained with audio data using a 16 kHz mono sample rate. Specifically, a spectrogram is computed using magnitudes of the Short-Time Fourier Transform (STFT) with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window [277]. In our case, as the time windows length is either 0.25 or 0.5 seconds, the audio input size contains either 24×64 or 48×64 log mel spectrograms. Each example covers 64 mel bands and 48 or 24 frames of 10 ms each. These inputs were adapted to fit the requirements of the proposed MATER framework.

7.3.2. FRAMEWORK'S COMPONENTS

MATER employs the attention mechanism of the Transformer, which was introduced in [104]. The Transformer is a neural network architecture that has an encoder-decoder structure. The architecture of the Transformer is explained in detail in Subsection 2.4.3. In our study, we employ the encoder part of the Transformer on each modality's time segments. Figure 7.2 shows the components of an encoder in the Transformer. The encoder consists of a Multi-Head Self Attention (MHSA) layer and is followed by an element-wise

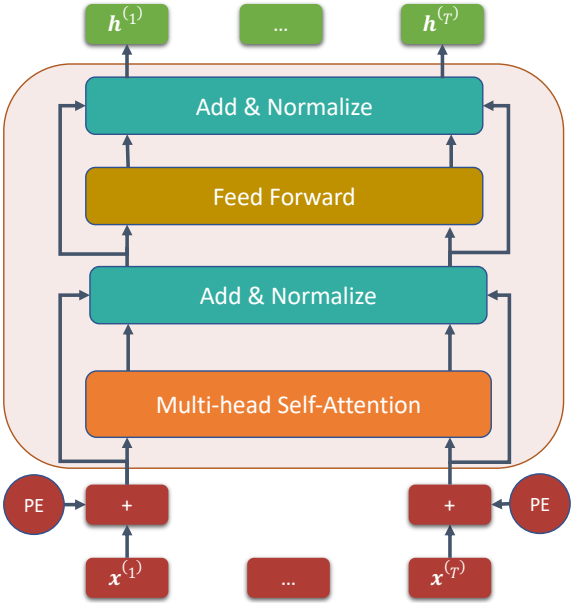


Figure 7.2: The encoder part of the Transformer is used for each modality. Note that, in our framework, we stack 6 encoder layers.

feed-forward layer. As suggested in the original work introducing the topology of Transformer mechanisms [104], we also use 6 stacked encoder layers. The following subsections detail the inputs of the audio-video encoders, explain the positional encoding operation, and then elaborate on the usage of MHSA within MATER.

AUDIO-VIDEO INPUTS

As input to each sub-network (audio and visual encoders), we consider audio-visual embeddings, $X^{(m)(t)}$, where m refers to a modality: $m \in \{a, v\}$, and t represents a time window: $t \in \{1, 2, ...T\}$. T is the number of time windows in a video clip (as shown in Figure 7.3). As a result, each sub-network of a modality has a sequence of embeddings,

$$X^{(m)} = \{\boldsymbol{x}^{(m)(0)}, \boldsymbol{x}^{(m)(1)}, ..., \boldsymbol{x}^{(m)(T)}\}$$

, as an input to its encoders, which attends to each time window "token" with a different weight. MHSA helps the model to learn representations from different time segments, in both modalities. Following the MHSA layer, each output is fed onto the position-wise feedforward layer, independently for each time window.

POSITIONAL ENCODING (PE)

As explained in Subsection 2.4.3, the Transformer does not make use of recurrence or convolutional operations; rather, it adopts Positional Encoding (PE) in order to make use of ordinal information in a sequence. In particular, "positional encodings" are added to the sequential input of the encoder, e.g., in our case, the embeddings of audio-visual

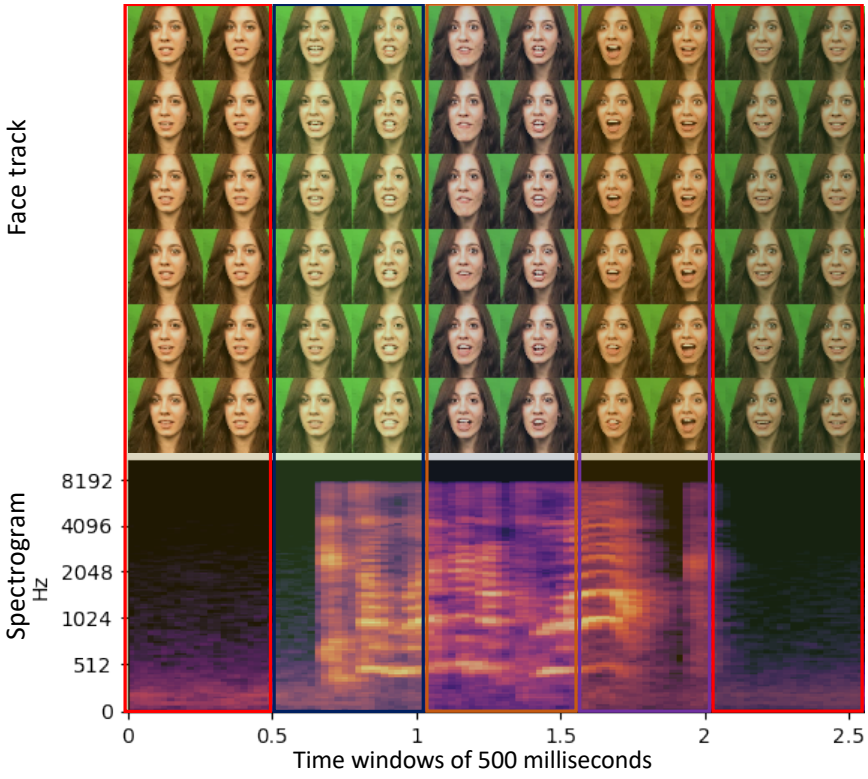


Figure 7.3: An example of non-overlapping time windows of 500 milliseconds.

time windows of a video clip. This addition (sum) operation is applied once, before the flow of the audio-visual inputs to the encoders. Besides, they have the same dimensions (d) as the input embeddings to facilitate their sum. The authors of the Transformer [104] proposed to employ PE using sine and cosine functions, which are fixed ones, with variant frequencies as follows:

$$\begin{aligned} pe_{(t,2i)} &= sin(t/10000^{2i/d}) \\ pe_{(t,2i+1)} &= cos(t/10000^{2i/d}) \end{aligned} \tag{7.1}$$

where t indicates a time window and i refers to a specific dimension in the embeddings of this time window.

MULTI-HEAD SELF-ATTENTION (MHSA)

The input of an encoder flows through a self-attention layer. The self-attention employed in the Transformer is used to assign a weighing score for each token (time window) in a time-series. In our case, within the proposed framework, the tokens are the given embeddings of each time window in each modality. In a video clip, self-attention focuses on specific time windows where emotion expression is strong by automatically

assigning activations (weights) to these time windows. In addition, as explained in section 2.4.3, attention mechanisms help DNN models to learn context related to time and proximity of sequential inputs, e.g. the audio-video time windows of a video-clip : $[\boldsymbol{x}^{(m)(1)}, ..., \boldsymbol{x}^{(m)(T)}]$. A self-attention layer aims to weigh these vectors with respect to each other and results in the following weighted outputs: $[\boldsymbol{h}^{(m)(1)}, ..., \boldsymbol{h}^{(m)(T)}]$, where, e.g., $\boldsymbol{h}^{(m)(2)}$ is a weighted vector over all the input sequence. For instance, as shown in Figure 7.3, the embeddings of the facial expressions in the second time window can be associated with the ones in the middle time windows due to their proximity.

As described in Subsection 2.4.3, we can use the concepts of queries, keys, and values from information retrieval when considering the computation of attention mechanisms. In particular, the computation of attention can be considered as mapping a set of target vectors (*queries*) with a set of candidate vectors (*keys*). Subsequently, the scores resulting from these mappings are used to compute the weighted combination of the *values*, where the scores indicate the compatibility (similarity) of each *key* with the *query*. In our case, a query can be embeddings at t time window. At the same time, the set of keys and the values are the whole sequences of the modalities embeddings (all the time windows). Moreover, the authors in [104] proposed using the "scaled dot-product attention", which is formulated as follows:

$$Attention(Q^{(m)}, K^{(m)}, V^{(m)}) = softmax(\frac{Q^{(m)} K^{(m)T}}{\sqrt{d_k}}) V^{(m)} \tag{7.2}$$

Note that, in equation (7.2), queries ($Q^{(m)}$), keys ($K^{(m)}$), and values ($V^{(m)}$) matrices are created from the same input in a sequence. This is due to the fact that the encoder part of the Transformer employs a self-attention mechanism, by attending to its input sequence, $X^{(m)}$. In addition, since we employ two encoders, for audio and video modalities separately, the scaled dot-product attention is applied on each modality accordingly.

Moreover, MHSA is a key component in the Transformer architecture. As illustrated in Figure 7.2, following the addition of the PEs to the audio and video sequence embeddings, the resulted embeddings are fed forward through the MHSA layer. Specifically, MHSA splits the learning loads to learn context information over several heads. In particular, for the queries, keys, and values, we learn linear projections with d_q, d_k d_v dimensions, respectively. In the encoder part of the Transformer, these dimensions are the same. For example, in our case, each time window's embeddings in a video clip, $\boldsymbol{x}^{(m)(t)}$, has 512 dimensions (i.e. $d = 512$), and we use 8 attention heads. As a result, in a head (i), the linear projection matrices are as follows: $W_i^{(q)(m)}$, $W_i^{(k)(m)}$, and $W_i^{(v)(m)} \in \mathbb{R}^{d_k \times d}$.

Practically, for each modality, MHSA is applied on the input of queries ($Q^{(m)}$), keys ($K^{(m)}$), and values ($V^{(m)}$), which are created (copied) from the modalities input matrices: $X^{(m)}$. Subsequently, the resulting outputs from different attention heads, with d_k dimensions each, are concatenated and projected again (with $W^{(o)(m)}$ linear projection) to obtain the final weighted matrices. These matrices are used in the following sublayer, namely: element-wise feedforward sublayer, which is explained in Subsection 2.4.3. To summarize, MHSA computations are performed as follows:

$$\begin{aligned} MHSA(X) &= W^{(o)(m)}(concatenate(head_1^{(m)}, ..., head_h^{(m)})) \\ \text{where } head_i^{(m)} &= Attention(W_i^{(q)(m)} Q^{(m)}, W_i^{(k)(m)} K^{(m)}, W_i^{(v)(m)} V^{(m)}) \end{aligned} \tag{7.3}$$

where $W_i^{(q)(m)}, W_i^{(k)(m)}$, and $W_i^{(v)(m)}$ are learnable linear transformations that help the self-attention mechanism to get stronger representations and exploit the context in a given sequence of audio-video time windows.

FUSION: PREDICTION LAYERS

On the final output of the last modalities' encoder layers, hidden representations $\boldsymbol{h^{(t)}}$ are obtained for each time window. Note that, in our framework, the last encoder layer is the sixth encoder layer, since we used six encoder layers for each modality. Subsequently, over the input sequence T, we apply a mean pooling for each modality, separately, in order to get the final audio and video representations:

$$\hat{\boldsymbol{h}}^{(v)} = \frac{1}{T}\sum_{t=0}^{T} \boldsymbol{h}^{(v)(t)} \text{ and } \hat{\boldsymbol{h}}^{(a)} = \frac{1}{T}\sum_{t=0}^{T} \boldsymbol{h}^{(a)(t)} \tag{7.4}$$

Two fully connected (FC) layers are then applied on the resulting audio ($\hat{\boldsymbol{h}}^{(a)}$) and video ($\hat{\boldsymbol{h}}^{(v)}$) representations as the prediction layers. The predictions from the two modalities are averaged, and the network is optimized accordingly,

$$predictions = \frac{1}{2}\sum_{m\in\{a,v\}} (W^{(m)})\hat{\boldsymbol{h}}^{(m)} + \boldsymbol{b}^{(m)}, \tag{7.5}$$

where $W^{(m)}$ and $\boldsymbol{b^{(m)}}$ are the parameters of a fully connected layer. These averaged predictions are normalized via softmax operation and are used to compute the cross-entropy loss. The late-joint fusion paradigm guides the optimization of the network to avoid overfitting one modality and helps the bi-modal encoders to learn at different paces.

7.4. RESULTS

This section presents the experimental setup, implementation details, and the general evaluation metrics and results of Multimodal Attention mechanism for Temporal Emotion Recognition (MATER). The proposed framework's efficiency is evaluated on two public multimodal emotion recognition datasets, namely Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [115] and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [52].

RAVDESS has two sets: speeches and songs subsets. We use the speech set as it is labeled with eight archetypal discrete emotions [15]: anger, happiness, disgust, fear, surprise, sadness, calmness and neutral. The dataset has 24 subjects, 12 males and 12 females, with an age range of [21,33]. It contains of short speech video-clips of an average of 3.82 ± 0.34 seconds. The total number of videos is 1444.

CREMA-D contains 7442 video clips of 91 subjects. The video clips' average duration is 2.63 ± 0.53 seconds. Each video is labeled with five basic Ekmanian emotions: anger, disgust, fear, happiness, and sadness, or neutral. In each video clip, emotion expression can have one of the following four different levels (intensities): low, medium, high, and unspecified. The dataset includes people with diverse backgrounds in terms of gender, ethnicities, and age. More details about the CREMA-D and RAVDESS datasets are discussed in Section 3.1.

7.4.1. TRAINING DETAILS

MATER was optimized during the training phase using Adam optimizer [257], which is a variant of Stochastic Gradient Descent (SGD). Cross-entropy loss is used in this optimization. We use a batch size of 64 and the framework was trained for 300 epochs. Initially, the learning rate (lr) was set to $1e^{-6}$ and it was reduced if it reaches a plateau state after 20 epochs. The detailed results are provided in the following sections.

Evaluation Protocols: For both datasets, we use subject disjoint k-fold cross-validation. To have an equal number of subjects per fold, RAVDESS and CREMA-D were divided into 12 and 10 folds, respectively. In each fold, a subject's samples are either in a testing or a training fold. This is applied to ensure that the MATER is optimized to learn based on emotion labels rather than over-fitting subjects and their emotions. In addition, training and evaluations have been conducted separately on each dataset. The reason for having different numbers of folds for cross-validation is to have balanced samples across the folds. For instance, there are 24 subjects in RAVDESS datasets, where it is best to divide the subjects among the 12 folds since ten folds yield an unbalanced number of samples across the folds. Moreover, similar to Chapter 6, we conduct our experiments on RAVDESS and CREMA-D. The RAVDESS and CREMA-D datasets provide extensive studies on human perception of emotions, forming a benchmark and a reference for our evaluations. The analysis of human perception is useful for our studies since this Chapter focuses on spotting the temporal dynamics of emotions within short video clips on providing a meta-analysis of automatic perception of emotions.

7.4.2. BASELINE MODELS AND RESULTS

To evaluate MATER, we built baseline models for analytical comparisons, which examine the role of attention in audio-visual (AV) emotion recognition. MATER consists of time windows based audio-visual embeddings, 6 stacked audio-visual encoders (in which, each one has positional encoding, multi-head self-attention, and feedforward layers with their residual connections). To check the impact of the components of MATER, MHSA, PE, or both are removed from the baseline models. The baseline models and the attention model (where MHSA and PE are kept) have the same number of layers and use the same settings in terms of audio-visual embeddings. The six stacked encoders' feedforward layers are kept which makes it a strong baseline and provides a fair comparison. In addition, the flow of the embeddings, the training, and optimization processes are similar across the experiments. In other words, the baseline represents the case of averaging time windows without weighing their importance for each modality.

The comparisons, which are presented in Table 7.1, aim to check the research's goal regarding the weighing mechanism that the attention scheme provides. In addition, it examines the role of PE in the framework, where PEs are added to the embeddings. These comparisons were tested on different numbers of time windows. Due to different lengths of video-clips in CREMA-D and RAVDESS, the number of windows was set differently. We use sets of {8, 16} and {6, 12} time windows for RAVDESS and CREMA-D, respectively. For fair comparisons, across all evaluation scenarios, the depth of the framework was kept the same. As shown in Table 7.1, we notice that the best performance on both datasets is achieved when using MATER with PE and attention (provided through MHSA), where the accuracy reaches 76.3% and 67.2% for RAVDESS and CREMA-D, respectively. PE en-

Table 7.1: Model's accuracies for various scenarios. RAVDESS and CREMA-D have averages of 3.82 ± 0.34, and 2.63 ± 0.53 seconds length video clips, respectively. Here, Multi-Head Self Attention (MHSA) represents the attention mechanisms within the proposed framework.

Dataset	#windows	Duration (seconds)	PE	MHSA	Average Accuracy (%)±std
RAVDESS	8	0.50	✓	✓	**76.3**±3.6
	8	0.50	✓	✗	70.6±6.2
	8	0.50	✗	✓	75.2±4.4
	8	0.50	✗	✗	69.4±5.4
	16	0.25	✓	✓	74.4±3.8
	16	0.25	✓	✗	68.8±5.6
	16	0.25	✗	✓	72.4±4.2
	16	0.25	✗	✗	65.2±6.0
CREMA-D	6	0.50	✓	✓	**67.2**±3.6
	6	0.50	✓	✗	64.4±3.6
	6	0.50	✗	✓	65.0±3.4
	6	0.50	✗	✗	61.8±3.2
	12	0.25	✓	✓	66.4±1.8
	12	0.25	✓	✗	62.3±3.6
	12	0.25	✗	✓	63.6±2.8
	12	0.25	✗	✗	58.3±3.3

Table 7.2: Audio-Video average accuracies (%) of MATER and other related work.

Approach	**CREMA-D**	**RAVDESS**
Human Perception: AV	*74.8*	*80.0*
Dual Attention with LSTM: AV [243]	65.0	58.3
Multimodal Emotion Recognition Metric Learning (MERML) [278] (Chapter 5)	66.5	66.3
Visual (I3D) [253] + Audio (SoundNet [135]) + LSTM + triplet loss [279] (Chapter 6)	**74.3**	70.1
Visual (I3D) [253] + Audio (SoundNet [135]) representations [279] (Chapter 6) + attention + softmax loss	74.1	74.6
MATER: $AV_{+PE+MHSA}$	67.2	**76.3**

hances the performance since it provides the topology with temporal information, where the improvement over using only the attention is at least 1%. This information is further utilized through the the Multi-Head Attention layer. Moreover, PE's impact is more obvious when the number of time windows is large. For example, in the case of RAVDESS, with $T = 16$, PE improved the performance by 2.6%.

In the baseline results of the framework, in case of not using both PE and attention, we observe that the performance drops by at least 5% and 3% for RAVDESS and CREMA-D, respectively. This gap increases when the number of time windows is doubled, where the improvement reaches at least 8%.

COMPARISONS TO OTHER METHODS

Previous work results in both datasets, including human performance, are presented in Table 7.2. MATER results in this table are obtained using the attention (of MHSA), with 8, and 6 time windows for CREMA-D and RAVDESS, respectively. In CREMA-D, our approach outperformed the recently published results in [243, 278]. However, it gave lower performance accuracies than the ones based on human-perception (through relative majority). Besides, in RAVDESS, the proposed topology resulted in higher accuracy than those in the literature but less than the recognition rate obtained through human perception. In [243], the performance (65.0% and 58.3.% accuracies, for CREMA-D and RAVDESS, respectively) was obtained by combining facial and audio temporal features with LSTM using Dual-Attention. In [278], a metric learning approach (MERML) was applied to fuse audio-video modalities. MATER's results show its efficiency for enhanced joint multimodal learning and fusion. Another reason behind these improvements is that MATER deals with the interaction of the multi-modal data over-time using the time windows segments. This makes the framework weigh and evaluate the importance of the two modalities per emotion across time.

In our previous work, in Chapter 6, the procedure applied an end-to-end learning from raw audio and visual data using SoundNet [135] for audio and I3D [253] for visual mapping. These paradigms achieved accuracies of 74.3% and 70.1% on CREMA-D and RAVDESS, respectively. Besides, when replacing Long-Short Term Memory (LSTM) by the Transformer's encoders and attention mechanism, and the triplet loss of the Deep Metric Learning (DML) by softmax loss function (in particular, we use cross-entropy), we obtained high performance with 74.1% and 74.6% accuracies on CREMA-D and RAVDESS, respectively. Nonetheless, we notice that replacing the audio-visual mappings of SoundNet [135] and I3D [253], with the extracted embeddings in this study has the following advantages:

- In this chapter, we employed smaller time windows, in comparison to our previous study. Also, SoundNet and I3D work best with larger time windows in the end-to-end learning paradigm. Nonetheless, the *interpretability* of the attention mechanism on top of these deep and large models was not feasible. As a result, we opt for replacing the audio and visual mappings with the extracted embeddings as elaborated in Subsection 7.3.1.

- MATER on the pre-extracted embeddings of small time windows offers *explainability* of attention mechanism for fusing audio-visual cues and spotting the important time segments. Subsequently, using MATER, it is attainable to analyze the obtained results extensively. In fact, this chapter presents detailed studies and meta-analysis findings, linking the outputs of our proposition to research from psychology (as will be analytically explained in the following sections).

ABLATION STUDY

Table 7.3 introduces the accuracies of the underlying Audio (A) and Video (V) modalities, within the framework. These results show the sub-modalities' contribution in the performance of the framework. They show that the accuracy is significantly increased when using both audio and video modalities. While attention has only small gains on

Table 7.3: Multimodal and individual performance of MATER with and without attention.

Dataset	Attention	Audio Modality	Video Modality	Audio and Video Modalities
RAVDESS	✓	59.2	58.2	76.3
	✗	60.7	56.0	69.4
CREMA-D	✓	57.5	51.7	67.2
	✗	56.0	49.0	61.8

the individual modalities, its strength is shown in the multimodal case. For instance, we can see improvements of at least 10%, over the uni-modal perception, and 6% over the baseline. Notably, attention helps in the multimodal fusion due to the weighing mechanism of the modalities overtime. This also highlights the essential role of multimodal perception in obtaining higher emotion recognition results.

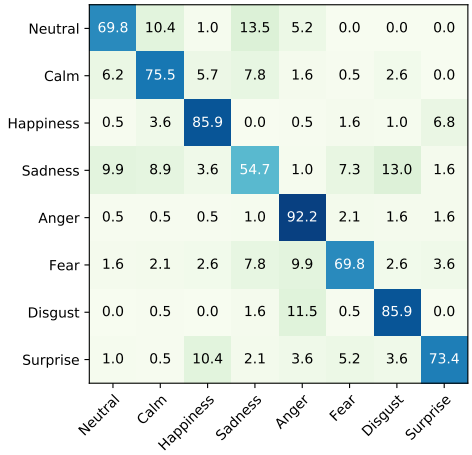
CONFUSION MATRICES

Confusion Matrices (CMs) displayed in Figure 7.4 show the achieved performance of our approach on RAVDESS and CREMA-D classes. The x-axis represents the intended emotions, and the y-axis shows the predicted emotions. Without exception, the diagonal elements have the highest accuracies, which indicates the high classification accuracy of the intended emotions. More importantly, the improvement margin over the baseline is more obvious in emotions such as anger and neutral.

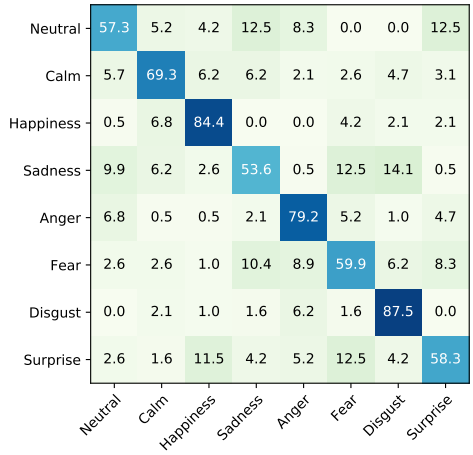
In terms of MATER performance on emotions, anger, neutral, disgust, and happiness have higher accuracy detection, compared to fear and sadness. While we notice that, e.g., fear and sadness were confused with other emotions in varied ratios. These results are also compatible with the reported human perception and confusion in [52, 115].

7.5. EXTENDED ANALYTICAL RESULTS

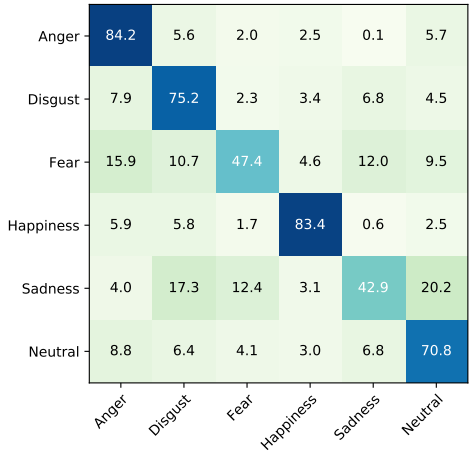
In the previous sections, the analysis focused on unimodal and multimodal emotion recognition. In this subsection, our aim is to study the impact of temporal information on emotion recognition accuracy, as well as the impact of attention mechanisms mostly from a qualitative point of view, instead of merely performance-driven experiments. In addition, we evaluate the performance in case of injecting noise during the evaluation to check its response when exposed to certain conditions during the training and the testing processes. Also, this section explores the reasons behind the disparity between the increase in performance for unimodal versus multimodal cases (see Table 7.3, where the increase using attention mechanism is significantly higher when both modalities are involved). Indeed, the proposed framework offers various aspects to study, and this section adds further evaluations, observations, and validations on its performance for emotion recognition. For example, the analytical results explore the interaction between the audio and video modalities regarding their complementarity and redundancy for bimodal emotion recognition. In addition, we study the temporal performance of audio-visual cues and the impact of their fusion over time. These evaluations are pre-



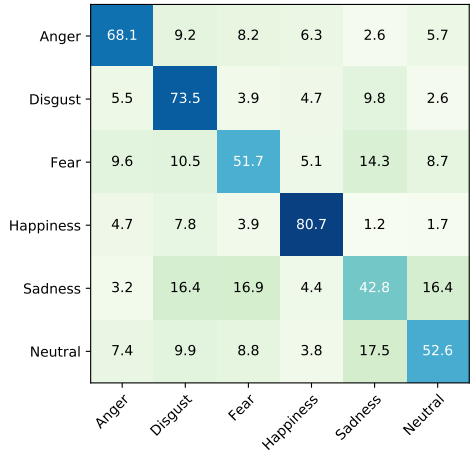
(a) RAVDESS: with attention



(b) RAVDESS: without attention



(c) CREMA-D: with attention



(d) CREMA-D: without attention

Figure 7.4: Confusion matrices between true and predicted labels.

sented for both cases, which depend on whether the attention-mechanism with PE is employed or not. In the following results, we refer to both approaches simply by "with attention" and "without attention" approaches, where the "without attention" implies the baseline as described in Subsection 7.4.2.

7.5.1. MIXTURE OF EMOTIONS

Emotion expression could be ambiguous, even for humans perception [3, 52]. The analysis, in this subsection, aims at evaluating this ambiguity per-emotion and per-modality.

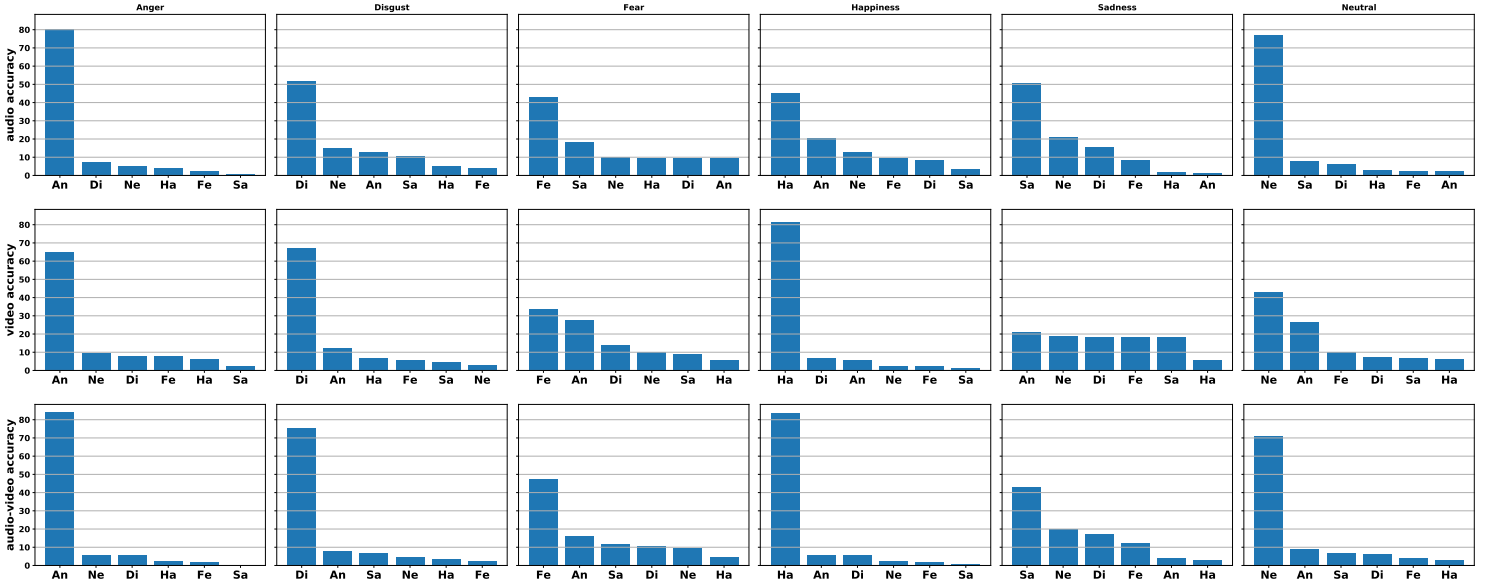


Figure 7.5: CREMA-D with attention and T = 6. This figure shows sorted confusion per emotion for each modality. Due to subfigures size, labels are referred to as follows: Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Neutral (Ne), and Sadness (Sa).

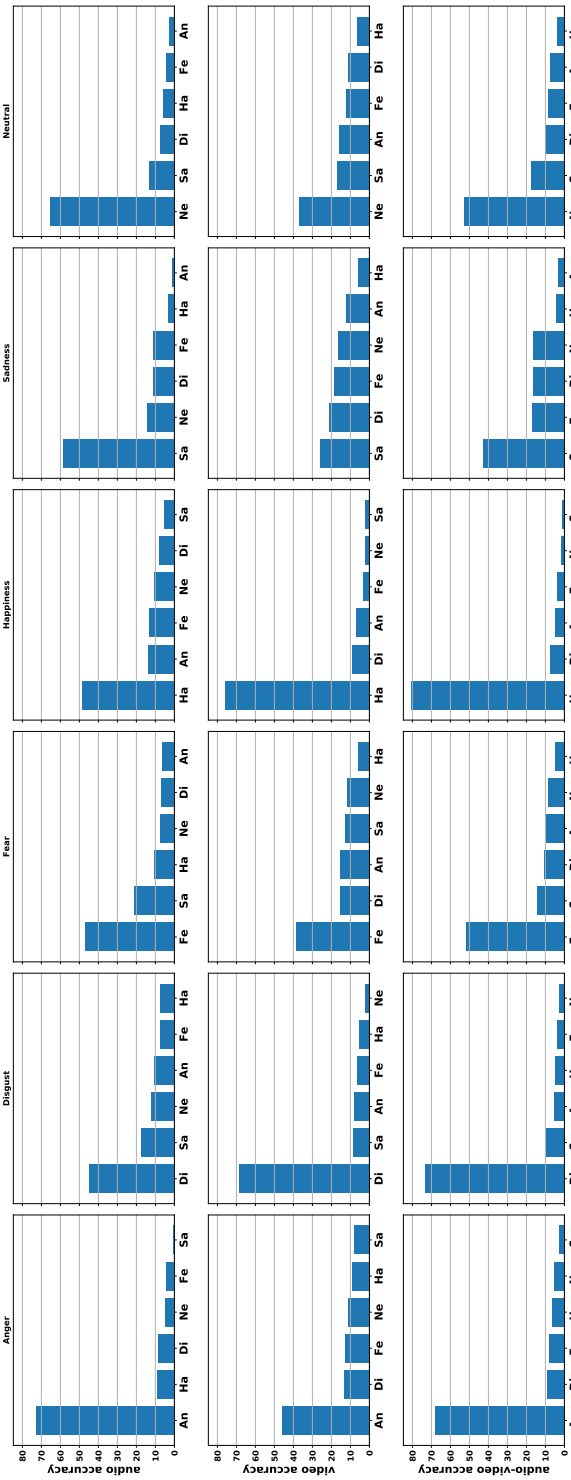


Figure 7.6: CREMA-D without attention and T = 6. This figure shows sorted confusion per emotion for each modality. Due to subfigures size, labels are referred to as follows: Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Neutral (Ne), and Sadness (Sa).

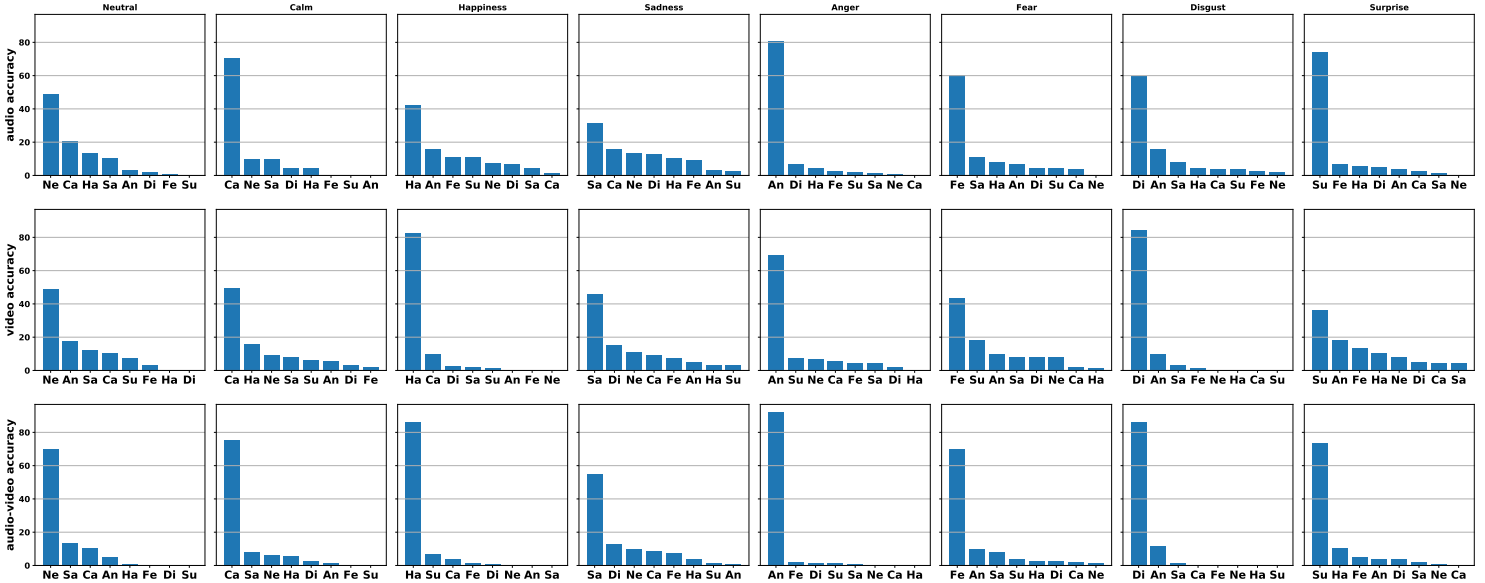


Figure 7.7: RAVDESS with attention and T = 6. This figure shows sorted confusion per emotion for each modality. Due to subfigures size, labels are referred to as follows: Anger (An), Calmness (Ca), Disgust (Di), Fear (Fe), Happiness (Ha), Neutral (Ne) Sadness (Sa), and Surprise (Su).

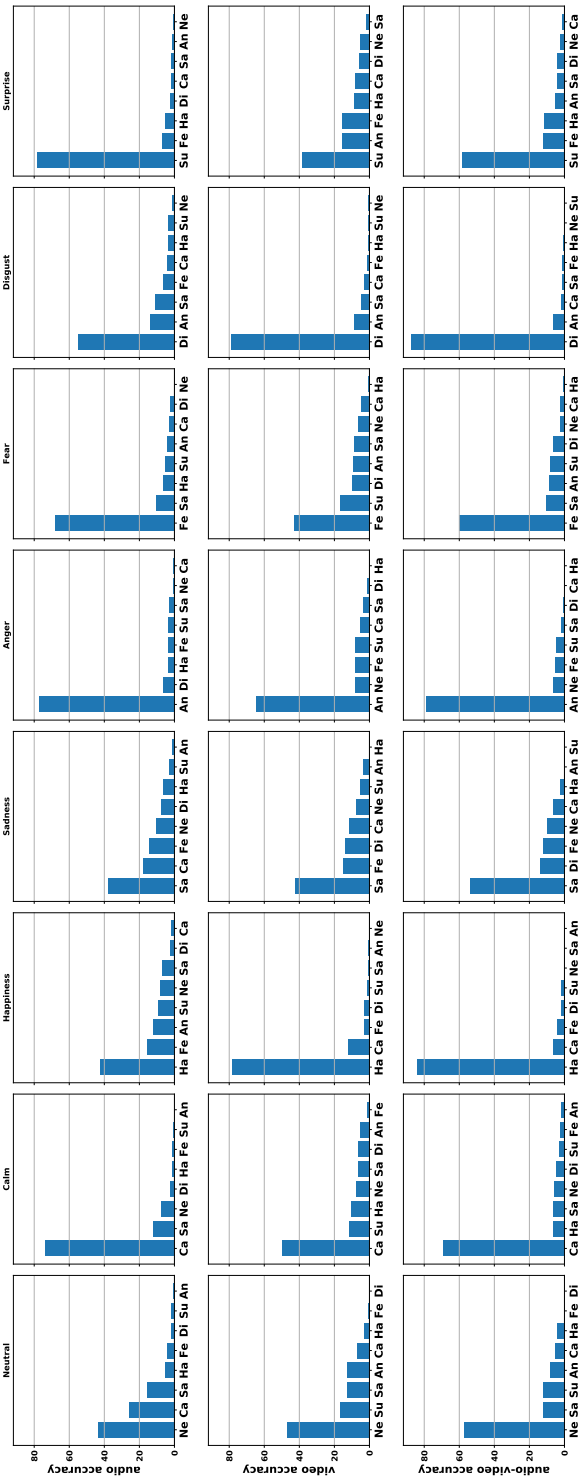


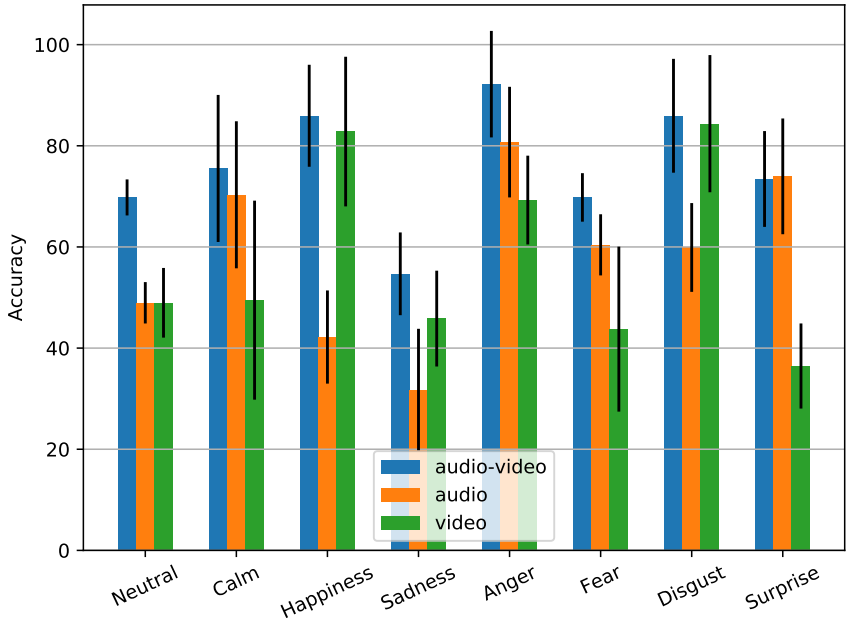
Figure 7.8: RAVDESS without attention and T = 6. This figure shows sorted confusion per emotion for each modality. Due to subfigures size, labels are referred to as follows: Anger (An), Calmness (Ca), Disgust (Di), Fear (Fe), Happiness (Ha), Neutral (Ne) Sadness (Sa), and Surprise (Su).

In particular, it examines the degree to which a certain emotion is confused with the rest in a given modality. Figures 7.5 and 7.6, and Figures 7.7 and 7.8 give a bar view of the confusion matrices (which are shown in Figure 7.4), for CREMA-D and RAVDESS, respectively. This summary is provided for each modality and emotion. For example, in Figures 7.5 and 7.6, we found out that anger is confused with disgust and vice versa. In addition, sadness was confused with neutral, disgust, and fear. These two observations are considerably similar to the ones reported in CREMA-D [52], which provides human raters' confusions. Nonetheless, in the performance of this framework, happiness was confused with anger and disgust, while human raters confused it mainly with neutral expressions. Also, the framework performance shows its confusion of neutral as anger when it uses only video modality. Human raters tended to identify neutral with high confidence in CREMA-D [52]. Moreover, in sadness, using audio-video fusion increased the confusion. This could be traced to the fact that video-modality performed lower on sadness which decreased the multimodal recognition rate.

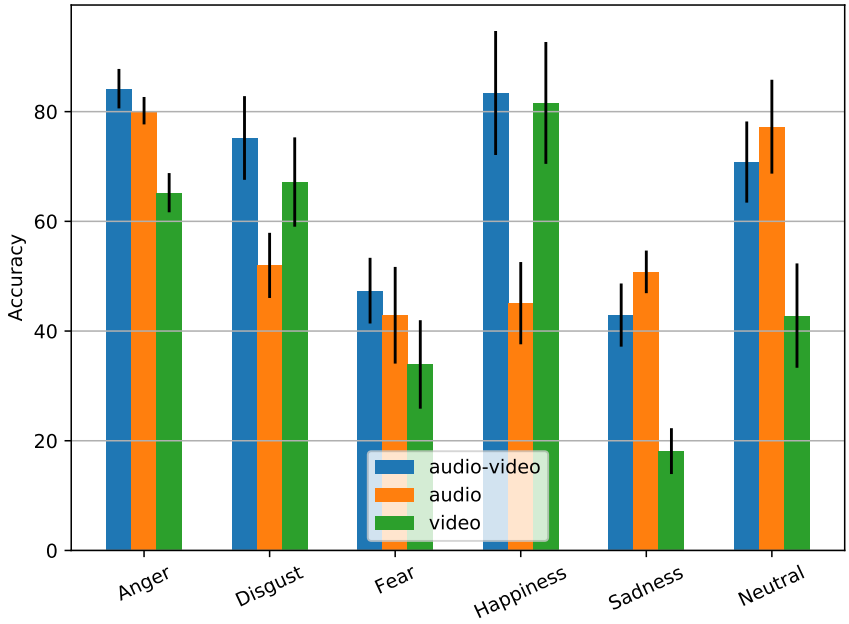
Also in RAVDESS [115], authors provide human annotations considering both modalities. It appears that human annotators, quite often, confused calmness for neutrality and vice versa. Similar to our proposed paradigm, Figures 7.7 and 7.8 indicate that neutral was confused with calm and sadness as well. In addition, similar to CREMA-D, disgust is confused as anger and sadness. To a lesser extent, our multimodal fusion's results show the following confusion: surprise with happiness, and happiness with surprise and calmness. However, unlike the reported multimodal perception of human-raters, the framework performance shows its confusion of sadness as calm and neutral. Furthermore, people rated fear as sadness or surprise, while in our case, fear was confused with anger and sadness. Finally, although the baseline and the framework with the attention differ in the multimodal accuracy, the trend in the emotion confusion in the two modalities and their fusion is quite similar.

7.5.2. OVERALL MODALITIES PERFORMANCE PER EMOTION

Figure 7.9 demonstrates the overall performance. In other words, it shows which emotions can be identified with one modality, and which ones require simultaneous audio-visual cues for accurate recognition. In both datasets, happiness and anger can be recognized with high accuracy via video and audio modality, respectively. However, disgust and fear benefited the most from the bimodal perception. In more details, for RAVDESS, audio-video fusion is giving the best recognition rates for most emotions. A closer look at Figure 7.9a reveals that audio contributes more than video in anger, calm and, fear. On the other hand, facial expressions are more crucial in the identification of happiness and disgust. Similar to RAVDESS, in CREMA-D, audio and video contributed more to the recognition of anger and happiness, respectively. Furthermore, human raters recognized happiness, disgust, and fear mainly through facial expressions. Human raters recognized happiness more accurately when perceiving facial expressions than when perceiving both audio-visual cues. Nonetheless, sadness and neutral were predicted more accurately using audio. On the other hand, sadness was the emotion with the least recognition rate, surprisingly, even by human raters [52]. These results show consistency between automatic emotion recognition and human raters when it comes to identifying extreme cases of positive (such as happiness), and negative (such as anger) emotions.

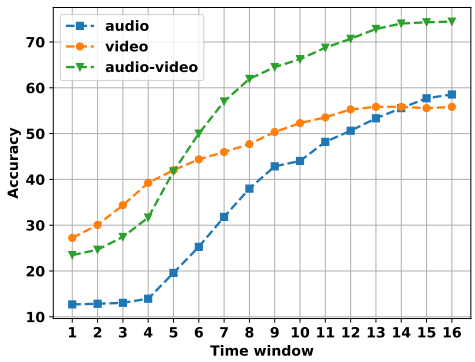


(a) RAVDESS with attention

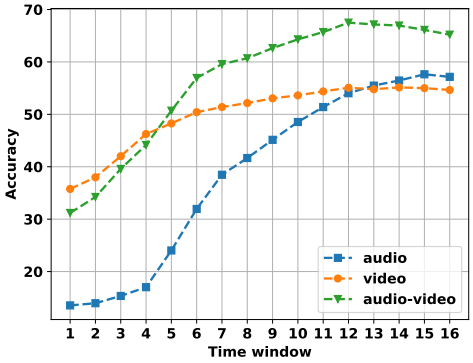


(b) CREMA-D with attention

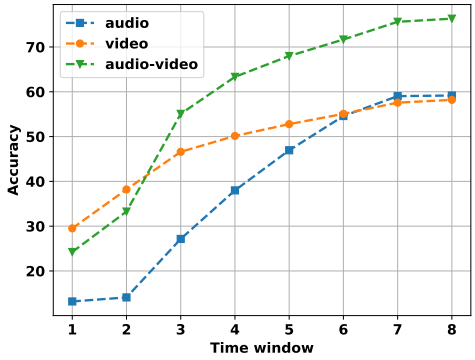
Figure 7.9: Performance of each modality per emotion (time windows are 8 and 6 for RAVDESS and CREMA-D, respectively). Error bars are reported using standard deviations.



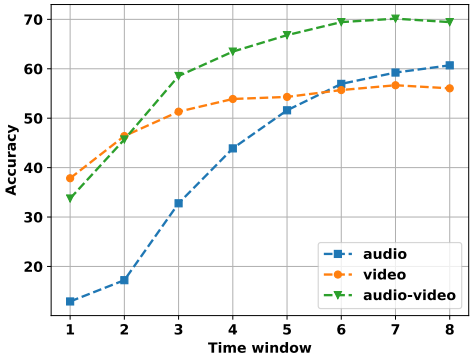
(a) RAVDESS with attention



(b) RAVDESS without attention



(c) RAVDESS with attention



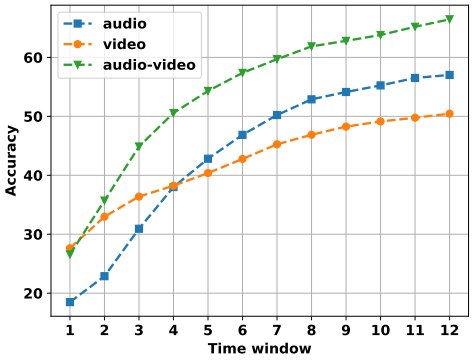
(d) RAVDESS without attention

Figure 7.10: RAVDESS: incremental performance results. These results show the incremental presentations of the embeddings to the framework, for audio-only, video-only, and their multimodal fusion. For example, when $T = 3$, it means that the corresponding accuracy shows the results when the first three time windows were used.

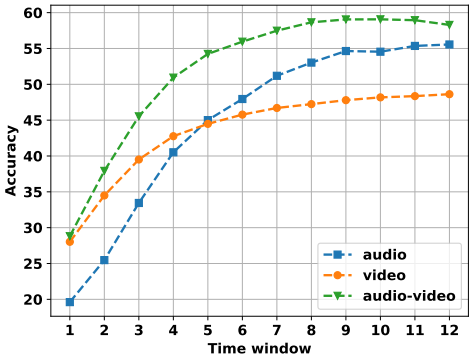
7.5.3. INCREMENTAL EMOTION PERCEPTION

Humans recognize emotions at different rates across modalities [52]. For example, the developers of CREMA-D [52] studied human raters' recognition speed of emotions. They found that the raters need more time to recognize emotions through vocal expressions than the time needed to recognize emotions through facial expressions. Motivated by these findings, in the experiments of this subsection, we study the differences between modalities response times in the MATER framework.

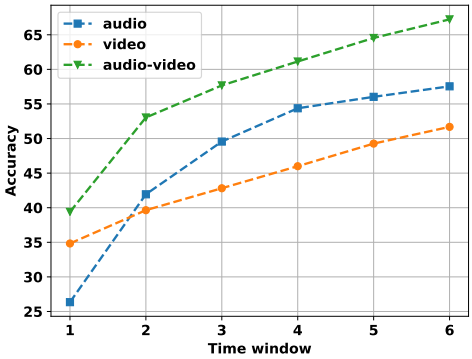
First, we examine how the proposed framework captures the temporal display of emotions through audio and video modalities. For instance, Figure 7.10 shows the results of the framework up until each time window (t). Specifically, in Figure 7.10a, the performance at time window 3 implies that the audio, video, or audio-video embeddings of the first, second, and third time windows were used in the framework. The x-axis, in



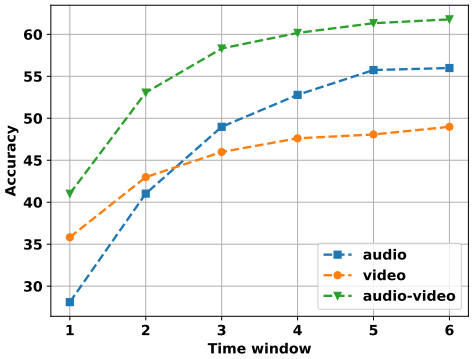
(a) CREMA-D with attention



(b) CREMA-D without attention



(c) CREMA-D with attention

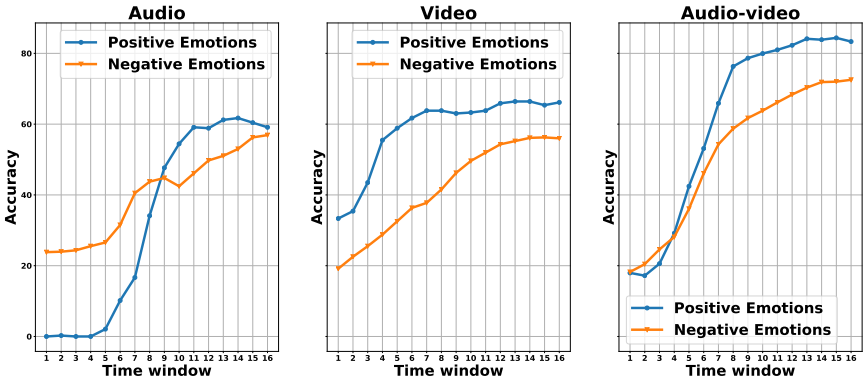


(d) CREMA-D without attention

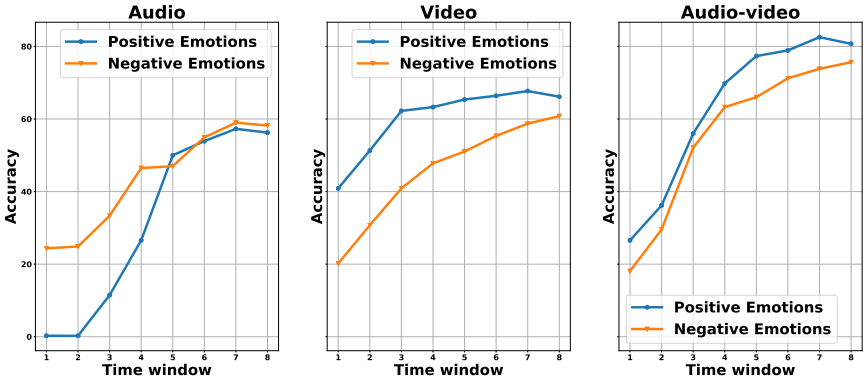
Figure 7.11: CREMA-D: incremental performance results. These results show the incremental presentations of the embeddings to the framework, for audio-only, video-only, and their multimodal fusion. For example, when $T = 3$, it means that the corresponding accuracy shows the results when the first three time windows were used.

the sub-figures, represents the time windows (T), and the y-axis indicates the achieved accuracy. We notice that the performance of MATER with PE and MHSA peaks at the last time window. It means that the framework could make use of the embeddings from all the time windows.

However, this evaluation shows that, in the baseline case, where PE and MHSA are not used, the performance drops prior to the last time windows. The fact that the performance peaks at the last time window, when using the attention mechanism, shows that these final time windows could be useful and the attention mechanism is able to capture their patterns for emotion recognition. This trend in the performance is observed in both datasets, using time windows with varying lengths and numbers. For instance, there was a drop in the performance when using $T = 16$ (illustrated in Figure 7.10b) and $T = 12$ (demonstrated in Figure 7.11b), in RAVDESS and CREMA-D, respectively. The



(a) RAVDESS with attention and T=16



(b) RAVDESS with attention and T=8

Figure 7.12: Incremental performance on positive vs. negative emotions. The titles of the sub-figures refer to the modality used in emotion recognition.

length of these time windows is 0.25 seconds. Furthermore, the presented results reveal that, in most of the cases, the video modality reached a plateau state earlier than the audio modality.

Moreover, there are other interesting findings in Figures 7.10 and 7.11 that are related to the recognition rates over time windows. The first observation is that the recognition rates of audio modality in the initial time windows are lower than the video modality ones. Specifically, the audio modality requires more time to achieve comparable performance (or even higher in some cases) with the video modality. Also, the recognition rates of emotions through video modality usually rise sooner than the audio modality ones. Moreover, as mentioned previously, there is usually a drop in the network's performance in case of not using the attention. However, the attention mechanism did better in bridging the gap between audio and video modalities' performances through the time windows. Subsequently, the reduced gap contributed to accurate bimodal emotion recognition.

POSITIVE VS. NEGATIVE EMOTIONS

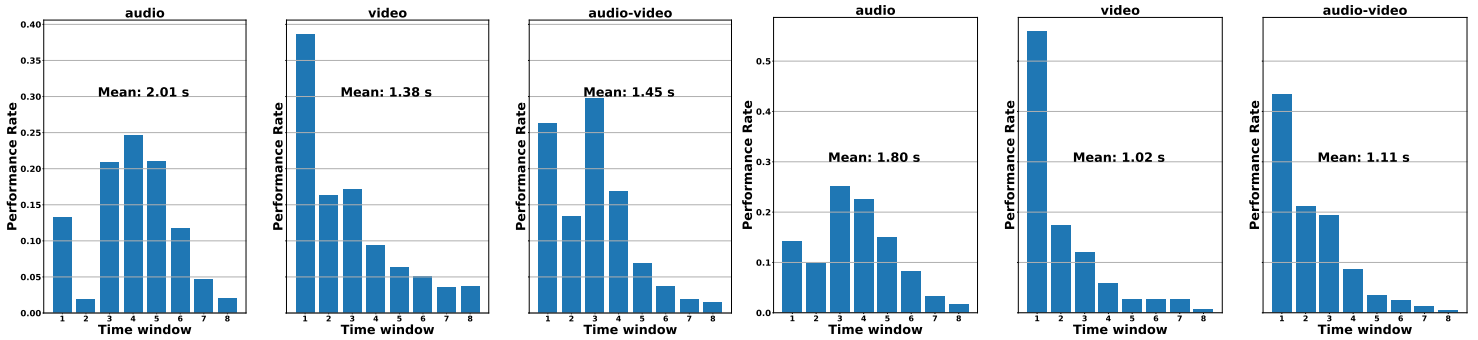
Emotion perception can vary as a function of time. Research in the literature suggests that time's impact can be different on negative and positive emotions [77]. RAVDESS has a reasonable number of positive and negative emotions. On the other hand, CREMA-D has mainly negative emotions and only one positive emotion, namely happiness. Consequently, we examine the behavior of the recognition rates overtime on negative emotions (namely: sadness, anger, and fear) and positive emotions (namely: calmness and happiness) on RAVDESS. Surprise was excluded since, as an emotion, it can be both positive or negative, depending on the context of the stimulus [280]. Moreover, many scholars consider surprise as a pre-affective state.

Figure 7.12 provides the accumulated performance for these two categories, focusing on the case of using the attention mechanism. Interestingly, we noticed that the recognition scores increase faster for positive emotions than for the negative ones. Indeed, the figure shows that the duration of an expressed negative emotion plays a larger role in recognizing it, whereas for positive ones, a plateau in recognition accuracy can be reached earlier in time. Furthermore, video is more influential in recognizing positive emotions. This is, especially, reflected in the performance of audio-video fusion on positive and negative emotions. This outcome is in alignment with previous findings in [279], even though audio and video representations were based on different models than VGG [276].

7.5.4. MODALITIES RESPONSE TIME

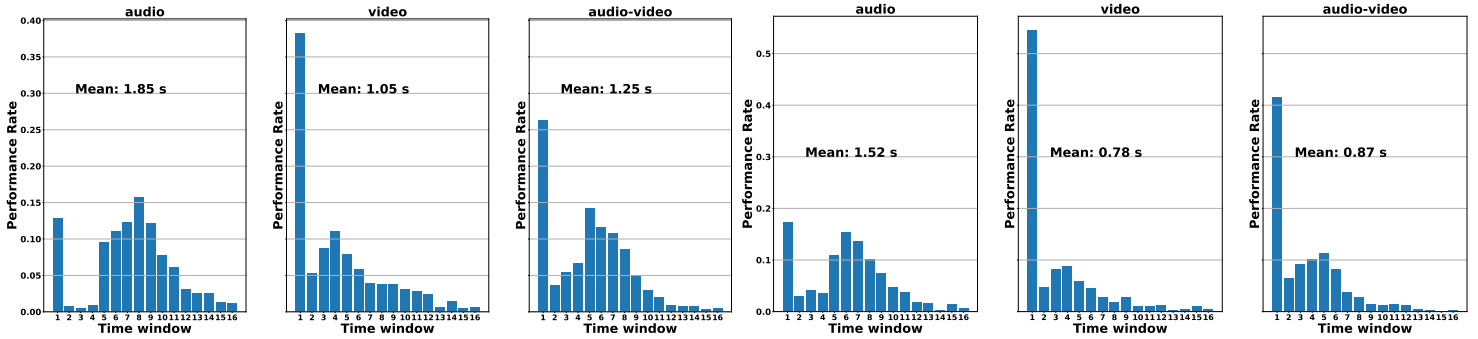
People's response time could vary according to the presented modalities [52, 73, 77]. In addition, response time could depend on the emotion's intensity and duration. Response time refers to the minimum duration required for emotion perception and recognition. For example, the authors in CREMA-D [52] studied the mean response time of human perception of emotions through audio only, video only, and their combination. The study reported that, on average, the mean response times of human raters for emotion recognition through audio only, video only, and the bimodal audio and video perception are 2.98, 2.05, and 1.95 seconds, respectively. The results show that the recognition speed through audio modality is at least one second lower than video only or both audio-visual cues. Furthermore, in RAVDESS [115], human raters' average response times are 1.55, 1.31, and 1.32 seconds, for audio only, video only, and the audio-video bimodal perception, respectively.

In human rating, when presenting video clips to people and asking them to identify emotions, the minimal duration can be measured by taking the time when the raters first identify an emotion of a video clip. However, in automatic emotion recognition, due to technical requirements, emotion perception could be measured with the smallest possible time window of a framework. In our case, this duration is either 0.5 or 0.25 seconds. Therefore, we measure the performance of each modality for emotion recognition across time by using the following protocol: since the performance mostly peaks at the end of a video clip, we check the response time by identifying the first time window in which a video-clip was classified correctly, among the correctly classified samples in the last time window. The last time window was taken as the performance reference, due to its nature of accumulating the information from the whole duration of a video-clip. In other words, considering only correctly classified samples, in the performance of the



(a) RAVDESS With Attention

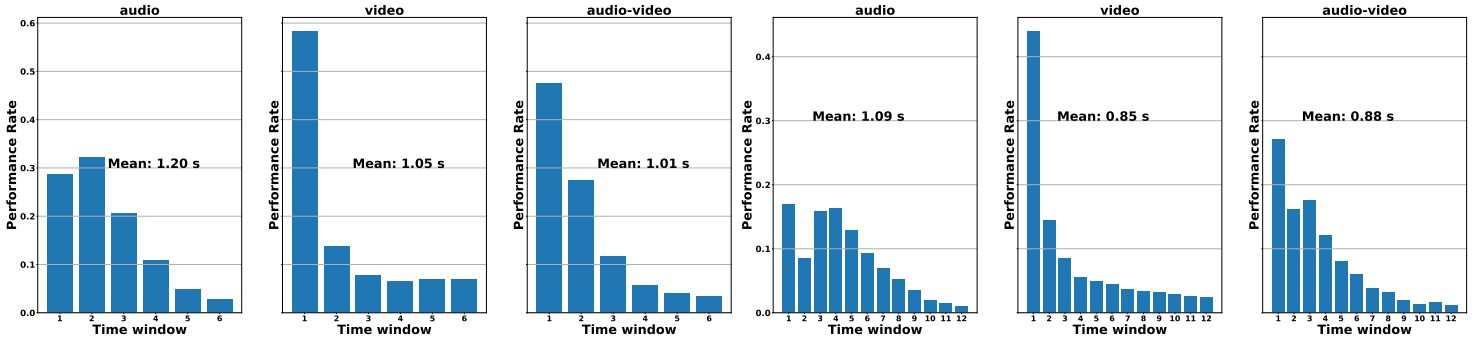
(b) RAVDESS without Attention



(c) RAVDESS with attention

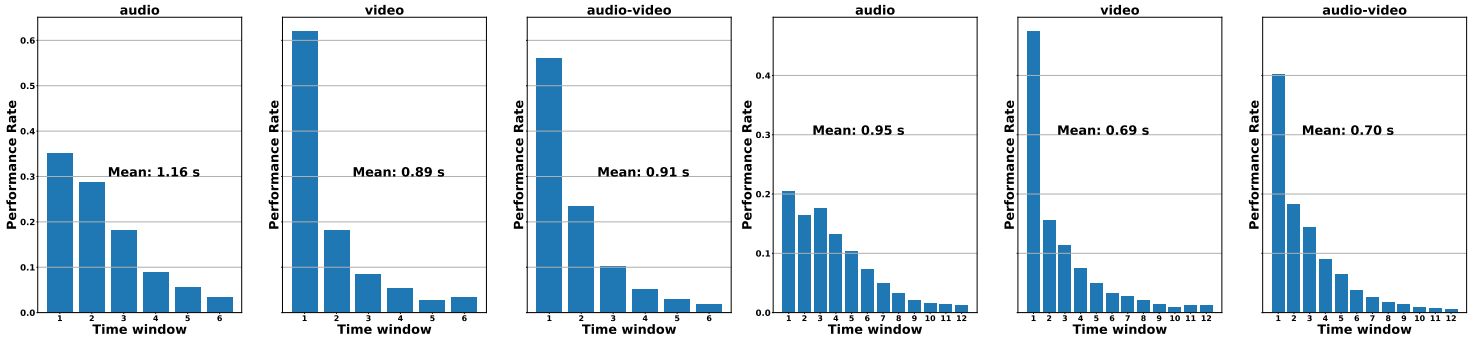
(d) RAVDESS without attention

Figure 7.13: RAVDESS: Average response time per-modality. The bar diagrams indicate the contribution ratios of each time window in emotion recognition for audio-only, video-only, and their fusion.



(a) CREMA-D With attention

(b) CREMA-D with attention



(c) CREMA-D without attention

(d) CREMA-D without attention

Figure 7.14: CREMA-D: Average response time per-modality. The bar diagrams indicate the contribution ratios of each time window in emotion recognition for audio-only, video-only, and their fusion.

Row	Dataset	MHSA	T	PE	A Acc.	V Acc.	AV Acc.	A Entropy	V Entropy	A-V Entropy Diff	KL-D
1	CREM-D	✓	6	✓	57.5	51.7	67.2	0.59	0.44	0.15	3.12
2		✗	6	✓	57.6	51.4	64.4	0.69	0.33	0.36	3.66
3		✓	12	✓	57.0	50.5	66.5	0.54	0.38	0.16	3.59
4		✗	12	✓	57.2	51.1	62.3	0.76	0.28	0.48	4.07
5		✓	6	✗	53.5	49.8	65.0	0.37	0.28	0.09	5.06
6		✗	6	✗	56.0	49.0	61.8	0.54	0.24	0.30	5.09
7		✓	12	✗	51.6	49.5	63.6	0.34	0.27	0.07	5.48
8		✗	12	✗	55.6	48.6	58.3	0.65	0.20	0.44	5.24
9	RAVDESS	✓	8	✓	59.2	58.2	76.3	0.41	0.32	0.09	4.95
10		✗	8	✓	61.6	55.3	70.6	0.73	0.25	0.47	4.84
11		✓	16	✓	58.5	55.8	74.4	0.40	0.30	0.10	5.30
12		✗	16	✓	59.0	56.2	68.8	0.90	0.25	0.65	4.92
13		✓	8	✗	55.4	54.4	75.2	0.27	0.25	0.02	6.74
14		✗	8	✗	60.7	56.0	69.4	0.60	0.21	0.39	5.46
15		✓	16	✗	54.7	54.2	72.4	0.26	0.23	0.02	6.97
16		✗	16	✗	57.2	54.7	65.2	0.77	0.20	0.57	5.68

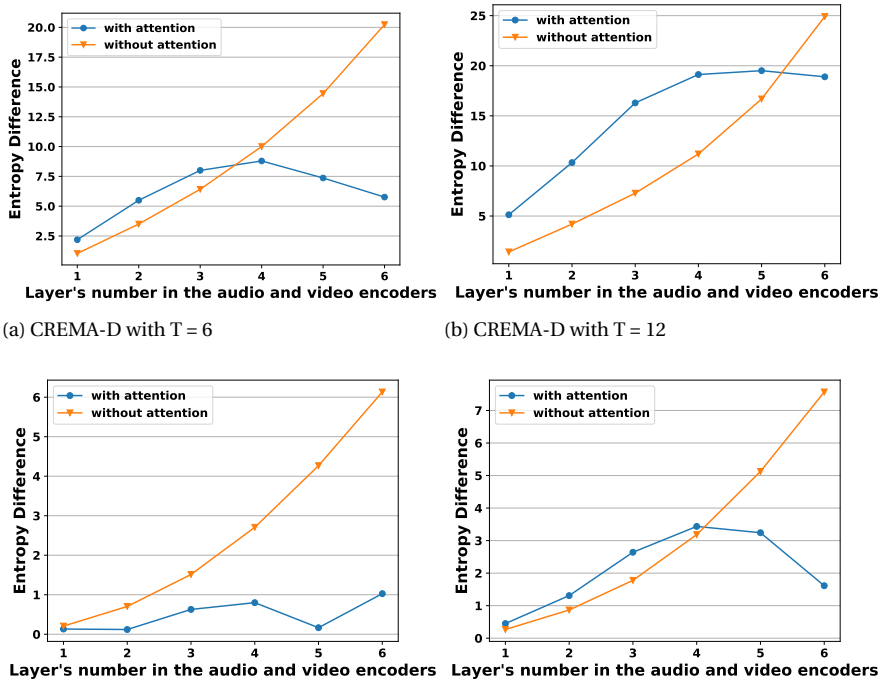
Table 7.4: A detailed performance analysis using different parameters, such as MHSA and PE, with different measures.

audio, video, and audio-video fusion, Figures 7.13 and 7.14 show performance rates of different time windows, related to their contributions (ratios) in recognizing emotions as a function of time. E.g. in sub-figure 7.13a (audio-video), emotions were identified correctly already since time window 1 in 26% of the cases, while they were correctly classified starting from the second time window in a 14% of the time.

According to Figures 7.13 and 7.14, the video modality relies heavily on the initial time windows. However, the middle time windows are important in the audio modality. Moreover, Figures 7.13 and 7.14 display the mean response time per-modality, for RAVDESS and CREMA-D. The figures illustrate differences in the average response time of audio and video modalities. For example, in RAVDESS, Figure 7.13a shows that the mean response times are 2.01, 1.38, and 1.45 seconds, for audio-only, video-only, and the audio-video fusion, respectively. This implies the fact that audio modality requires more time than video modality for emotion recognition. This observed tendency, in the response time of modalities, can be generalized for the rest of the scenarios and approaches. Moreover, these observations are similar to those obtained through the response time of human perception of emotion. For example, we also notice that the recognition speed of emotions through the audio modality is lower than the video modality. In the automatic bimodal audio-video emotion recognition, we observed that the delay in the audio mean response time slightly impacts the bimodal one, making the response time of the video modality the shortest one among the three settings.

7.5.5. INSPECTION OF THE DISCREPANCY BETWEEN UNIMODAL AND MULTIMODAL PERFORMANCES

We observed that the differences in performance with and without attention in the unimodal case are much lower than the corresponding figures in the multimodal case. For example, for CREMA-D, as we see in Table 7.3, using only audio with and without attention yields 57.5% and 56.0%, respectively. However, when employing both modalities,



(a) CREMA-D with T = 6

(b) CREMA-D with T = 12

(c) RAVDESS with T = 8

(d) RAVDESS with T = 16

Figure 7.15: Entropy differences between audio and video embeddings in the case of the baseline and the framework with attention.

the corresponding performances are 67.2% and 61.8%, with and without attention, respectively. This disparity led us to conduct a further examination of the framework's performance to check the reason behind the variant improvements.

During the evaluation of the test samples, we adopted the entropy to check the underlying agreement in audio and video scores. This is because entropy measures the average level of uncertainty (information) in variables' outcomes. In particular, we calculated the entropies of the audio and video predictions. Next, we measured the differences between the entropies of each of the audio and video samples. Besides, we compute the KL-divergence between the two modalities' probabilities (predictions) of each sample. It is important to note that the entropy values are obtained from each modality separately (as displayed in Table 7.4). Table 7.4 details the average entropies of audio and video modalities (in columns 9 and 10), as well as their average differences for each sample (column 11). On the other hand, KL-D measurement considers the predictions of both audio and video modalities by calculating the KL-D between their predictions (KL-D values are reported in the last column of Table 7.4).

We observed that the lower the difference between the entropies, the higher the performance is. This also indicates the increase of the bimodal certainty on the obtained predictions, and; therefore, it yields good performance. For instance, we notice that

the attention mechanism helps in bringing the entropies of both modalities closer, and; hence, it reduces the uncertainty when they are fused. For example, the entropy differences between audio and video outcomes are 0.15 and 0.30 for attention (row 1) and the baseline (without attention and PE (row 6)), respectively. These differences are also reflected in the two approaches' performances, where the accuracies of the attention and without attention are 67.2% and 61.8%, respectively. Similarly, there is a negative correlation between the value of the KL-D and the bimodal performance. The negative correlation between the bimodal audio-video recognition rates and the KL-D is noticeable when using the Positional Encoding. According to the results in Table 7.4, PE contributes to improving the bimodal fusion accuracy. The improvement can be seen from the consistently higher accuracies in all experiments where PE was involved.

ENTROPIES OF LAYERS EMBEDDINGS

Furthermore, we measured the entropies of the embeddings, at each layer of the framework (the encoder consists of $l = 6$ stacked blocks, as explained in Subsection 7.3.2). As a result, we first apply softmax normalization on the audio-visual representations, and then the average entropies ($H^{m\in\{audio(a),video(v)\}}$) of $d-$dimensional audio-visual representations across the layers were calculated as follows:

$$H^{(m)} = -\sum_{j=0}^{l}\sum_{i}^{d} \boldsymbol{h}_i^{(j)(m)} \log(\boldsymbol{h}_i^{(j)(m)}) \tag{7.6}$$

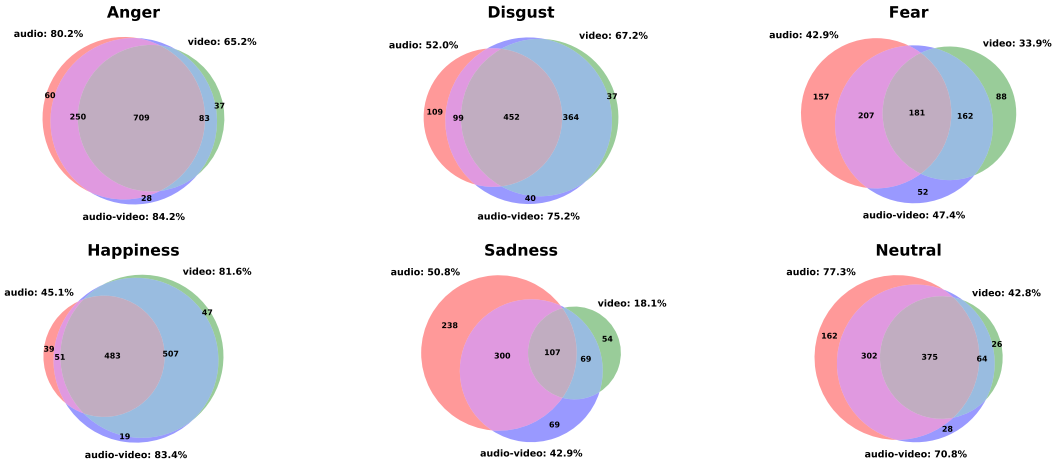
For example, in CREMA-D models, we observed that the entropy of the last layer, prior to the linear prediction layer, has the following values: entropies of audio and video, in the baseline, are around ≈ 4.7 and ≈ 4.5 respectively. However, the entropies, in the case of attention, are ≈ 4 and ≈ 3.9, for audio and video, respectively. Notice that the values and the gap between them are lower in the case of using the attention mechanism. Moreover, Figure 7.15 depicts a detailed comparison between the entropies of audio and video embeddings across the framework layers. The figure reveals that the usage of MHSA helped to decrease the difference between audio and video entropies, while in the baseline, the difference is increasing in the later layers. This led to a smaller gap in audio and video modalities' entropies and, consequently, a more accurate multimodal performance. To summarize, attention mechanisms helped the framework bring entropy measurements of audio and video embeddings and predictions closer to each other. As a result, these low differences between these entropies enhanced the bimodal certainty and improved the performance of audio-visual fusion.

7.5.6. MULTIMODAL INTERACTION

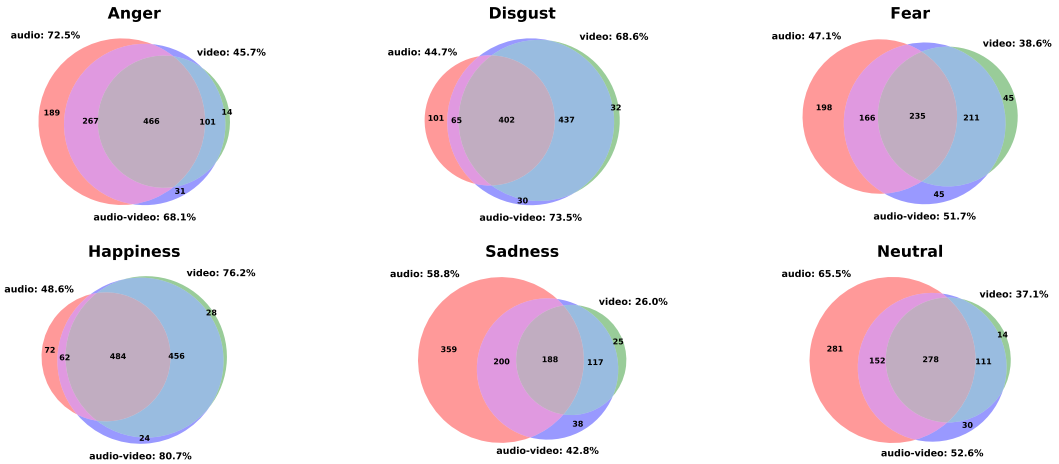
Further insights can be obtained, regarding the interaction between the two modalities. For a better understanding of the contribution of the audio and video modalities, we check the agreement (overlap) in emotion prediction, based on audio and video, compared to the audio-video perception.

MULTIMODAL INTERACTION PER-EMOTION

Figures 7.16 and 7.17 present the Venn diagrams of clips predicted correctly based on each combination of modalities, for CREMA-D and RAVDESS, respectively. These results are taken from the multimodal framework which was trained jointly with audio

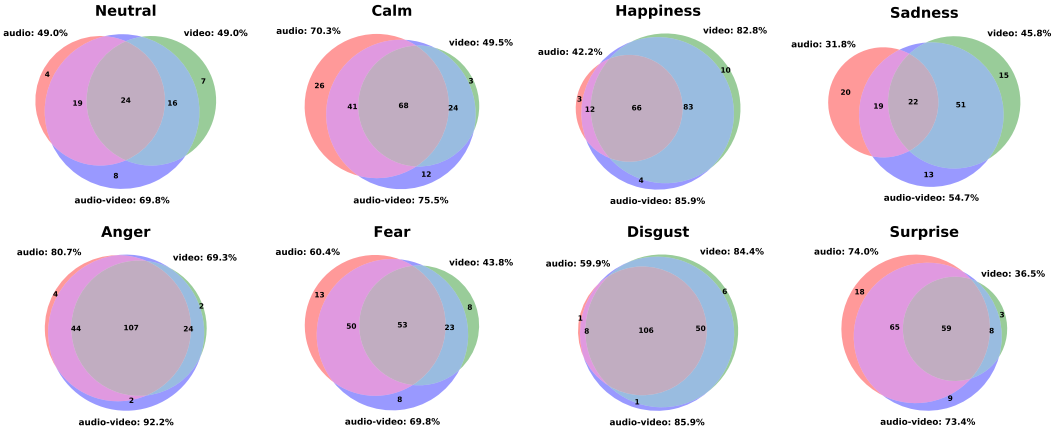


(a) With attention and T = 6

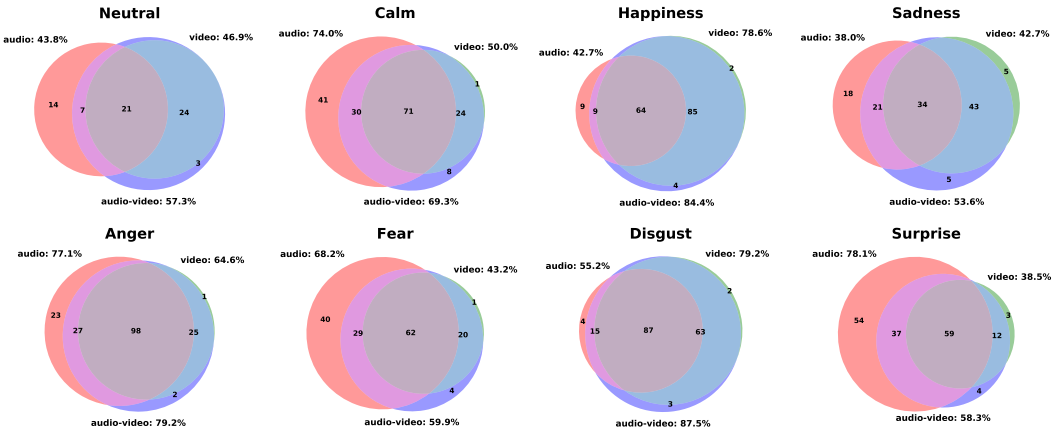


(b) Without attention and T = 6

Figure 7.16: Venn diagram per-emotion in CREMA-D, to show the multimodal and unimodal interactions between audio-only (A), video-only (V), and audio-video (AV) recognition. For example, the number in the central overlap indicates the number of samples that are recognized correctly in all combinations. The outer values show the number of samples recognized correctly only in one modality.



(a) With attention and T = 8



(b) Without attention and T = 8

Figure 7.17: Venn diagram per-emotion in RAVDESS, to show the multimodal and unimodal interactions between audio-only (A), video-only (V), and audio-video (AV) recognition. For example, the number in the central overlap indicates the number of samples that are recognized correctly in all combinations. The outer values show the number of samples recognized correctly only in one modality.

and video signals, and are analyzed separately for audio, video, and audio-video fusion. For each emotion, there is a Venn diagram where, for each combination, the number in the circle is the total counts of clips correctly recognized as the emotion of the Venn diagram title. For example, in Figure 7.17, the happiness expression, the correctly classified videos are 573 (39+51+483), 1007 (483+507+19), and 1037 (483+507+47) clips for audio-only, video-only and the bimodal audio-visual recognition, respectively.

In a Venn diagram figure, if a diagram has a high number in the non-overlapping lower circle (representing audio-video fusion), the high number suggests that the bimodal recognition is advantageous over the unimodal ones. For example, the audio-video fusion numbers (in the diagrams of Fear, Neutral, and Sadness in Figure 7.17 and Figure 7.16) show the importance of the bimodal perception for these emotions. Also, In Figure 7.16, the greater the agreement between audio-video fusion and the two submodalities, the less crucial the multimodal perception is. In other words, this means that the multimodal perception is redundant since unimodal perception achieves good numbers (without the need for the other modality). For example, the recognition rates of happiness and anger are high with video only and audio only, respectively. Moreover, the overlaps between audio-only, video-only, and their fusion diagrams indicate the unimodal constraint to the bimodal recognition rates. For example, the audio contribution is more significant than the facial expressions for recognizing anger, whereas video contributes to recognizing happiness more than audio.

OVERALL MULTIMODAL INTERACTION

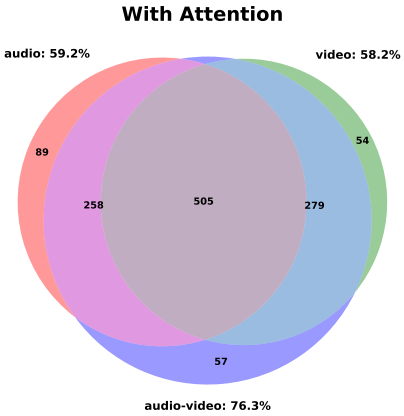
Figure 7.18 presents Venn diagrams to show the overall interaction between audio and video modalities with audio-video fusion. In both datasets, for the baseline and the case of using attention, the agreement (overlap) between the video-only and audio-video cases is higher than the one between audio-only and the audio-video cases. More importantly, video performance is higher with the attention, a fact which led to higher gain in the multimodal performance compared to the baseline method. For instance, in the CREMA-D dataset, there are 236 and 198 samples that are recognized correctly in the multimodal fusion using MHSA and the baseline, respectively.

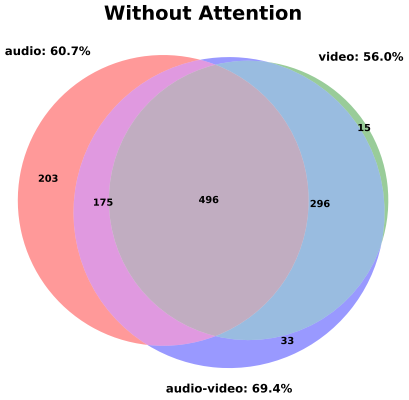
ATTENTION VS. BASELINE INTERACTION PER MODALITY

Finally, Figure 7.19 presents the agreement between the baseline and the framework with the attention, for each modality. Although there is a large agreement between the two approaches, the figure shows the fact that they differ in many samples. For example, regarding the RAVDESS dataset, in the case of the audio-video fusion, the attention and the baseline share 932 samples, however, attention correctly recognized 167 different samples while the baseline classified the other 68 videos correctly.

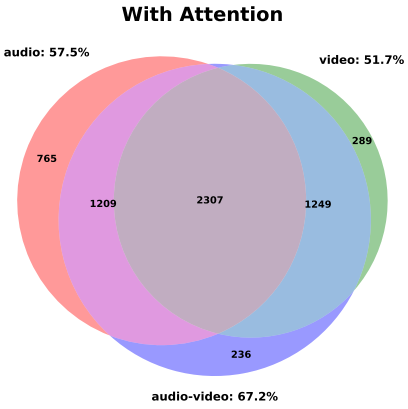
7.6. HANDLING NOISY DATA

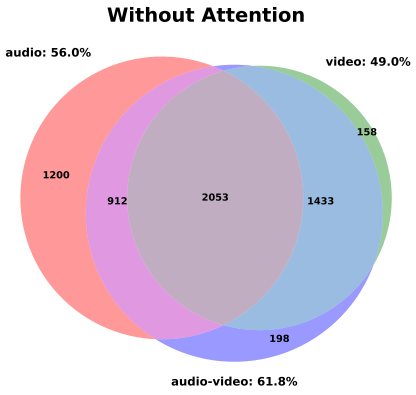
The evaluations in this section aim to investigate the robustness of the framework when trained and evaluated with noisy data. The analysis offers an insightful view of the proposed framework's performance. This is an interesting facet since audio and visual cues in real scenarios could be noisy, which makes data representation a challenging task, especially for the purpose of emotion recognition due to the nature of emotions (which is





(a) RAVDESS



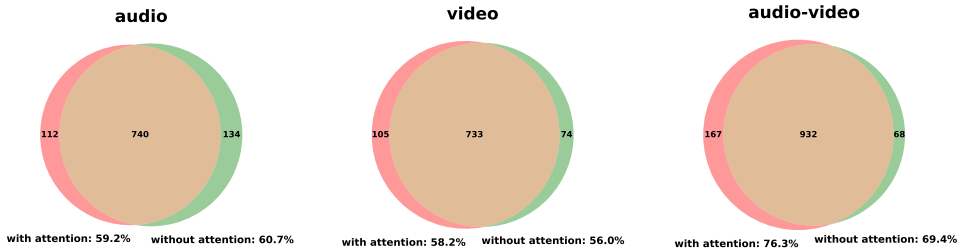


(b) CREMA-D

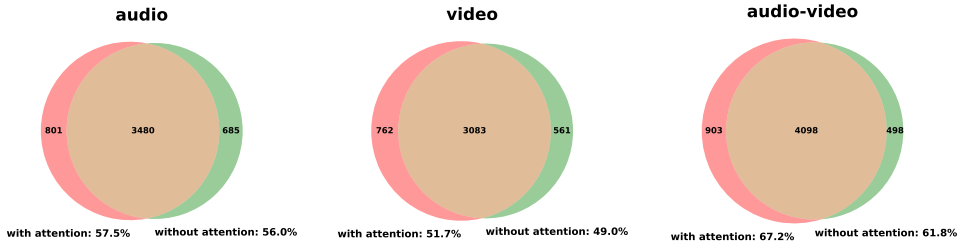
Figure 7.18: Venn diagram per approach. These diagrams show how the multimodal fusion is overlapping with the underlying audio and video modalities. The more overlap, the higher the agreement is between the unimodal and multimodal fusion.

discussed in Subsection 1.1.1) [3]. Moreover, in the real world, face tracks of facial expressions can have challenges such as varying illumination and head poses. Also, audio signals can be noisy, and that can be challenging.

In the following subsections, we present two types of evaluations. The first set of experiments examines the models that were trained with *"noise-free"* but are evaluated with noisy embeddings of some or all time windows. The second set of experiments introduces the notion of re-training and also evaluating the MATER framework with noisy data.



(a) RAVDESS



(b) CREMA-D

Figure 7.19: Venn diagram per modality with and without attention. These diagrams show the overlap between the two approaches (with and without attention). The aim here is to show how both approaches differ in terms of emotion recognition, in the cases of audio-only, video-only, and audio-video fusion.

7.6.1. EVALUATING NOISE FREE MODELS

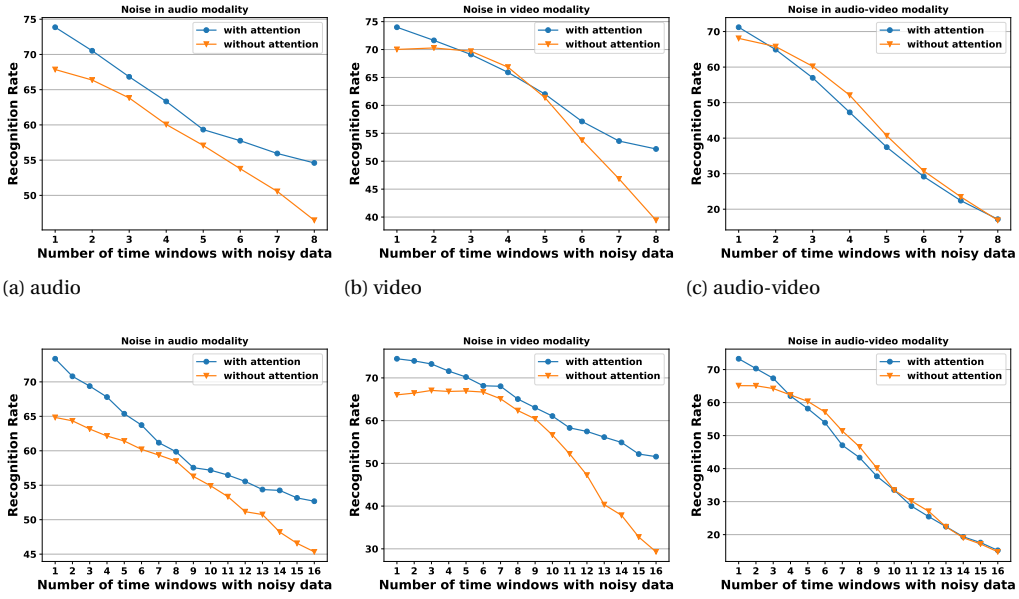
In this experiment, we assess the baseline and the framework with attention and PE (as detailed in Section 7.4). The two models, with different lengths and numbers of time windows, were trained using *"noise-free"* data.

NOISE INJECTION SCENARIO

During the evaluation, the noise was injected at different numbers of time windows in the three settings, i.e. audio only, video only, and audio-video fusion. For instance, a number of 1 and up to T time windows were replaced with noise in audio-only, video-only, and on both audio-video fusion. More importantly, since noise is a random signal and was injected in random places in the overall data, we opted for repeating the experiment a number of times (here, 10), in order to establish reliable figures with regards to performance. Hence, the following reported results are the average results of the 10 evaluations. The reason behind these repeated evaluations is to obtain a representative performance since each one resulted in different values. It is important to note that, in these settings, noise injection refers to replacing time windows' embeddings with random signals sampled from Gaussian.

EVALUATION RESULTS

Figures 7.20 and 7.21 demonstrate the multimodal recognition rates of the aforementioned scenario. We notice that when both modalities are employed in the case of using the attention mechanism, the framework's performance is degraded with noisy data



(a) audio (b) video (c) audio-video

(d) audio (e) video (f) audio-video

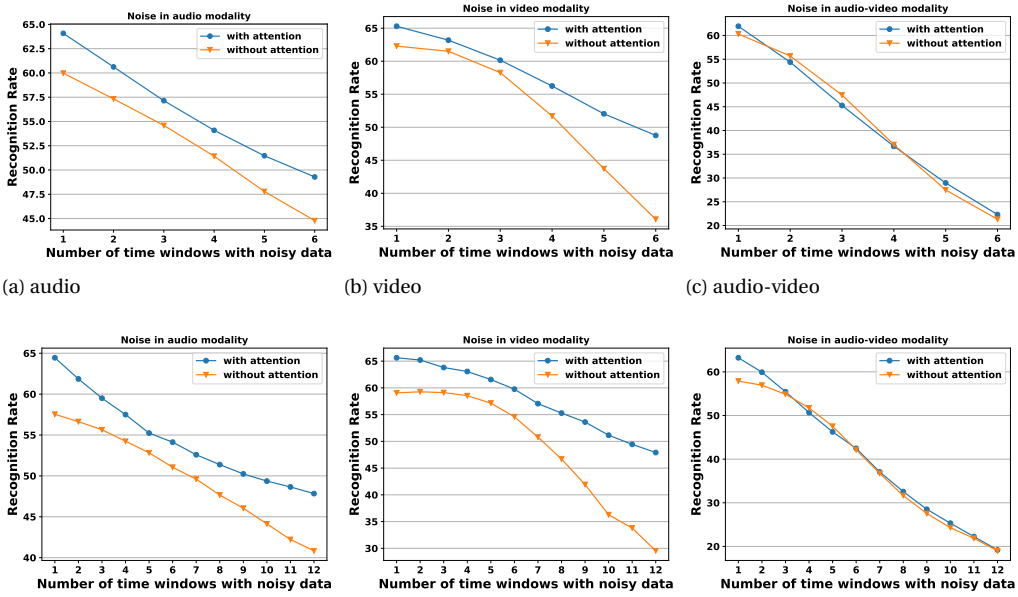
Figure 7.20: RAVDESS: The results of evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The adopted models in these figures were trained with both audio and video modalities jointly. Nonetheless, in the evaluation phase, as indicated by each subfigures title, the noise was placed in audio-only, video-only, or both modalities' embeddings.

quicker than the baseline, as displayed in Figures 7.20c, 7.20f, 7.21c, and 7.21f. However, this is an expected outcome, since attention might focus on the noise window, while baseline always takes all windows with equal contribution. In other words, in the case of using MHSA, time windows are interacting with each other (as explained in Subsection 7.3.2), while in the baseline, the multimodal prediction mainly relies on the time windows' aggregated predictions.

In addition, an interesting observation of this evaluation is that when contaminating with the noise in one modality, the framework could still depend on the other modality without losing its unimodal performance. For instance, when placing noise in all the time windows of video modality (as shown in Figures 7.20b, 7.20e, 7.21b, and 7.21e), the multimodal performance of the framework was similar to the reported unimodal results in Table 7.3. In other words, this observation indicates that the framework depends on the none noisy modality when one of them is random.

7.6.2. RETRAINING THE FRAMEWORK WITH NOISY TIME WINDOWS

The motivation behind the retraining of the framework with noise is to check its performance when it is exposed to similar conditions, during the training and the evaluation processes. This apparatus has several interesting aspects, as it allows us to monitor the



(a) audio (b) video (c) audio-video

(d) audio (e) video (f) audio-video

Figure 7.21: CREMA-D: The results of evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The adopted models in these figures were trained with both audio and video modalities jointly. Nonetheless, in the evaluation phase, as indicated by each subfigures title, the noise was placed in audio-only, video-only, or both modalities' embeddings.

validity and the robustness of its performance under harder settings.

RETRAINING SCENARIOS

The framework was retrained (from the scratch) using noise in audio-only, video-only, and both audio-video modalities, separately. In other words, we obtain three models from the proposed method, each model has noisy in audio, video, or audio-video embeddings. Another aspect of this process is the placement of the noise in the time windows. With regard to this, we put the following restriction: the maximum number of noisy time windows is set to half the total number of the time windows. For instance, if T(the total number of time windows) is 8, the number of the noisy time windows might range between 1 and 4, and the specific number is chosen randomly at each iteration during the training process. Moreover, their positions are scattered randomly across the time windows. These restrictions were placed in order to allow the framework to converge during the training process, in the case of employing noisy embeddings. Algorithm 4 summarizes the noise injection scenarios in the training processes.

EVALUATION SCENARIOS

The baseline and the framework with the attention mechanism, each, have three models trained with noise, for each modality. In the evaluation process, if a model was trained

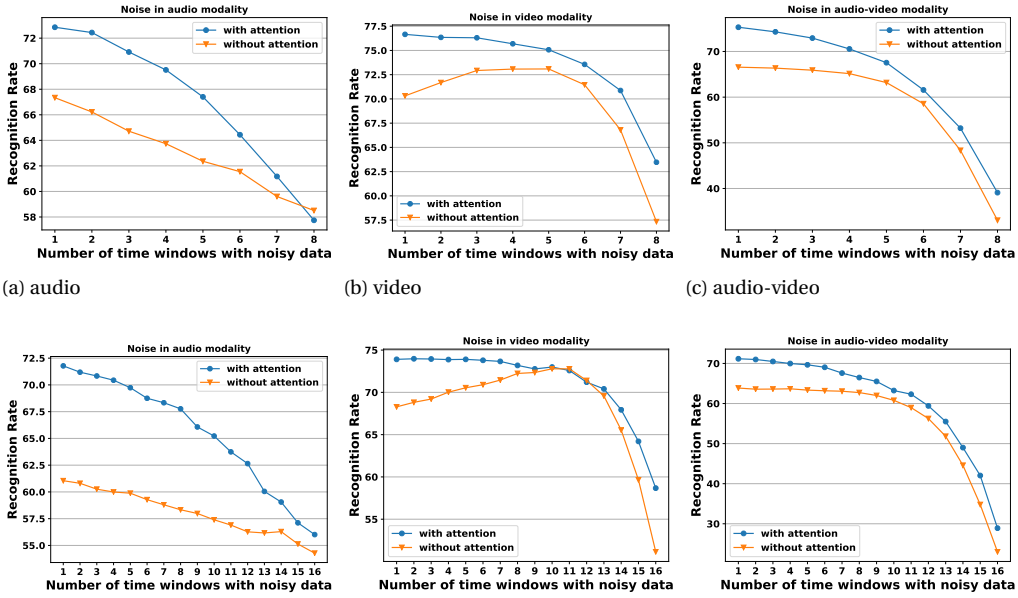
Algorithm 4 Noise Injection Algorithm.

procedure RETRAINING MATER WITH NOISY DATA($\mathbb{D}$)
Inputs:
    Formulate the method as shown in Figure 7.1 and described in Section 7.3
    Audio ($f^{(a)}(\boldsymbol{x}^{(a)})$) and visual ($f^{(v)}(\boldsymbol{x}^{(v)})$) embeddings
Initialization:
    Number of time windows (T)
    Noisy modalities: M ∈ {audio, video, both audio-visual}
    Training and evaluation parameters as described in Subection 7.4.1
Training:
    for `m = 1:M` **do**
        Re-train MATER (with attention and the baseline models) using noise in m
        Use the same parameters and details as described in Subection 7.4.1
        In each iteration during the training, pick a random number: $T_{noisy} \in \{1, ..., \frac{T}{2}\}$, and replace the resulted T_{noisy} time windows with noise
    end for
Evaluation:
    Evaluate the obtained baseline and attention models using noise in m modality (with the same settings)
end procedure

with noise, e.g. in audio, during the testing process, the noise was also used only in the same modality. The settings, in terms of the number of noisy time windows for noise injection and their positions, were similar to the ones during the training process. Moreover, as the used embeddings contain noise, the models' evaluations were performed 10 times, and the reported results are the average results of these evaluations.

Figures 7.22 and 7.23 illustrate the results of the retraining and testing schemes. Without exception, we notice that when the framework is trained with noisy data, and tested in a similar manner, its performance is robust, especially in the case of using the attention mechanism. This is in contrast to the case of evaluating the models which were obtained using *"noise-free"* data during training. In the latter case, the framework performance was degraded due to noise injection during the testing phase.

This consistency implies the adaptability of the method to more challenging settings. In addition, it demonstrates that the attention mechanism is potentially able to handle the instances of noisy time windows due to the re-training procedure. In addition, interestingly, Figures 7.22b, 7.22e, 7.23b, and 7.23e reveal how the video modality can recover from noise. This suggests that there is not a specific range or number of time windows that are more important than others. A further interpretation could be drawn that the attention mechanism played a bigger role in audio modality than the one in video modal-



(a) audio (b) video (c) audio-video

(d) audio (e) video (f) audio-video

Figure 7.22: RAVDESS: The results of retraining and evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The title of each sub-figure indicates the modality in which noise was used.

ity, as its strength is concentrated in the middle of video clips.

NOISE PLACEMENT AT SPECIFIC TIME WINDOWS

Finally, we examine the impact of the noise when placed at certain positions (time windows) in the existing modalities. For this reason, we conducted an analysis where noise replaced the following embedding: the two-first, the two-middle, and the last-two time windows. This additional scrutiny offers an opportunity to inspect the system performance in the three critical intervals. Also, in this section, we present the results in the case of injecting noise to 0 to $\frac{T}{2}$ of the time windows, which are selected from the entire sequence. We refer to this last scenario as half-global.

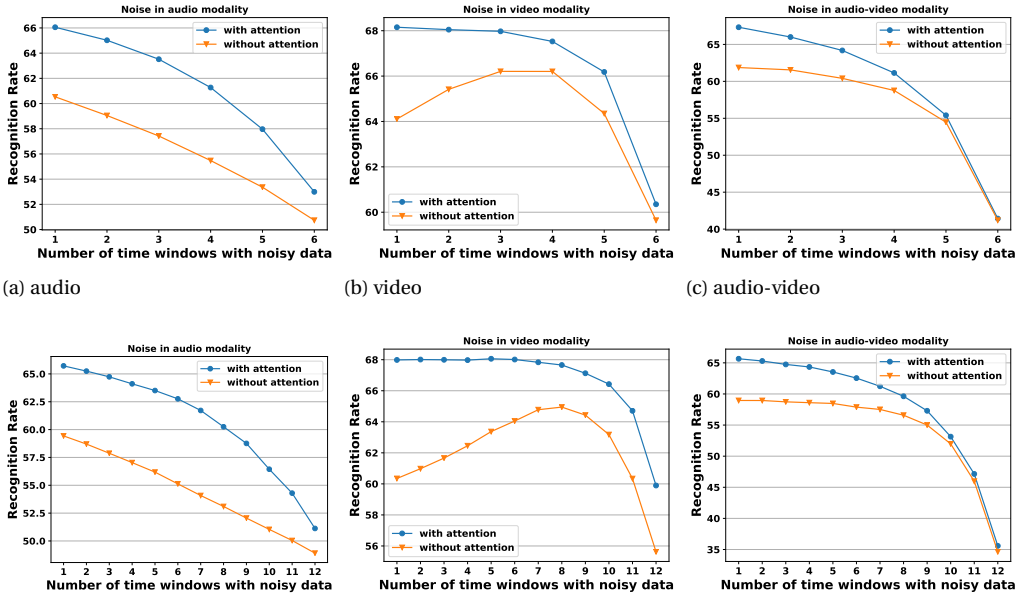
Tables 7.5 and 7.6 present the detailed average results of these scenarios. In both tables, results indicate that allocating the noise in the middle time windows of audio modality reduces the multimodal performance in comparison to placing them in the end or in the initial time windows. In fact, this is in good agreement with the observations of the average time-responses of each modality (as described in Subsection 7.5.4), a fact where the middle time windows of audio modalities were critical in its performance. Nonetheless, noise in video modality has little or no impact on the performance. According to our inspection, the noise in video modality has a regularization factor, especially when it is used in a few time windows. Finally, the performance tendency in audio-video

MHSA	Noisy Input	# Time Windows	Mean Accuracy ± std	Noise Position
✓	audio	6	64.3 ± 0.10	two-first
✓	audio	6	64.2 ± 0.20	two-middle
✓	audio	6	65.6 ± 0.10	two-end
✓	audio	6	65.1 ± 0.27	half-global
✗	audio	6	58.8 ± 0.11	two-first
✗	audio	6	58.1 ± 0.16	two-middle
✗	audio	6	59.9 ± 0.05	two-end
✗	audio	6	59.3 ± 0.30	half-global
✓	audio	12	65.5 ± 0.10	two-first
✓	audio	12	64.9 ± 0.15	two-middle
✓	audio	12	65.6 ± 0.11	two-end
✓	audio	12	64.5 ± 0.23	half-global
✗	audio	12	59.6 ± 0.50	two-first
✗	audio	12	57.6 ± 0.11	two-middle
✗	audio	12	59.6 ± 0.06	two-end
✗	audio	12	57.7 ± 0.28	half-global
✓	video	6	69.0 ± 0.06	two-first
✓	video	6	68.1 ± 0.10	two-middle
✓	video	6	67.3 ± 0.14	two-end
✓	video	6	68.1 ± 0.19	half-global
✗	video	6	65.7 ± 0.11	two-first
✗	video	6	65.3 ± 0.06	two-middle
✗	video	6	64.4 ± 0.09	two-end
✗	video	6	64.8 ± 0.16	half-global
✓	video	12	68.5 ± 0.08	two-first
✓	video	12	68.0 ± 0.09	two-middle
✓	video	12	67.6 ± 0.07	two-end
✓	video	12	68.0 ± 0.12	half-global
✗	video	12	61.3 ± 0.05	two-first
✗	video	12	60.8 ± 0.12	two-middle
✗	video	12	60.6 ± 0.04	two-end
✗	video	12	61.9 ± 0.25	half-global
✓	audio-video	6	66.1 ± 0.20	two-first
✓	audio-video	6	65.3 ± 0.18	two-middle
✓	audio-video	6	65.8 ± 0.09	two-end
✓	audio-video	6	66.3 ± 0.32	half-global
✗	audio-video	6	61.9 ± 0.07	two-first
✗	audio-video	6	59.9 ± 0.22	two-middle
✗	audio-video	6	61.6 ± 0.10	two-end
✗	audio-video	6	61.5 ± 0.28	half-global
✓	audio-video	12	66.4 ± 0.14	two-first
✓	audio-video	12	65.2 ± 0.13	two-middle
✓	audio-video	12	65.1 ± 0.14	two-end
✓	audio-video	12	64.3 ± 0.25	half-global
✗	audio-video	12	60.4 ± 0.10	two-first
✗	audio-video	12	58.1 ± 0.11	two-middle
✗	audio-video	12	59.5 ± 0.04	two-end
✗	audio-video	12	58.6 ± 0.15	half-global

Table 7.5: CREMA-D: The results of the re-trained models with global noise and evaluation them with noise at specific time windows.

MHSA	Noisy Input	# Time Windows	Mean Accuracy ± std	Noise Position
✓	audio	8	73.2 ± 0.28	two-first
✓	audio	8	70.5 ± 0.61	two-middle
✓	audio	8	72.1 ± 0.27	two-end
✓	audio	8	71.4 ± 0.80	half-global
✗	audio	8	68.1 ± 0.35	two-first
✗	audio	8	63.8 ± 0.32	two-middle
✗	audio	8	66.8 ± 0.26	two-end
✗	audio	8	66.0 ± 0.43	half-global
✓	audio	16	71.5 ± 0.18	two-first
✓	audio	16	71.1 ± 0.34	two-middle
✓	audio	16	71.5 ± 0.17	two-end
✓	audio	16	69.9 ± 0.65	half-global
✗	audio	16	61.4 ± 0.31	two-first
✗	audio	16	60.4 ± 0.24	two-middle
✗	audio	16	61.4 ± 0.22	two-end
✗	audio	16	59.8 ± 0.44	half-global
✓	video	8	76.6 ± 0.15	two-first
✓	video	8	76.9 ± 0.31	two-middle
✓	video	8	75.9 ± 0.17	two-end
✓	video	8	76.3 ± 0.37	half-global
✗	video	8	71.1 ± 0.20	two-first
✗	video	8	71.9 ± 0.30	two-middle
✗	video	8	71.3 ± 0.20	two-end
✗	video	8	71.3 ± 0.55	half-global
✓	video	16	74.3 ± 0.16	two-first
✓	video	16	74.1 ± 0.16	two-middle
✓	video	16	73.7 ± 0.18	two-end
✓	video	16	73.7 ± 0.28	half-global
✗	video	16	68.4 ± 0.17	two-first
✗	video	16	69.0 ± 0.20	two-middle
✗	video	16	68.4 ± 0.10	two-end
✗	video	16	69.9 ± 0.50	half-global
✓	audio-video	8	75.7 ± 0.29	two-first
✓	audio-video	8	71.3 ± 0.38	two-middle
✓	audio-video	8	74.0 ± 0.33	two-end
✓	audio-video	8	74.0 ± 0.62	half-global
✗	audio-video	8	67.0 ± 0.29	two-first
✗	audio-video	8	64.4 ± 0.42	two-middle
✗	audio-video	8	66.5 ± 0.19	two-end
✗	audio-video	8	66.4 ± 0.86	half-global
✓	audio-video	16	71.4 ± 0.15	two-first
✓	audio-video	16	70.2 ± 0.35	two-middle
✓	audio-video	16	70.9 ± 0.20	two-end
✓	audio-video	16	69.5 ± 0.72	half-global
✗	audio-video	16	64.5 ± 0.13	two-first
✗	audio-video	16	62.9 ± 0.20	two-middle
✗	audio-video	16	64.3 ± 0.17	two-end
✗	audio-video	16	63.4 ± 0.66	half-global

Table 7.6: RAVDESS: the results of the re-trained models with global noise and evaluation them with noise at specific time windows.



(a) audio (b) video (c) audio-video

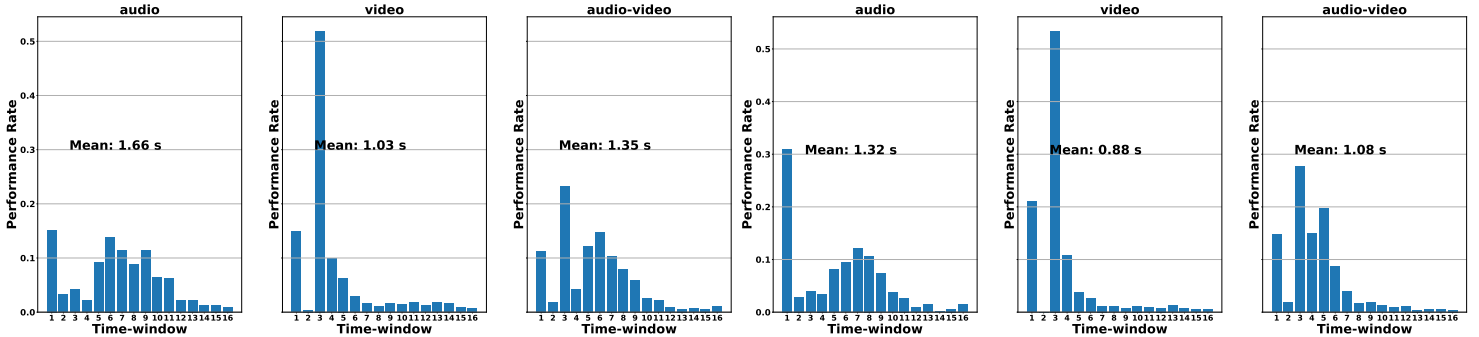
(d) audio (e) video (f) audio-video

Figure 7.23: CREMA-D: The results of retraining and evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The title of each sub-figure indicates the modality in which noise was used.

fusion, in this scheme of noise positions, follows the patterns of the underlying modalities, a reason why the performance dropped when noise was placed in the middle time windows.

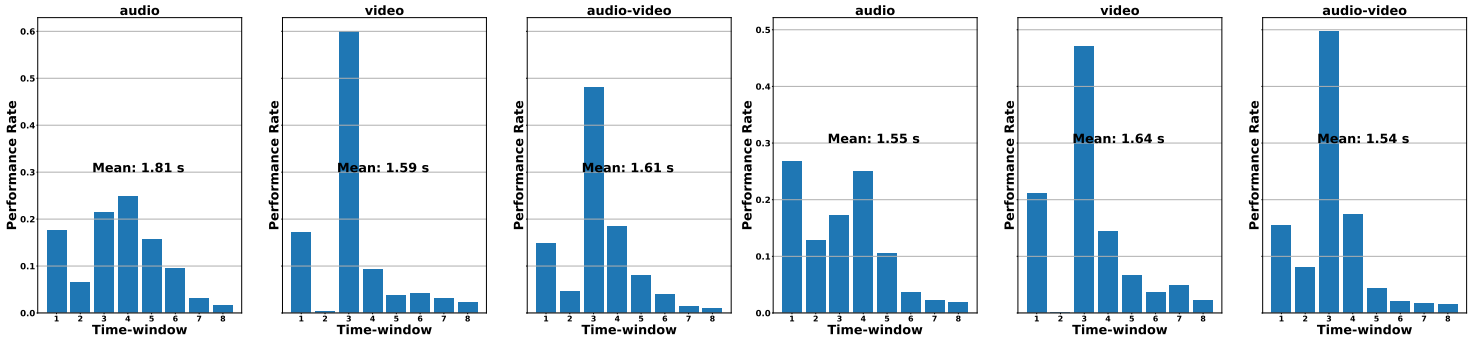
A meta-analysis of the CREMA-D results in Table 7.5, of the four different scenarios, shows that, for $T \in \{6, 12\}$, placing the noise in audio-only, video-only, and audio-video-modalities resulted in following accuracy changes (+ improvement or − deterioration): $\{-2.6\%, -0.46\%\}$, $\{+2.1\%, +2.3\%\}$, $\{-0.96\%, -0.15\%\}$, respectively. Similarly, the meta-analysis on RAVDESS results in Table 7.6 reveals that injecting the noise in audio-only, video-only, and audio-video inputs, for $T \in \{8, 16\}$, resulted in the following accuracy changes: $\{-3.9\%, -4.8\%\}$, $\{+1.1\%, +0.7\%\}$, $\{-2.9\%, -3.6\%\}$, respectively. These results show two outcomes, first, when the number of the time windows is increased, the case where their length is smaller, the performance difference is less. Second, placing the noise in audio modality has a greater impact on the multimodal performance, than using noisy embedding on video embedding.

Finally, as our observations in Subsection 7.5.4 suggest the importance of the initial time windows for video modality, Figures 7.24 and 7.25 examine the behavior of video performance in terms of response time when noise is added at the first two time windows. We notice that the mean response time of video-modality increased due to its reliance on the later time windows. This implies the fact that, in the overall performance of video modalities, time windows have equal contributions to the performance of the



(a) RAVDESS with attention

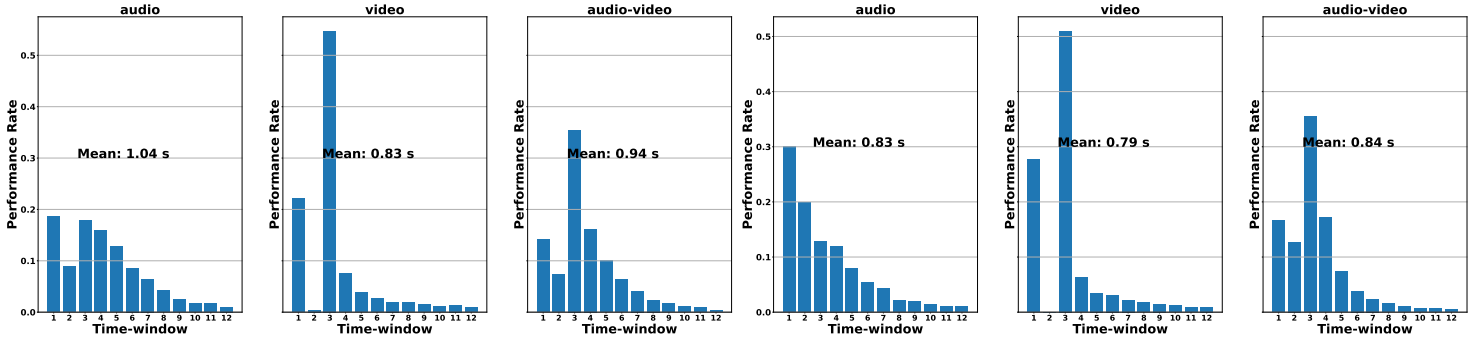
(b) RAVDESS without attention



(c) RAVDESS with attention

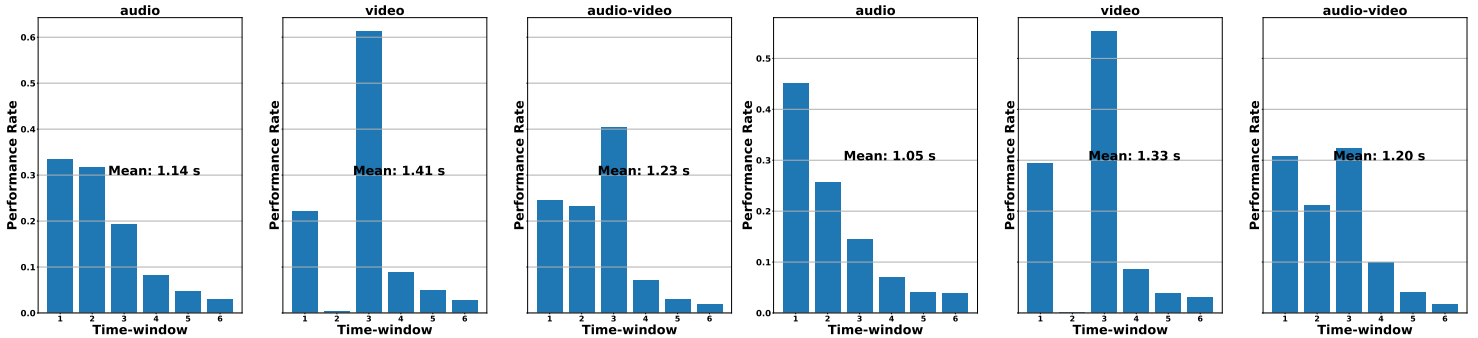
(d) RAVDESS without attention

Figure 7.24: RAVDESS: average response time when the proposed method is re-trained with noise in video modality and evaluated by injecting (using) noise in the first two-time windows of the framework. We notice that the response time of video modality was shifted to the later time windows, a reason why its performances were not deteriorated by the noise.



(a) CREMA-D with attention

(b) CREMA-D without attention



(c) CREMA-D with attention

(d) CREMA-D without attention

Figure 7.25: CREMA-D: average response time when the proposed method is re-trained with noise in video modality and evaluated by injecting (using) noise in the first two-time windows of the framework. We notice that the response time of video modality was shifted to the later time windows, a reason why its performances were not deteriorated by the noise.

framework. Furthermore, there is not a difference between the usage of the MHSA and the baseline, other than the considerable difference between the two approaches' response time across modalities. This difference is not related to the usage of the noise (more implications and explanations are discussed in Subsection 7.5.4).

7.7. CONCLUSIONS

The research in this chapter highlights the importance of exploiting audio-video signals' temporal strength for emotion recognition. We utilize the attention mechanism on audio-visual embeddings over time windows to leverage their properties for emotion recognition. Evaluation on two datasets shows that the proposed method with the transformer attention mechanism significantly improves the performance over the baseline (which is the same topology without the attention mechanism's selective character). Our results indicate the importance of weighing the contribution, not only of each modality separately but also of different time windows. In addition, the proposed framework provides a close interpretation and insights regarding multimodal emotion recognition, making use of visual and audio cues. Specifically, the meta-analysis performed (Section 7.5 and Section 7.6), led to the following conclusions:

- The attention mechanism is able to utilize embeddings from all time windows and to capture how video and audio modalities behave across time. For example, the examination of the framework showed that the attention mechanism helps in capturing the incremental presentation of the audio-video embeddings, resulting in the best performance in the last time window.
- The temporal perception of audio-visual cues shows that recognition rates increase faster for positive emotions than negative ones. For example, the duration of an expressed negative emotion plays a larger role in recognizing it. Whereas, for positive ones, a plateau in recognition accuracy can be reached earlier in time. These findings are in alignment with the ones in Chapter 6 and those coming from perceptual studies of how humans rate emotions across time.
- The analysis of the multimodal interaction showed that the contribution of the video modality in the multimodal fusion is greater than the one of the audio modality. This is inline with human raters when they were presented with audio and video modalities. Besides, attention mechanisms efficiently integrated the audio modality in the bimodal perception for enhanced emotion recognition.
- The joint modeling of audio-visual cues using the attention mechanism helped to bring the entropies of the audio and video modalities closer. This modeling increased the certainty in the multimodal predictions, which consequently enhanced the multimodal perception, compared to the baseline model.
- Evaluating the framework with noise demonstrated that the method is robust when it is exposed to similar conditions during the training and testing procedures. Finally, injecting noise into the framework showed that the audio modality is more vulnerable to noisy data, while the video modality is more robust against noise.

8

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation addressed the automatic multimodal emotion recognition using audio and visual cues. It approached this problem through several methods and proposed techniques to exploit the complementary and supplementary information of audio and video channels. The studies conducted in this dissertation targeted several aspects of multimodal emotion recognition, such as investigating various representations of audio and visual cues, proposing fusion algorithms, utilizing the temporal information in a video clip to integrate both modalities, and studying the patterns and behavior of audio-video signals for emotion recognition as a function of time. This chapter summarizes the findings of the research in this dissertation (in Section 8.1). Next, in Section 8.2, it highlights potential directions for the research in automatic emotion recognition.

8.1. CONCLUSIONS OF RESEARCH QUESTIONS AND OBJECTIVES

IN this dissertation, as proposed in Chapter 1, four research questions were formulated to guide the flow of conducted studies. These questions were addressed in the content chapters and their connections to address the joint modeling of audio-visual cues for Multimodal Emotion Recognition (MER) are discussed. This section gives an overview of the answers to each of the listed research questions.

8.1.1. OBJECTIVE 1: OBTAINING ROBUST DATA MODELING AND REPRESENTATION FOR EMOTION RECOGNITION

Automatic emotion recognition relies on the model of emotions, data, and mathematical modeling for data representations and automatic inference [2]. These steps might vary according to applications of emotion recognition. Various algorithms can be developed to handle different phases in a pipeline of a system in affective computing. Having this perspective, we formulated the following question:

First research question: *How to extract and fuse robust features and which is their contribution to automatic emotion recognition?*

The fourth chapter addresses this question in two studies. The first one deals with emotion recognition in video clips, where audio and visual cues are the primary information for emotion perception. It presented a framework for multimodal hierarchical emotion recognition and was evaluated on two datasets. The proposed research employed Fisher vector-based representations to capture the discriminative and temporal information across the frames in a video sample. This encoding was applied to different types of audio and visual features (e.g. Dense Scale-Invariant Feature Transformation (SIFT), geometric, Convolutional Neural Network (CNN), audio) enabling mapping them into a common space, where feature level fusion is performed. The strategy of employing information gain principles, for selecting the best combination of features to be fused, achieved more accurate performance than when all the features vectors are concatenated. The decision-level fusion approach on top of the best features was designed to optimize the modality weights for each emotional state using a genetic search algorithm, which led to results that surpassed the baselines and the unimodal performances.

The second study investigated the link between students' self-reported affective states, according to the theory of flow, and their interactions with learning materials. This study designed a framework to track contextual information and interaction features during learning activities. It utilized a standard tracking tool, the xAPI framework, for learning analytics. This research made use of data collected through a serious game, as the result of work of other colleagues in the research team, described in [222]. Using the Theory of Flow as an affect model, the students self-reported their experienced affective state during their interaction with the platform. The conducted evaluations highlighted the generalization ability of the system across students with different profiles. This system can be especially useful in boosting the adaptive nature of the learning process to improve the outcome of the learning experience, maximizing the learners' knowledge acquisition, and enabling personalization.

8.1.2. OBJECTIVE 2: BUILDING EFFICIENT FUSION OF AUDIO-VISUAL REPRESENTATIONS

The study in Chapter 4 elaborated on various data representations and their usability for emotion inference, in two different scenarios. Nonetheless, there should be dedicated solutions to target an efficient fusion of audio-visual cues for emotion recognition. The implications of an efficient fusion of this multi-sources information inspired us to ask the following research question:

Second research question: *What is the impact of multimodal learning and fusion on emotion recognition?*

Chapter 5 introduced a modality-specific joint multimodal metric learning for audio-video emotion recognition. The proposed Multimodal Emotion Recognition Metric Learning (MERML) was applied to improve the latent-space of audio-visual data representation. This approach exploited successfully the dependencies and the complementary information of audio and video modalities in the context of emotion recognition,

as their representations are well structured in the newly learned subspace, and their capacity for optimized emotion recognition is maximized. The conducted quantitative and qualitative evaluations of the method on several datasets, utilizing distinct pairs of visual and audio representations, demonstrated the significant contribution of the method to increased classification accuracy. Furthermore, the comparison with baseline metric learning algorithms showed the benefits of our method, which is efficiently learning the two modalities and optimizes their contribution to enhanced performance.

8.1.3. OBJECTIVE 3: EXPLOITING THE TEMPORAL DYNAMICS OF EMOTION EXPRESSION AND PERCEPTION

The previous study proposed a global solution for audio-visual emotion recognition. However, emotion display and perception are sophisticated processes. Both can vary as a function of time, and this depends on the given modality and the intended/perceived emotion [77]. In addition, one possible direction of MERML is to improve the audio and visual representations, through an end-to-end learning scheme, starting from the raw audio-visual data to emotion labels. This can be done using a deep similarity learning to handle the highly non-linear relationships between audio and video modalities and to group the similar data together and the dissimilar ones apart [71], based on the emotion categories. These facts and ideas led us to formulate the following question:

> **Third research question:** *What is the role of temporal dynamics in audio visual cues, in automated emotion recognition?*

In chapter 6, we proposed an end-to-end multimodal and temporal Deep Metric Learning (DML) framework for Audio-Video Emotion Recognition (AVER). The novel method embeds audio-visual cues diachronically, taking the advantages of the temporal display of emotions. The procedure employs Long-Short Term Memory (LSTM)s between time windows for incremental perception and applies the gating paradigm [247]. The framework showed efficiency in modeling the temporal context of multimodal emotion recognition. In addition, within this framework, algorithms for triplet sets' mining and data augmentation were developed. The developed method and the associated techniques, such as Multi-Window Triplet Sets Mining (MWTSM), contributed significantly to the stability and the performance of the framework. Also, the obtained results are significantly higher than the baseline results and boost the performances for two public datasets. The evaluations showed that the incremental perception of both audio and visual cues improves the recognition rates overtime. We noticed that increasing the number and the length of the time windows improves the accuracy of emotion recognition. Additionally, the temporal differences of the recognition speed for positive and negative emotions differ, where positive emotions are recognized faster than the negative ones.

8.1.4. OBJECTIVE 4: PRODUCING AN ATTENTIVE SYSTEM TO MULTIMODAL AND TEMPORAL EXPRESSIONS OF EMOTIONS

Studying the temporal nature of emotions and the contributions of audio-visual signals to identify this phenomenon, showed us the importance of the temporal and multimodal

information. Emerging techniques in machine learning, such as attention mechanisms [104], are intriguing and applicable in the domain of temporal emotion recognition. Chapter 7 employed these techniques to address the following research question:

> **Fourth research question:** *How can we capture the contributions of the temporal dynamics of affect display using attention mechanisms?*

The presented method examines the importance of audio-video signals' temporal strength for emotion recognition. It employed an attention mechanism on audio-visual embeddings over time windows to leverage their properties for emotion recognition. The proposed method improved the performance over the baseline (which is the same topology without the selective character of the attention mechanisms) significantly. The results highlighted the importance of weighing the importance, not only of each modality separately but of different time windows, as well. Besides, the framework gave more insights with regards to the multimodal interaction and presentation of audio-visual cues. For instance, the examination of the framework showed that the attention mechanism helps in capturing the temporal presentation of the audio-video embeddings, resulting in the best performance in the last time window. In addition, the temporal perception revealed that the duration of an expressed negative emotion plays a larger role in recognizing it, whereas, for positive ones, a plateau in recognition accuracy can be reached earlier.

Moreover, this study showed that the video modality contribution in multimodal fusion is more significant than the one of the audio modality. Another finding of this study was that the joint modeling of audio-visual cues, via the proposed method, using the attention mechanism helped increasing models' certainty of the bimodal perception. This observation is validated by estimating entropies on the unimodal outputs, where attention mechanisms reduced the gap between the audio and video modalities' entropies. As a result, this modeling enhanced the multimodal perception compared to the baseline model. Finally, applying noise injection into the time windows' embeddings during the evaluation or/and the training phases demonstrated that the method is accurate when noise is employed in both training and testing processes. In a broader aspect, attention mechanisms increased the framework's robustness when exposed to noisy conditions during the training and the evaluation phases. Finally, our evaluation shows that the audio modality is more prone to noise injection than the video modality. In contrast, the video modality can recover from noisy data if enough time windows are applicable. This conclusion also shows that individual frames in a video sequence of facial expressions can be more informative than audio time windows with a small duration.

8.2. DISCUSSION AND FUTURE DIRECTIONS

Due to the nature of emotions, their descriptions and representations are challenging [3, 6]. In the field of affective computing, this is reflected in various aspects, such as data collection, annotation, and, naturally, the automatic inference of emotions. For example, obtaining large scale datasets is a notoriously difficult task due to the ambiguity of emotion labels [78]. Advances in the field of artificial intelligence can allow us to address these challenges. This section summarizes potential directions for research to adapt and

investigate the developments of our understanding of multimodal and temporal emotion expression, perception, as well as emotion representations.

AFFECTIVE DATA FOR AFFECTIVE COMPUTING

Computational methods in Affective Computing benefit from different fields such as computer vision, machine learning, pervasive computing, and psychology. Affective Computing domains, e.g. emotion recognition and personality computing, usually rely on data-driven approaches. These approaches require representative data and accurate annotations. As discussed in Section 3.1, one of the challenges in Affective Computing is the limited availability of labeled data. This is because of the complexity of generating a multimodal corpus and due to the fact that emotions have an ambiguous nature, which makes instance labeling hard. On the other hand, with the recent technological developments, diverse sensors can be used to collect affective data. Multimodal data has been shown to provide complementary and supplementary information to enhance emotion recognition through automatic systems. Future work should focus on producing large, diverse, and naturalistic (spontaneous) corpora that have physiological, audio, and video modalities with variant affect measures and models [281]. The resulting multimodal and naturalistic corpora will help the Affective Computing goals in understanding human emotion expression and recognition [281]. Furthermore, having access to representative and larger corpora will help the field of Affective Computing in producing robust and automatic recognition systems, which, subsequently, will accelerate the field's effort towards achieving its goals, i.e. recognizing, simulating, and inducing emotions.

Moreover, emotions are multilayer subjective affective states that can vary according to many factors, such as personal and environmental context, mood, personality, and culture. Context, personality, and subjective experiences should be included in efforts towards affective data gathering and labeling. Such information can be included through self-assessment. Self-assessment can be beneficial in reporting emotions and providing context-related information [6]. Models for self-reports include Positive and Negative Affect Schedule (PANAS) [282] and The State-Trait Emotion Measure (STEM) [283], which was recently framed to assess negative and positive emotions at the workplace. Besides, incorporating demographic data, personal characteristics, cultural background, and context information will enable data-driven methods to achieve technical breakthroughs with accurate performances. Moreover, having such information, in addition to multimodal and naturalistic data, will enable Affective Computing systems to produce explainable decisions. For example, utilizing affective data with rich context, efforts towards emotion recognition should emphasize explaining the relationship between emotional expressions and the cognitive appraisals behind generating these behavioral responses. Finally, affect and context-rich data can contribute to Human-Computer Interaction's personalization, enabling human-centered automatic systems.

EXPLORING EMOTION LATENT REPRESENTATIONS VIA SIMILARITY LEARNING

Metric learning aims to group data with similar labels while putting the dissimilar ones further apart [71]. Emotions can be influenced by context such as subject, gender, and culture [284]. In addition, according to The Wheel of Emotion model [16], emotions themselves can be placed hierarchically as well be grouped according to their closeness to each other. For example, in this model, anger and disgust are closely related, while

they are in the opposite position to joy and happiness. This might be a cultural and semantic representation of these emotions, however, it is interesting to explore these dynamics between emotions in the latent space of multimodal data representations. In fact, information such as culture, subject, and the relationship between emotions themselves can be embedded and exploited to make their features reflect these semantic relationships.

PATCH BASED AUDIOVISUAL FUSION

Facial expressions are a central part of human-human communications. They carry a lot of information and are one of the main channels in social interactions. Facial Action Coding System (FACS) [46] suggests that physical movements of facial muscles can assist in understanding emotions. Multimodal and temporal perception of emotions can employ techniques that exploit facial patches and correlate them to the audio channel. Our study in Chapter 4 found that the contribution of mouth temporal representations, when used with audio, is far more significant than other facial patches. An intriguing research question is to handle these findings, such that an automatic system can be built to fuse the temporal display of emotions through facial patches and align them with audio cues to enhance emotion perception.

SELF-SUPERVISED LEARNING FOR SPATIO-TEMPORAL MULTIMODAL EMOTION RECOGNITION

Emerging learning systems based on contrastive losses can benefit from spatial and temporal information for self-supervised learning [285]. This learning paradigm can be embedded in a multimodal context for building audio and visual representations in an unsupervised manner. Techniques such as data augmentation, time information of video clips, as well as the spatial nature of facial expressions can be used to develop automatic recognition systems, and reduce the requirement of big data. Besides, the knowledge obtained from facial expressions tracks can be transferred through cross-modal knowledge distillation to improve audio representations, as the two modalities have strong correlations for emotion predictions [78].

IMPACT PARAGRAPH

In this addendum, a discussion is presented to introduce the scientific and social impact of the conducted research in this dissertation, its results, and the proposed methods. The research in this dissertation can be transferred to different tasks in Human-Computer Interaction (HCI) and Affective Computing (AC), as well as be implemented in various applications. These tasks and applications have enormous social and economical interests in the current society and the future. According to Maastricht University's "Regulations for obtaining the doctoral degree Maastricht University" [1]*, the scientific impact includes short-term and long-term contributions of the conducted research and its results to shifting insights and stimulating science, methods, results, theory, and applications. On the other hand, the social impact is the short-term and long-term contributions of the conducted research to changes in or development of social sectors and to social challenges. This paragraph is addressing the drafted four questions in the given regulations, which are related to the main objective of the research and its relevance, target group, and activity.*

> **Research:** *what is the main objective of the research described in the thesis and what are the most important results and conclusions?*

The main objective of this dissertation is to address a fundamental research problem in Affective Computing (AC): Multimodal Emotion Recognition (MER) from audio and visual cues. It approaches the problem from different perspectives with various methods to enhance the performance of emotion recognition in video clips. Chapter 2 introduces the state-of-the-art approaches in Artificial Intelligence (AI) which are used in the fields of Human-Computer Interaction (HCI), Affective Computing (AC), and, in particular, in this dissertation. Furthermore, Chapter 3 presents state-of-the-art technologies, datasets, applications, modalities' representations, learning schemes, and fusion techniques for MER. Besides, the subsequent chapters introduce the proposed methods, their findings, and conclusions as follows:

- In two studies, the research in Chapter 4 demonstrates the importance of multimodal features, their fusion, as well as an application of automatic emotion inference in educational settings. The first study, in this chapter, shows the impact of fusing different feature representations from audio and video modalities for emotion perception. It proves that these two modalities are complementary to each other, and their features improve performance significantly. The second study examines the link between the self-reported affective states by students and their interactions with learning materials. It provides an example of utilizing learning analytic technologies and machine learning techniques for understanding student's

[1] `https://www.maastrichtuniversity.nl/support/phds`, retrieved on November 28, 2020.

emotions, which is an important aspect in the future of e-learning. The study finds a correlation between students' affects and their interaction parameters.

- The study in Chapter 5 focuses on improving the fusion algorithm to make use of audio-visual cues, efficiently. The proposed method, namely, Multimodal Emotion Recognition Metric Learning (MERML), shows the potential of powerful approaches such as metric learning for multimodal learning and fusion, which improved the latent space of audio-visual representations and subsequently the performance of emotion recognition.

- Chapter 6 follows the metric learning approach to produce joint multimodal and temporal feature representations for Audio-Video Emotion Recognition (AVER) using Deep Metric Learning (DML). This study exploits the temporal dynamics of audio and video signals using an end-to-end learning paradigm. In this chapter, the research demonstrates the importance of time information in the incremental presentation of audio-visual signals for emotion recognition. The proposed method was further adjusted to address another area in Affective Computing (AC), namely, personality recognition from bodily expressivities using motion and context information. Both studies prove the usability of Deep Metric Learning (DML), along with the proposed frameworks, for capturing multimodal data and modeling the temporal information in the field of Affective Computing (AC).

- Finally, Chapter 7 revisits the research problem of capturing the temporal dynamics of audio-visual modalities and further focuses on the idea of attending to informative time segments in these two modalities' cues. The research in this chapter employs attention mechanisms to address the research objective. The results highlight the importance of automatically attending to informative time slices. Furthermore, the study introduces a meta-analysis, linking the research findings and propositions to research in psychology. The results offer an insightful perspective of the performance of the audio-visual cues over time, such as multimodal recognition speed on positive and negative emotions, the contribution of each modality in the multimodal fusion, the importance of their joint learning via the attention mechanisms, and the robustness of the proposed framework in more challenging environments.

Relevance: *what is the (potential) contribution of the results from this research to science, and, if applicable, to social sectors and social challenges?*

The research in this dissertation aims at contributing to the goal of Affective Computing (AC), where the focus is to enhance emotionally incapable machines with emotional intelligence to improve human-machine interaction [164]. Each part of the conducted studies demonstrates the ability of the proposed solutions to perform Multimodal Emotion Recognition (MER), efficiently. Besides, the presented methods can be adapted to other tasks of Affective Computing (AC), as demonstrated in Chapter 6 with personality recognition, or other domains in Artificial Intelligence (AI) in general, where spatio-temporal and multimodal information persist, such as action recognition, person identification, and multimedia retrieval.

Moreover, many regard the current progress in Artificial Intelligence (AI) as a crucial part of the Fourth Industrial Revolution (Industry 4.0). The term Fourth Industrial Revolution was first referred to by Klaus Schwab in 2015, and published in [286]. It refers to the automation of manufacturing pipelines using the technological advances in the fields of Artificial Intelligence (AI), quantum computing, nanotechnology, the internet of things, etc. In this new era, where the focus is on machine-machine and human-machine communications, it is important to keep humans in the loop [287]. For this reason, within the advances of Artificial Intelligence (AI), a key factor to consider is emotional intelligence. Emotional intelligence refers to a set of skills that contribute to correct appraisal and expression of emotions, emotions modulation and regulation, and subsequently, the efficient usage of emotions in planning, working, and communications [288]. Emotional intelligence is essential in business, relationships, education, and other life aspects. Furthermore, the new era, brought by Industry 4.0, will lead to enormous consequences, that will shape our interactions with each other and the way we live, work, and learn [286, 287]. A key part of Affective Computing (AC) is emotion recognition. Indeed, obvious signals such as facial expressions and vocal utterances, which are addressed in this dissertation, can contribute to equipping machines with emotional intelligence. In this way, AI advancements are accompanied and supported with features to keep humans in new communication loops, namely, the machine-machine and human-machine communication channels.

Target group: *to whom are the research results interesting and/or relevant? And why?*

The primary target groups of the studies in this thesis are researchers in Human-Computer Interaction (HCI) and Affective Computing (AC) fields, and the field of Artificial Intelligence (AI) in general. For example, the proposed multimodal and temporal architectures can be applied in other tasks, which include, but are not limited to gesture recognition, personality computing, action recognition, and multimedia retrieval. Moreover, as discussed in Section 3.4, the applications of AC and Multimodal Emotion Recognition (MER) can range from education [13, 18, 19], automatic vehicle driving [22–26], health-care [4, 20, 21], to entertainment [27–30]. These applications are of great interest in our societies with a tremendous social and economic impact.

For example, the developed techniques can be used in education where Technology Enhanced Learning (TEL) brings new kinds of educational and learning experiences [164]. According to R. W. Picard *et al.* , TEL systems should incorporate the emotional aspect of the learning process, in addition to the cognitive process [164]. As a result, human emotional needs are considered, beyond aspects that address merely productivity and efficiency. Indeed, in an educational context, emotions experienced by a learner directly affect the learning outcome [165, 166]. In fact, in the course of this dissertation, we considered emotion understanding from facial expressions, vocal utterances, and students' interaction with learning materials. These three modalities can be part of an integrated learning platform that has affective capabilities to recognize learners' emotions and to respond to their individual needs. In other words, accurate automatic multimodal emotion recognition can be useful in enhancing the learning outcomes by providing personalized and adaptive educational processes according to students' emotions, as well

as other performance indicators related to productivity and cognitive skills. Another interesting area for the applications of Affective Computing (AC) and Multimodal Emotion Recognition (MER) is the recognition of drivers' affective states in automatic vehicle driving. For example, the developed methods within this dissertation can be used to fuse various sensorial data to infer driver's attention and stress-level. The sensorial data can include facial expressions, gaze, and physiological measurements. An affective system can ensure drivers' and other people's safety.

> **Activity:** *in what way can these target groups be involved in and informed about the research results, so that the knowledge gained can be used in the future?*

The studies in Chapters 4, 5, 6, and 7 have been published in various peer-reviewed conference proceedings and high-impact journals. At the beginning of each chapter, the papers which are parts of the corresponding chapter are listed. Moreover, throughout the course of the Ph.D. research, the proposed methods and the conclusions of their findings have been presented in the respective scientific venues. Furthermore, the Convolutional Neural Network (CNN) representations of the study in Chapter 4 were part of an interactive demo that has been used to demonstrate its abilities to recognize facial expressions in real-time. Besides, the study of correlating students' affective states with their interactions with the learning materials was one of the modalities in MaTHiSiS, which is a learning platform developed within the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS)[2].

[2] http://mathisis-project.eu/

BIBLIOGRAPHY

REFERENCES

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[2] S. D'Mello, A. Kappas, and J. Gratch, "The Affective Computing Approach to Affect Measurement," *Emotion Review*, vol. 10, no. 2, pp. 174–183, 2018.

[3] S. K. D'mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems," *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, 2015.

[4] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, Eva-Maria Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," *AVEC 2019 - Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, co-located with MM 2019*, no. Avec, pp. 3–12, 2019.

[5] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019.

[6] A. Kappas, "Social regulation of emotion: messy layers," *Frontiers in Psychology*, vol. 4, p. 51, 2013.

[7] R. W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker, "Affective learning—a manifesto," *BT Technology Journal*, vol. 22, no. 4, pp. 253–269, 2004.

[8] G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud, "Virtual agent multimodal mimicry of humans," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 367–388, 2007.

[9] R. W. Picard, *Affective computing*. MIT press, 2000.

[10] M. Pantic, *Facial expression analysis by computational intelligence techniques*. PhD thesis, TU Delft, 2002.

[11] M. Minsky, *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster, 2007.

[12] Y. N. Harari, *Sapiens: A brief history of humankind*. Random House, 2014.

[13] R. W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker, "Affective learning - a manifesto," *BT Technology Journal*, vol. 22, no. 4, pp. 253–269, 2004.

[14] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[15] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion.," *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[16] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, and other, "Multimodal deep learning," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML-11)*, pp. 689–696, 2011.

[18] E. Ghaleb, M. Popa, E. Hortal, S. Asteriadis, and G. Weiss, "Towards Affect Recognition through Interactions with Learning Materials," *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pp. 372–379, 2019.

[19] C.-H. Wu, Y.-M. Huang, and J.-P. Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.

[20] F. Alvarez, M. Popa, V. Solachidis, G. Hernández-Peñaloza, A. Belmonte-Hernández, S. Asteriadis, N. Vretos, M. Quintana, T. Theodoridis, D. Dotti, *et al.*, "Behavior analysis through multimodal sensing for care of parkinson's and alzheimer's patients," *IEEE Multimedia*, vol. 25, no. 1, pp. 14–25, 2018.

[21] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *International Symposium on Visual Computing*, pp. 368–377, Springer, 2012.

[22] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[23] C. D. Katsis, G. Rigas, Y. Goletsis, and D. I. Fotiadis, "Emotion recognition in car industry," *Emotion Recognition: A Pattern Analysis Approach*, pp. 515–544, 2015.

[24] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2014–2027, 2015.

[25] G. Moore, "Emotional Drive Wearing your heart on your car," *HCI 2017: Digital Make Believe - Proceedings of the 31st International BCS Human Computer Interaction Conference, HCI 2017*, vol. 2017-July, pp. 1–4, 2017.

[26] L.-l. Chen, Y. Zhao, P.-f. Ye, J. Zhang, and J.-z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, 2017.

[27] G. N. Yannakakis and J. Togelius, *Artificial intelligence and games*, vol. 2. Springer, 2018.

[28] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[29] K. Karpouzis, G. N. Yannakakis, N. Shaker, and S. Asteriadis, "The platformer experience dataset," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 712–718, IEEE, 2015.

[30] P. M. Blom, S. Bakkes, C. T. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers, "Towards personalised gaming via facial expression recognition," in *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[31] S. Marsella, J. Gratch, P. Petta, *et al.*, "Computational models of emotion," *A Blueprint for Affective Computing-A Sourcebook and Manual*, vol. 11, no. 1, pp. 21–46, 2010.

[32] R. Reisenzein, "A short history of psychological perspectives on emotion," in *The Oxford Handbook of Affective Computing*, pp. 21–37, Oxford University Press UK, 2015.

[33] R. S. Lazarus, "Progress on a cognitive-motivational-relational theory of emotion.," *American Psychologist*, vol. 46, no. 8, p. 819, 1991.

[34] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.

[35] W. McDougall, *An introduction to social psychology*. Psychology Press, 2015.

[36] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.," *Journal of Personality and Social Psychology*, vol. 76, no. 5, p. 805, 1999.

[37] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial expressions of emotion.," *Handbook of Emotions*, p. 211–234, 2008.

[38] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[39] A. Mehrabian, *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*, vol. 2. Oelgeschlager, Gunn & Hain Cambridge, MA, 1980.

[40] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[41] N. Dael, M. Mortillaro, and K. R. Scherer, "Emotion expression in body action and posture.," *Emotion*, vol. 12, no. 5, pp. 1085–1101, 2012.

[42] N. Dael, M. Goudbeek, and K. R. Scherer, "Perceived gesture dynamics in nonverbal expression of emotion," *Perception*, vol. 42, no. 6, pp. 642–657, 2013.

[43] K. M. Loewenthal and C. A. Lewis, *An introduction to psychological tests and scales*. Psychology press, 2018.

[44] H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, "Automatic emotion recognition based on body movement analysis: a survey," *IEEE Computer Graphics and Applications*, vol. 34, no. 6, pp. 35–45, 2014.

[45] G.-B. Duchenne and G.-B. D. de Boulogne, *The mechanism of human facial expression*. Cambridge university press, 1990.

[46] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[47] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: The manual on cd rom," *A Human Face, Salt Lake City*, pp. 77–254, 2002.

[48] B. De Gelder and J. Vroomen, "The perception of emotions by ear and by eye," *Cognition & Emotion*, vol. 14, no. 3, pp. 289–311, 2000.

[49] A. S. Walker-Andrews, "Intermodal emotional processes in infancy," *Handbook of Emotions*, pp. 364–375, 2008.

[50] J.-A. Bachorowski and M. J. Owren, "Vocal expressions of emotion," *Handbook of Emotions*, vol. 3, pp. 196–210, 2008.

[51] J. A. Bachorowski and M. J. Owren, "Vocal expression of emotion: Acoustic Properties of Speech Are Associated With Emotional Intensity and Context," *Psychological Science*, vol. 6, no. 4, pp. 219–224, 1995.

[52] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[53] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[54] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handbook of Affective Sciences*, pp. 433–456, 2003.

[55] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, no. 2, pp. 53–57, 1999.

[56] K. R. Scherer, "Adding the affective dimension: a new look in speech analysis and synthesis.," in *ICSLP*, Citeseer, 1996.

[57] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[58] V. Aubergé and M. Cathiard, "Can we hear the prosody of smile?," *Speech Commun.*, vol. 40, pp. 87–97, Apr. 2003.

[59] D. Keltner and D. T. Cordaro, "Understanding multimodal emotional expressions: Recent advances in basic emotion theory," *The Science of Facial Expression*, pp. 57–75, 2017.

[60] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

[61] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[62] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *Int. Conf. on Pattern Recognition (ICPR)*, pp. 4125–4128, 2010.

[63] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, "Evaluating aam fitting methods for facial expression recognition," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–8, IEEE, 2009.

[64] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2015.

[65] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, N. Vidrascu, L. K. Amir, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Proceedings of IS-LTC*, pp. 240–245, 2006.

[66] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 494–501, ACM, 2014.

[67] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 517–524, ACM, 2013.

[68] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, and Others, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, pp. 543–550, ACM, 2013.

[69] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[70] S. Chen, X. Li, Q. Jin, *et al.*, "Video emotion recognition in the wild based on fusion of multimodal features," in *Proc. of the 18th ACM International Conference on Multimodal Interaction*, pp. 494–500, ACM, 2016.

[71] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[72] Z. Xu, K. Q. Weinberger, and O. Chapelle, "Distance metric learning for kernel machines," *ArXiv preprint arXiv:1208.3422*, 2012.

[73] Y. Kim and E. M. Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 92–99, ACM, 2016.

[74] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, 2017.

[75] K. Edwards, "The face of time: Temporal cues in facial expressions of emotion," *Psychological Science*, vol. 9, no. 4, pp. 270–276, 1998.

[76] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS One*, vol. 6, no. 11, p. e27256, 2011.

[77] P. Barkhuysen, E. Krahmer, and M. Swerts, "Crossmodal and incremental perception of audiovisual cues to emotional speech," *Language and Speech*, vol. 53, no. 1, pp. 3–30, 2010.

[78] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 292–301, 2018.

[79] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[80] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[81] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[82] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, IEEE, 2006.

[83] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. of the 24th Int. Conf. on Machine learning*, pp. 209–216, 2007.

[84] P. H. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in *Proc. of the 33rd Int. Conf. on Machine Learning (ICML)*, pp. 19–24, 2016.

[85] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3025–3032, 2013.

[86] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*, pp. 566–579, Springer, 2012.

[87] B. Kulis, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[88] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[89] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[90] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[91] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[92] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[93] A. Karpathy *et al.*, "Cs231n convolutional neural networks for visual recognition." `https://cs231n.github.io/`, 2016. Accessed: 27-04-2021.

[94] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[96] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 41.1–41.12, BMVA Press, September 2015.

[97] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *ArXiv Areprint arXiv:1405.3531*, 2014.

[98] C. Olah, "Understanding lstm networks." `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`, 2015. Accessed: 27-04-2021.

[99] M. Chen, G. Ding, S. Zhao, H. Chen, Q. Liu, and J. Han, "Reference based lstm for image captioning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[100] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[101] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional cnn-lstm model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 225–230, 2016.

[102] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *International Conference on Learning Representations (ICLR)*, 2015.

[103] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[104] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[105] K. P. Murphy, *Probabilistic Machine Learning: An introduction.* MIT Press, 2021.

[106] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[107] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

[108] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3104–3112, 2014.

[109] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1875–1882, 2014.

[110] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92, Springer, 2015.

[111] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, pp. 1857–1865, 2016.

[112] I. Newton, "Letter from Sir Isaac Newton to Robert Hooke," *Historical Society of Pennsylvania. Available at: http://digitallibrary.hsp.org/index.php/Detail/Object/Show/object_id/9285*, 2016.

[113] J. Kossaifi, B. W. Schuller, K. Star, E. Hajiyev, M. Pantic, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, and A. Toisoul, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[114] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE Int. Conf. on*, pp. 1–8, IEEE, 2013.

[115] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, 2018.

[116] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," 2006.

[117] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emoti W 2016: Video and group-level emotion recognition challenges," *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 427–432, 2016.

[118] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[119] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[120] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.

[121] M. Yanagimoto and C. Sugimoto, "Recognition of persisting emotional valence from eeg using convolutional neural networks," in *2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA)*, pp. 27–32, IEEE, 2016.

[122] J. Li, Z. Zhang, and H. He, "Implementation of eeg emotion recognition system based on hierarchical convolutional neural networks," in *International Conference on Brain Inspired Cognitive Systems*, pp. 22–33, Springer, 2016.

[123] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pp. 95–108, 2016.

[124] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[125] C.-C. Lee, J. Kim, A. Metallinou, C. Busso, S. Lee, and S. S. Narayanan, "Speech in affective computing," in *The Oxford Handbook of Affective Computing*, pp. 170–183, Oxford University Press, 2014.

[126] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[127] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Int. Conf. on Affective Computing and Intelligent Interaction*, pp. 415–424, 2011.

[128] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[129] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 8, pp. 4898–4901, IEEE, 2005.

[130] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, IEEE, 2016.

[131] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pp. 801–804, 2014.

[132] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231, IEEE, 2017.

[133] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-January, pp. 1537–1540, 2015.

[134] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *ArXiv*, vol. abs/1609.08675, 2016.

[135] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, pp. 892–900, 2016.

[136] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multicultural dimensional emotion recognition in dyadic interactions," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 37–45, 2019.

[137] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 19–26, 2019.

[138] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[139] R. T. Ionescu, M. Popescu, and C. Grozea, "Local learning to improve bag of visual words model for facial expression recognition," in *Workshop on Challenges in Representation Learning at ICML*, 2013.

[140] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 6, pp. 915–928, 2007.

[141] M. Popa, L. Rothkrantzy, P. Wiggers, and C. Shan, "Assessment of facial expressions in product appreciation," *Neural Network World*, vol. 27, no. 2, p. 197, 2017.

[142] R. Shbib and S. Zhou, "Facial expression analysis using active shape model," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 1, pp. 9–22, 2015.

[143] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, "Effective geometric features for human emotion recognition," in *IEEE 11th Int. Conf. on Signal Processing (ICSP)*, vol. 1, pp. 623–627, 2012.

[144] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in *Proc. of the Int. Conf. on Multimodal Interaction*, pp. 459–466, 2015.

[145] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 19–27, 2015.

[146] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.

[147] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[148] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[149] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[150] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. how individual face parts contribute to successful emotion recognition," *PloS One*, vol. 12, no. 5, 2017.

[151] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[152] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445–450, ACM, 2016.

[153] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[154] F. Lingenfelser, J. Wagner, J. Deng, R. Brueckner, B. Schuller, and E. Andre, "Asynchronous and Event-Based Fusion Systems for Affect Recognition on Naturalistic Data in Comparison to Conventional Approaches," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 410–423, 2018.

[155] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3444–3453, IEEE, 2017.

[156] C. Athanasiadis, E. Hortal, and S. Asteriadis, "Audio–visual domain adaptation using conditional semi-supervised generative adversarial networks," *Neurocomputing*, 2019.

[157] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 3687–3691, IEEE, 2013.

[158] Z. Wu, X. Zhang, T. Zhi-Xuan, J. Zaki, and D. C. Ong, "Attending to emotional narratives," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 648–654, IEEE, 2019.

[159] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, pp. 1–6, 2013.

[160] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 985–990, IEEE, 2017.

[161] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 596–609, 2018.

[162] G. C.D.Katsis, N.Katertsidis and D.I.Fotiadis, "Toward emotion recognition in carracing drivers: A biosignal processing approach," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2008.

[163] J. Fan, J. W. Wade, A. P. Key, Z. E. Warren, and N. Sarkar, "Eeg-based affect and workload recognition in a virtual driving environment for asd intervention," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 43–51, 2017.

[164] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: theoretical and practical implications," *Interacting with Computers*, vol. 14, no. 2, pp. 141–169, 2002.

[165] D. Goleman, *Emotional intelligence*. Bantam, 2006.

[166] D. Goleman, "The brain and emotional intelligence: New insights," *Regional Business*, vol. 94, 2011.

[167] N. Vretos, P. Daras, S. Asteriadis, E. Hortal, E. Ghaleb, E. Spyrou, H. C. Leligou, P. Karkazis, P. Trakadas, and K. Assimakopoulos, "Exploiting sensing devices availability in ar/vr deployments to foster engagement," *Virtual Reality*, vol. 23, no. 4, pp. 399–410, 2019.

[168] E. Ghaleb, M. Popa, E. Hortal, and S. Asteriadis, "Multimodal fusion based on information gain for emotion recognition in the wild," in *2017 Intelligent Systems Conference (IntelliSys)*, pp. 814–823, IEEE, 2017.

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

[170] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, no. 12, pp. 1615–1618, 2003.

[171] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Webbased database for facial expression analysis," in *Proc. of the IEEE Int. Conf. on Multimedia and Expo (ICME)*, pp. 317–321, 2005.

[172] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, vol. 2, 2011.

[173] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[174] R. Kullback, S.and Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[175] Z. Meng, S. Han, M. Chen, and Y. Tong, "Feature level fusion for bimodal facial action unit recognition," in *IEEE Int. Symposium on Multimedia (ISM)*, 2015.

[176] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," in *Proc. of the IEEE Int. Conf. on Pattern Recognition (ICPR)*, 2014.

[177] D. Wu, L. Pigou, P.-J. Kindermans, L. Nam, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[178] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the ACM Multimedia (MM)*, pp. 835–838, 2013.

[179] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems*, pp. 2222–2230, 2012.

[180] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 105, no. 3, pp. 222–245, 2013.

[181] P. Viola, M. Jones, *et al.*, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, no. 34-47, p. 4, 2001.

[182] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 532–539, IEEE, 2013.

[183] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. of the British Machine Vision Conference (BMVC)*, vol. 2, p. 4, BMVC, 2013.

[184] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman, "A compact and discriminative face track descriptor," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1693–1700, 2014.

[185] I. J. Goodfellow, D. Erhan, P. L. Carrier, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Int. Conf. on Neural Information Processing*, pp. 117–124, Springer, 2013.

[186] D. Johnson and S. Sinanovic, "Symmetrizing the kullback-leibler distance," *IEEE Trans. on Information Theory*, 2000.

[187] J. H. Holland, "Genetic algorithms and the optimal allocation of trials," *SIAM Journal on Computing*, vol. 2, no. 2, pp. 88–105, 1973.

[188] G. Chetty, M. Wagner, and R. Goecke, "A multilevel fusion approach for audiovisual emotion recognition," *Emotion Recognition: A Pattern Analysis Approach*, pp. 437–460, 2015.

[189] T. Gehrig and H. K. Ekenel, "Why is facial expression analysis in the wild challenging?," in *Proc. of the 2013 on Emotion Recognition in the Wild Challenge and Workshop*, pp. 9–16, 2013.

[190] M. Paleari, R. Chellali, and B. Huet, "Features for multimodal emotion recognition: An extensive study," in *Cybernetics and Intelligent Systems (CIS), 2010 IEEE Conf. on*, pp. 90–95, 2010.

[191] D. Nguyen, K. Nguyen, S. Sridharan, *et al.*, "Deep spatio-temporal features for multimodal emotion recognition," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pp. 1215–1223, 2017.

[192] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proc. of the 18th ACM Int. Conf. on Multimodal Interaction*, pp. 506–513, ACM, 2016.

[193] X. Ouyang, S. Kawaai, E. G. H. Goh, *et al.*, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 577–582, ACM, 2017.

[194] J. Schwan, "Towards Emotion Recognition in an HCI Setting using ComputerVision Techniques," Master's thesis, Maastricht University, the Netherlands, 2017.

[195] B. P. Woolf, "Ai and education: Celebrating 30 years of marriage," in *Workshop on Les Contes du Mariage: Should AI stay married to Ed?*, p. 38, 2015.

[196] S. D'Mello, N. Blanchard, R. Baker, J. Ocumpaugh, and K. Brawner, "Affect-sensitive instructional strategies," *Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management*, vol. 2, p. 35, 2014.

[197] K. Bahreini, R. Nadolski, and W. Westera, "Towards real-time speech emotion recognition for affective e-learning," *Education and Information Technologies*, vol. 21, no. 5, pp. 1367–1386, 2016.

[198] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, and V. Shute, "Temporal generalizability of face-based affect detection in noisy classroom environments," in *International Conference on Artificial Intelligence in Education*, pp. 44–53, Springer, 2015.

[199] K. Bahreini, R. Nadolski, and W. Westera, "Towards multimodal emotion recognition in e-learning environments," *Interactive Learning Environments*, 2016.

[200] M. B. Ammar, M. Neji, A. M. Alimi, and G. Gouardères, "The affective tutoring system," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3013–3023, 2010.

[201] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM International Conference on Multimedia*, pp. 677–682, ACM, 2005.

[202] Z. A. Pardos, R. S. Baker, M. O. San Pedro, S. M. Gowda, and S. M. Gowda, "Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 117–124, ACM, 2013.

[203] R. S. d Baker, S. M. Gowda, M. Wixon, J. Kalka, A. Z. Wagner, A. Salvi, V. Aleven, G. W. Kusbit, J. Ocumpaugh, and L. Rossi, "Towards sensor-free affect detection in cognitive tutor algebra.," *International Educational Data Mining Society*, 2012.

[204] M. Cscikszentmihaly, *Beyond boredom and anxiety: Experiencing Flow in Work and Play*. Josey-Bass Publishers, 1975.

[205] M. E. Seligman and M. Csikszentmihalyi, *Positive psychology: An introduction*. Springer, 2014.

[206] M. O. Z. San Pedro, R. S. d Baker, S. M. Gowda, and N. T. Heffernan, "Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system," in *International Conference on Artificial Intelligence in Education*, pp. 41–50, Springer, 2013.

[207] L.-F. Liao, "A flow theory perspective on learner motivation and behavior in distance education," *Distance Education*, vol. 27, no. 1, pp. 45–62, 2006.

[208] D. J. Shernoff, M. Csikszentmihalyi, B. Schneider, and E. S. Shernoff, "Student engagement in high school classrooms from the perspective of flow theory," in *Applications of Flow in Human Development and Education*, pp. 475–494, Springer, 2014.

[209] J. Chen, "Flow in games (and everything else)," *Communications of the ACM*, vol. 50, no. 4, pp. 31–34, 2007.

[210] R. Berta, F. Bellotti, A. De Gloria, D. Pranantha, and C. Schatten, "Electroencephalogram and physiological signal analysis for assessing flow in games," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 5, no. 2, pp. 164–175, 2013.

[211] D. Johnson and J. Wiles, "Effective affective user interface design in games," *Ergonomics*, vol. 46, no. 13-14, pp. 1332–1345, 2003.

[212] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 357–366, ACM, 2014.

[213] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Estimation of behavioral user state based on eye gaze and head pose application in an e-learning environment," *Multimedia Tools and Applications*, vol. 41, no. 3, pp. 469–493, 2009.

[214] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "A game-based corpus for analysing the interplay between game context and player experience," in *Affective Computing and Intelligent Interaction*, pp. 547–556, Springer, 2011.

[215] G. N. Yannakakis and J. Hallam, "Real-time game adaptation for optimizing player satisfaction," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 1, no. 2, pp. 121–133, 2009.

[216] C. Coutrix and N. Mandran, "Identifying emotions expressed by mobile users through 2d surface and 3d motion gestures," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 311–320, ACM, 2012.

[217] J. L. Santos, K. Verbert, J. Klerkx, E. Duval, S. Charleer, and S. Ternier, "Tracking data in open learning environments," *Journal of Universal Computer Science*, vol. 21, no. 7, pp. 976–996, 2015.

[218] Á. Del Blanco, Á. Serrano, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "E-learning standards and learning analytics. can data collection be improved by using standard data models?," in *Global Engineering Education Conference (EDUCON), 2013 IEEE*, pp. 1255–1261, IEEE, 2013.

[219] C. Glahn, "Using the adl experience api for mobile learning, sensing, informing, encouraging, orchestrating," in *Next Generation Mobile Apps, Services and Technologies (NGMAST), 2013 Seventh International Conference on*, pp. 268–273, IEEE, 2013.

[220] A. Corbi and D. B. Solans, "Review of current student-monitoring techniques used in elearning-focused recommender systems and learning analytics: The experience api & lime model case study," *IJIMAI*, 2014.

[221] M. Megliola, G. De Vito, R. Sanguini, F. Wild, and P. Lefrere, "Creating awareness of kinaesthetic learning using the experience api: current practices, emerging challenges, possible solutions," in *CEUR Workshop Proceedings*, vol. 1238, pp. 11–22, 2014.

[222] C. Athanasiadis, E. Hortal, D. Koutsoukos, C. Z. Lens, and S. Asteriadis, "Personalized, affect and performance-driven computer-based learning.," in *CSEDU (1)*, pp. 132–139, 2017.

[223] C. Athanasiadis, M. Amestoy, E. Hortal, and S. Asteriadis, "e-learning: A dataset for affect-driven adaptation of computer-based learning," *IEEE MultiMedia*, vol. 27, no. 1, pp. 49–60, 2019.

[224] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[225] L. A. Feldman, "Valence focus and arousal focus: Individual differences in the structure of affective experience.," *Journal of Personality and Social Psychology*, 1995.

[226] R. W. Picard, "Building an affective learning companion," in *Intelligent Tutoring Systems*, pp. 811–811, 2006.

[227] J. Hamari and J. Koivisto, "Measuring flow in gamification: Dispositional flow scale-2," *Computers in Human Behavior*, vol. 40, pp. 133–143, 2014.

[228] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work?–a literature review of empirical studies on gamification," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pp. 3025–3034, IEEE, 2014.

[229] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[230] P. Ekman, "Facial expression and emotion.," *American Psychologist*, vol. 48, no. 4, p. 384, 1993.

[231] P. Wu, S. C. Hoi, P. Zhao, C. Miao, and Z.-Y. Liu, "Online multi-modal distance metric learning with application to image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 454–467, 2016.

[232] Y. Ying, K. Huang, and C. Campbell, "Sparse metric learning via smooth optimization," in *Advances in Neural Information Processing Systems*, pp. 2214–2222, 2009.

[233] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S.-F. Chang, "Low-rank similarity metric learning in high dimensions.," in *AAAI*, pp. 2792–2799, 2015.

[234] H. Zhang, V. M. Patel, and R. Chellappa, "Hierarchical multimodal metric learning for multimodal classification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3057–3065, 2017.

[235] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Asian Conference on Computer Vision*, pp. 252–267, Springer, 2014.

[236] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval.," in *AAAI*, AAAI, 2013.

[237] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE Int. Conf. on*, pp. 878–883, 2011.

[238] P. Xie and E. P. Xing, "Multi-modal distance metric learning.," in *IJCAI*, pp. 1806–1812, Citeseer, 2013.

[239] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1867–1874, IEEE, 2014.

[240] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[241] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[242] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[243] R. Beard, R. Das, R. W. Ng, P. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proc. of the 22nd Conf. on Computational Natural Language Learning*, pp. 251–259, 2018.

[244] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, IEEE, 2014.

[245] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, 2016.

[246] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 553–560, ACM, 2017.

[247] F. Grosjean, "Gating," *Language and Cognitive Processes*, vol. 11, no. 6, pp. 597–604, 1996.

[248] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, pp. 1988–1996, 2014.

[249] J. Lee, S. Abu-El-Haija, B. Varadarajan, and A. P. Natsev, "Collaborative deep metric learning for video understanding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, (New York, NY, USA), pp. 481–490, ACM, 2018.

[250] Y. Li, M. Yang, and Z. M. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[251] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4014–4024, 2016.

[252] M. Abavisani and V. M. Patel, "Deep multimodal subspace clustering networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1601–1614, 2018.

[253] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[254] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, PMLR, 07–09 Jul 2015.

[255] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, 2016.

[256] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv preprint arXiv:1703.07737*, 2017.

[257] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[258] J. C. S. J. Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. Van Gerven, R. Van Lier, *et al.*, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, 2019.

[259] D. Dotti, M. Popa, and S. Asteriadis, "Behavior and personality analysis in a nonsocial context dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2354–2362, 2018.

[260] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th International Conference on Multimodal Interfaces*, pp. 53–60, ACM, 2008.

[261] D. Dotti, E. Ghaleb, and S. Asteriadis, "Temporal triplet mining for personality recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 379–386, IEEE, 2020.

[262] J. Block and J. H. Block, "The role of ego-control and ego-resiliency in the organization of behavior," in *Development of Cognition, Affect, and Social Relations*, pp. 49–112, Psychology Press, 2014.

[263] D. Dotti, M. Popa, and S. Asteriadis, "Being the center of attention: A person-context cnn framework for personality recognition," *ACM Trans. Interact. Intell. Syst.*, vol. 10, Nov. 2020.

[264] H. Coskun, D. Joseph Tan, S. Conjeti, N. Navab, and F. Tombari, "Human motion analysis with deep metric learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 667–683, 2018.

[265] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou, and A. Shamir, "Deep motifs and motion signatures," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–13, 2018.

[266] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[267] D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments.," in *VISIGRAPP (5: VISAPP)*, pp. 210–217, 2017.

[268] W. James, "The perception of reality," *Principles of Psychology*, vol. 2, pp. 283–324, 1890.

[269] M. S. Pápai and S. Soto-Faraco, "Sounds can boost the awareness of visual events through attention without cross-modal integration," *Scientific Reports*, vol. 7, p. 41684, 2017.

[270] T. Afouras *et al.*, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[271] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention lstm networks for video captioning," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 537–545, 2017.

[272] C. Hori *et al.*, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2352–2356, IEEE, 2019.

[273] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *INTERSPEECH*, pp. 3569–3573, 09 2019.

[274] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

[275] Y. Shi, N. Siddharth, B. Paige, and P. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Advances in Neural Information Processing Systems*, pp. 15718–15729, 2019.

[276] S. Albanie and A. Vedaldi, "Learning grimaces by watching tv," *BMVC*, 2016.

[277] S. Hershey *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, IEEE, 2017.

[278] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning based multimodal audio-visual emotion recognition," *IEEE MultiMedia*, 2019.

[279] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 552–558, IEEE, 2019.

[280] M. K. Noordewier and S. M. Breugelmans, "On the valence of surprise," *Cognition & Emotion*, vol. 27, no. 7, pp. 1326–1334, 2013.

[281] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3438–3446, 2016.

[282] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *Journal of Personality and Social Psychology*, vol. 54, no. 6, p. 1063, 1988.

[283] E. L. Levine, X. Xu, L.-Q. Yang, D. Ispas, H. D. Pitariu, R. Bian, D. Ding, R. Capotescu, H. Che, and S. Musat, "Cross-national explorations of the impact of affect at work using the state-trait emotion measure: a coordinated series of studies in the united states, china, and romania," *Human Performance*, vol. 24, no. 5, pp. 405–442, 2011.

[284] N. Chodorow, *The power of feelings: Personal meaning in psychoanalysis, gender, and culture*. Yale University Press, 1999.

[285] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, (Virtual), pp. 1597–1607, PMLR, 13–18 Jul 2020.

[286] K. Schwab, *The fourth industrial revolution.* Currency, 2017.

[287] R. Campa, "Fourth industrial revolution and emotional intelligence: A conceptual and scientometric analysis," *Changing Societies & Personalities. 2020. Vol. 4. Iss. 1*, vol. 4, no. 1, pp. 8–30, 2020.

[288] P. Salovey and J. D. Mayer, "Emotional intelligence," *Imagination, Cognition and Personality*, vol. 9, no. 3, pp. 185–211, 1990.

SUMMARY

Emotions are a key component in human-human communications, with a highly complex socio-psychological nature. A great amount of affective information is displayed through facial expressions, gestures, speech, and other means. Recently, Multimodal Emotion Recognition (MER) has gained a notable amount of research interest. It aims to recognize the displayed affective information through techniques and methods from the fields of Affective Computing (AC) and Artificial Intelligence (AI). Furthermore, the recent technological advancements brought interactivity between people and digital devices to a completely different level, making computers and mobile phones an important part of our daily lives. Besides, AI is a rapidly improving field, offering us new mathematical methods for data representations and classification procedures. Therefore, there is an increased interest in the Human-Computer Interaction (HCI) field towards enhancing digital devices with emotion recognition abilities for obtaining a more natural HCI experience. However, HCI still lacks elements of emotional intelligence to enable a more human-centered interaction. Human-centered computation through affective computing can help in recognizing emotions, and generate proper actions to achieve richer and human-like communications in settings like HCI. Automatic systems with affective capabilities can be essential in many applications of affective computing, which range from education, autonomous driving, entertainment to health-care.

Nonetheless, the facts that emotions are multifaceted, socio-psychological, and biological concepts, make automatic emotion recognition a challenging task. This dissertation addresses various problems in multimodal emotion recognition, which is an important task towards achieving Affective Computing (AC) goals. Chapter 1 motivates the research problem and introduces its theoretical foundations. In particular, the research of this dissertation focuses on the two primary forms of emotion expressions and modulations: the face and the voice. In human-human communications, these two modalities are the most expressive and perceived channels. They are widely used in our daily communication for our social interactions. Although information obtained from non-obvious signals of emotion expression (e.g. heart beating, sweating, and respiration) could be informative as well, people rely on the apparent cues in sensing others' emotions. Besides, sensing emotions from auditory and visual channels is not invasive. As a result, this dissertation aims to predict emotions through audio-visual cues, which consequently can lead to enhanced interactions between humans and robots and machines in general. It employs and proposes progressive research towards audio-visual emotion recognition, coming from state-of-the-art techniques in the field of Artificial Intelligence, which are presented in the technical Chapter 2.

Chapter 3 introduces an extensive literature review of Affective Computing (AC) and Multimodal Emotion Recognition (MER). Earlier research in emotion recognition targeted, either individual modalities (such as facial expressions and acoustic-prosodic cues) or global multimodal emotion recognition. On the other hand, the research in this

dissertation focuses on multimodal recognition and exploits temporal interactions between audio-visual channels. It aims to capture modalities' strengths for emotion recognition to utilize their complementary and supplementary information. It adopts recent advances in Affective Computing (AC) such as Deep Neural Networks (DNNs), Deep Metric Learning (DML), end-to-end learning, and the attention mechanism for Audio-Video Emotion Recognition (AVER). Also, over the course of this research, the literature was lacking in-depth analyses regarding automatically extracted, dynamic interactions between audio and video signals in emotionally rich contexts. This dissertation presents studies that investigate the temporal relationships of both modalities and exploits their strength for emotion recognition. Besides, it employs state-of-the-art methods, such as Deep Metric Learning (DML) to perform similarity learning for multimodal emotion recognition.

In this dissertation, four research questions and objectives are introduced to address the joint modeling of audio-visual cues for Multimodal Emotion Recognition (MER). The objective of Chapter 4 is related to data modeling and producing robust multimodal representations for emotion recognition. In two studies, this chapter addresses the first research question: *How to extract and fuse robust features and which is their contribution to automatic emotion recognition?*. The first one deals with emotion recognition in video clips, where audio and visual cues are the primary information for emotion perception. It presents a hierarchical framework for multimodal emotion recognition. The proposed research employs Fisher Vectors (FVs) representations to aggregate frame-level features in a video sample. This encoding is applied on different types of audio and visual features (e.g. Dense Scale-Invariant Feature Transformation (SIFT), geometric, Convolutional Neural Network (CNN), audio), enabling mapping them into a common space, where feature level fusion is performed. It then uses a strategy of employing information gain principles, for selecting the best combination of features to be fused. Finally, a decision-level fusion approach on top of the best features is applied to optimize modalities' weights for each emotional state using a genetic search algorithm. The experimental results show that the two fusion schemes on the employed modalities and their features improve the accuracy of emotion prediction compared to unimodal emotion recognition. The second part of the chapter studies the correlation between students' self-reported affective states, according to the Theory of Flow (ToF), and their interactions with learning materials. This study designs a framework to track contextual information and interaction features during learning activities. It utilizes a standard tracking tool, the xAPI framework, for learning analytics. The conducted evaluations highlighted the potential usage of interaction parameters with learning materials as a useful channel for measuring affective states.

Chapter 5 focuses on the objective of efficient fusion for audio-visual representations. It addresses the second research question: *what is the impact of multimodal learning on emotion recognition?*. It introduces a modality-specific Multimodal Emotion Recognition Metric Learning (MERML). This method is applied to improve the latent-space of audio-visual data representations. It successfully exploits the complementary information of audio and video modalities for emotion recognition. As a result of this approach, audio-visual representations are well structured in the newly learned subspace, and their capacity for optimized emotion recognition is maximized. The conducted

quantitative and qualitative evaluations of the method demonstrated the contribution of the method to increased classification accuracy. Chapter 6, benefits from the findings of the MERML framework, and builds an end-to-end Deep Metric Learning (DML) with triplet loss for audio-visual temporal emotion recognition. In addition, it aims at the objective of exploiting the temporal dynamics of emotion display and perception, and also at answering the third research question: *What is the role of temporal dynamics in audio-visual cues, in automated emotion recognition?*. In the study of Chapter 6, inspired by the gating paradigm, we investigate how introducing multimodal cues with increasing durations impacts the recognition rates of positive and negative emotions. The procedure employs Long-Short Term Memory (LSTM)s between time windows for incremental perception to mimic the gating paradigm. The proposed framework embeds audio-visual cues overtime, taking advantage of the temporal display of emotions. It also checks the contribution of audio, visual, and audio-visual fusion in emotion recognition. The framework showed efficiency in modeling the temporal context of multimodal emotion recognition. Besides, within the introduced framework, algorithms to tackle the challenges of triplet sets' mining and the convergence of Deep Metric Learning (DML) are proposed. The developed approach and the associated techniques, such as Multi Window Triplet Sets Mining (MWTSM), contributed significantly to the stability and the performance of the framework for Multimodal Emotion Recognition (MER). In addition, the evaluations proved the benefits of the incremental perception of both audio and visual cues in the recognition rates overtime. Additionally, the temporal differences of the recognition speed for positive and negative emotions differ, where positive emotions are recognized faster than the negative ones.

Chapter 7 targets the research objective of attending to informative time segments in temporal audio and video signals. It addresses the fourth research question: *How can we capture the contributions of the temporal dynamics of affect display using attention-mechanisms?*. The study of this chapter employs attention mechanisms on audio-visual embeddings over time windows to capture their temporal properties for emotion recognition. The evaluation of the proposed method, namely Multimodal Attention mechanism for Temporal Emotion Recognition (MATER), highlights the importance of weighing the time windows in audio-visual cues. The presented method offers interpretability and explainability of the attention mechanisms for temporal and multimodal fusion. Furthermore, MATER presents extensive studies and meta-analysis findings, linking the outputs of our proposition to research from psychology. For example, it gives more insights with regards to the multimodal interaction and presentation of audio-visual cues. It examines how the attention mechanisms helps in joint modeling of multimodal cues and subsequently enhances their performance. Moreover, this study shows that the contribution of the video modality in multimodal fusion is greater than the one of the audio modality. Finally, it applies noise injection into the time windows embeddings, during the evaluation or/and the training phases, to demonstrate the robustness of the method and its ability to adapt to challenging conditions.

Finally, Chapter 8 concludes the conducted research, highlights its findings, and points out some directions for future work in affective computing and multimodal emotion recognition.

CURRICULUM VITAE

Esam A.H. Ghaleb was born in January 1989 in Taiz, Yemen. In 2013, he completed his Bachelor of Science in the Department of Computer Engineering at Istanbul Technical University (ITU), where he also earned his Master of Science degree in Computer Vision and Machine Learning (2015). In his Master's thesis, he studied the aging effect on face recognition utilizing Harry Potter movies. Early academic work experiences include his position as a research assistant in the Smart Interaction, Mobile Intelligence, and Multimedia Technologies (SiMiT) Research Group at ITU for the EU project Collaborative Annotation of multi-MOdal, multI-Lingual and multi-mEdia documents (CAMOMILE), from June 2013 through February 2016. Furthermore, he conducted two internships in the Computer Vision for Human-Computer Interaction (CVHCI) laboratory at Karlsruhe Institute of Technology (2014 & 2015).

In March 2016, Ghaleb started his Ph.D. in the Department of Data Science and Knowledge Engineering (DKE) at Maastricht University, under the supervision of Associate Prof. Stelios (Stylianos) Asteriadis. He has been working in the Affective & Visual Computing Lab (AVCL), which was previously embedded in the Robots, Agents and Interaction (RAI) group. His research interests include computer vision, machine learning, Explainable AI (XAI), emotion recognition and human behavioral analysis. During the course of his Ph.D., he carried out research on bimodal emotion recognition through audio-visual cues. His work utilized facial expressions and speech signals in video clips and proposed progressive computational methodologies to capture their complementary information for emotion recognition. Ghaleb's Ph.D. research was supported by the Horizon 2020 project Managing Affective-learning THrough Intelligent atoms and Smart InteractionS (MaTHiSiS)[3]. Besides his research and project work, he was a teaching assistant for the computer vision course in DKE at Maastricht University. Since January 2020, he has been working in the Horizon 2020 project PeRsOnalized Integrated CARE Solution for Elderly facing several short or long term conditions & enabling a better quality of LIFE (ProCare4Life)[4]. In the ProCare4Life project, his research focuses on human behavioral analysis, mainly for people with neurodegenerative diseases. His research has been published in various peer-reviewed conference proceedings and high-impact journals, and finally, this dissertation.

[3] http://mathisis-project.eu/
[4] https://procare4life.eu

LIST OF PUBLICATIONS

1. **E. Ghaleb**, J. Niehues, and S. Asteriadis, "Joint modelling of audio-visual cues using attention mechanisms for emotion recognition", under submission.

2. **E. Ghaleb**, J. Niehues, and S. Asteriadis, "Multimodal attention mechanism for temporal emotion recognition," in 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020, pp. 251–255.

3. D. Dotti, **E. Ghaleb**, and S. Asteriadis, "Temporal triplet mining for personality recognition," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020, pp. 379–386.

4. **E. Ghaleb**, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019, pp. 552–558.

5. **E. Ghaleb**, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," IEEE MultiMedia, vol. 27, no. 1, pp. 37–48, 2019.

6. **E. Ghaleb**, M. Popa, E. Hortal, S. Asteriadis, and G. Weiss, "Towards affect recognition through interactions with learning materials," in 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2018, pp. 372–379.

7. **E. Ghaleb**, M. Popa, E. Hortal, and S. Asteriadis, "Multimodal fusion based on information gain for emotion recognition in the wild," in 2017 Intelligent Systems Conference (IntelliSys). IEEE, 2017, pp. 814–823.

8. J. Schwan, **E. Ghaleb**, E. Hortal, and S. Asteriadis, "High-performance and lightweight real-time deep face emotion recognition," in 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). IEEE, 2017, pp. 76–79.

9. N. Vretos, P. Daras, S. Asteriadis, E. Hortal, **E. Ghaleb**, E. Spyrou, H. C. Leligou, P. Karkazis, P. Trakadas, and K. Assimakopoulos, "Exploiting sensing devices availability in ar/vr deployments to foster engagement," Virtual Reality, vol. 23, no. 4, pp. 399–410, 2019.

SIKS DISSERTATION SERIES

2011

1. Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
2. Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
3. Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
4. Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
5. Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
6. Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
7. Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
8. Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
9. Tim de Jong (OU), Contextualised Mobile Media for Learning
10. Bart Bogaert (UvT), Cloud Content Contention
11. Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12. Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13. Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14. Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15. Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16. Maarten Schadd (UM), Selective Search in Games of Different Complexity
17. Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18. Mark Ponsen (UM), Strategic Decision-Making in complex games
19. Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20. Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21. Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22. Junte Zhang (UVA), System Evaluation of Archival Description and Access
23. Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24. Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25. Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26. Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27. Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28. Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29. Faisal Kamiran (TUE), Discrimination-aware Classification
30. Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31. Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32. Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33. Tom van der Weide (UU), Arguing to Motivate Decisions
34. Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35. Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36. Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37. Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38. Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39. Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40. Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41. Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42. Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43. Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44. Boris Reuderink (UT), Robust Brain-Computer Interfaces
45. Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46. Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47. Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48. Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49. Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

1. Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
2. Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
3. Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
4. Jurriaan Souer (UU), Development of Content Management System-based Web Applications
5. Marijn Plomp (UU), Maturing Interorganisational Information Systems
6. Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
7. Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
8. Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
9. Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10. David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11. J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12. Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13. Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14. Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15. Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16. Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17. Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18. Eltjo Poort (VU), Improving Solution Architecting Practices
19. Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20. Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21. Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22. Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23. Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24. Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25. Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26. Emile de Maat (UVA), Making Sense of Legal Text
27. Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28. Nancy Pascall (UvT), Engendering Technology Empowering Women
29. Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30. Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31. Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32. Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33. Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34. Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35. Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36. Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37. Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38. Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39. Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40. Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41. Sebastian Kelle (OU), Game Design Patterns for Learning
42. Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43. (Withdrawn)
44. Anna Tordai (VU), On Combining Alignment Techniques

45. Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46. Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47. Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48. Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49. Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50. Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51. Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013

1. Viorel Milea (EUR), News Analytics for Financial Decision Support
2. Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
3. Szymon Klarman (VU), Reasoning with Contexts in Description Logics
4. Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
5. Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
6. Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
7. Giel van Lankveld (UvT), Quantifying Individual Player Differences
8. Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
9. Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10. Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11. Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12. Marian Razavian (VU), Knowledge-driven Migration to Services
13. Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14. Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15. Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16. Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17. Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18. Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19. Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20. Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21. Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22. Tom Claassen (RUN), Causal Discovery and Logic
23. Patricio de Alencar Silva (UvT), Value Activity Monitoring
24. Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25. Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26. Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning

27. Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28. Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29. Iwan de Kok (UT), Listening Heads
30. Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31. Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32. Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33. Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34. Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35. Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36. Than Lam Hoang (TUe), Pattern Mining in Data Streams
37. Dirk Börner (OUN), Ambient Learning Displays
38. Eelco den Heijer (VU), Autonomous Evolutionary Art
39. Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40. Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41. Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42. Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43. Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014

1. Nicola Barile (UU), Studies in Learning Monotone Models from Data
2. Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
3. Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
4. Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
5. Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
6. Damian Tamburri (VU), Supporting Networked Software Development
7. Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
8. Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
9. Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10. Ivan Salvador Razo Zapata (VU), Service Value Networks
11. Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12. Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13. Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14. Yangyang Shi (TUD), Language Models With Meta-information
15. Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16. Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17. Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18. Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19. Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20. Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21. Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22. Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23. Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24. Davide Ceolin (VU), Trusting Semi-structured Web Data
25. Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26. Tim Baarslag (TUD), What to Bid and When to Stop
27. Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28. Anna Chmielowiec (VU), Decentralized k-Clique Matching
29. Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30. Peter de Cock (UvT), Anticipating Criminal Behaviour
31. Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32. Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33. Tesfa Tegegne (RUN), Service Discovery in eHealth
34. Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35. Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36. Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37. Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38. Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39. Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40. Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41. Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42. Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43. Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44. Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45. Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46. Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47. Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015

1. Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
2. Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

3. Twan van Laarhoven (RUN), Machine learning for network data
4. Howard Spoelstra (OUN), Collaborations in Open Learning Environments
5. Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
6. Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
7. Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
8. Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
9. Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10. Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11. Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12. Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13. Giuseppe Procaccianti (VU), Energy-Efficient Software
14. Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15. Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16. Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17. André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18. Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19. Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20. Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21. Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22. Zhemin Zhu (UT), Co-occurrence Rate Networks
23. Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24. Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25. Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26. Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27. Sándor Héman (CWI), Updating compressed colomn stores
28. Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29. Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30. Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31. Yakup Koç (TUD), On the robustness of Power Grids
32. Jerome Gard (UL), Corporate Venture Management in SMEs
33. Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34. Victor de Graaf (UT), Gesocial Recommender Systems
35. Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

1. Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
2. Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

3. Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
4. Laurens Rietveld (VU), Publishing and Consuming Linked Data
5. Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
6. Michel Wilson (TUD), Robust scheduling in an uncertain environment
7. Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
8. Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
9. Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10. George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11. Anne Schuth (UVA), Search Engines that Learn from Their Users
12. Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13. Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14. Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15. Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16. Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17. Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18. Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19. Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20. Daan Odijk (UVA), Context & Semantics in News & Web Search
21. Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22. Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23. Fei Cai (UVA), Query Auto Completion in Information Retrieval
24. Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25. Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26. Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27. Wen Li (TUD), Understanding Geospatial Information on Social Media
28. Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29. Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30. Ruud Mattheij (UvT), The Eyes Have It
31. Mohammad Khelghati (UT), Deep web content monitoring
32. Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33. Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34. Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35. Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36. Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37. Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38. Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39. Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40. Christian Detweiler (TUD), Accounting for Values in Design
41. Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42. Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43. Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44. Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45. Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46. Jorge Gallego Perez (UT), Robots to Make you Happy
47. Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48. Tanja Buttler (TUD), Collecting Lessons Learned
49. Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50. Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

1. Jan-Jaap Oerlemans (UL), Investigating Cybercrime
2. Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
3. Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
4. Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
5. Mahdieh Shadi (UVA), Collaboration Behavior
6. Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
7. Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
8. Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
9. Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10. Robby van Delden (UT), (Steering) Interactive Play Behavior
11. Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12. Sander Leemans (TUE), Robust Process Mining with Guarantees
13. Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14. Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15. Peter Berck (RUN), Memory-Based Text Correction
16. Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17. Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18. Ridho Reinanda (UVA), Entity Associations for Search
19. Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20. Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21. Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22. Sara Magliacane (VU), Logics for causal inference under uncertainty
23. David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24. Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25. Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26. Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27. Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28. John Klein (VU), Architecture Practices for Complex Contexts
29. Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30. Wilma Latuny (UvT), The Power of Facial Expressions
31. Ben Ruijl (UL), Advances in computational methods for QFT calculations
32. Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33. Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34. Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35. Martine de Vos (VU), Interpreting natural science spreadsheets
36. Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37. Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38. Alex Kayal (TUD), Normative Social Applications
39. Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40. Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41. Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42. Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43. Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44. Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45. Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46. Jan Schneider (OU), Sensor-based Learning Support
47. Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48. Angel Suarez (OU), Collaborative inquiry-based learning

2018

1. Han van der Aa (VUA), Comparing and Aligning Process Representations
2. Felix Mannhardt (TUE), Multi-perspective Process Mining
3. Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling,

Model-Driven Development of Context-Aware Applications, and Behavior Prediction

4. Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
5. Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
6. Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
7. Jieting Luo (UU), A formal account of opportunism in multi-agent systems
8. Rick Smetsers (RUN), Advances in Model Learning for Software Systems
9. Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10. Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11. Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12. Xixi Lu (TUE), Using behavioral context in process mining
13. Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14. Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15. Naser Davarzani (UM), Biomarker discovery in heart failure
16. Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17. Jianpeng Zhang (TUE), On Graph Sample Clustering
18. Henriette Nakad (UL), De Notaris en Private Rechtspraak
19. Minh Duc Pham (VUA), Emergent relational schemas for RDF
20. Manxia Liu (RUN), Time and Bayesian Networks
21. Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22. Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23. Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24. Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25. Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26. Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27. Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28. Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29. Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30. Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

1. Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
2. Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
3. Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data
4. Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
5. Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
6. Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
7. Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

8. Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
9. Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10. Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11. Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12. Jacqueline Heinerman (VU), Better Together
13. Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14. Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15. Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16. Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17. Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18. Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19. Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20. Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21. Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22. Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23. Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24. Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25. Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26. Prince Singh (UT), An Integration Platform for Synchromodal Transport
27. Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28. Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29. Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30. Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31. Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32. Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33. Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34. Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35. Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36. Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37. Jian Fang (TUD), Database Acceleration on FPGAs
38. Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

1. Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
2. Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
3. Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

4. Maarten van Gompel (RUN), Context as Linguistic Bridges
5. Yulong Pei (TUE), On local and global structure mining
6. Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
7. Wim van der Vegt (OUN), Towards a software architecture for reusable game components
8. Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
9. Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10. Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11. Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models
12. Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13. Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14. Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15. Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16. Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17. Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18. Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19. Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20. Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21. Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22. Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23. Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24. Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25. Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26. Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27. Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28. Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29. Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30. Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31. Gongjin Lan (VU), Learning better – From Baby to Better
32. Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33. Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34. Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35. Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

1. Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
2. Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
3. Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
4. Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
5. Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
6. Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
7. Armel Lefebvre (UU), Research data management for open science
8. Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
9. Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10. Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11. Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12. Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13. Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14. Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15. Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm