Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Modeling Affection Mechanisms using Deep and Self-Organizing Neural Networks

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
der Universität Hamburg

eingereicht beim Fach-Promotionsausschuss Informatik von
Pablo Vinicius Alves de Barros

Hamburg, August 2016

Day of oral defense:
Wednesday, 28th September, 2016

The following evaluators recommend the admission of the dissertation:

Prof. Dr. Emilia Barakova
Dept. of Industrial Design
Technische Universiteit Eindhoven, Neetherlands

Prof. Dr. Frank Steinicke (chair)
Dept. of Computer Science
Universität Hamburg, Germany

Prof. Dr. Stefan Wermter (advisor)
Dept. of Computer Science
Universität Hamburg, Germany

Aos meus pais...
... pois se cheguei tão longe foi por ter
me apoiado sobre ombros de gigantes.

Abstract

Emotions are related to many different parts of our lives: from the perception of the environment around us to different learning processes and natural communication. They have an important role when we talk to someone, when we learn how to speak, when we meet a person for the first time, or to create memories about a certain experience in our childhood. Because of this crucial role in a human's life, studies on emotions date from the first centuries of written history and until today it is a very popular research field involving a lot of different disciplines: from neuroscience and psychology to artificial intelligence and robotics.

The research field of affective computing introduces the use of different emotional concepts on computational systems. Imagine a robot which can recognize spontaneous expressions and learn with it how to behave in a certain situation, or yet it uses emotional information to learn how to perceive the world around it. This is among the hardest challenges in affective computing: how to integrate emotion concepts in artificial systems to improve the way they perform a task, like communication or learning. One of the most important aspects of affective computing is how to make computational systems recognize and learn emotion concepts from different experiences, for example in human communication. Although several types of research were done in this area in the past two decades, we are still far away from having a system which can perceive, recognize and learn emotion concepts in a satisfactory way.

This thesis addresses the use of three models for emotion perception, recognition, and learning. The models proposed here use different computational concepts to solve each of these problems and implement solutions which proved to enhance the performance and generalization when recognizing emotion expressions. We evaluate our models using different databases with multimodal and spontaneous emotional information and proceed with a detailed analysis of each model. We also developed a novel database for emotion behavior analysis, the KT Emotion Interaction Corpus, which contains interactions from different human-human and human-robot scenarios.

The first of our models, named Cross-channel Convolution Neural Network (CCCNN), uses deep neural networks to learn how to represent and recognize spontaneous and multimodal audio-visual expressions. We implement modality specific channels to introduce particular feature representation and use shunting inhibitory neurons to generate robust expression representations. We present the Cross-channel architecture for high-level multimodal integration which makes the model not only an expert on single-modality data, but also on multimodal information. We evaluate our model using different corpora, for each modality and in complex multimodal scenarios. During our experiments, we also show that our model can deal with spontaneous expressions and performs better than state-of-the-art approaches in the same tasks. We also introduce the use of different mechanisms to visualize the learned knowledge of the network, showing how the use of the shunting inhibitory fields, modality-specific channels, and cross-channel integrations affect expression representations.

Our second model uses self-organizing layers in conjunction of our CCCNN in a way to learn different emotion concepts in an unsupervised manner. This improves the recognition and generalization capabilities of the model and introduces the ability to learn new expressions. In this model, we extend our CCCNN with the capability to create neural clusters which identify similar emotion concepts and show how these concepts relate to categorical and dimensional views on emotions. Also, we show how our model learns new emotion clusters and how it can be used for describing emotional behaviors in different scenarios.

Finally, our third model introduces concepts from emotional attention and memory as modulators for the learning and representation models presented before. Such modulators improve the capability of the model to recognize expressions, introduce visual selective attention for detecting emotion expressions in a large visual field, and make use of different memory mechanisms to adapt the model's knowledge at various situations. We also propose a unified Emotional Deep Neural Circuitry which integrates selective attention, emotion representation and recognition, learning of emotion concepts and storage of different affective memories. This system works on an online unsupervised learning manner, adapting its internal representation to different human-human and human-robot scenarios.

The models proposed and discussed in this thesis contribute to the field of affective computing by introducing a unified solution for selective attention, emotion recognition, and learning. These models are competitive in each of these tasks, and also provide an overview of learning mechanism which adapts its knowledge according to a given situation. We also develop a novel interaction dataset with different spontaneous human-human and human-robot interactions and use it in the evaluation of our models. This thesis introduces and discusses novel mechanisms which inspire different research on affective computing and provide an adaptive solution for various emotion tasks in a way that was not done before, and thus serves as the basis for upcoming research.

Zusammenfassung

Emotionen begegnen uns in vielerlei Lebensbereichen: von der Wahrnehmung unserer Umwelt bis hin zu verschiedenen Lernprozessen und natürlichsprachlicher Kommunikation. Sie spielen eine bedeutende Rolle, wenn wir eine Konversation führen, wenn wir das Sprechen lernen, wenn wir das erste Mal einer Person begegnen oder wenn wir uns an ein Ereignis aus unserer Kindheit erinnern. Vorhandene historische Studien sind schriftliche Zeugen der bedeutenden Rolle die Emotionen im Leben der Menschen spielen, und bis zum heutigen Tag sind sie ein anerkanntes, interdisziplinäres Forschungsgebiet, welches die Gebiete der Neurowissenschaften, Psychologie, Künstlichen Intelligenz und Robotik vereint. Die Forschung innerhalb des sogenannten "Affective Computing" beschäftigt sich mit der Verwendung emotionaler Konzepte in computergestützten Systemen. So kann zum Beispiel ein Roboter spontane emotionale Ausdrucksweisen erkennen und darauf basierend lernen, wie er sich in einer bestimmten Situation verhalten kann, oder die emotionale Information nutzen, um etwas über die umgebende Welt zu erfahren. Die größte Herausforderung in "Affective Computing" ist, emotionale Konzepte so in künstliche Systeme zu integrieren, dass diese in der Lösung von Aufgaben unterstützt werden, z.B. in der Kommunikation und dem Lernen. Einer der wichtigsten Aspekte in diesem Zusammenhang ist, computergestützte Systeme auf Grundlage verschiedener Erfahrungen, z.B. in der zwischenmenschlichen Kommunikation, zu befähigen jene emotionalen Konzepte zu erkennen und zu lernen. Obwohl diesbezüglich bereits viel Forschungsarbeit in den letzten zwei Jahrzehnten geleistet wurde, sind wir noch immer weit davon entfernt ein hinreichend zufriedenstellendes System zu haben, welches emotionale Konzepte wahrnehmen, erkennen und lernen kann.

Die vorliegende Dissertation beschreibt drei Modelle, die die beschriebenen Problematiken der Emotionswahrnehmung, der Emotionserkennung und des Lernens addressieren. Die vorgeschlagenen Modelle implementieren verschiedene Berechnungsverfahren, welche in geeigneter Weise die Probleme lösen und zeigen, wie sich die Performanz und Generalisierungsfähigkeit zur Erkennung emotionaler Ausdrücke damit erhöhen lässt. Zur Evaluation unserer Modelle verwenden wir diverse Datenbanken, welche multimodale und spontane emotionale Informationen beinhalten, und geben außerdem eine detaillierte Analyse unsere Modelle. Wir entwickelen außerdem eine neue Datenbank zur Analyse emotionalen Verhaltens, den "KT Emotion Interaction Corpus", der unterschiedliche Interaktionsszenarien zwischen Menschen und zwischen Mensch und Roboter enthält.

Unser erstes Modell, welches wir "Cross-channel Convolution Neural Network" (CCCNN) nennen, verwendet neuronale Netze mit einer underschiedliche verschiedener Anzahl an versteckten Schichten, die lernen, wie spontane und multimodale, audio-visuelle Äußerungen repräsentiert und erkannt werden. Dazu wurden modalitätsspezifische Kanäle zur Bestimmung spezieller Merkmalsrepräsentationen implementiert, sowie inhibitorische Neuronen zur Generierung robuster Repräsentationen der emotionalen Ausdrucksweisen verwendet. Wir stellen unsere "Cross Channel" Architektur zur multimodalen Integration vor und evaluier unser Modell anhand verschieden er Datensätze, die sowohl einzeln Modalitäten beinhalten wie

auch komplexere, multimodale Szenarien. Unsere Experimente zeigen, dass unser Modell spontane Ausdrucksweisen bewältigen kann und außerdem eine ingesamt bessere Performanz erzielt als bisherige Ansätze zur gleichen Aufgabe. Wir führen außerdem eine Visualisierung trainierter Netze ein um aufzuzeigen, wie sich die Verwendung von inhibitorischen Feldern und modalitätsspezifischen Kanälen und die Integration aus den "cross channels" auf das Wissen im Netz bezüglich der Ausdrucksrepräsentationen auswirkt.

Das zweite hier vorgestellte Modell verwendet das Konzept selbstorganisierender Karten in Verbindung mit dem eingeführten CCCNN, sodass mehrere emotionale Konzepte unüberwacht, d.h. ohne a priori Wissen, gelernt werden können. Dies verbessert die Erkennung und Generalisationsfähigkeit des Modells und bietet die Möglichkeit auch neue Ausdrucksformen zu erlernen. In der Konsequenz wird das CCCNN um die Fähigkeit erweitert, neuronale Cluster zu generieren, die ähnliche emotionale Konzepte identifizierbar machen und aufzeigen, wie sich diese Konzepte zur kategorischen und dimensionalen Perspektive auf Emotionen verhalten. Wir zeigen zusätzlich, wie unser Modell neue Gruppen emotionaler Ausdrucksweisen lernt und wie sie benutzt werden können, um emotionales Verhalten in verschiedenen Situationen beschreiben zu können. Zum Schluß führen wir ein drittes Modell ein, das die Konzepte von Aufmerksamkeit und Gedächntisleistung zur Modulierung des Lernens und der Repräsentation aufgreift. Diese Modulatoren verbessern die Fähigkeit des Modells zur Emotionserkennung, behandeln visuelle selektive Aufmerksamkeit zur Bewegungsdetektion in einem großen rezeptiven Feld und verwenden verschiedene Arten von Gedächtnis um die Adaptivität des Modells an neue Situationen zu gewährleisten. Wir schlagen ein vereinheitliches "Emotional Deep Neural Circuitry" Modell vor, welches selektive Aufmerksamkeit, Emotionsrepräsentation und Emotionserkennung, das Lernen von emotionalen Konzepten und das Speichern verschiedener affektiver Erinnerungen integriert. Dieses System arbeitet im sogenannten online-Modus und unüberwacht, welches ermöglicht dass interne Repräsentationen auf Grundlage einer Reihe von Mensch-zu-Mensch oder Mensch-zu-Roboter Interaktionen adaptiert werden. Die in dieser Dissertation vorgeschlagenen und beschriebenen Modelle steuern einen wichtigen Beitrag im Bereich des "Affective Computing" bei, in dem erstmals Erkenntnisse aus der Forschung der selektiven Aufmerksamkeit mit den Aufgaben der Emotionserkennung und des Lernens von Emotionen vereinheitlicht werden. Die Modelle sind jeweils performant zur gegebenen Aufgabe und bieten einen überblick Über Lernmechanismen die das Wissen adaptiv zur Situation nutzen. Wir haben außerdem eine neue Datenbank entwickelt die spontane Mensch-zu-Mensch und Mensch-zu-Roboter Interaktionen enthält und unsere Modelle anhand derer evaluiert.

Die vorliegende Dissertation stellt neuartige Mechanismen vor und diskutiert diejenigen, welche im Bereich des "Affective Computing" zu inspierenden Forschungsfragestellungen führen könnten.Die Arbeit bietet adaptive Lösungen fr die diversen Aufgaben der Emotionserkennung, dabei kann diese Dissertation durch den dargestellten, neuartigen Ansatz als Basis für weiterführende Forschung dienen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The most necessary skills of human-human communication are the capability to perceive, understand and respond to social interactions, usually determined through affective expressions [96]. Therefore, the application of emotion expression recognition in robots can change our interaction with them [246]. A robot capable of understanding emotion expressions can increase its own capability of solving problems by using these expressions as part of its decision-making process, in a similar way as humans do [10]. A robot that develops this judgmental capability based on human interaction observation can realize complex tasks, enhance its interaction skills and even create a certain discernment about the information it is receiving.

Although much research was done in automatic emotion recognition and interpretation in the past decades, still some problems exist. Most of the works on emotion recognition are restricted to a limited set of expressions, do not take into consideration spontaneous reactions and cannot be easily adapted to other users or situations. Also, most of the research stops at the perception of expressions, but much more is necessary to have a deep understanding and application of emotions in HRI.

1.1 Motivation

How to give a robot the capability of recognizing spontaneous expressions in interactions with a human? There is no consensus in the literature to define emotional expressions [36]. However, Ekman et al. [84] developed a study that shows that emotion expressions are universally understood, independent of gender, age and cultural background. They established the six universal emotions: "Disgust", "Fear", "Happiness", "Surprise", "Sadness" and "Anger". Although they show that these emotions are commonly inferred from expressions by most people, the concept of spontaneous expressions increases the complexity of the expression representation. Humans usually express themselves differently, sometimes even combining one or more characteristics of the so-called universal emotions. Furthermore, several researchers built their own categories of complex emotional states,

with concepts such as confusion, surprise, and concentration [3].

To define spontaneous emotions, the observation of several multimodal characteristics, and among them, facial expressions, movement and auditory signals, has been shown to be necessary [170]. It was shown that face expression alone may contain misleading information, especially when applied to interaction and social scenarios. The observation of different modalities, such as body posture, motion, and speech intonation, improved the determination of the emotional state of the subjects.

Another problem of most HRI research is that it is restricted to a certain set of emotional concepts, such as the six universal emotions. Humans have the capability to learn emotion expressions and adapt their internal representation to a newly perceived emotion. This is explained by Hamlin [120] as a developmental learning process. Her work shows that human babies perceive interactions into two very clear directions: positive and negative. When the baby is growing, this perception is shaped based on the observation of human interaction. Eventually, concepts such as the six universal emotions are formed.

The developmental aspect of the emotion perception is also the focus of different works [125, 188, 242], and the correlation of perceiving visual and auditory emotion expressions and developing them through childhood is evident [115]. It was shown that these modalities complement each other and are one of the foundations of recognizing and understanding unknown emotional expressions.

Besides emotion perception and learning, attention and memory mechanisms showed to be important for processing emotional information. There is a strong selective attention mechanism which focuses on emotional events [295, 97], which produces an attention modulation that improves spatial perception [229, 233]. Affective memory is also an important part of perception, recognition and learning process [250, 46], and is shown to modulate how these processes work. Such systems are part of a larger emotional circuitry, which affects most of the cognitive processes in the human brain.

The emotional learning mechanisms presented in this thesis are related to these three systems: perception, attention, and memory. Although very well studied, such systems are very complex and affect and are affected by many other mechanisms in the human brain. This is a multi-interdisciplinary study field involving philosophy, psychology, neuroscience and recently, computer science. Studies on decision making [63], emotion estimation [260], wisdom evaluation [137] and artificial intuition [7] have been made, and still present many open topics.

In computer science, several models for expression recognition [45], emotion representation [284], affective states estimation [41], mood determination[66], and empathy measurement [198] were proposed. Most of these works are complementary but do not integrate the developmental aspect of emotion learning, both in relation to multimodal expressions and emotional concepts, with mechanisms such as emotional memory and attention.

To have a complete artificial affective system we need to achieve three goals: recognize multimodal emotion expressions, represent these expressions into emotional concepts, which can be learned without constraints, and integrate memory

and attention mechanisms as modulators for the learning framework. Each of these problems is difficult enough alone, and thus the solutions presented so far were very domain-dependent or not suitable for integration in a complete scenario due to computational limitations, such as sensors, algorithms, and robust representation.

1.2 Objectives

This thesis proposes an artificial affective system based on the developmental learning aspects of human emotion perception. Such a system uses different neural architectures to represent different behaviors of emotional learning, and it is built in three steps: perception, learning, and modulation.

The first step is to create, with a deep neural network, a perception model for different modalities that preserves the information of each individual modality, but also models the correlations within them. Such model should be robust enough to deal with spontaneous expressions, and adaptive enough to be able to recognize expressions from different users.

The second step builds a self-organizing network for developmental emotional perception and gives the system the capability to adapt its own perception mechanisms to different persons and expressions. Such a model uses the unsupervised learning characteristics to learn different emotional concepts based on the previous model's multimodal representations.

The last step builds an attention system and different emotional memory mechanisms to modulate what the network learned. Such mechanisms are implemented as growing neural networks and deep localization models and contribute to making the learning mechanism more adaptable to different subjects, situations, and environments.

This thesis aims to address the following research questions:

- Can a deep neural network represent multimodal spontaneous human expressions?
- How to learn different emotional concepts from multimodal spontaneous expression representations?
- How to adapt attention and memory mechanisms as modulators for emotion perception and learning?

In contrast to existing research, the models described in this thesis aim to demonstrate how different neural computational techniques can be implemented and trained in a similar way as the human developmental process to identify and learn emotional concepts.

The proposed models implement neural-inspired methods and are integrated into a complex emotional neural circuitry. A series of experiments, motivated by different neural-cognitive and psychological studies, are performed and analyzed.

These experiments range from learning how to classify spontaneous expressions to evaluating the emotional framework in different interaction scenarios.

1.3 Research Methodology

The work presented in this thesis is neurally inspired but only from a functional point of view. No attempts are made to produce a detailed biological model.

The first step of our model deals directly with data representation. The most successful way to represent data is the one done by the human brain [2]. The human brain recognizes emotional expressions from visual and auditory stimuli, correlating information from different areas. The brain also correlates past experiences, movements and face expressions with perceived sounds and voices. It is capable of integrating this multimodal information and generates a unique representation of the visual and auditory stimuli. The simulation of this process in computer systems can be achieved by neural models, particularly ones which are able to create a hierarchy of feature representations such as Convolutional Neural Networks (CNNs) [179].

The second step implements a self-organizing layer on top of the learned features in order to establish separation boundaries to the perceived expressions. Our self-organizing layer gives the model the capability to learn new expressions by creating different emotional clusters. This approach allows us to validate how representative the learned features are and gives us a powerful tool to understand how different emotions are categorized.

The third step implements two different modulation mechanisms: First an attention model is implemented with a deep neural network to improve the expression representation. This model uses shared representation to modulate what was perceived in the perception model. The second mechanism implements growing self-organizing networks to represent different memory modulations, which affect how the model learn different emotional concepts.

The focus of this research is to use the proposed model in the evaluation of different communication scenarios, with and without the presence of robots. Each of the presented steps contains its own roles and constraints, where the first one is used to identify the perceived expression, the second to model to learn emotional concepts and the third to modulate the learning.

To help in the evaluation of the proposed models we make use of a set of corpora presented and used in the literature. However, these corpora do not incorporate interactions between humans and robots, therefore we created a new interaction corpus. This corpus implements human-human and human-robot interactions and we present several different analyses on different aspects of the corpus.

We also use different visualization techniques to demonstrate that our model has a hierarchical emotion expression representation, where regions of neurons represent specific characteristics of each modality. Also, we visually demonstrate that in the self-organizing layers, each neuron codes for different emotional concepts, and how each region represents different ideas, such as perceived emotions, inner

emotional representation, and affective states.

1.4 Contribution of the Work

The neural networks implemented in this thesis use concepts such as supervised and unsupervised learning for emotion expression representations and emotional concepts, respectively. Our models implement deep neural networks for perception and localization and growing neural models for memory mechanisms. Such combination of models, architectures and concepts contribute to artificial intelligence and machine learning as a whole, while the application of such model in learning emotional concepts introduces novelty in fields as Human-Robot Interaction (HRI) and affective computation.

Besides the proposed model, deeper analysis, statistical measures and neural visualization introduce different novelties in the understanding of different neural networks. The design, recording, and processing of a novel emotional behavior analysis corpus also contribute to the field of automatic emotion recognition and introduces the use of such scenarios in an HRI environment. The main contributions of this work can be listed as follows:

- A new deep neural model based on Convolution Neural Networks for learning multimodal emotion expressions is proposed. This algorithm applies shunting inhibitory neurons in order to learn specific visual representations and the concept of cross-learning to generate robust filters for different emotional modalities. It is explained how the model creates a hierarchical emotion representation and how this contributes to the final expression representation.

- A self-organizing-based model is proposed to create emotional concepts based on perceived emotion expressions. It is demonstrated how this model represents different emotions in a non-categorical view and how these representations enhance emotion recognition tasks. This model is also used for behavioral analysis based on perceived expressions and it has the capability to identify how different expressions are represented and what these representations mean in a behavioral context.

- An emotional attention mechanism is proposed as a deep Convolution Neural Network. Such networks are commonly used for classification tasks, however, we adapt it for localization, and specify our architecture for emotional attention. Such a model is integrated into our first model as a modulator and improves the recognition and localization of different expressions. Also as a modulator, we implement attention mechanisms with growing self-organizing networks and introduce the use of such memories to improve emotional concepts learning.

- A novel emotional expression analysis corpus is designed and recorded. The corpus implements different scenarios for Human-Human- and Human-Robot-Interaction, and we perform several analyses and statistics on the data. The

corpus was designed to be used for different emotion-related tasks involving Human-Human and Human-Robot interactions.

1.5 Structure of the Thesis

This thesis is structured into 8 chapters. The initial chapters place this thesis within the field of emotion recognition in humans and in Human-Robot Interaction scenarios. They provide an overview of the broad fields touched on by this thesis.

The current chapter, Chapter 1, introduces the motivation of this work and provides the scope and objectives of the mentioned experiments.

Chapter 2 presents the conceptual and neural-biological foundations of emotion perception and recognition in humans. These include basic mechanisms for perception in different modalities and emotional concepts representation. Chapter 3 extends the discussion and describes complementary emotional concepts, such as attention and memory, and shows the psychological concepts behind emotional learning. At the end of chapter 3, the application of some of the presented concepts, and the state of the art of artificial intelligence-based models are provided.

Chapter 4 introduces the neural network concepts necessary for the understanding of the proposed models and the corpora used for the experiments. The novel corpus is presented and the details of its design, recording and analysis are presented. In Chapter 5, the emotion perception model based on deep neural networks is introduced and evaluated in different scenarios. A discussion of the results and the model itself are presented. In Chapter 6, the self-organizing architecture for learning emotional concepts is presented. The idea of how the model understands different expressions is introduced in the discussions of this chapter. Chapter 7 introduces the emotional attention and different memory mechanisms, which modulate the learning of the model.

A general discussion is provided in Chapter 8 resuming not only the outcomes of the individual chapters, but also the contribution of this thesis in the field of cognitive robots.

Chapter 2

Emotion Perception

Emotions are part of human life and have received attention since the first philosophers started to study the human behavior. In one of the earlier references on emotions, Plato [239] defined that the human soul consists of three basic energies: reason, emotion, and appetite, where reason should rule and control the emotions if a person wants to have a balanced life. In his allegory, a chariot, representing the journey of the soul, is driven by reason and pulled by two winged horses: a white one, representing positive passions (or emotions) and a black one representing negative ones. Similarly, philosophers like Aristotle [166], Spinoza and Humes [217], and Descartes [102] created theories about emotions. Through the centuries, emotions were discussed and explained as feelings [145], intentions [168], morality modulators [59] and cognitive mechanisms [59]. However, it was not until the 20th century that the study of emotions, both as a biological and psychological mechanism, became very prominent and several important types of research were made which changed how we understand the role of emotions in human life [255, 280, 221, 219, 48].

In this chapter, the concepts of emotional representation and perception will be discussed. Firstly, several philosophical concepts of emotions and how to represent them will be exhibited. Then, the basic principles behind unimodal and multimodal emotion perception in humans will be discussed in the light of neural aspects.

2.1 Emotion Representation

There is no consensus in the literature to define emotions. According to Dixon et al. [71], the term emotion replaced the idea represented by the word passion around the 16th century. Depending on different researchers emotions can be defined as intense feelings directed at someone or something [140], the state of the mind of a person [95] or even as responses to internal and external events which have a particular significance for the organism, as described by Fox et al.[98].

In their work, Fox et al. differentiate emotions into three constructs:

- **Feelings** are a subjective representation of emotions which are experienced

by one individual and are short-longing and intense.

- **Moods** are affective states, which last longer than feelings, but are less intense.
- **Affect** is a term which relates feelings and moods to persons, objects, events or memory in general.

These constructs relate to the perception mechanisms only, and how to categorize different concepts of emotions. The feelings concept is usually the one which receives the most attention, as the other two are closely related to it. The representation of these feelings is another big problem, as there is no consensus on how to group or identify different feelings. One of the first ones to deal with that in a scientific manner was Descartes [102], who stated that feelings can be described as the combination of a few basic emotional concepts, such as irritation or excitement.

Contemporary psychologists base their work on the concept which here we name Categorical Models and was described by Descartes. Others describe every feeling as irreducibly specific components divided into finite dimensions, as intensity, pleasure, self-directness among others, which we name here Dimensional Models. Yet other models were evaluated as for example the ones based on the evolutionary psychology which relates emotions to the fulfillment of basic needs, such as mating, affiliation, defense and avoidance of predators [223, 54]. In this section, we will discuss two of these views: the categorical models and the dimensional ones, which are the two most common in several approaches and present valid theories on emotion perception and learning, which are the basis of this thesis.

2.1.1 Categorical Models

In the past fifty years, many researchers tried to identify and categorize emotions. One of the most important works in this area was done by Ekman and Friesen [81]. They identified certain emotions which appeared to be universally recognized, independent of cultural or geographical background, which they called universal emotions. Only the face expression was used to create these emotional concepts, however, they evaluated their research with persons belonging to different cultures, including subjects which have no access to any kind of media, giving their evidence a strong claim. They found six universal emotions: “anger”, “disgust”, “fear”, “happiness”, “sadness” and “surprise”, as illustrated in Figure 2.1.

The concept of universal emotions from Ekman and Friesen successfully identified some cross-cultural characteristics on emotion perception, but still some emotional concepts are too complex to be understood easily. Based on their work, Robert Plutchik [240] developed the Wheel of Emotions. He suggested eight primary emotions aligned in two axes: a positive and a negative one. Differently from Ekman and Friesen, he states that the emotions are not only the feeling but the mood and affect as well. This way, he defines his eight basic emotions as “joy“, “trust”, “fear”, “surprise”, “sadness”, “anticipation”, “anger”, and “disgust”. In

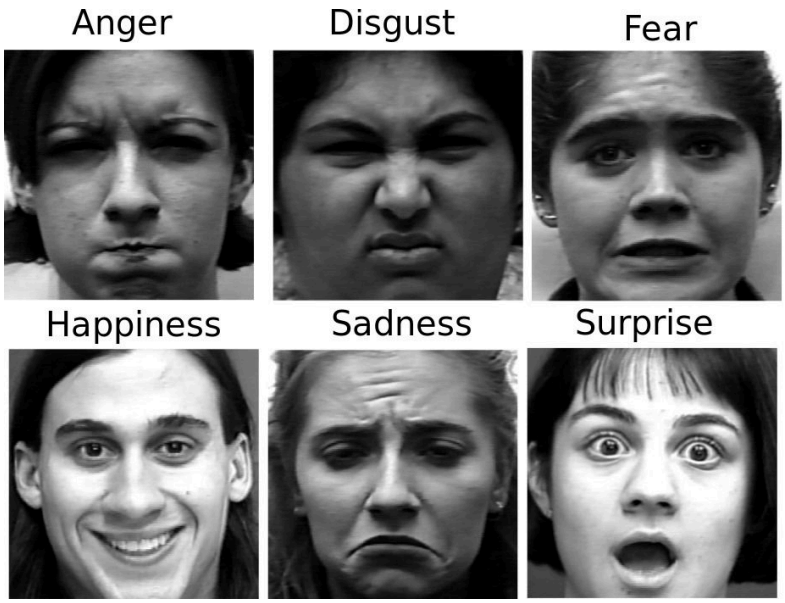

Figure 2.1: Six universal emotions described by Ekman and Friesen [81]. According to their research, these emotions could be perceived and understood independently of the person's cultural background. Based on Ekman and Friesen [81].

his Wheel of Emotions, "joy" is opposite to "sadness", "fear" to "anger", "anticipation" to "surprise" and "disgust" to "trust". Figure 2.2 illustrates the Wheel of Emotions.

Ekman and Friesen's model identifies what we perceive from what another person is feeling according to one's individual perception. Plutchik's model goes further and identifies an emotional concept, which could be specified or generalized depending on different contextual situations. As an example, in his model "happy" could be a state of "joy" or "happiness". In his work, he describes emotions as an evolving mechanism, which does not only adapt but evolve based on one's own perception, life experience, and even environment. The Emotion Wheel has important characteristics, which describe the emotional aspect of human behavior:

- **Basic emotions.** Similarly to Ekman and Friesen's model, the Plutchik model uses the concept of basic emotions. These concepts are the ones which have the most probability to be identified or felt by any person, independent of their cultural background.

- **Combinations.** The combination of the basic emotions generate all other emotions, which is a concept defended by Descartes. In this case, "love" could be expressed as a combination of "trust" and "joy".

- **Idealized states.** The basic emotions are idealized states, which means that it is not possible that they exist by themselves. Only through the observation of several different pieces of evidence (perception mechanisms, context, other emotions) it is possible to describe them.

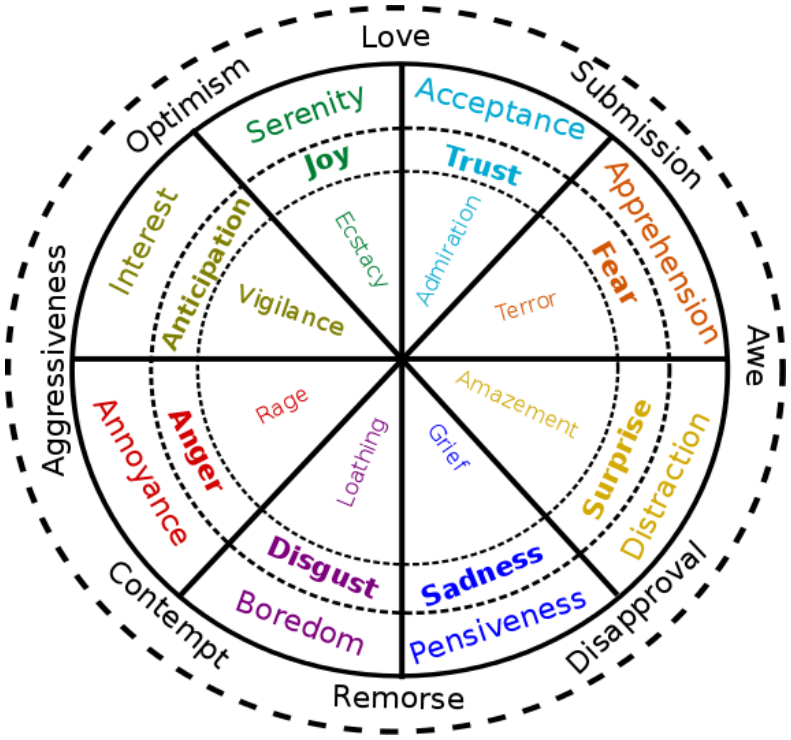

Figure 2.2: Wheel of Emotions proposed by Plutchik [240]. In this model, there are eight basic emotions which are aligned with a positive-negative axis creating opposite relations such as "joy" and "sadness", "fear" and "anger", "anticipation" and "surprise", and "disgust" and "trust". Based on Plutchik [240]

- **Opposites.** The primary emotions have opposite axes, so "joy" and "sadness" are different instances of the same emotion.

- **Similarity.** All emotions have different degrees of similarity to one another, meaning that border concepts of "sadness" and "disgust" can be blended as "remorse", for example.

- **Intensity.** Each basic emotion can vary in intensity, besides the positive and negative axis. In the Emotion Wheel, the intensity increases as you move towards the center. That means that "boredom" can be understood as a less intense "disgust" and "loathing" as a very intense "disgust".

The contribution of Ekman and Friesen's model is enormous because they introduce the idea that every human can understand a set of emotions. The work of Plutchik developed this concept and extended the way we can categorize very specific emotions. With the Wheel of Emotions, it is possible to identify very abstract concepts, like love or optimism, and very basic instincts like rage or terror. Other models were proposed, with several variants of these two models, but they tend to be more complex and more specific, pushing away from the idea of universal description from Ekman and Friesen.

These models are supported by researchers which state that the basic emotions are not learned, but produced by dedicated circuits in the brain, although they are the result of an evolutionary process [182]. That explains why persons with a different cultural background can identify the basic emotions and why they can learn to identify or describe different emotions. Following this theory, the Wheel of Emotions can be expanded infinitely, depending on the person's own experience.

2.1.2 Dimensional Models

One of the problems of the categorical models is that different persons can identify and relate emotions in different ways. Some of them can relate optimism with joy and surprise or with joy and anticipation. Besides that, it is not possible to measure how interconnected these emotions are, and the Wheel of Emotions will change depending on the person who is describing them based on personal experiences or even the current mental state of the person [144].

A different way to represent these emotions is to identify and give values to components which the emotions are made of. One of the most influential works in this area is the work of Russel et al. [261]. In their model, an emotional experience is described by two dimensions: valence and arousal. Valence measures how positive or negative that experience feels, and arousal how active the experience was. These two dimensions create a 2D coordinate system, which can describe feelings, moods and affect.

These two dimensions are the basis to identify the core affect [262]. The core affect is the main component of the conceptual act model of emotion, proposed by Barret [12]. This theory tries to solve what was called the emotion paradox: How to measure, with consistent validity, how a person describes his or her emotional experiences? In her experiment, several persons tried to describe an emotional experience using categorical models. No consistency was found, and in addition to this, the same person described the same experience differently in different time periods.

The conceptual act model of emotion claims that the perception of emotional experiences is not discrete. An opposite effect happens when describing colors. The physical colors are continuous, but when a person describes a color as blue, he or she is using his or her knowledge of colors to give the perceived wavelength. What differs is that independent of other factors, the same wavelength will always be perceived as blue by the person. With emotions, this perception is different. In her experiments, Barret found out that a person will change the category of an emotional experience (from excitement to fear when seeing a snake, for example) depending on her mood and affect. That means that instead of having dedicated circuits in the brain for the basic emotions, the brain identifies some aspects of what is being perceived and how (the most important of them is the core affect) and based on that approximates to the person's own experience.

Based on the conceptual act model of emotions, if we can identify the core affect properly, we can identify an emotion. Using the two dimensions described by Russel, the core affect could be measured easily. Figure 1 illustrates the arousal

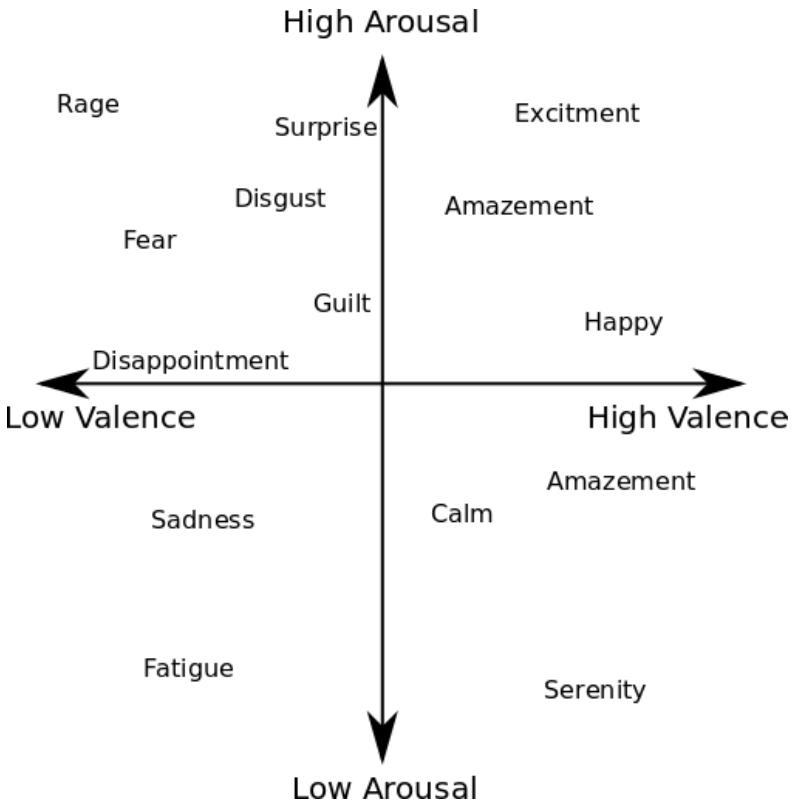

Figure 2.3: Dimensional representation of the core affect into two dimensions: Arousal and Valence. The core affect is the main component when identifying an emotional event. By determining the core affect precisely, it is possible to generate emotions based on the person's own emotional knowledge. Based on Barret et al. [13]

and valence coordinate system representing the core affect.

Using the idea of describing the core effect, two different persons can describe an emotional experience the same way, but give different names to it. As an example, if a person sees someone crying, what could identify this emotional experience as a very negative valence (crying) and with very positive arousal (high intensity), but identify it as a sad emotion. Another person would identify the valence and arousal the same way, but interpret it as a surprised emotion.

Several other researchers introduced different dimensional models, including dimensions such as self-directness and power were developed. However, most of them introduce an extra complexity in the development and description. Also, most of these models do not show any relation with neural-biological finds [258] and the arousal/valence model still showed to be the most reliable one, with strong neural-biological evidence.

2.1.3 Cognitive Emotions

The relation between cognition and emotion is still not clear. Historically, they were treated separately, but in the past two decades this area received a lot of attention and many researchers describe different integrative models.

Cognition can be understood as the mental action or process of acquiring knowledge and understanding through experience and the senses [292]. It comprises processes like memory, attention, language, problem-solving, planning, judgment and reasoning. Many of these processes are thought to involve sophisticated functions and mechanisms which are still not fully understood, including emotional processes.

Most of the cognitive processes happen in the cortical regions of the brain, connected directly to a higher evolutionary state, and some of them are found mainly in primates [103]. On the other hand, some researchers believe that many emotional processes are related directly to subcortical regions, such as the amygdala, the hypothalamus and the ventral striatum, which are often considered primitive in an evolutionary point of view [103], and are present in other mammals, for example. These regions are described as being responsible for some emotional processes such as the ones driven by rewards and punishment [253], the basic, or primary, emotions [80, 240] and unconscious body reactions [58]. For example, a dog could be conditioned to react to an action based on an emotional punishment (fear, for example), but will still have the white of the eyes very prominent.

Although some emotional processes are subcortical, the cognitive processes like perception, attention, learning, and memory have been connected with emotional characteristics [181, 228, 180]. Current thinking emphasizes the interdependence of emotional and cognitive processes, and the view of the cortical-cognitive and subcortical-emotional area is now viewed as largely simplified especially when the brain is looked at in detail.

Based on the interdependence view between emotion and cognition, the idea of cognitive appraisal has been developed in the past decades. This theory explains why persons react differently to the same things. The works of Magna Arnold [6] and Richard Lazarus [178] model the idea that the first step of an emotion is an appraisal of the situation, that means that the person's environment, current mental state, and memory will determine how he or she will describe the emotional experience.

Lazarus explains the appraisal theory using a structural model. In this model, emotions involve a relational, a motivational and a cognitive aspect [177]. The relational aspect describes the relation between the person and the environment, mostly using memory modulated by current perception. The motivational aspect deals with the person's goal, and how important the emotional experience is for the person to achieve the goal. The cognitive aspect evaluates how important the emotional experience is for the person's life, and how the person behaved in a similar experience in the past. This way, the same emotional event can be experienced differently if the person is in a good mood, or has good memories related to the situation, for example.

Lazarus' structural model is also divided into two categories: the primary appraisal and the secondary appraisal. The primary appraisal is where the person evaluates the motivational relevance and the motivational congruence. The motivational relevance indicates how relevant this situation is to the person's own needs, and the motivational congruence evaluates if the situation is consistent with the person's goals. The secondary appraisal evaluates the person's resources and options for coping with the situation. It involves the determination of who should be held accountable for the experience, the person itself, another person or entity or a group of persons, and this is determined by blame or credit values. The person also determines the coping potential and separates it in problem-focused or emotion-focused. Problem-focused coping refers to the person's ability to change the situation to be congruent to the person's goal, while emotional-coping refers to the ability of the person to deal with the situation if it cannot be changed to be congruent to the person's goal.

The structural model received some critics, especially for failing to capture the dynamic nature of emotions. To deal with that, the model was transformed into a cyclic model: after the secondary appraisal, a reappraisal mechanism was included in the attempt to capture long-term emotional responses [281]. Still, the model fails to capture the rapid or automatic emotional responses [205]. To solve that, several models based on dynamic emotional updates were proposed, the most prominent among them was the multi-level sequential process model of Scherer et al. [270].

The multi-level sequential process model describes an emotional experience in three processing levels: innate, learned and deliberate. They describe a strictly ordered step-by-step progression, in which these processes are carried out:

- **Cognitive Appraisal** evaluates events and objects, giving the personal experience an individual value.
- **Bodily Symptoms** define the physiological component of emotion experience, comprising neural and bio-chemical mechanisms.
- **Action Tendencies** describe the motivational component, giving a context of direction and motor responses.
- **Expressions** exhibit the internal intentions of an individual, using facial expressions, vocalization and body movements.
- **Feelings** describe how the subject experiences the emotional state once it has occurred, related to emotional memory.

There are various evaluation checks throughout the processes, but four of them have an important role: a relevance check, to define novelty and relevance of the experience; implication check, measure the cause, urgency and how it affects the goal; coping check, which determines how to deal with the situation and finally the check for normative significance, which evaluates how the experience is compatible with the person's standards, including moral and survival ones.

To illustrate the difference between these two models, imagine the following situation: a student is giving a speech for the first time in his life, while he is talking he looks at the audience and sees someone laughing. Dealing with this situation with the structural model from Lazarus [177] the following would happen: First, the primary appraisal mechanisms identify the motivational relevance and congruence of the situation. The student identifies that someone in the audience does not like him and this will affect his speech. The secondary appraisal mechanisms then derive that the student is not good enough (he is to blame for the situation), and as coping mechanism his mouth gets dry. In the reappraisal cycle, this situation will be always related to discomfort and failure.

In the multi-level sequential process model from Scherer et al. [270], first the situation will pass through a cognitive appraisal check, and a sequence of processes are carried out. In the cognitive appraisal check, it is perceived that the situation is new (relevance check), someone laughs (implication check), that the student himself is not good enough (coping check) and it is the first time that this situation happens (normative check). The cognitive appraisal will drive the other processes starting with the innate bodily symptom, basically, the perception that someone is laughing and the attention focus on that person. The learned process of action tendency will indicate that the student will look at the person who is laughing. Then, the expression process will be triggered, and it will make the student's mouth dry. Finally, this experience will be connected with a bad memory or even a trauma.

Figure 2.4 illustrates both the structural model and multi-level sequential process model. It is possible to see that the structural model derives information early on, and thus does not have the capability to adapt to changes happening in the process, while the multi-level sequential process model can adapt to different things happening. For example, if someone else is laughing in the same situation, the bodily symptoms process will depict that as well, and the whole process gets updated.

2.2 Emotional Experience Perception

Emotional experiences are perceived through visual, auditory and physiological sensory processes, however, mechanisms such as memory modulate what was perceived [72]. This idea was different for many years when research believed that perception and emotions were separate study domains. Only recently relevant studies were made in this area, and nowadays the consent is that emotions modulate perception, and perception influences directly the emotional experience [8, 304].

The human perception system integrates diverse senses and different brain areas. Cortical regions usually deal from low to high-level information, integration and also memory. Sub-cortical regions, such as the amygdala, have an important role on localization and unseen emotion determination, meaning experiences which are perceived but not yet processed by cognitive regions, like extreme fear or anger. All these processes start with sensory perception, for example when a person sees

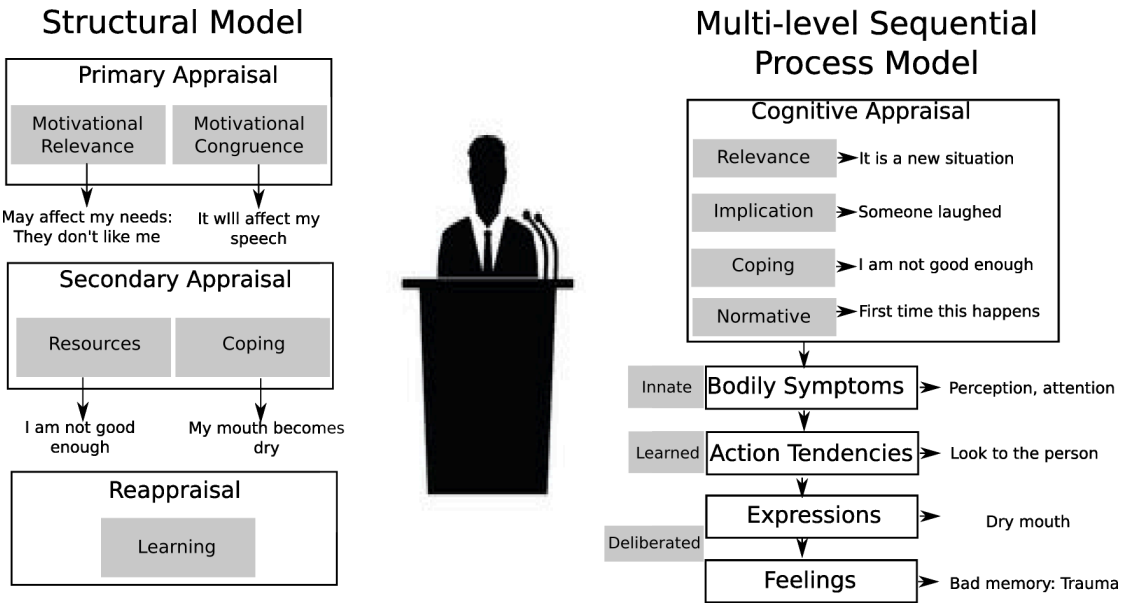

Figure 2.4: Illustration of how the two different emotional appraisal theories, structural [177] and multi-level sequential process model[270], deal with the same scenario: a student is giving a speech for the first time and someone in the audience is laughing.

and listens to someone crying, the whole emotional experience system starts with sensory processing.

Emotions are perceived with many human senses, but two of them are predominant: visual and auditory systems. Many types of research show that with these two systems humans can perceive [275] and experience [305] many emotional situations. This section exhibits how humans perceive emotions through the visual and auditory cortex, and how they are integrated into different cortical and sub-cortical brain regions.

2.2.1 Visual Pathway

The visual processing system in the brain is part of the central nervous system and processes information coming from the eyes. The whole visual system is very complex and not fully understood, but it involves all processing from the capture of the light by the eyes to the response of motor behavior and memory association. The visual information is usually processed through the visual cortex, which is the largest area in the human brain. The visual cortex is located in the rear part of the brain, above the cerebellum and both hemispheres of the brain contain a visual cortex. However, the left hemisphere is responsible for processing the right visual field and the right hemisphere the left visual field.

The visual cortex processes sensory information in a hierarchical way, and different regions have neurons reacting to different visual concepts. The information first flows through the primary cortex, composed of the V1 region and go into the

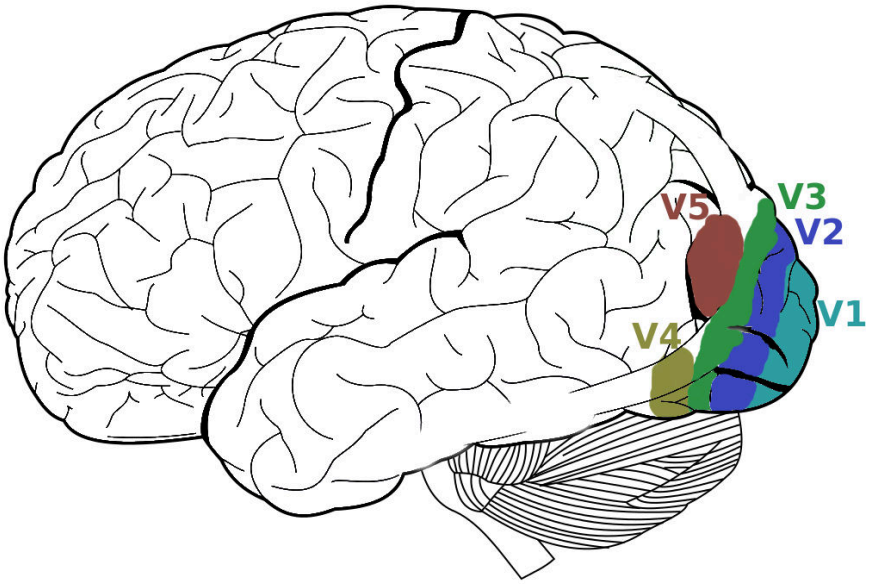

Figure 2.5: Illustration of the location of the visual cortex in the human brain in the rear part of the cerebrum, above the cerebellum. It is possible to see the regions from V1 to V5. Modified from [53] based on [109].

deeper V2, V2, V4 and V5 regions. The neurons in the primary cortex usually respond to different line segments and orientations, while neurons in V4, for example, react to complete objects or movement. This hierarchical processing allows the information to be shared through all these areas, and each of them reacts and processes different levels of abstraction. Figure 2.5 illustrates the visual cortex regions. All these regions are driven by feedforward connections, however, they are modulated by feedback and lateral interactions.

The primary cortex, or V1 area, is the most studied area in the brain. It is also the simplest and probably the earliest area of the visual cortex to develop, and it is highly specialized for processing of static objects and simple pattern recognition [185]. The neurons in V1 tend to have a strong response to a small set of stimuli, which happens because the V1 area has the smallest receptive field size in the visual cortex. Meaning that the neurons in the V1 area tend to react to small changes in orientation, spatial frequencies and colors [11]. The information encoded by the V1 neurons are basically edge detectors, representing the local contrast between different small structures and colors on the visual field. This region has straight-forward connections with the other regions, providing this fast and simple processing to deeper and more complex structures [4, 278]. Recent research [11] shows that feedback connections change also the properties of the V1 neurons over time. At first, the neurons in this region detect the small structures and information, but after this information is processed, feedback connections to the V1 neurons make them sensitive to the more global organization of the scene, such as macro disturbances and movements.

The V2 region is the second major area in the visual cortex and it receives strong feedforward connections from the V1 neurons. The neurons in V2 encode orientation, spatial frequency, and color, as the V1 area, but they have a larger receptive field. That means that the neurons in V2 identify small objects and complex patterns in multiple orientations and in different regions in the visual field [103]. These neurons are strongly modulated by orientation and binocular disparity and thus can identify background and foreground information [245]. Also the neurons in this region code a small attentional modulation, identifying macro focus regions, such as a person's shape.

The neurons in the V3 region are generally associated with the processing of global motion [27]. They receive feedforward connections from the V2 and V1 areas and are known to cover the complete visual field [200]. Usually, they encode coherent motion of large patterns, showing an understanding of what the movement means. They are mostly associated with the perception of gestures and body movements [254].

The area known as V4 receives strong feedforward connections from V2 and weak connections from V1. These neurons usually encode space relations between different objects, orientation, and color. Different from V2, the neurons in V4 encode mostly patterns with small complexity, like general shapes (circles, squares). Some research [109, 252] states that V4 is responsible for dealing with color processing, especially spatial contrast defined by different colored objects, for example, background-foreground identification based on different colors. Also, the neurons in V4 are strongly modulated by attention mechanisms [212], which have a strong influence on the firing behavior of the neurons. This behavior illustrates how subcortical mechanisms influence the visual processing.

The V5 area is also known as the middle temporal region (MT) and plays a major role in the perception of motion, integration of local motion in the global view and connections with the motor area, mostly for eyes movement. The V5 neurons receive connections from the V1, V2, and V3 neurons, and although the strongest connections are coming from V1 neurons [24], studies show that visual information reaches the V5 area even before it reaches V1 neurons [75]. The neurons in this region encode speed and direction of movements in the whole visual field, integrating local movements into the whole scene.

Therefore, we can see that the visual cortex regions process different visual information: some of them relate to spatial relation between objects and some to movement. Based on that, Milner and Goodale [111] propose the two-streams hypothesis. This hypothesis states that the visual systems process information in two brain pathways: the ventral and the dorsal stream. They exhibit anatomical, neurophysiological and behavioral evidence that the ventral stream participates in the visual cognition process, determining information about what the person is visualizing. The dorsal stream, on the other hand, is involved in the recognition and processing of where the object is, related to space. The dorsal stream processes the spatial information of what the person is visualizing, for example, the distance of the object to the person. Regions V3 and V5 are directly associated with the dorsal stream, while regions V2 and V4 are placed in the ventral stream. The

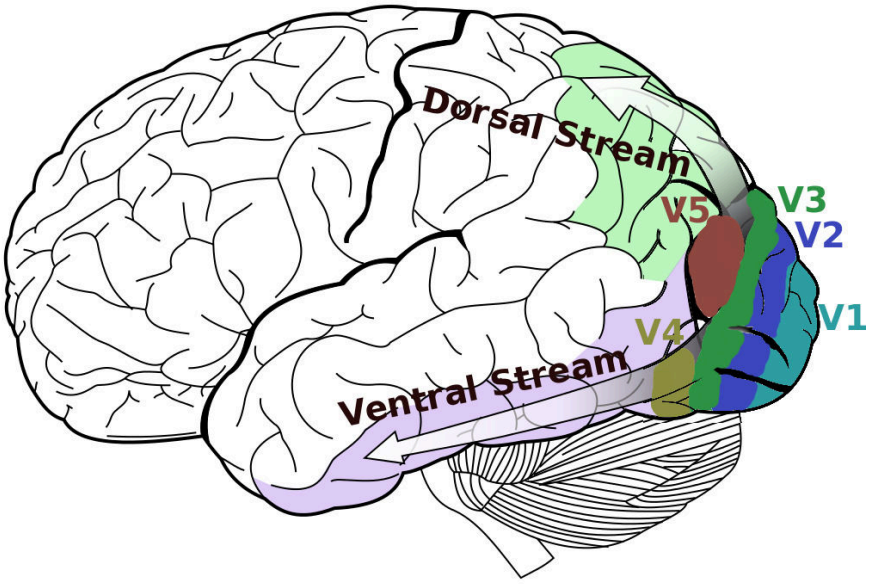

Figure 2.6: Illustration of the ventral and dorsal streams together with the visual cortex. Modified from [53] based on [109].

V1 neurons are usually connected to all of the other regions and serve as a first processing step for most of the visual cortex processing. Figure 2.6 illustrates the ventral and dorsal streams.

The ventral stream is directly associated with object and form recognition. Some research names it the "what" stream. The neurons in this region increase their receptive fields size in the deeper regions, which increases the complexity of objects recognized in the visual field. Attention and memory have a large influence on the processing of the ventral stream, giving this region a strong role in the judgmental significance of the visual field. It was shown, for example, that the damages in the ventral stream cause the inability of a person to recognize facial expressions and identifying emotional experiences [110].

The neurons in the dorsal stream region are connected directly to the V1 neurons and are known to be involved in the guidance of action and recognition of where some objects are in space. This explains the "where stream" name which is often given to the dorsal stream. The neurons in the ventral stream are directly connected with the motor system and have interconnections with the ventral stream. The neurons in this region encode two distinctive things: a detailed spatial map of the visual field and the detecting of movements. They are responsible for the perception of body movements and gestures, identifying speed, orientation, and direction of these movements. Damages in the dorsal region can lead to an inability to perceive motion and description of complex scenes, focusing only on single objects.

Both ventral and dorsal streams contribute to the perception of emotional experiences. Focusing on the identification of emotion expressions, the processing

of facial movements and body postures gives us the capability to perceive what others are expressing. The visual cortex is a very complex region and yet not fully understood, but what we know so far about it enables us to understand better how visual emotion expressions are perceived.

2.2.2 Auditory Pathway

The auditory cortex is responsible for the processing and understanding of auditory stimuli and it is located in both brain hemispheres, roughly at the upper side of the temporal lobe. This region receives and sends information from the ears via subcortical auditory systems, and it is connected with other parts of the cerebral cortex. The auditory cortex is separated into the primary cortex and secondary projection areas which are responsible for auditory pre-processing. The main processing and stronger connections with other cerebral areas are in the primary cortex [237].

The neurons in the primary cortex are organized according to the frequency of sounds they process, with neurons that react to low frequencies in one extreme of the cortex and neurons which react to high frequencies in the other. There are many auditory areas which can be distinguished anatomically, like the visual cortex, and they process the audio in a similar hierarchical way [18], where earlier regions process only some frequencies and deeper regions process a complete frequency map.

The auditory cortex represents the higher abstraction processing in the auditory system. It has a very important role in understanding language, dealing with auditory conflict and semantic processing, and despite the recent interest in understanding it, we still do not know much about it [67].

The auditory cortex is also encapsulated in the ventral and dorsal stream hypothesis [226]. Studies focused on language show evidence that auditory interpretation from phonemes to syntax understanding also occurs in the ventral and dorsal stream [248]. Figure 2.7 illustrates the auditory cortex in the human brain.

While in the visual system, the ventral stream is responsible mostly for the understanding of complex patterns and its relation to attention and memory, in the auditory system the ventral stream neurons process auditory information and relate it to semantics. First, the signals proceeding from the ears are transformed into information through the subcortical auditory pathway. Then, this information is processed in phonemes and later on recognized as words. This information then enters the ventral stream and the individual words are related with language concepts, and later on into large semantical ideas, such as phrases and scene description. Several mechanisms are involved in this process, including memory, attention, and even the visual cortex.

The dorsal stream has a very different role in the auditory pathway. Hickock and Poeppel [130] propose that the dorsal stream is directly related to articulatory motor representations. They claim that learning to speak and understand language is directly related to motor behavior, especially mouth movements. The auditory information enters the dorsal stream earlier than the ventral stream. Here, the

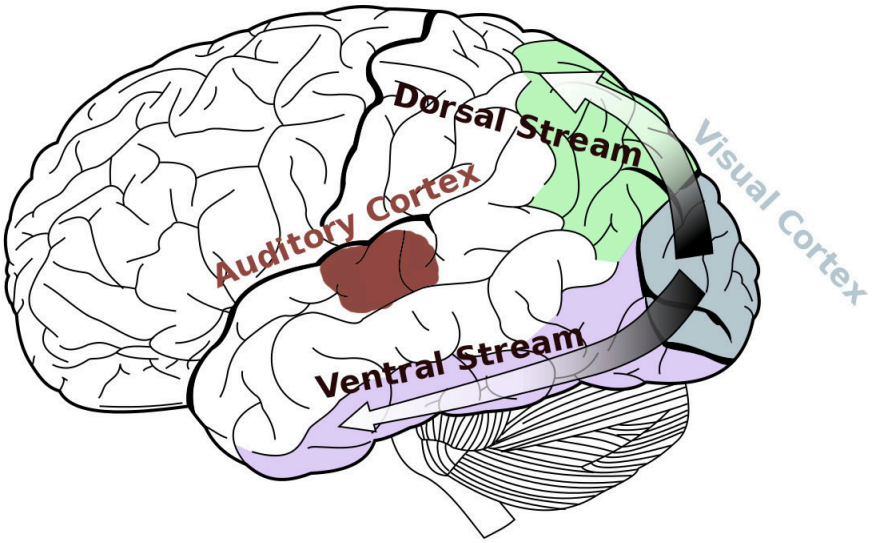

Figure 2.7: Illustration of the auditory cortex location in the brain. Modified from [53] based on [109].

information is not yet processed semantically but represented as phonetic acoustics. The first step of the dorsal stream is the sensorimotor interface for phonetic reproduction. That means that the information is paired with the ability of the person to reproduce that sound. Then, it is related to the short-term memory of the articulatory mechanisms for phonological processes [132].

The ventral and dorsal streams are also interconnected for the auditory pathway. The integration of other mechanisms, such as memory, are important for the learning and understanding of language [268]. The subcortical pathway has a very important role in the processing of auditory signals and especially attention, which modulates the cortical auditory pathway [202]. Although language is one of the most studied fields in auditory pathway understanding, the cortical regions are also related to interpretation and understanding of prosodic stimuli. Someone screaming will be processed first in the subcortical area, but the auditory pathway will have an important judgmental role on the stimuli understanding [285].

Emotional experiences can be expressed in both language and prosodic auditory stimuli. Someone talking about his or her past can produce the same sad experience as listening to someone crying. The two-streams theory helps us to understand how stimuli are processed, but still, many other complex structures are involved. For both, vision and auditory systems, the concept of independent emotional processing is becoming less evident [244, 116], and the integration of these areas with other brain regions helps us to understand how difficult emotion processing really is.

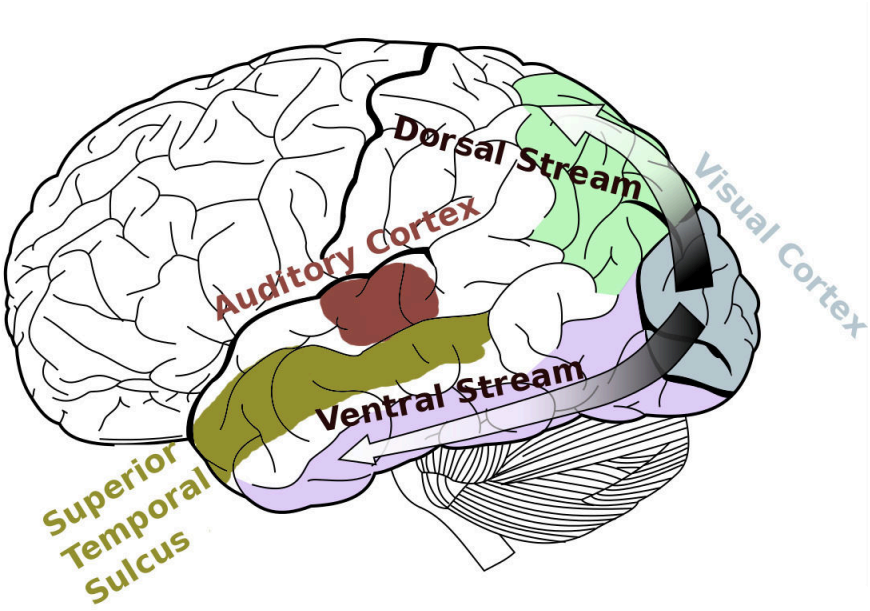

Figure 2.8: Illustration of the Superior Temporal Sulcus (STS) in the brain. Modified from [53] based on [109].

2.2.3 Multimodal Integration

Emotional experiences are perceived by a large number of sensors and processing brain regions. The visual and the auditory cortex are two of these regions, and they process unimodal information in a complex hierarchical way, with forward and feedback connections. These two regions also communicate between each other and are directly connected to other brain regions, such as the superior temporal sulcus (STS). The neurons in the STS encode several different information and include multisensory processing [276]. Figure 2.8 illustrates the STS in the brain.

The STS is directly related to social perception [19] and so it integrates several sensory information, including signals coming from the auditory and visual cortices. The neurons in the STS react strongly for semantic understanding [289], selective attention [39], emotional experience determination [114] and language recognition and learning [133]. All these tasks are highly correlated with very complex visual and auditory stimuli.

The STS is the responsible for the interpretation of the vocal input coming from the auditory cortex [132]. The neurons in the STS react to phones that compose words[289] and show very weak activation when environment sounds are present. The strong connections between the auditory cortex and the STS are mostly feedforward, but feedback connections are also found and modulate the perception of certain phones when certain words are recognized [133]. This behavior indicates that the STS is part of the modulation of the perception of words in a dialogue, taking into consideration what was spoken before. That helps in the prediction of certain words which usually come after the other, helping in a dialogue.

In a particular manner, the neurons in the STS react to semantic informa-

tion when images are present [126]. The interpretation of these images, and the construction from singular objects to high cognitive concepts - such as semantic parsing - have been shown to influence strongly in the STS neuron's firing behavior [19]. The integration of visual and auditory information for scene parsing and social understanding, including language and emotion experience determination, happens at the highest level in the STS and is sent as feedback to many other brain regions.

The STS has a strong connection to what was described with face neurons [127], which are brain neurons that react strongly when faces are present in the visual field. Neurons in the STS respond strongly when the face is moving in the visual field [30], but not when an object is moving. That indicates that after faces and objects are recognized and their movements described in the visual cortex, the STS integrates this information and modulates the attention into the most semantically important concept. This effect implicates a very strong role of the STS in two distinct concepts: semantical integration [265] of the two visual cortex streams and a modulator in the attention mechanisms [216].

Experimental results [39] show that the STS has an important role in joint attention mechanism. When a group of persons is looking at the same place, the STS is responsible for processing, identify and also drive the attention of the eyes to this place. This is a very complex task and involves the sensory perception, attention modulation, short- and long-term memory and motor integration. This is a very important aspect of understanding other person's emotions, as the social behavior of emotional experiences is a very complex task. Also, the joint attention mechanism modulates the person's own emotional perception, which is reflected in the way that the visual and auditory cortices process emotional stimuli.

2.3 Summary on Emotion Perception

Different research has been done in emotion understanding in the past centuries, and yet there is no consensus on how to define emotions. Mainly because emotions are part of many different research fields, as seen in this chapter. This chapter discussed different ways to model emotions: from categorical concepts to dimensional abstract spaces and cognitive processes. It is important to note that among the various views presented above, there is not a right or wrong emotion representation concept, and all of them are important to understanding different aspects of emotions and how they affect human lives. The research described in this thesis benefits from these different representations and use distinct properties of each of these representations to perceive, model and learn emotional concepts in different tasks.

Besides emotion representation, this chapter discussed emotion perception in the human brain. Understanding how the visual and auditory systems in the brain process emotional information help us to abstract some characteristics and mechanisms which served as inspiration for our models.

Chapter 3

Towards Emotional Behavior and Affective Computing

Recent studies show that the multisensory nature of emotional experiences is processed in several different brain regions [38]. Besides perception, which is already a very hard and complex task, mechanisms like attention, short- and long-term memory, motor representation, and even the most primitive reaction systems are involved in understanding emotional experiences [73].

Traditionally, the understanding of emotional experiences has been studied only with one modality [31, 106], but recent works study the effect of cross-modal emotional stimuli [208, 92] in emotion recognition. It was shown that cross-modal interpretation occurs in different brain levels with different mechanisms but also that the presence of cross-modal stimuli created different behavioral responses in the brain [61]. Studies show that this cross-modal modulation occurs even if the persons are asked to base their analysis only on one of the modalities in a scenario where different modalities are present [92].

The integration of these mechanisms with the perception systems creates a complex information and processing network which is not fully understood. This research helps to understand how humans perceive and understand the world around them, but also how we can solve complex problems such as emotional interpretation. This chapter exhibits insights on concepts such as attention and memory, and introduces a summary of the brain's emotional circuitry. Also, it shows an overview of the aspects of emotions which are studied by psychologists and how they help us, in addition to philosophers and neuroscientists, to understand the role of emotions in our lives. Finally, different computational models for emotion perception, learning and integration are discussed, introducing the emotional concepts from the perspective of artificial systems.

3.1 Emotional Attention

Attention is one of the most important mechanisms in the brain. It allows humans to process relevant areas of the whole stimuli perception field while suppressing

irrelevant ones, and many neural regions have been identified to be involved in spatial attention during perception [74].

Spatial attention allows us to use what is described as selective attention: to use processed stimuli to modulate the perceptive field. One of the most important cases of selective attention is emotional attention, where the affective components of the stimuli affect the perceptive field [52].

Studies with visual emotional perception show that affective estimation has a direct influence on attention. It was also shown that neural processes responsible for determining interest regions receive strong feedback connections from emotion perception regions, such as the STS and the visual cortex [293].

Many behavioral studies show that people pay more attention to emotional rather than neutral stimuli and that this behavior is often reflexive and involuntary [78, 294, 300]. Richard et al. [249] show that there is a difference in the time perception of emotional and non-emotional stimuli even when emotions are not part of the task, such as in the Stroop Task. This study shows us the importance of emotional attention in the perception processes in the brain, modulating several spatial attention tasks.

The attention modulation is the focus of different studies [229, 233], which show that low contrast stimuli improved if emotional cues are present, suggesting that emotional-driven mechanisms improve the spatial attention. Additional studies [295, 97] propose that in the case of limited attentional resources, emotional-related information is prioritized over information without any emotion connection.

One of the most important attention-related regions in the brain is the superior colliculus (SC). This region is a subcortical structure, mostly responsible for integrating low-level audiovisual stimuli and motor responses [288]. The SC plays a crucial role in general attention, and it is an important part of the attentional emotional processing [169].

Another brain region which is directly related to emotion processing is the amygdala. The amygdala is responsible for being involved in low-level decision-making (mostly involuntary), memory and emotional reactions. Research show that the amygdala has a very important role in creating an associative memory for emotional events [214], and it is also associated with reflectional processing as fear conditioning [22] and rage [42].

The amygdala and the SC are said to be part of the emotional attention processing mechanism [175]. The two-pathway hypothesis [183] suggests the role of these two structures in the processing of emotional events. This theory states that there is a partial parallel processing of the stimuli and a feedback connection between the subcortical structures - the amygdala and the SC - and the early stages of the cortical ones.

In the subcortical regions, the stimuli are first processed by the SC, which integrates the audiovisual information and has a strong response to simple emotion expressions, such as face expressions or body posture. Studies show that when presented with face expressions, neurons in the SC react to the ones which are expressing an emotion [213], demonstrating an unseen emotional filter effect. Later on, this stimulus is processed in the amygdala, which associates the filtered stim-

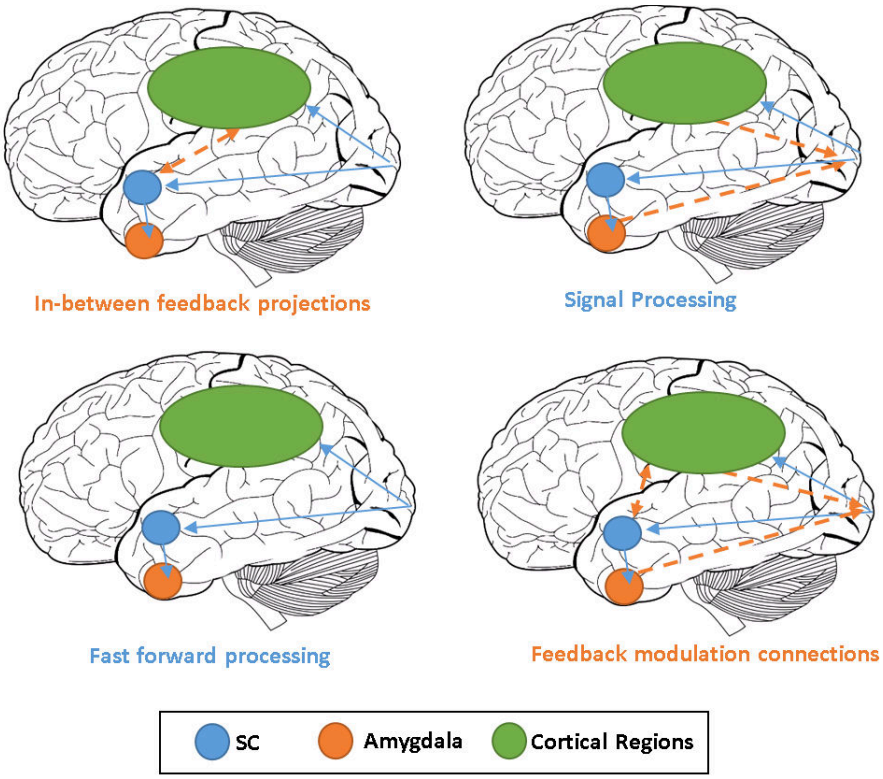

Figure 3.1: Illustration of the role of the superior colliculus (SC), amygdala (here seen through the cortical surface, although they are subcortical structures) and cortical regions in the two-pathway [183] and two-stage [32] hypothesis for emotional attention perception, respectively in the first and second row. It is possible to see the main difference of these theories, which is when the feedback between cortical and subcortical structures happens. For the two-pathway theory, the feedback happens after the primary processing of the superior colliculus (SC) and the amygdala. The two-stage theory, on the other hand, states that the feedback happens during the process. Modified from [53] and based on [293].

uli coming from the SC with emotional memories, and sends this information to cortical regions as a feedback connection. In this theory, the cortical regions send in-between feedback projections to the SC, which acts as cortical modulators.

Alternatively, the two-stage hypothesis [32] states that there is a feedback communication between high-level cortical regions and the subcortical structures after the processing of emotional attention. This theory suggests that there is a fast full stimuli processing, without feedback, from both subcortical and cortical structures, and after that, the feedback connections act as modulators for both regions.

Both theories state that feedback connections are important for processing of emotional attention and that emotional attention has a strong role in perception modulation. The main difference between them is how this processing occurs: while the two-stage hypothesis states that there are only feedback connections at the end of the subcortical processing, the two-pathway hypothesis claims that this

connection occurs in the middle of the process. Figure 3.1 illustrates both theories and show how the feedback process happens. In both cases, it is important to note the role of the SC and amygdala as a crucial modulation for the cortical areas showing how complex and important the attention modulation is.

3.2 Emotion and Memory

Memory is processed in the brain in four stages [76]: encoding, consolidation, storage and retrieval. The encoding stage is the perceptual processing of the instantaneous or short-time stimuli. The consolidation is the stage where the memory is retained in the brain [28] and happens in two steps: the synaptic consolidation, which is associated with short-term memory [172], and the system consolidation, associated with long-term memory and learning [282]. Storage is a passive stage, where the memory is stored in the brain as short-term or long-term memory. The last stage involves the recollection of these memories, usually for modulation of perception or learning.

In the early stages of evolution, emotional concepts related to survival instincts, like fear or feeling safe, followed several different experiences. These experiences were directly attached to these emotional concepts, and over time an emotional association was created with this memory [269]. Through evolution, this process of learning became genetically embedded in most animals, which has been referred to as instinctive behavior [142].

During the life of a human, many events are perceived and to each of these events, an affective concept is associated, and usually represented as an emotional dimensional concept [262], composed of arousal and valence. Studies show that usually the arousal dimension is related to memory association [37], but the valence dimension affects how an event will be remembered [174, 218]. That means that usually a person will remember exciting events, but how this event will be remembered depends on the valence giving to this event. This is directly related to goal achievements, a high arousal and high valence event [153], and traumas, a high arousal, and low valence event [47].

Emotional aspects are known to be part of the processing of three of the four memory stages: encoding, consolidation, and retrieval [250, 46]. During the encoding stage, emotions are known to modulate the perception itself and it was shown that emotional events, ones with high arousal, receives a stronger attention than non-emotional ones, with low arousal [233]. Also, it was shown that events with a high emotional arousal are more likely to be processed when attention is limited [153]. An example of this behavior is a modulation in the magnitude of the visual field when an object attached with an emotional relevance is present in the scene, meaning that if a person fears spiders, he or she will spot a spider faster than other objects.

The consolidation stage is modulated by the emotional aspects captured in the encoding stage, but also by the arousal and valence of the memory itself [174]. During consolidation, the emotion arousal appears to increase the likelihood of

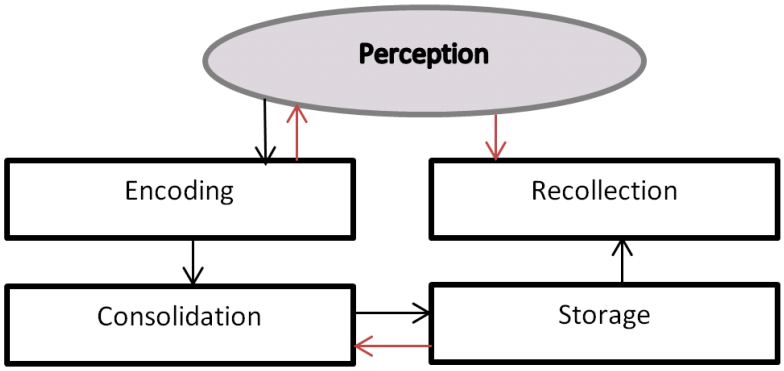

Figure 3.2: Illustration of the forward and feedback modulation connections in the four stages of the emotional memory scheme [76]. The red arrows represent the modulation, while the black arrows the forward connections.

the event being retained in long-term memory, explaining behavioral effects like punishment and reward learning [93]. Studies [47] show that the event with high arousal will receive more focus than peripheral information. That is why a person usually can remember the content of a very difficult test, but not what he or she was wearing that day.

In the last two stages, storage and recollection, the modulation comes from two different paths: from the perceived event, which triggered the recollection, and from the memory itself [153]. For each event stored in memory, an emotional concept is associated. There is a consensus in the field [26, 189] that the mood of the person, meaning the evaluation of the current emotional state the person is in, affects the recollection of an event in the memory. Watkins et al. [299] show that a person in a depressive, low valence mood tends to remember negative events. This effect was named as mood congruence. Other studies [187] show that the emotional concept associated with a specific perceived event changes depending on the person's mood, usually reflecting the mood itself. That means that a person with a depressive mood will associate more low valence labels with the perceived event, while a happy person will tend to look for the high valence elements. Figure 3.2 illustrates the forward and feedback modulation connections in the emotional memory scheme.

3.3 Early and Late Emotion Recognition

Besides describing the perceived emotional stimuli, the cortical regions of the brain play a strong role in processing and recognizing what was perceived. This recognition happens in different regions and associates the perceived stimuli, emotional attention, and interpretation of the semantic concept behind the stimuli [92]. On the other hand, the subcortical regions, like the superior colliculus and the amygdala, have a strong role in reflexive movements and association with primitive emotional behavior as self-preservation [213].

The ventral and dorsal streams communicate with each other through several levels of feedforward and feedback connections, as in the emotional attention mechanisms, but also process the stimuli in distinct ways, such as the perception and reflexive reactions of fear [73]. When a human perceives fear and to be in a threat situation, the heart-beat accelerates, pumping more blood to the brain, the face becomes paler, retracting the blood from the surface of the skin among other effects. This mechanism is still not fully understood, but it is part of a more primitive emotional experience perception: the involuntary reflexive processing [22].

Among other structures, the amygdala is one of the most important involuntary reflexive structures in the brain [42]. The neurons in the amygdala are connected with the primary sensors of stimuli perception, like the V1 region in the visual cortex, and the motor system. Studies [42] show that the amygdala is responsible for, among other things, fear to condition, connecting experiences such as a scary encounter with a monster with a scream. Other studies [227] show that the amygdala is also involved in the positive reward mechanisms, like connecting food with a positive reward (satiating the hungry). Yet, other studies relate the amygdala with memory association in the most primitive way: traumas, phobias, and emotional conditioning events, like remembering a good friend when looking at a flower.

The amygdala is also associated with memory modulation [22]. Long-term memory is formed over time, probably for a lifelong time. That means that many experiences happen until this memory is formed. Recent studies [232] show that the amygdala acts as an emotional modulator in the formation of these memories. The emotional arousal following the perceived event influences the impact of the event in the memory consolidation, and the amygdala is directly associated with this process.

In the cortical regions, emotions are recognized as a complex event which involves low- and high-level stimuli processing and memory modulation. Areas as the visual and auditory cortices and the STS process what the person is perceiving, in this case, auditory and visual information, integrate this with other senses, as temperature, smell and self-movement. This whole process is modulated by memory and by the subcortical structures.

Although very complex, emotional concepts are present in several brain regions. The emotional modulations happen from perception to memory storage and affect the daily life of every human being. The recognition of an emotion is directly associated with a perceived event, and humans are expert in labeling emotional concepts to different objects, scenes and time where the event happened. The full integration of all these mechanisms is still not fully understood, however understanding the functioning of these structures would give us a very important overview on how to recognize emotions in different perceived events and memories, and also how to make decisions, do actions and determine the person's own mood. A simplified version of the brain emotional circuitry discussed in this chapter is illustrated in Figure 3.3.

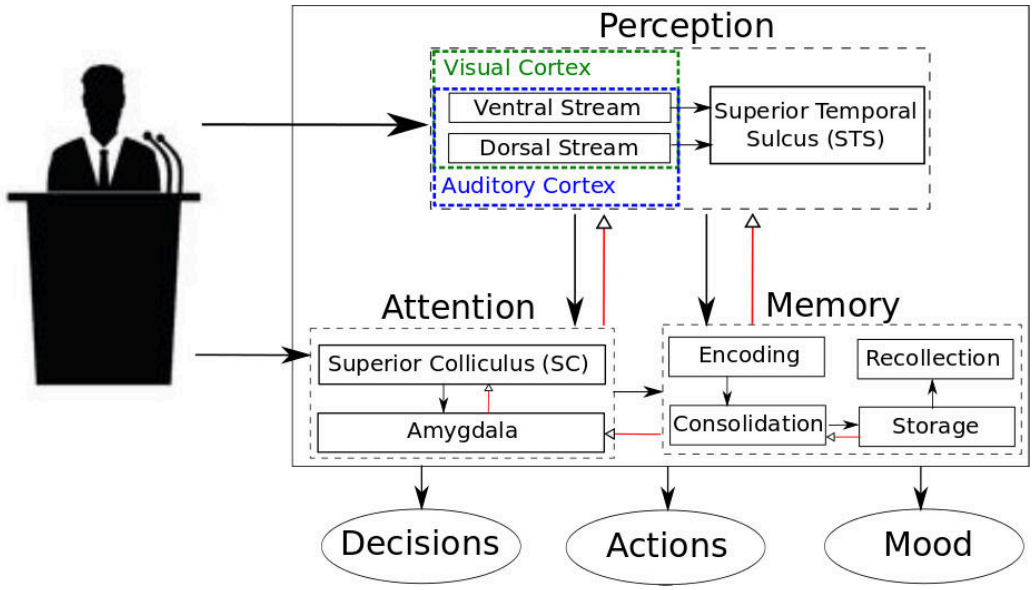

Figure 3.3: Illustration of a simplified version of the brain emotional circuitry discussed in this chapter. The red lines represent the modulation, while the black lines the forward connections.

3.4 Emotional Behavior

The cycle of perceiving an event, understanding it and associating it with different memory levels and actions is a basic mechanism and happens from birth onwards. For the whole neural system involved in perceiving, understanding and using emotions, several researchers show different psychological and behavioral aspects of emotions.

Emotions are sometimes referred to as the window to the soul [303]. In the last 100 years, the study of several psychological aspects of emotions gave us very important and different insights on how we use emotional concepts in our life. The cognitive appraisal theory tries to correlate the neural aspects of emotion perception and understanding and some psychological findings. However, many questions are not addressed by this theory, such as the emotion expression mechanisms and learning behaviors. This section discusses these two concepts.

3.4.1 Expressing Emotions

One of the most important ways to convey emotions is through emotion expressions. Humans and animals are known to show physical changes which reflect the current mood. These expressions usually involve the whole body and different modalities, such as speech, face expressions, and body movement. We usually can relate these expressions with an emotional concept, using the before-mentioned mechanisms: perception, representation, recognition, and memory.

Different persons will express differently, depending on several factors: the gender, the cultural background, the current mood, the context situation and so

on. However, humans are able to, in most cases, identify these expressions and the emotional concept behind them. In a study about expressions itself [82] it is shown that when a single emotion expression occurs, it lasts between 300 milliseconds and 2 seconds. These expressions usually carry a broader meaning, such as the six universal expressions, and are usually synchronized within different modalities. That means that an angry person would scream at the same time as his or her hands are moving frenetically in the air; and the moment the face changes, the movements also tend to adapt to it. This introduces a concept of co-occurrence of different modalities which are naturally synchronized.

Ekman and Friesen [85] introduced a way to describe face expressions in a deterministic way. The Facial Action Coding System (FACS) is a taxonomic system to describe human facial movements. The FACS defines Action Units (AU), which are contractions or relaxations of one or more muscles, and the combination of AUs can represent a facial expression. In the latest version of FACS, a total of 98 AUs is present. This system was used by psychologists and animators in the past three decades and is still a standard method of analyzing facial expressions. Figure 3.4 illustrates an example of different FACS codings for eyebrow movements.

Although humans convey their mood, most of the time, with clear expressions, a deeper analysis shows that there is more to see than the obvious. The concept of micro expressions started to be analyzed in the beginning of the behavioral studies [60], and today is one of the most important factors when trying to determine which emotions a person is conveying. Micro expressions are involuntary movements of the body, face, or change on the voice tone that do not reflect the macro expression and convey important detailing information about the person's mood or intention.

Ekman [83] demonstrates that facial micro-expressions last from 40 to 300 milliseconds, and are composed of an involuntary pattern of the face, sometimes not related to the expressions being performed. He also shows that micro expressions are too brief to convey an emotion, but usually are signs of concealed emotions, giving the expression a different meaning. For example, facial micro-expressions are usually the way to distinguish when someone is angry while using a happy sarcastic expression. In this case, the addition of facial micro-expressions as an observable modality can enhance the capability of the model to distinguish spontaneous expressions, but the observation of the facial micro-expression alone does not carry any meaning.

The interpretation of expressions and micro expressions is an important tool for human communication and social life. The understanding of such expressions has also a strong role in the learning mechanisms, and in the contextual interpretation of the environment, helping also in decision-making and character building mechanisms.

3.4.2 Developmental Learning of Emotions

There are two perspectives when talking about innate perception: the traditional one says that emotional perception and morality are learned from scratch, while others believe that some of these aspects are built-in, meaning that human babies

Figure 3.4: Example of seven different Action Units (AU), presented in the Facial Action Coding System (FACS), for eyebrow: inner brow raised (AU1), outer brow raised (AU2), brow lowered (AU4), upper lid raised (Au5), cheek raised (AU6), and lid tightened (AU7). Adapted from [16].

have already some innate judgmental capability.

In the past century, psychologists performed several experiments trying to prove that emotional perception is learned during childhood [161, 234]. They state that babies are born completely amoral, processing no or very basic moral information about what happens around them, and then during childhood and several developmental processes develop a mature perspective. This also explains that persons growing in different cultural regions have different moral standards.

The traditional perspective claims that our moral perspective is always updating, even through adulthood, and that it would be possible to train a person to follow a certain behavior, as long as the training starts in early ages. The concept of moral behavior, in this case, is directly associated with emotional perception and memory association [79]. An action could be considered good or bad, and thus shape the perception of a certain experience, for example hunting an animal. Some people will perceive the scary face of the animal and understand the whole hunting experience as a cruelty, while others will understand it as a sport event. All the experiences we have are associated with one or more emotional labels, which according to the researchers in the traditional perspective, shape our morality.

On the other hand, other researchers [119, 122] address a different question: all the aspects of the morality are learned and shaped during childhood? In the experiments performed by the traditional stream researchers this question was neglected, and they only evaluated the whole idea of innate perception or not innate perception.

Recent research [148, 211] shows that the understanding of certain conceptual domains emerges even with the absence of certain experiences. For example, a person does not necessarily have to be helped or harmed in a particular situation to understand when another person is being helped or harmed. This moral core would remain intact during the person's life.

The concept of innate moral goodness is growing among those who believe humans have some innate characteristics for emotional perception [207]. Sagi and Hoffman [263] show that from birth, newborns show very rudimentary emotional reactions to other's suffering. Hamlin [121] discusses how babies, when growing up, adapt these reactions towards different social behaviors like comforting someone in

distress [77], helping others to achieve goals [298] and sharing their own resources [107].

Based on these concepts, Hamlin [121] proposes an experiment with toddlers. A scenario with two entities, a box, and a ball is presented to the toddler. In this scenario, two actions happen: in the first one, the ball helps the box to climb a hill. In the other, the ball prevents the box to climb the hill. The toddler is then asked to choose between one of the entities, and most of the toddlers create an empathy towards the helper. This experiment shows that the toddlers could identify good and bad actions, based on a simple task. This research showed that the toddler could give a valence (between positive and negative) to the experience.

Other research [143, 266] confirms also that humans have an innate capability to evaluate an event with a valence domain. That means that humans are born with an innate emotional identification, which can be expanded throughout childhood. This is the basis for developmental research [144, 273] on emotion acquisition, and paired with the cognitive appraisal theory creates a robust framework for understanding the learning of processing new emotions.

Based on the theory that emotion perception is innate, several researchers [125, 188] state a developmental aspect of learning emotional concepts. They state that as babies we shape our emotional perception first based on two valences: positive and negative. As we grow up, the experiences are perceived in more complex ways, with the inclusion of different arousal states.

Grossman [115] shows the correlation of perceiving visual and auditory emotion expressions and developing them through childhood. They show that these modalities complement each other and are one of the foundations of recognizing and understanding unknown emotional expressions. Also, there is evidence [23] that this mechanism could be one of the most important in the development of intuition, and thus play a very important role in decision-making [63].

3.5 Affective Computing

In the past decades, many researchers approached the issue of automatic emotion recognition. With the progress in Human-Machine Interaction research, in particular, Human-Robot Interaction (HRI), the necessity of having a natural way of communication emerged.

In the nineties, the term affective computing was developed [235]. As a definition, affective computing is the area which develops computational systems that can recognize, interpret, process and simulate emotion and affective states. It is an interdisciplinary topic, and involves psychologists, computer, and cognitive scientists.

Early works on affective computing focused on trying to create universal descriptors for emotion perception, mainly on visual [62, 192] and auditory streams [90, 230]. However, given the complexity of human emotional behavior, many of these solutions were very domain-specific or neglected important aspects such as multimodal perception or emotional learning.

Recent researchers approach the topic from neuroscience by inspiring their models on how the human brain processes and interprets emotions [45, 251, 196], and from the recent findings of psychology, especially on the emotional learning and behavior. These models can deal with complex tasks, such as recognition of emotions in the wild, but also the use of emotional concepts to improve communication and learning in artificial systems.

In the next sections, the most relevant of these approaches, models and systems are presented and discussed.

3.5.1 Emotional Descriptors

Early works on automatic emotion recognition systems proposed the use of descriptors [62, 90, 230] to model the expression, and mostly stochastic [241] or rule-based classifiers [192] to label it. The contributions of these works are on representing emotion expressions, and most of them were based on Ekman's assumption of universal emotions.

Most of the early works dealt with two separately streams: face expressions and speech. These works applied different feature descriptors in a way to describe an expression. Most of these feature descriptors were based on general image descriptors, when applied to face expressions, or general sound descriptors, when applied to speech. A review of the field from Chellappa et al. [44] exhibits a broad range of different methods used to recognize faces and shows that most works were still at an early stage and not close to being used in real-world scenarios.

In early research, some common methods for face expression recognition involve the use of motion descriptors [152, 89] and template matching [20]. All these works were computationally expensive, not able to deal with real-time processing, and had problems with generalization. For example, if the pictures were captured under different lighting condition or from different persons, these models did not show good performance.

Some of the works for face expression, however, were inspired by the psychological studies of Ekman and Friesen and the FACS system was used as inspiration for many computational models [90, 191, 51]. These models use mostly template-matching approaches to identify the Action Units (AUs) in a person's face and use simple rule-based classifiers [51] or even simple neural networks [257] to identify the expressions.

The works using the FACS models solved a big problem in the field: they could recognize expressions with a fair degree of generalization based on the evidence of Ekman's research. However, these models faced a new problem: they had to identify and describe, perfectly, the AUs, otherwise the classification would not work due to the fact that similar AUs are involved in completely different emotions [191].

Several approaches proposed better ways to map the AUs, including the temporal analysis of the face movement [224], geometric deformation analysis [167], and profile contour analysis [225]. These methods are strongly dependent on the AUs to describe expressions and were able to identify dynamic and static ones. Cowie

et al. [57] exhibit different solutions for identifying and describing the AUs, and discuss how effective they are for generalizing the expression for different persons.

The use of explicit descriptors, mostly related to FACS, introduced an increase of studies for automatic emotion recognition systems. However, the FACS model, and models based on common explicit descriptors show a big problem: it is difficult to represent spontaneous expressions with them [86, 307]. The purpose of FACS is to describe muscle movements and classify them into emotional concepts. However, emotion expressions are spontaneous, and different persons will express the same emotional concept in different ways. The works based on FACS and explicit descriptors cannot deal with this nature, and are mostly used for basic emotion recognition, such as the six universal emotions [290].

In the past decade, the introduction of Convolutional Neural Networks (CNNs), among other deep learning models, provided an important evolution on image recognition tasks. Such models use the concept of implicit feature descriptors to classify complex images. Instead of using a set of pre-defined descriptors, CNNs learn a particular descriptor which will perform best on the classification task which it was applied. This was shown to be efficient in several different image recognition tasks [171, 151, 146].

Given the success of CNNs in several different tasks, it was largely applied for emotion recognition tasks [94, 209, 149], showing an improvement on generalization. Different architectures were applied for different tasks. Fasel et al. [94] propose an architecture which is head-pose invariant. Their model evaluates a CNN trained with different expressions presented with different head-poses and this model was able to generalize the learned features for different subjects.

In the approach of Matsugu et al. [209], they use a rule-based mechanism in between the convolution layers to improve the generalization of the detected features. This rule-based mechanism identifies if the learned features are related to different face structures, such as eyes or mouth, and use this information to apply a selective mask in the filters. This is used to create a face detection mechanism, firstly, and then to identify face expressions. They show that their system increases the recognition of the six basic emotions when compared to common FACS-based systems.

Karou et al. [149] use a combination of general and specific convolution channels to recognize emotions in videos. Their architecture has different training steps, first based on general image descriptors, and later on specific emotion expressions. They use a video aggregation scheme to recognize emotions in a sequence, where N-frames are aggregated in one representation by summing the frames' own representations. They show that their network could be used in real-life emotion recognition tasks by using their model to recognize emotions in movies and series clips.

The use of CNNs to describe facial emotions showed also a big improvement for spontaneous expressions recognition [302, 186, 283]. Although using CNNs, these works rely on heavily pre-processed data [302], complex tree and rule-based descriptors identification [186] or even in different explicit feature descriptors used to complete the final facial representation [283].

Although the use of CNNs to learn face expressions presented a better performance when compared to FACS-based systems, recent studies show a certain correlation on what the network learns and the AUs of the FACS, as presented by Khorrami et al. [154]. In their work, they discuss that the learned features of a CNN trained with facial expressions approximate some AUs, showing that the network actually learns how to detect AUs without any pre-defined rule. This shows that the FACS system could be limiting the recognition of spontaneous expressions, however, describing expressions with muscular movements show to be a robust facial expression descriptor.

Similarly as facial expressions, the development of auditory emotional descriptors evolved in the past decades from explicit descriptors [64] to implicit ones [1]. The FACS system does not have any auditory emotional description, which led to a wide range of different descriptors and acoustic models used in emotion recognition.

Most of the works on speech emotion recognition are based on popular auditory descriptors, which were mostly developed to describe the human voice. Early works used a combination of simple acoustic descriptors [230], such as vocal energy, pitch, speech rate, among others to identify mostly the six universal emotions. These models rely heavily on a very clean input data, mostly an expression of a word or a short sentence.

Different kinds of descriptors were developed and applied to speech recognition. The most successful were the ones based on the Mel Frequency Cepstral Coefficients (MFCC), which proved to be suitable for speech representation [264]. MFCCs are described as the coefficients derived from the cepstral representation of an audio sequence, which converts the power spectrum of an audio clip into the Mel-scale frequency. The Mel scale was shown to be closer to human auditory system's response than the linear frequency.

Each auditory feature descriptor carries its own information, changing the nature of the audio representation. For the same clip of sounds, very distinct information can be extracted for different tasks. Thus, Madsen et al. [203] use a set of three different descriptors, namely chroma features, loudness, and MFCC, to represent distinct auditory information. They also obtain different temporal/non-temporal representations for each descriptor, using sliding windows for discrete representations or Hidden Markov Models for temporal dependencies. A total of 83 feature representations is obtained. After the feature extraction, they use a hierarchical non-parametric Bayesian model with a Gaussian process to classify the features. They show that their approach has a certain degree of generalization, but the exhaustive search for tuning each parameter of the model for multiple feature representations and feature combinations is not a viable option.

Similar to the work of [203], several approaches [251, 147, 195] use an extensive feature representation strategy to represent the audio input: they extract several features, creating an over-sized representation to be classified. The strength of this strategy relies on redundancy. The problem is that usually, it is not clear how well each of these descriptors actually represents the data, which can lead to not capturing the essential aspects of the audio information, and decreasing the gener-

alization capability [134]. The use of over-sized auditory feature representations is the focus of heavy critics [307, 87] because of the incapability of these descriptors to describe the nuances of emotion expressions in speech.

In a similar way as facial expressions, implicit descriptors methods, mostly CNNs, were used recently to describe emotions in human speech [138, 204]. These systems use a series of pre-processing techniques to remove mostly the noise of the audio signal, and let the CNN learn how to represent the data and classify it in emotion concepts [264]. In this strategy, the CNN is able to learn how to represent the auditory information in the most efficient way for the task. Initial research was done for music [190], and speech recognition [272] and was shown to be successful in avoiding overfitting. The problem with this approach is that it needs an extensive amount of data to learn high-level audio representations, especially when applied to natural sounds, which can include speech, music, ambient sounds and sound effects [65].

Although there are models to describe systematically the emotional components in speech [56], there is no consensus on how to identify affective components in a human voice, and thus the evolution of emotional auditory descriptors is limited and much more work is necessary.

Many other models which are not based on face expressions or speech were proposed. The use of several physiological signals, coming from muscle movements, skin temperature, and electrocardiograms were investigated [236, 156, 155] and delivered good performance, however, used uncomfortable and specific sensors to obtain the signals, not suitable for be used in real-world scenarios. Mechanisms such as bracelets or localized electrodes were used.

In a similar approach, the use of Electroencephalogram (EEG) signals was used as emotion descriptors [25, 194, 197]. Using non-invasive EEG mechanisms, the brain behavior was captured when different emotion experiences were presented to a human. The initial results showed poor accuracy and generalization but open one new approach on emotion descriptors.

Another way to describe emotions is using body movement and posture. There are studies which show the relation between different body postures and movements with emotional concepts [55], and even with micro-expressions [83]. Most of the computational models in this area use the body shape [43] or common movement descriptors [40] to represent the expression and showed good accuracy and a certain level of generalization. However, the best performance used different sensors such as depth cameras or gloves to capture the movements [158].

3.5.2 Emotion Recognition and Learning

Once the emotion expression is described, another problem arises: how to classify them into emotional concepts. Many approaches deal with this problem by applying general classification techniques such as ruled-based systems and tree structures [192], stochastic methods [241] or neural-based models [43].

Rule-based systems, as decision trees [184], were applied in many automatic emotion recognition systems. Many of them use strong descriptors, such as the

FACS [192], and use if-then rules to identify emotion expressions. Such systems present a simple and efficient scheme to identify pre-defined expressions. Most of these works are used to recognize the six universal emotions and usually does not show good generalization due to the fact that the emotional concepts must be very clearly separable and identifiable. Applications of such systems in speech emotion recognition delivered good performance when used in very restricted word scenarios [301], however, achieved poor performance when used in more complex or natural cases, such as real-world interaction [68].

The models based on stochastic approaches usually deal with sequence problems. Among these models, the ones based on Hidden Markov Models (HMM) became very popular in the last decades. Such models use different descriptors to represent the expression and use them to feed one [193] or several HMMs [50]. Each HMM introduces a sequence dependency processing, creating a Markovian chain that will represent the changes in the expression and use it to classify them as an emotional concept. Such models are very popular for speech emotion recognition, due to the good performance delivered in speech recognition tasks [173]. The biggest problem with these models is that they are limited to the amount of information they can learn, and they do not generalize well if new information is present. Also tend to be computationally expensive, especially for larger datasets.

Another popular method for emotion recognition are neural networks. Since the first approach in this field, neural networks have been used to recognize emotional concepts [159, 257]. These models are inspired by the human neural behavior, and were approached from very theoretical [105] to practical approaches [272]. Neural networks were used as single-instance classifiers [141] to sequence processing and prediction, using recurrent neural networks [271]. These models tend to be more complex to design and understand, and thus limit the implementation and development of applications. Usually, a large amount of data is necessary to train neural networks, and generalization can be a problem for some architectures.

With the advance of deep learning networks, most of the recent work involves the use of neural architectures. The ones with the best performance and generalization apply different classifiers to different descriptors [193, 45, 147], and there is no consensus on a universal emotion recognition system. Most of these systems are applied to one-modality only and present a good performance for specific tasks. By using one modality, either vision or audition in most of the cases, these systems create a domain-specific constraint and sometimes are not enough to identify spontaneously and/or natural expressions [274].

Multimodal emotion recognition has been shown to improve emotion recognition in humans [35], but also in automatic recognition systems [118]. Usually, such systems use several descriptors to represent one expression and then one or several classifiers to recognize it [45]. Gunes et al. [118] evaluate the efficiency of face expression, body motion and a fused representation for an automatic emotion recognition system. They realize two experiments, each one extracting specific features from face and body motion from the same corpus and compare the recognition accuracy. For face expressions, they track the face and extract a series of features based on face landmarks. For body motion, they track the position of the

shoulders, arms and head of the subject and extract 3522 feature vectors, using dozens of different specific feature extraction techniques. These feature vectors are classified using general classification techniques, such as Support Vector Machines and Random Forests. At the end, they fuse all feature vectors extracted from both experiments and classify them. The results obtained when fusing face and motion features were better than when these modalities were classified alone.

The same conclusion was achieved by Chen et al. [45]. They apply a series of techniques to pre-process and extract specific features from face and body motion, similarly to Gunes et al. Differences are that they use fewer features in the final representation and the time variance representation is different in both approaches. Gunes et al. use a frame-based-classification, where each frame is classified individually and a stream of frames is later on scored to identify which emotional state is present. Chen et al. analyze two temporal representations: one based on a bag-of-words model and another based on a temporal normalization based on linear interpolation of the frames. Both approaches use the same solution based on manual feature fusion, which does not take into consideration the inner correlation between face expression and body motion, but fused both modalities using a methodological scheme.

The observation of different modalities, such as body posture, motion, and speech intonation, improved the determination of the emotional state of different subjects, increasing the generalization of the models. This was demonstrated in the computational system of Castellano et al. [43], where they process facial expression, body posture, and speech, extracting a series of features from each modality and combining them into one feature vector. Although they show that when different modalities are processed together they lead to a better recognition accuracy, the extraction of each modality individually does not model the correlation between them, which could be found when processing the modalities together as one stream.

The same principle was also found for visual-only modalities [118, 45], and audio-only modalities [251, 147, 195]. However, all these works deal with a set of restricted expression categorizations which means that if a new emotion expression is presented to these systems, they must be re-trained and a new evaluation and validation of the whole system need to be done.

3.6 Summary on Emotion Learning and Affective Computing

Many types of research were done on understanding how emotions affect humans on many different levels. From neural modulation in perception and learning systems to behavioral aspects of emotional concepts and affective computing, the most important notions and methods were discussed in this chapter. In each of these areas, different researchers discussed, proposed and described many different mechanisms, models, and theories, and yet we are far from a unified model of emo-

tions. This can be explained by the wide range of systems and mechanisms were emotions have an important role, and thus indicates how important is to continue understanding and to research in the field.

This chapter also presented an overview of several approaches and models for different emotional tasks in affective computing. Although these models offer solutions for various tasks, none of them led to an integrative solution for emotion recognition, processing, and learning, which is the primary focus of this thesis. Inspiring some solutions present in our models in the neural-psychological mechanisms presented in this chapter allows us to address important aspects of our research questions, and contribute to the field of affective computing.

Chapter 4

Neural Network Concepts and Corpora Used in this Thesis

This chapter discusses the most important neural network concepts and techniques used for the development of the models proposed in this thesis. The techniques are exhibited in their original models and any necessary modification is indicated in each of the model's own chapters.

To evaluate the proposed models a number of different corpora are necessary and hey are described in section 4.6. During the execution of this work, a new corpus was recorded and all details involving the design, collection and analysis of the recorded data is presented in section 4.7.

4.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models which are inspired by the behavior of the neurons. The first of these models was the perceptron [256], which simulates a single neuron and it is the elemental computing unit of an ANN. The perceptron consists of a weight vector (w), representing synaptic connections, that is multiplied by the input signal (x) and a bias unit (b), which usually has a value of -1. The values resulting from this operation are then summed. In the last step, these summed values are fed as input to an activation function(y), also known as a transfer function, which will produce the output signal (o). Figure 4.1 illustrates the structure of a perceptron unit.

Stacking several perceptron units together in one layer and connecting these stacks sequentially, without connections between the neurons in the same layer, produces what is known as a multilayer perceptron (MLP) neural network. The MLPs are the main component of most of the neural networks applications, and were applied to several different tasks in the past 60 years. Figure 4.2 illustrates the high-level common architecture of an MLP. Each neuron in an MLP is usually fully connected with all the neurons in the next layer, with its own set of weights. The first layer of an MLP is usually the input signal and the last layer is called output layer. The layers between the input and output layers are called hidden

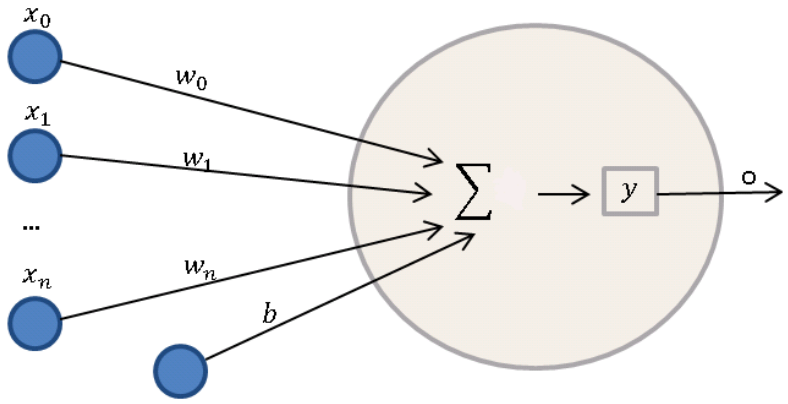

Figure 4.1: Perceptron illustration, representing the input signal (x), the weights (w), the activation function (y) and the output signal (o).

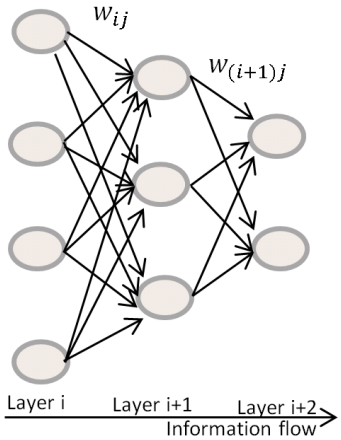

Figure 4.2: Multilayer perceptron illustration. Each layer contains several perceptron units, which are then connected to units in the subsequent layer.

layers, because its representation is not known and not important for the network designer.

The knowledge of ANNs is stored directly in its weights. Their weights represent the transformation of certain patterns in the input signal into a processed representation, which can be related to another concept, as for example in a classification task. This means that ANNs can be trained to create a separation space between the input signal and the desired output concept. Depending on the way the network is designed, it can be tuned to create different representations for different parts of the input stimuli, or integrate several different stimuli into one single representation.

The idea of using several ANNs to solve a complex task is somehow an abstract method inspired by how information is processed in brain circuits. Because of that, the ANNs have similar properties to the brain's neural circuits, such as being able to have parallel processing, adaptation over time, be robust against noise, and most importantly, being capable of generalizations.

An ANN can be trained to adapt its weights to a task. The parameters such as the weights, the number of layers, the number of units, the activation function among others can be chosen or updated by certain rules in order to obtain an expected result. The automatic update of these parameters, mostly the connection weights, give the network the power to learn and thus adapt to the problem it was applied to.

There are many learning strategies that can be used to adapt, or train, the weights of an ANN. These strategies can be separated into two categories: supervised and unsupervised methods [199]. A supervised strategy uses the concept of a teacher, which will guide the ANN during the training. This teacher will indicate to the ANN how the task has to be done by updating the weights based on the error between the current network output and the desired output. In this strategy, the teacher knows what the outputs are which the network should learn.

On the other hand, the unsupervised strategy has no teacher. This strategy relies on an underlying probabilistic distribution of the data to update the weights. The unsupervised strategies are used mostly to create an ANN capable of describing the data, reduce the dimensionality and complexity of the problem or to increase generalization.

4.2 Supervised Learning with Backpropagation

Backpropagation [259] is one of the most used supervised learning strategies for training multilayer perceptron networks. It became so popular because of two characteristics: the simplicity of the concept and the efficiency in the tasks it was used for so far.

The main idea of backpropagation is to minimize the error E between the network output and the desired output, also known as target or teaching signal, t. The algorithm updates the weights of the neuron connections, and thereby tries to minimize E. The most common way to calculate the error E is through the sum-of-squares error function. Each output unit (k) error is summed into one value using the following rule:

$$E = \frac{1}{2}\sum_{k=1}^{n}(t_k - y_k)^2 \tag{4.1}$$

where the factor 1/2 is used to simplify the computation and has no major effect on learning [199]. After calculating the error, each unit has its weights updated using the following rule:

$$w_{t+1} = w_t - \eta\frac{\partial E}{\partial w_t} \tag{4.2}$$

where w_{t+1} is the updated connection or weight, w_t is the current value of the weight, and η is the learning rate, a parameter which modulates learning speed. $\frac{\partial E}{\partial w}$ represents the change of the error with respect to the weight w_t.

The error is then backpropagated through the entire network, and the layers are updated in relation to the connections in the previous layer. As an example of this operation, imagine a network with the same structure as the one in Figure 4.2: one input layer, one hidden layer and one output layer. The activation of each output unit (Y_k^3) can be calculated as:

$$Y_k^3 = f\left(\sum_j w_{jk} Y_k^2\right) \tag{4.3}$$

where Y_k^2 represents the activation of the $k-$th unit in the 3rd layer, w_{jk} is the connection weight between the k-th unit in the current layer and the j-th unit in the previous layer. Because the error calculation shown in Equation 4.1 depends on the derivative of the error, the activation function f needs to be differentiable.

In the same way, we can calculate the activation for each unit in the hidden layer (Y_k^2):

$$Y_k^2 = f\left(\sum_j w_{ij} x_i\right) \tag{4.4}$$

where w_{ij} represents the connection of each j-th unit in the hidden layer with the k-th unit in the input layer, and x_i is the i-th input unit.

To proceed with the weight updated in the output layer, we have to calculate $\frac{\partial E}{\partial w_{jk}}$. To do that, we derive the Equation 4.1 after we replace the network output y_k by Equation 4.3. Applying the chain rule we obtain:

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial Y_k^3}\frac{\partial Y_k^3}{\partial w_{jk}} = \left(t_k - Y_k^3\right) f'\left(\sum_l w_{lk} Y_l^2\right) Y_j^2 \tag{4.5}$$

The same rule can be applied to calculate the derivative of the weights in the hidden layer, just replacing Y_j^2 in Equation 4.5 by Equation 4.4.

Theoretically, backpropagation could be used to update any network independent of the number of layers, although a network with two hidden layers is enough to approximate any function with arbitrary accuracy [199]. Important to note that backpropagating the error rapidly becomes ineffective to update the weights in deep multilayer perceptrons, especially in the layers closer to the inputs [179], which is referred as vanishing or exploding gradient problem. This happens because the derivative of the cost function in the first layer is lower than in a deeper layer, and the contribution of the error for each weight connection is reduced. To improve the capability of backpropagation to train arbitrary networks, some regularization techniques are used. In the next sections the L1 and L2 normalization, momentum and dropout techniques will be explained.

4.2.1 L1 and L2 Regularization

Regularizing the weights during network training helps to avoid problems such as the exponential decrease or increase of the error gradient. Regularization also

prevents that the weights memorize the input data, which is known as overfitting. An ANN should be able to process data which was never presented to it during the training, generalizing its knowledge to new information, so a network that just memorizes the training data is not ideal.

The L1 and L2 regularization rules add terms to the weights updated in order to avoid them to memorize the data. The difference between the L1 and L2 is how these added terms are calculated. For L1, the sum of the weights is used to calculate the regularization term. Adding the L1 regularization term to the weight updated rule represented in Equation 4.2 can be expressed as:

$$w_{t+1} = w_t - \eta \frac{\partial E}{\partial w_t} + \lambda \sum_{k}^{i} \|w_t\| \tag{4.6}$$

where λ represents a parameter which controls the relative importance of the regularization term. The L2 regularization term can be defined as the sum of the square of the weights, and in a similar way as in Equation 4.6, it can be expressed as:

$$w_{t+1} = w_t - \eta \frac{\partial E}{\partial w_t} + \lambda \sum_{k}^{i} w_t^2 \tag{4.7}$$

4.2.2 Momentum Term

The backpropagation algorithm is used to train the network and minimize the error, trying to find the global minimum of the function. The global minimum is the optimal value of the weights which will produce the minimal error, independent of the inputs. Finding the global minimum is difficult most of the time, and because of the nature of the backpropagation, the algorithm can be stopped by what is known as local minimum. The local minimum is an intermediate solution, which minimizes the error of the network but not optimally.

To avoid local minima is a very difficult task, and it can be achieved by using different regularization techniques, changing the network topology or preprocessing the data, among other solutions. To help the network to avoid local minimum, several algorithms are proposed, including the momentum term. This algorithm is probably the most used one to avoid local minima and produces good results for most of the applications.

The concept behind the momentum term is that, to avoid local minima, the update of the weights should be enhanced when the update is larger. This algorithm introduces an effect that increases the size of the weight change when the gradient keeps pointing the same direction, making the weight change faster when the network is following an optimization path. On the other hand, when the direction of the gradient keeps changing, the momentum term will smooth out the variation. This property helps when the network finds several local minima, which will change the direction of the error. We can update Equation 4.2 to add the momentum term and express it as:

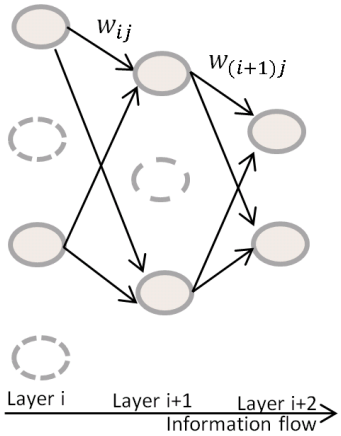

Figure 4.3: Illustration of the dropout algorithm applied to Figure 4.2 during one training cycle. It is possible to see that some of the neurons have dropped out during the training, reducing the number of connections.

$$w_{t+1} = w_t - \eta \frac{\partial E}{\partial w_t} + \mu w_t \tag{4.8}$$

where μ represents the momentum term, which should be larger than 0 and smaller than 1. The momentum term adds a fraction of the previous weight value to the current update.

4.2.3 Dropout

Deep MLPs tend to overfit easily. Because of the many numbers of weights to update, the network usually ends up memorizing the data instead of generalizing. A way to prevent this is using the algorithm known as dropout [134]. The dropout algorithm tries to prevent the co-adaptation of the weights of the network to the training data by omitting some units in the hidden layers during training.

The algorithm follows the concept that whenever a training sample is presented to the network, each hidden unit has a chance of been turned off. That means that all the connections to this unit and departing from this unit are also turned off along with it. This behavior is temporary, and it is only valid for one training cycle. That is achieved by using a probability of dropping out for a unit, which is usually around 50% [134]. Figure 4.3 illustrates the architecture depicted in Figure 4.2 when dropout is used during one training cycle. Note that for each training cycle, a different set of neurons can be dropped out, meaning that for each training cycle a different sub-set of neurons is trained.

Dropout could be compared to training a lot of different neural networks within the same training cycle. The difference is that, training and testing many different networks would be very expensive, and they would be completely independent of each other. What dropout does is that a new subset of the neural network is trained during every training cycle, however, the units in this subset can be present in another subset.

The resulting effect of dropout is that the training time of the network will increase, but also improve the generalization. By training the hidden units with different input units, the algorithm is making that specific unit robust to different input data entirely. The dropout gives each hidden unit a robustness against incomplete or useless information, and also avoids that the unit memorizes the input data.

Computationally speaking, dropout can be seen as a mask applied to the network topology. Giving the activation function described in Equation 4.4, dropout could be expressed as:

$$Y_k^2 = d * f\left(\sum_j w_{ij} x_i\right) \tag{4.9}$$

where d is a binary mask filled with zeros and ones which are randomly distributed in a way that summed they reach the dropout probability factor. The mask d is used only when training the network, and for every training cycle the values of d are recalculated. When using the network for classification tasks, the mean of the network topology is used. The mean is calculated over all the hidden units, however, it has to compensate for the fact that during testing roughly twice as many hidden units is used because of the dropout probability factor of 50%. In this case, the weights are then divided by 2.

4.3 Unsupervised Learning with Hebbian Learning

The concept of unsupervised learning differs most from supervised learning in one property: the absence of a teacher. In the supervised learning techniques, it is necessary to update the network related to an error which is calculated by comparing the network output with the desired output. On the other hand, unsupervised approaches use the data's own distribution to update the network. In other words, the network learns how to update itself based on the data it is receiving.

One of the most popular and efficient unsupervised approaches is Hebbian Learning [129]. Hebbian learning is a simple, but powerful mechanism to encode the input stimulus in a way that resembles memory formation and associative learning in the brain.

The idea behind the Hebbian learning is that neurons which encode similar information will fire at the same time. That means that the update of the weights of each neuron is proportional to the difference between the input stimuli and the neuron information. This can be expressed with the following rule:

$$w_t = \begin{Bmatrix} W_{t-1} + \varepsilon Y_t Y_r & , if a_s > \bar{a}, \\ W_{t-1} & otherwise, \end{Bmatrix} \tag{4.10}$$

where w_t is the weight update, ε is the learning rate Y_t is the activation of the

current unit, and Y_r the activation of a similar unit, and a the mean activation of the current layer.

4.4 Convolutional Neural Network

A Convolutional Neural Network (CNN) is composed of several layers of convolution and pooling operations stacked together. These two operations simulate the responses of simple and complex cell layers discovered in visual cortex area V1 by Huben and Wiesel [139]. In a CNN, the abstraction of the simple cells is represented by the use of convolution operations, which use local filters to compute high-level features from the input stimuli. The pooling operation creates a higher level of abstraction of the complex cells and increases the spatial invariance of the stimuli by pooling simple cell units of the same receptive field in previous layers.

Every layer of a CNN applies different filters, which increases the capability of the simple cells to extract features. Each filter is trained to extract a different representation of the same receptive field, which generates different outputs, also known as feature maps, for each layer. The complex cells pool units of receptive fields in each feature map. These feature maps are passed to another layer of the network, and because of the complex cells' pooling mechanism, each layer applies a filter in a receptive field which contains the representation of a larger region of the initial stimuli. This means that the first layer will output feature maps which contain representations of one region of the initial stimuli, and deeper layers will represent larger regions. At the end, the output feature map will contain the representation of all stimuli.

Each set of filters in the simple cell layers acts in a receptive field in the input stimuli. The activation of each unit $u_{n,c}^{x,y}$ at (x,y) position of the nth feature map in the cth layer is given by

$$u_{n,c}^{x,y} = max\left(b_{nc} + S, 0\right), \tag{4.11}$$

where $\max(\cdot, 0)$ represents the rectified linear function, which was shown to be more suitable than non-linear functions for training deep neural architectures, as discussed by [108]. b_{nc} is the bias for the nth feature map of the cth layer and S is defined as

$$S = \sum_{m=1}^{M} \sum_{h=1}^{H} \sum_{w=1}^{W} w_{(c-1)m}^{hw} u_{(c-1)m}^{(x+h)(y+w)}, \tag{4.12}$$

where m indexes over the set of filters M in the current layer, c, which is connected to the input stimuli on the previous layer (c-1). The weight of the connection between the unit $u_{n,c}^{x,y}$ and the receptive field with height H and width W of the previous layer $c-1$ is $w_{(c-1)m}^{hw}$. Figure 4.4 illustrates this operation.

A complex cell is connected to a receptive field in the previous simple cell, reducing the dimension of the feature maps. Each complex cell outputs the maximum activation of the receptive field $u(x, y)$ and is defined as:

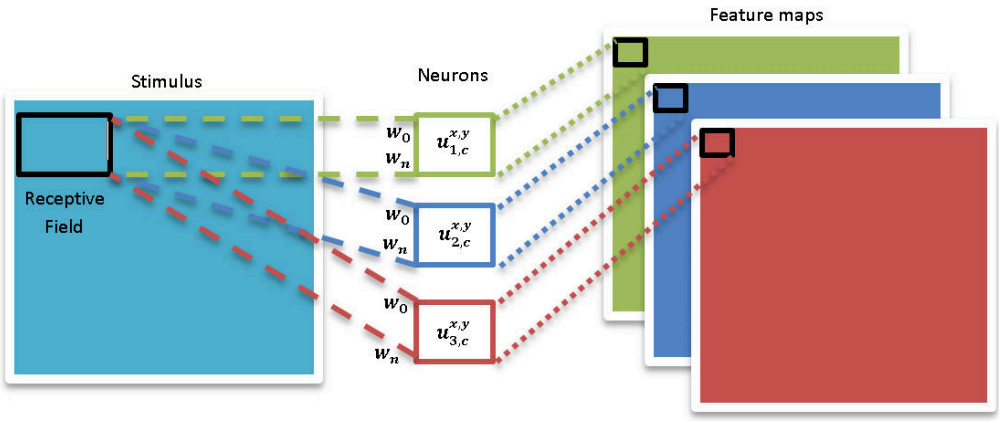

Figure 4.4: Illustration of the convolution process. Each neuron u is connected to a receptive field in the input stimuli by a set of weights w, which represents the filters, and is affected by a bias b, which is the same for all the filters in the same layer. Each filter produces a feature map, composed of several neurons which are then passed to the next layer.

$$a_j = max_{n \times n} \left(u_{n,c}(x, y) \right), \tag{4.13}$$

where $v_{n,c}$ is the output of the simple cell. In this function, a simple cell computes the maximum activation of the receptive field (x, y). The maximum operation down-samples the feature map, maintaining the simple cell structure. Figure 4.5 illustrates this operation.

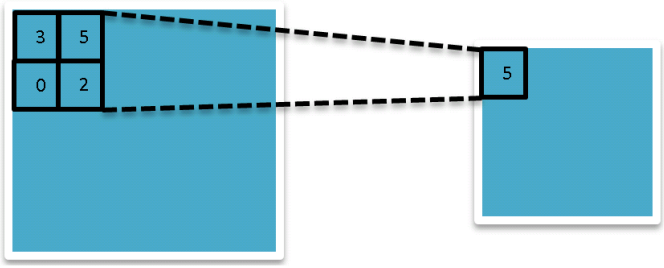

Figure 4.5: Illustration of the pooling process. Each unit of the complex cell is connected to a receptive field of the feature map, and applies a maximum operation, resulting in one activation per receptive field.

4.4.1 Cubic Receptive Fields

In a CNN, each filter is applied to a single instance of the stimuli and extracts features of a certain region. This works well for an individual stimulus, but does not work when a certain sequence dependency is necessary, as in the cases of gestures, speech or even emotion expressions. In this case, if the filter extracts the

same features in each snapshot of the sequence, it will not take into consideration that a hand is moving towards one direction, or that a smile is being displayed.

To introduce sequence dependency cubic receptive fields are used [146]. In a cubic receptive field, the value of each unit(x,y,z) at the nth filter map in the cth layer is defined as:

$$u_{n,c}^{x,y,z} = max(b_{nc} + S_3, 0) \tag{4.14}$$

where $\max(\cdot, 0)$ represents the rectified linear function, b_{cn} is the bias for the nth filter map of the cth layer, and S_3 is defined as

$$S_3 = \sum_{m} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{r=1}^{R} w_{(c-1)m}^{hwr} u_{(m-1)}^{(x+h)(y+w)(z+r)}, \tag{4.15}$$

where m indexes over the set of feature maps in the (c-1) layer connected to the current layer c. The weight of the connection between the unit (h,w,r) and a receptive field connected to the previous layer $(c-1)$ and the filter map m is $w_{(c-1)m}^{hwr}$. H and W are the height and width of the receptive field and z indexes each stimulus; R is the number of stimuli stacked together representing the new dimension of the receptive field.

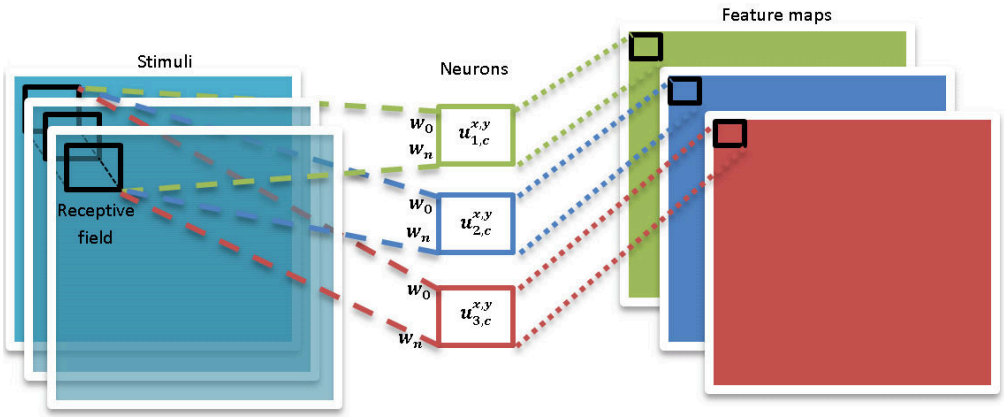

Figure 4.6: Illustration of the cubic convolution process. Different from the common convolution, each neuron u is connected to a receptive field in all of the stimuli at the same time. This way, each neuron has R filters represented by the weights w, where R is the number of input stimuli.

4.4.2 Shunting Inhibition

To learn general features, several layers of simple and complex cells are necessary. Which leads to a large number of parameters to be trained. This, put together with the usual necessity of a large amount of data, which is one of the requirements for the filters to learn general representations, is a big problem shared among deep neural architectures. To reduce the necessity of a deeper network, we introduce

the use of shunting inhibitory fields [99], which improves the efficiency of the filters in learning complex patterns.

Shunting inhibitory neurons are neural-physiological plausible mechanisms that are present in several visual and cognitive functions [113]. When applied to complex cells, shunting neurons can result in filters which are more robust to geometric distortions, meaning that the filters learn more high-level features. Each shunting neuron S_{nc}^{xy} at the position (x,y) of the n^{th} receptive field in the c^{th} layer is activated as:

$$S_{nc}^{xy} = \frac{u_{nc}^{xy}}{a_{nc} + I_{nc}^{xy}} \tag{4.16}$$

where u_{nc}^{xy} is the activation of the common unit in the same position and I_{nc}^{xy} is the activation of the inhibitory neuron. The weights of each inhibitory neuron are trained with backpropagation. A passive decay term, a_{nc}, is a defined parameter and it is the same for the whole shunting inhibitory field. Figure 4.7 illustrates shunting neurons applied to a complex cell layer.

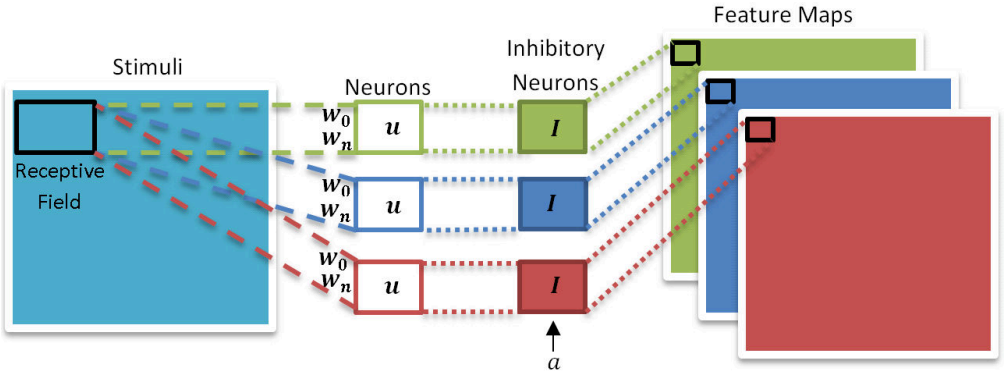

Figure 4.7: Illustration of the shunting inhibitory neuron in complex cells. Each neuron u has an inhibitory neuron, I, attached to it. Each inhibitory neuron has its own set of weights, that connect the inhibitory neuron to the common neuron, and a passive decay a, which is the same for all the neurons in a layer.

The concept behind the shunting neurons is that they will specify the filters of a layer. This creates a problem when applied to filters which extract low-level features, such as edges and contours. When applied to such filters, the shunting neurons specify these filters causing a loss on the generalization aspects of the low-level features. However, when applied to deeper layers, the shunting neurons can enhance the capability of the filters to extract strong high-level representations, which could only be achieved by the use of a deeper network.

4.4.3 Inner Representation Visualization in CNNs

CNNs were successfully used in several domains. However, most of the work with CNNs does not explain why the model is so successful. As CNNs are neural networks that learn a representation of the input data, the knowledge about what

the network learns can help us to understand why these models perform so well in different tasks. The usual method to evaluate the learned representations of neural networks is the observation of the weights matrices, which is not suited for CNNs. Each filter in the convolution layers learns to detect certain patterns in the regions of the input stimuli, and because of the pooling operations, the deeper layers learn patterns which represent a far larger region of the input. That means that the observation of the filters does not give us a reliable way to evaluate the knowledge of the network.

Zeiler and Fergus [306] proposed the deconvolutional process, which helps to visualize the knowledge of a CNN. In their method, they backpropagate the activation of each neuron to an input, which helps to visualize which part of the input the neurons of the network are activated for. This way, we can determine regions of neurons that activated for similar patterns, for example, neurons that activate for the mouth and others for the eyes.

In the deconvolution process, to visualize the activation of a neuron a in layer l (a^l), an input is fed to the network and the signal is forwarded. Afterward, the activation of every neuron in layer l, except for a, is set to zero. After that, each convolution and pooling operation of each layer are reversed. The reverse of the convolution, named filtering, is done by flipping the filters horizontally and vertically. The filtering process can be defined as

$$Sf = \sum_{m=1}^{M} \sum_{h=1}^{H} \sum_{w=1}^{W} wf_{(c-1)m}^{hw} u_{(c-1)m}^{(x+h)(y+w)}, \tag{4.17}$$

where $wf_{(c-1)m}^{hw}$ represents the flipped filters.

The reverse of the pooling operation is known as unpooling. Zeiler and Fergus [306] show that it is necessary to consider the position of the maximum values in order to improve the quality of the visualizations, so these values are stored during the forward-pass. During the unspooling, the values which are not the maximum are set to zero.

Backpropagating the activation of a neuron will cause the reconstruction of the input in a way that only the region which activates this neuron is visualized. Our CNNs use rectified linear units, which means that the neurons which are activated output positive values, and zero represents no activation. That means that in our reconstructed inputs, bright pixels indicate the importance of that specific region for the activation of the neuron.

In a CNN, each filter tends to learn similar patterns, which indicates that those neurons in the same filter will be activated to resembling patterns. Also, each neuron can be activated for very specific patterns, which are not high-level enough for subjective analysis. To improve the quality of our analysis, we apply the concept of creating visualizations for all neurons in one filter, by averaging the activation of each neuron in that filter. That allows us to cluster the knowledge of the network in filters, meaning that we can identify if the network has specialized filters and not specialized neurons. Also, visualizing filters on all layers help us to understand how the network builds the representation and helps us to demonstrate

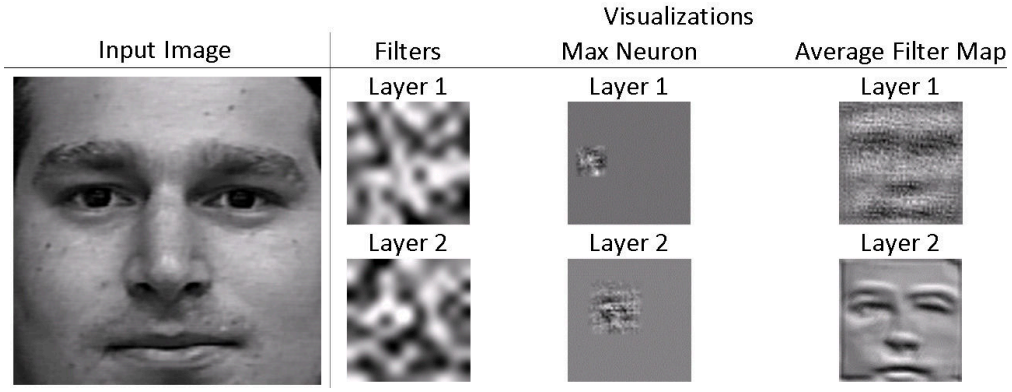

Figure 4.8: For an input image containing a face, it is possible to see the application of different filters. The filters themselves do not contain any information, but by visualizing the activation of the neurons it is possible to identify patterns like eyes and mouth, in the first layer, and a full face representation in the second layer.

the hierarchical capabilities of CNNs.

To demonstrate the advantages of the visualizations, Figure 4.8 illustrates examples of filters, single neurons and mean of filter visualizations. It is possible to see that the filters themselves do not contain any indication about what stimuli structure the neurons activated for. The neuron activation gives us a rough indication of which regions of the image the neuron activated for, but still the concept is too abstract, especially in the first layers. The visualization by mean of neurons in a filter allows us to cluster the representation of similar neurons, and with that, we can see regions of interest. It is possible to see that each layer of the network builds a different representation, starting with very simple edges and contours, going to mouth and eyes and finally a full face representation.

The visualizations are a very powerful tool that helps us to have an important insight about the knowledge learned by the network. With them, we can validate the parameters of our model, understand what the model learns and illustrate the advantages of using concepts such as the inhibitory fields and the cross-channels. We also use the visualizations to illustrate which are the most important features, from the network's perspective, for emotion expression, and how they are combined in different modalities.

4.5 Self-Organizing Maps

Self-Organizing Maps (SOMs) [163] are artificial neural networks which use unsupervised learning techniques to represent the input data in a low-dimensional space. The SOM is usually represented as a grid of neurons which are fully connected with the input data, and through a competitive mechanism, encode the stimuli. The concept behind the SOM is that each neuron encodes a different input representation, and neighbor neurons have a similar encoding. Thus, it is possible to represent the high-dimensional input stimuli using the 2D projection

of the SOM's topology.

The training of the neurons in the SOM rely heavily on the Hebbian learning structure, where neurons which are neighbors fire together and thus should be updated toward the same goal. The neurons in the SOM are fully connected with the input data, so each neuron will be updated to resemble one or a set of input stimuli which are closely related.

The SOM uses a competitive learning mechanism based on the Euclidean Distance. After a random initialization of the neurons, one sample is presented to the network. The distance between the input sample and the neurons is calculated, and the unit which has the smaller distance is selected. This unit is commonly known as the best matching unit (BMU). During the update cycle, only the weights of the BMU and the neurons neighboring this unit are adjusted using the input stimuli as the goal. The update rule for the neuron in the SOM can be expressed as:

$$w_t = w_{t-1} + \theta(b, n)\alpha(x - w_t), \tag{4.18}$$

where w_t is the updated weight of the current unit, w_{t-1} is the current weight, $\theta(b, n)$ is the neighboring function between the current unit, n and the BMU b, α is the learning coefficient, a parameter which decreases during training, and x is the input. This function calculates the distance between the current unit and the BMU, and can be implemented in several different ways. The most common one is to define the function as 1 if the neuron is adjacent to the BMU and 0 if not.

During training, it is possible to see an emerging effect: some neurons will be updated to reflect the distribution of the input data. This effect forms a cluster, where certain regions in the SOM can be associated with certain concepts in the input data distribution. For example, if the SOM is being used to classify human faces, it is possible to cluster male and female faces in different places of the grid. Figure 4.9 illustrates a SOM connected to the input layer, the BMU and the clustering effect which is represented by different colors in the SOM structure.

4.5.1 Growing When Required Networks

The Growing When Required (GWR) [206] networks are an extension of SOMs which does not have the concept of a 2D grid in the neuron's distribution. In a SOM the neurons are constructed in a way that they are disposed of in a 2D structure, where each neuron has a known number of neighbors. The number of neurons and their disposition in the grid is one of the decisions that should be made before starting to build the SOM. That gives the model the capability to reduce the dimensionality of the input data but restricts the network with respect to the amount of information it can learn.

The GWR was proposed to solve the fixed topology problem in the SOM. In this model, neurons are added only when necessary, and without any pre-defined topology. This allows this model to have a growing mechanism, increasing and decreasing the number of neurons, and their positions, when necessary. This makes

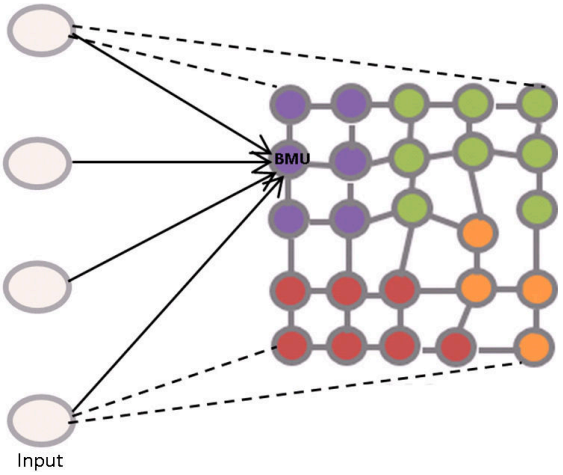

Figure 4.9: Illustration of a SOM. The input is fully connected with all the SOM units, and the best matching unit (BMU) is calculated. The different colors in the grid indicate neurons which encode similar information, introducing the idea of different clusters in the SOM.

the model able to represent data with an arbitrary number of samples and introduces the capability of dealing with novelty detection.

The neurons in the GWR have, besides the weight vector which connects the neuron to the input, the concept of edges. The edges connect the neurons, giving them the concept of neighbors. This allows the model to grow to match the topological distribution of the data, in contrary to the SOM which transforms the data topology into a 2D topology.

Differently from a SOM, the GWR starts only with two neurons, which are created based on two random samples in the input data. Then, as more data is presented to the network, the algorithm decides based on some constraints and the activation behavior of each neuron when and where to add a new neuron. That means that the network can create different clusters of neurons, which represents similar concepts, in different spatial regions. Imagine if you have a network trained with data representing two colors, blue and red, and suddenly a new color, green, is presented. The network will decide to create a new region of neurons to represent the new color. Theoretically, there is no limit to adding new neurons, meaning that the network can learn an arbitrary number of new concepts. The edges maintain the concept of similarity between the nodes, therefore clusters can have connections through edges, exhibiting that these clusters have certain properties in common.

To train this network, the main concept of the SOM training is kept: to find a BMU among the neurons in the model. The difference is that after finding the BMU, the activation of the BMU and its neighbors is calculated. In the GWR the activation can be represented as a function applied to the distance between the neurons and the input. Based on this distance, the network can identify if an input is too far from the knowledge stored in the neurons of the network, if that is the case, a new node is added and an edge between the closer node and the new

node is created.

Each of the nodes in the GWR is also equipped with a function to identify how often it has fired, meaning how often the distance between the neuron and the input was larger than a certain threshold. This mechanism modulates the creation of new nodes by creating a priority in updating neurons which have not been fired in a long time instead of creating new neurons. That gives the network a forgetting mechanism. This mechanism allows the network to forget useless information, that means forget representations which are not important to represent the data.

Together with that, each edge has an associated age that will be used to remove old connections. That means that if a new cluster is created during training, and suddenly is not related to the main neurons anymore, it should be deleted. In the end of each training iteration, the nodes without connections are removed. That makes the model robust against outliers.

The behavior of the GWR when iterating over a training set shows the emergence of concepts. In the first epochs the network will have an exponential grow in the number of neurons, but after achieving a topology that models the data, it will mostly converge. This behavior will change when a new set of training samples is presented to the network. If that new set does not match with some particular region of the network, the model will adapt around the new data distribution, forgetting and removing old neurons when necessary, and creating new ones. That gives the model a similar behavior found in the formation and storage of memory in the brain.

Figure 4.10 illustrates a GWR in two different steps of training. In the left side, Figure 4.10a, it shows the GWR in the second training cycle. In this example, each dot represents a neuron, and the color of the neuron represents an emotional concept. The image on the right, Figure 4.10b shows the model after 100 training cycles are completed. It is possible to see that the network created a non-uniform topology to represent the data. Also, it is possible to see that neurons with similar concepts stay together, creating the idea of clusters. Also, it is possible to identify that some emotional concepts, mostly the black ones, are merged with the others. The black concepts represent the neutral emotions, which are related to all others in this example.

4.6 Corpora

Different corpora were used to evaluate the models proposed in this thesis as each model has its own properties and necessities. All the corpora used are related to emotion expressions, however, differ on the data nature. While the bi-modal FAce and BOdy benchmark database (FABO) has visual expressions, the Surrey Audio-Visual Expressed Emotion (SAVEE) corpus has multimodal audio-visual expressions. To evaluate the models using complex and real-world data, the Emotion-Recognition-In-The-Wild-Challenge dataset is used. This dataset contains excerpts from different movies, and its known to be one of the most challenging corpora available.

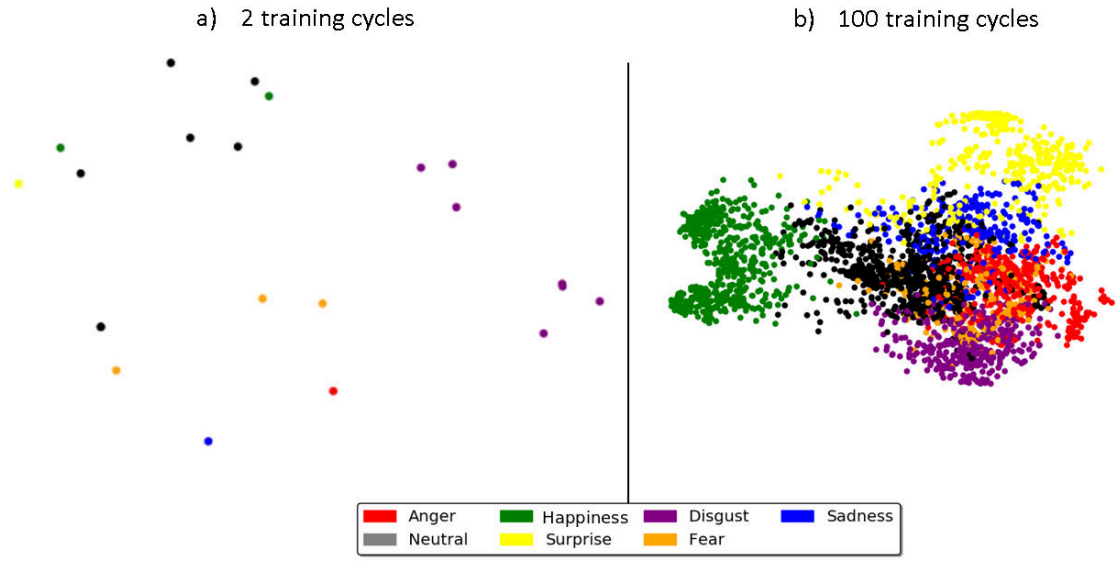

Figure 4.10: The neurons of a GWR during the training on emotional data, showed in the 2 first training cycles (a) and in the 100 first ones(b). Each color represents a different emotional concept associated with that neuron.

To evaluate the attention models proposed in this thesis, we introduce a new emotional attention corpus, based on the FABO dataset. This corpus contains data used for emotional attention tasks, which are not present in the previously mentioned datasets.

Most of the corpora used in training and evaluating affective computing models are related to single emotion expressions, and none contains data of Human-Robot-Interaction (HRI). To evaluate the capability of the models proposed in this thesis to cope with HRI scenarios a novel dataset, named KT Emotional Interaction Corpus, with emotional behavior observations was designed and recorded. This corpus was designed to be challenging in different tasks: spontaneous emotion expressions recognition, emotional attention identification, and emotional behavior analysis.

The following sections detail the FABO and SAVEE corpus, the emotional attention corpus and also the new KT Emotional Interaction Corpus.

4.6.1 Emotion Expression Corpora

The Bi-modal face and body benchmark database

For our experiments we use four corpora. The first one is the bi-modal FAce and BOdy benchmark database (FABO), introduced by Gunes and Piccardi [117]. This corpus is composed of recordings of the upper torso of different subjects performing spontaneous emotion expressions. This corpus contains a total of 11 expressions performed by 23 subjects of different nationalities. Each expression is performed in a spontaneous way, where no indication was given of how the subject must per-

Figure 4.11: Examples of images with an angry expression in the FABO corpus.

Figure 4.12: Faces with an angry expression in the SAVEE corpus.

form the expression. A total of 281 videos were recorded, each one having 2 to 4 of the following expressions: "Anger", "Anxiety", "Boredom", "Disgust", "Fear", "Happiness", "Surprise", "Puzzlement", "Sadness" and "Uncertainty". Each expression starts with a neutral phase, and continues until the apex phase, where the expression is at its peak. We use the neutral phase for each expression to create a 12th "Neutral" class in our experiments. Figure 4.11 illustrates images present in a sequence of an angry expression in the FABO corpus.

Surrey Audio-Visual Expressed Emotion

The second corpus is the Surrey Audio-Visual Expressed Emotion (SAVEE) Database, created by Haq and Jackson [123]. This corpus contains speech recordings from four male native English speakers. Each speaker reads sentences which are clustered into seven different classes: "Anger", "Disgust", "Fear", "Happiness", "Neutral", "Sadness" and "Surprise". These classes represent the six universal emotions with the addition of the "Neutral" class Each speaker recorded 135 utterances, with 30 representing "Neutral" expressions and 15 for each of the other emotions. All the texts are extracted from the TIMIT dataset and are phonetically balanced. Each recording contains the audio and face of the participant, with facial markers. The markers are present to be used for systems that need them, and unfortunately belong to the image. Figure 4.12 illustrates faces of a subject while performing an angry expression in the SAVEE corpus.

Emotion-Recognition-In-the-Wild-Challenge Dataset

The third corpus is the database for the Emotion-Recognition-In-the-Wild-Challenge (EmotiW) [69]. This corpus contains video clips extracted from different movies

Figure 4.13: Example of an angry sequence in the EmotiW corpus.

and organized into seven classes: “Anger”, “Disgust, “Fear”, “Happiness”, “Neutral”, “Sadness” and “Surprise”. A total of 1000 videos with different lengths is available, separated into training and validation sets. The test set is available, but without any label, and includes 700 extra videos. Therefore, we only evaluate our model on the validation set. This challenge is recognized as one of the most difficult tasks for emotion recognition because the movie scenes contain very cluttered environments, occluded faces, speech, music, sound effects, more than one speaker and even animals. Figure 4.13 illustrates some frames of an angry expression in the EmotiW corpus.

4.6.2 Emotional Attention Corpus

To evaluate our model with spontaneous expressions in emotional attention tasks, we adapted a dataset based on the bi-modal FAce and BOdy benchmark database (FABO) corpus introduced by [117]. To build our dataset, we extracted the face expression and body movement of the FABO corpus and located it in a random position in a meeting scene background. The expressions always had the same size, while the positions in the meeting scene were selected using a randomization algorithm. We created a dataset with 1200 expressions composed of different sequences from the FABO corpus using the happy and neutral expressions only. That means that the same expression could be selected more than once, but displayed at a different position in our meeting scene. We created three different categories of sequences: one with only the background, without any expressions, one with a single expression (either neutral or happy), and one with both a neutral and a happy expression. A total of 400 sequences for each category were created. Figure 4.14 illustrates one frame of a sequence of each category.

4.7 KT Emotional Interaction Corpus

The corpora presented previously were focused mostly on human interpretation of emotional concepts. To evaluate our model properly, we created a new corpus based on human-human and human-robot interaction.

The data recording happened in two different scenarios. The first one recorded human-human interactions, and the second one human-robot interactions. We annotate the data using dimensional and categorical representations, similar to

Figure 4.14: Examples of images with no expressions, one expression (happy) and two expressions (happy and neutral), used to train and evaluate our model.

other datasets in the field such as the IEMOCAP [34]. Our corpus differs from the other datasets in the area by introducing a scenario in which the interaction between humans and robots can be evaluated and compared with human-human interaction.

In both scenarios, two subjects conducted a fictitious dialogue based on a certain topic. The subjects were seated at a table across from each other. Images from the subjects faces and torsos, and audio were recorded. In one of the scenarios, both subjects are humans, however, in the second scenario, one of the subjects is replaced by a robot. The whole process of conceptual design, recording procedure, pos-processing and annotation are presented in the next sections.

4.7.1 Recording Setup

To collect the data we used the Robot Interaction Laboratory of the Knowledge Technology Group at University of Hamburg [17]. The subjects were placed inside the half-circle environment. A white canvas covers the whole environment, which means that the participants are separated from the instructors. This way, we assure that the instructors will not bias the recordings, and we let the participants have the dialogue as natural as possible. Figure 4.15 shows the half-circle environment.

Two Logitech HD Pro C920 cameras were placed in a position that it captured the torso of a person seated on the chairs. Each camera recorded a video at a resolution of 1024x768 and a framerate of 30FPS. Each participant had a Bluetooth Sennheiser EZX 80 microphone attached on their shirt, allowing to record an individual audio channel for each participant. Figure 4.16 exhibits an example of the recordings, showing one subject with the microphone attached to his shirt.

4.7.2 Data Collection

Both our scenarios, the human-human and human-robot, had the same basic setup: two subjects are seated across each other on a table and they talk about a topic. Before the experiment started, we explained to both subjects that they would be part of an improvised dialogue, and they would be given a topic to discuss. We did not let the subjects know that our goal was to analyze their emotional

Figure 4.15: Picture of the half-circle environment used for the data recording.

Figure 4.16: Picture of one example of the recordings, showing one subject and the microphone position.

reaction. The data collection happened in into two steps: the instruction step and the recording step.

Instruction

In the instruction step a consent form was given to each subject, which can be found in Appendix A. In the consent form the subject had information about which kind of data would be collected, audio and visual data and that the data would not be correlated to his personal identification. The subject was also asked to identify if

Figure 4.17: Picture of the instruction step of the data collection, where the subjects were informed about the scenarios and the consent form was presented and signed.

the collected data could be used in future experiments, and if it could be published or not. Figure 4.17 shows a picture of the instruction step, where the instructor gives the consent form and directions about the scenario to two participants.

During this step, we collected also some demographic information about each subject. Each subject was asked to range their age into a group age between 0-12 years, 13-18 years, 19-49 years and 50+ years, to state his or her gender, the mother tongue and the city/country of birth. This information helps us to create a distinction between the data, and if to identify gender, age or place of birth influence in the interactions.

All the instruction, recordings, and data labeling happened in English, although persons from different countries participated in all these steps.

Recording

In both scenarios, the Human-Human Interaction (HHI) and the Human-Robot Interaction (HRI), two subjects interacted with each other, however, in the HHI scenario both are humans and in the HRI one human is replaced by a robot. We invited different students from the informatics campus of the University of Hamburg, using an open call distributed via email to different student lists.

For both scenarios, we created two roles: an active and a passive subject. Before initiating each dialogue session, we gave to the active subject a topic, and he or she should introduce it during the dialogue. The passive subject was not aware of the topic of the dialogue, and both subjects should improvise. The subjects were free to perform the dialogue as they wanted, with the only restriction of not standing up nor changing places. The following topics were available:

- Lottery: Tell the other person he or she won the lottery.
- Food: Introduce to the other person a very disgusting food.
- School: Tell the other person that his/her school records are gone.
- Pet: Tell the other person that his/her pet died.
- Family: Tell the other person that a family member of him/her is in the hospital.

These topics were selected in a way to provoke interactions related to at least one of the universal expressions each: "Happiness", "Disgust","Anger", "Fear", and "Sadness". To none of the subjects any information was given about the nature of the analyses, to not bias their expressions.

In the HHI scenario, both participants were humans. One of them was chosen to be the first active subject, and the topic was presented only to him. The topics were printed on a paper and shown to the active person, in a way that all the active persons received the same instruction. Figure 4.18 shows the topic assignment moment for an active participant.

After each dialogue session, the role of the active subject was given to the previously passive subject and a new topic was assigned. For each pair of participants, a total of five dialogue sessions was recorded, one for each topic, and each one lasting between 30 seconds and 2 minutes. Although the topics were chosen to provoke different expressions, it was the case that in some dialogues none of the previously mentioned emotional concepts were expressed.

The HRI scenario followed the same pattern that the HHI scenario had but replaced the active subject with a humanoid iCub robot [210] head. In this scenario, the robot was always the active subject. As in the previous scenario, the passive subject was not informed of the theme of the dialogue and had to improvise a discussion.

The iCub head has three degrees of freedom for head movement and 21 different eyebrows and mouth positions, for face expressions. Figure 4.19 illustrates the iCub robot used in our experiments. We captured the pictures from the iCub perspective placing a camera just in front of it. The iCub has a built-in camera, however, the resolution and frame rate are very low and incompatible with the ones necessary for our scenario.

The robot was remote controlled, which means that movements, face expressions and what the robot spoke were typed by an instructor. The participants

Figure 4.18: Picture of the topic assignment. One of the participants is chosen as the active subject, and one of the five topics is given to him/her.

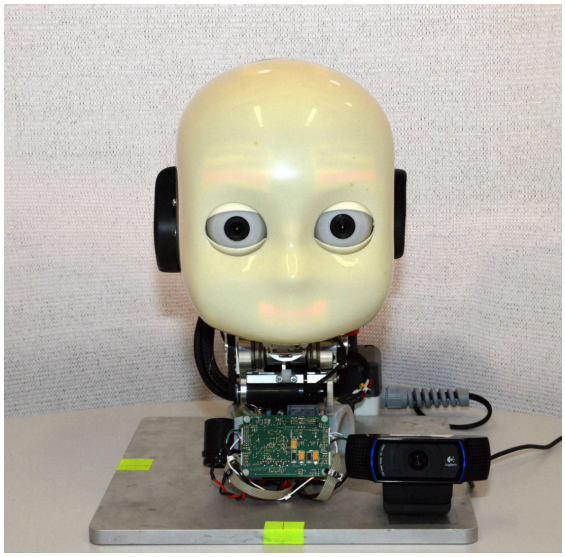

Figure 4.19: Picture of the iCub robot used in our experiments.

were not informed about that, and the instructor controlled the robot from another room. This way, only the robot and the participant were in the room during the recordings. Each dialogue session did not take more than 4 minutes, taking longer than those in the first scenario, mostly due to the delay in the robot's response. Figure 4.20 illustrates the HRI scenario.

Figure 4.20: Picture of the HRI scenario. A person is seated in front of the iCub robot, which is the active subject.

4.7.3 Data Annotation

To give the corpus a human ground truth analysis, we needed to label each interaction in a way that we can compare them to our computational models. To do that, we selected a dimensional [261] and a categorical [81] approach to describe each interaction. Instead of using emotional concepts, such as happiness or sadness, a dimensional model allows us to describe the interaction itself. This allows us to identify interactions which were similar but presented in different emotional situations, such as smiling and moving the arms at the same time.

Using a dimensional model, allows us to better deal with spontaneous expressions which were present in our recorded data. The dimensional model also allows us to identify the core affect [262] of the interaction which makes it possible to create a consistent and valid measure of an interaction [12], reducing the contextual bias of the analyzer, such as current mood, environment or cultural differences.

We choose the dimensional model proposed by Russel et al. [262], and updated by Rubin et al. [258] with the addition of dominance, besides the usual arousal and valence. Dominance gives the intensity attribution to the expression itself, differently from arousal which gives us the intensity of the person's behavior. According to Rubin et al., arousal evaluates how the person behaved, and dominance relates to the strength of the expression, not the person. By labeling each interaction using these three dimensions (arousal, valence, and dominance), we can identify more fine-tuned information about the expressions.

Besides the three dimensions, we decided to include a categorical value for each interaction. We use the six universal emotions to identify the expression, and although this information can be very difficult to identify in our spontaneous interactions, it was helpful to situate the annotator into an emotional concept. Also,

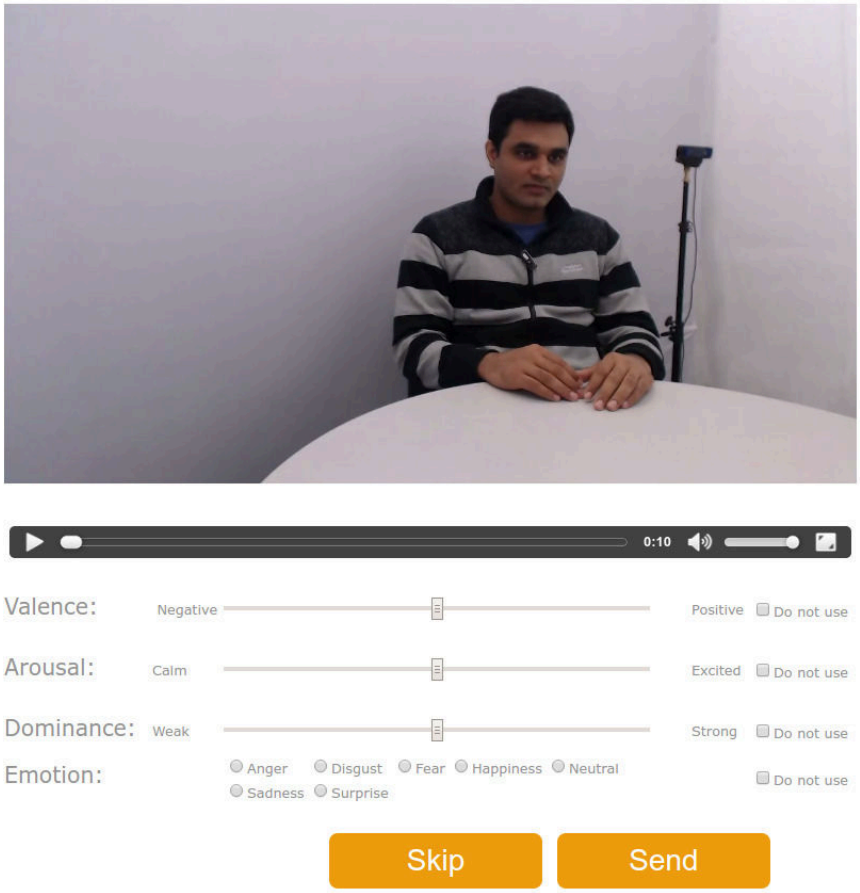

Figure 4.21: Picture of the labeling collection framework.

by collecting a categorical concept, we can use it to identify different emotional categories within our three-dimensional attributes.

Data Annotation Framework

To avoid biased labels, we use anonymous annotators. For that, we developed an online framework, which allows annotators to watch the interactions and give the labels. We let our annotation framework online, and distributed the access link to student and professionals in the informatics campus of the University of Hamburg. When the annotators access the framework, a first page is shown with instructions about the task. Then, after agreeing with the instructions, the annotator is sent to the labeling page. This pages shows a video and asks the annotator to evaluate the interaction using an interactive slider to range each attribute between two extreme concepts. For valence, the limits were negative and positive, arousal had limits ranging from calm to excited, and dominance from weak to strong. At the end, an emotional concept was asked. Figure 4.21 illustrates the labeling session.

As each interaction was based on a whole dialogue session, it could happen that several different expressions were present. To make possible that each annotator

watched and evaluated different expressions, we split each dialogue session into videos with 10s. This generated a problem: parts of the video which were temporally close to each other could have very different labels. We solved that by using many annotators to evaluate the same video parts, creating an averaged measure for each attribute.

To make the transitions between different video parts better, we introduce the use of a Gaussian distribution to represent each video label. Each attribute was evaluated in a range of two concepts, where 0 represents the minimal extreme and 100 the maximum. For example, on this scale, 0 represents the most negative valence and 100 the most positive. We then proceed to create a Gaussian distribution with the chosen value as mean and the standard deviation of all the annotations for that particular video clip as distribution spread.

The use of Gaussian distributions as labels allow us to create a robust label for the videos, removing outliers. As all the video clips could have very different annotators, the idea of using a distribution gives us a better view on the expression on each clip, instead of analyzing each annotation individually. Also, by using a distribution, we can smooth the transitions of each video clip better, and represent the real expression transitions with more naturality.

4.7.4 Recorded Data

The data recording happened in two different days: one for the HHI scenario and other for the HRI scenario. For each scenario, we recorded the image and audio, in a video format, for each of the subjects.

The HHI scenario had a total of 7 sessions, with 14 different subjects, two participating in each session. Each session had 6 dialogues, one per topic and an extra one where the two subjects introduced each other, using a fake name. Each subject only participated in one session, meaning that no subject repeated the experiment. A total of 84 videos were recorded, one for each subject in each dialogue, with a sum of 1h05min of recordings. Each video had a different length, with the longer one having 2 minutes and 8 seconds and the shorter one with 30 seconds.

The HRI scenario had one session more than the HHI, totaling 9 sessions, and each had one different subject. In the HRI scenario, each session had 5 dialogues, one per topic, without the introduction dialogue. A total of 45 videos were recorded, summing 2h05min of videos. As happened with the HHI, each video has a different length and the longer one had 3min40s and the shorter one with 1min30s. It is possible to see already, that the interactions with the robot produced longer dialogues than the only-human ones. That was expected, giving that the robot has a longer reaction time than humans. Table 4.1 summarize the number and duration of the videos in each scenario.

Table 4.1: Number and duration of videos for each scenario experiment.

Scenario	Videos	Subjects	Total Duration
HHI	84	14	1h05min
HRI	45	9	2h05min

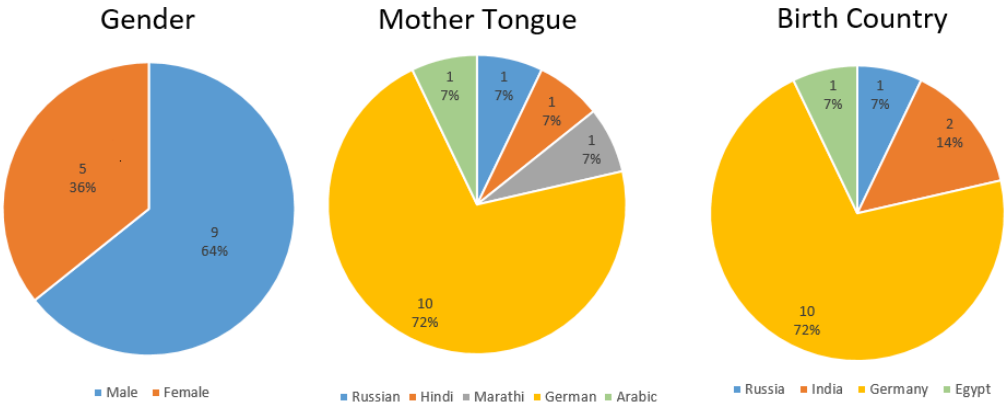

Figure 4.22: Demographic data summary for the HHI scenario, showing the total of subjects and the associated percentage for gender, mother tongue and birth country distributions.

Demographic data

The demographic data shows us how the subjects are clustered with respect to their gender, mother tongue and country of birth. All of our subjects were in the age group of 19-49, so we did not have any variation. However, for the other three characteristics we obtained different measures.

For the HHI scenario, most of the subjects, 64 %, were male and only 36% female. The subjects were from three different countries, and had five different mother languages. The majority of the subjects came from Germany (72%), but participants from Russia, India and Egypt were also present. Similarly, most of the participants had German as a mother tongue (also 72 %). Figure 4.22 illustrates a summary of the demographic data for the HHI scenario. This shows that our dataset has persons with different cultural background, and with different mother tongue. As all the experiments were performed in English, this information could help us to identify different behaviors which could be related to the subjects own cultural background.

Differently from the HHI, most of the participants of the HRI scenario were female (67%). Again, most of the participants came from Germany and had German as mother tongue (78% each characteristic), however, different countries were present as well, with Malaysian and Chinese as a mother tongue. Figure 4.23 illustrates a summary of the demographic data for the HRI scenario.

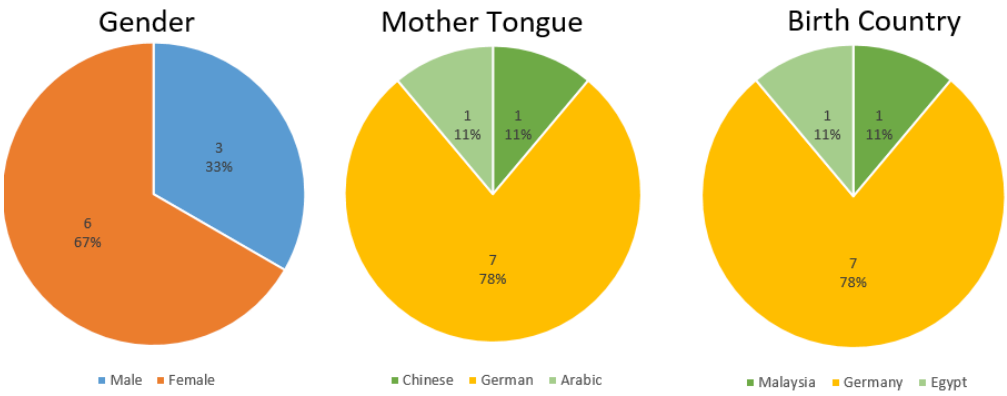

Figure 4.23: Demographic data summary for the HRI scenario, showing the total of subjects and the associated percentage for gender, mother tongue and birth country distributions.

4.7.5 Data Analysis

The annotation framework was made available for two weeks. A total of 39 annotators contributed. Each annotator labeled an average of 3 whole clips, and we obtained a total of total of 2365 annotations over all the 10s video clips. The annotations show us some interesting aspects of the dataset, and we clustered the analysis into three categories: general, per topic and per subject. The general analysis shows us how the data is distributed across the dataset and which information we can have about recordings as a whole. The analysis per topic shows us how each topic was perceived by each subject pair, giving us indications about how the subjects behaved on the different topics. Finally, the per subject analysis gives us individual information on how a certain subject performed during the whole dialogue sessions, showing how different subjects react to same scenarios.

We also cluster our analysis into the two scenarios, to gives us the possibility of understanding how the subjects performed when interacting with a human or with a robot. With these analyses, we intend to quantify the difference between human-human- and human-robot-interaction.

Analyzing the general distributions for the whole corpus gave us a perspective of what was expressed and an indication of what expressions are present in the interactions. Also, comparing the HHI and HRI scenarios, gave us a general indication on how people behaved. Figure 4.24 shows the histogram for the valence, arousal, dominance and the emotional labels for all the annotations in both scenarios. It is possible to see that the annotations for all of these dimensions are normally distributed, showing a strong indication that most of the interactions were not so close to the extremes. The exception is the lower extreme, which always showed a larger amount of data. That means that many of the interactions were evaluated as negative (valence), calm (arousal) and weak (dominance). The emotional concepts indicate a similar effect, where the neural expressions were mostly present, followed by angry expressions, which can explain the number of negative valences.

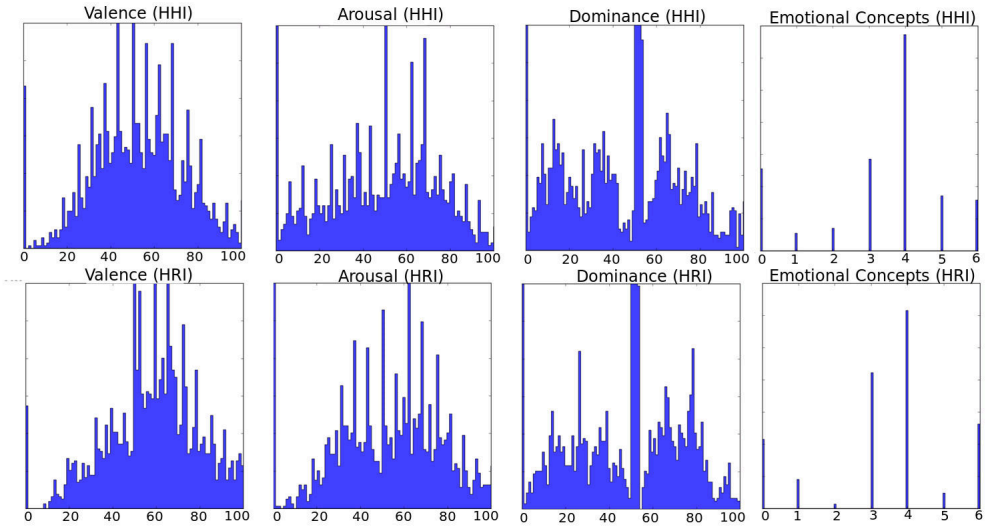

Figure 4.24: These plots show the histogram of the annotations for all the dataset. For the emotional concepts histogram, the x axis represents the following emotions: 0 - Anger, 1 -Disgust, 2 - Fear, 3 - Happiness, 4 - Neutral, 5 - Sadness and 6 - Surprise.

It is also possible to see that for both scenarios the distribution of the labels have some important differences: the valence of the HRI scenario is more distributed to the right when compared to the one in the HHI scenario, indicating that the subjects tend to be more positive with the robot than with a human. The arousal also shows that in the robot scenario the subjects tend to be less calm, the same for the dominance. It is also possible to see that there were more "Happiness" annotations and less "Sadness" ones in the HRI scenario than in the HHI scenario.

Dominance and arousal have a similar behavior in the histogram. To show how they are correlated we calculated the Pearson correlation coefficient [176], which measures the linear relationship between two series. It takes a value between -1 and 1, where -1 indicates an inverse correlation, 1 a direct correlation and 0 no correlation. The coefficient for dominance and arousal for the HHI scenario is 0.7, and 0.72 for the HRI scenario showing that for both scenarios there is a high direct correlation. These values are similar to other datasets [34] and indicate that arousal and dominance are influenced by each other.

The analysis per topic gives shows how the chosen topics produced different interactions. Figure 4.25 illustrates two topics from both scenarios: lottery and food. It is possible to see how the annotations differ for each topic, showing that in the lottery videos a lot of high valences is presented, while in the food videos the data presented mostly high arousal. The dominance is rather small when the arousal is also small for both videos. Comparing the difference between the HHI and HRI scenarios, it is possible to see that the HRI scenario presented more negative valence than the HHI scenario.

It is possible to see also how some emotional concepts are present in each scenario. While in the food scenario, many "Disgust" annotations are present, in

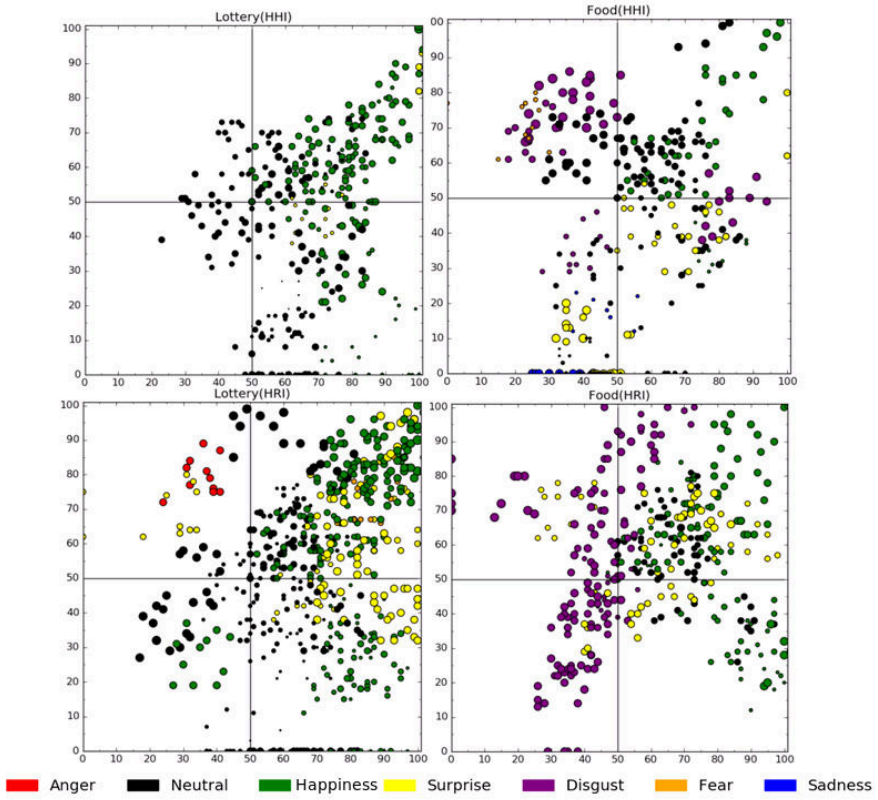

Figure 4.25: This plot shows the spread of the annotations for the dataset separated per topic. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

the food scenario the interactions are labeled mostly as "Happiness" or "Surprise". It is also possible to see that for the HRI scenario, some persons behaved with "Angry" in the lottery topic and that most of the "Surprise" annotations in the food scenario have higher arousal in the HRI scenario than in the HHI one.

To provide the analysis with an inter-rater reliability measure, we calculated the interclass correlation coefficient [160] for each topic. This coefficient gives a value between 0 and 1, where 1 indicates that the correlation is excellent, meaning that most of the annotators agree, and 0 means poor agreement. This measure is commonly used for other emotion assessment scenarios [33, 21, 29] and presents an unbiased measure of agreement. Table 4.2 exhibits the coefficients per topic for the HHI scenario. It is possible to see that the lottery scenario produced a better agreement in most cases, and the food scenario the worst one. Also, the dominance variable was the one with the lowest agreement coefficients, while the emotional concepts had the highest.

Differently from the HHI scenario, the interclass coefficient for the HHI scenario shows a higher agreement of the annotators. Although dominance still shows a lower agreement rate, valence and arousal present a higher one.

Table 4.2: Interclass correlation coefficient per topic in the HHI scenario.

Characteristic	Lottery	Food	School	Family	Pet
Valence	0.7	0.5	0.3	0.6	0.4
Arousal	0.5	0.6	0.6	0.5	0.4
Dominance	0.4	0.5	0.4	0.5	0.4
Emotion Concept	0.7	0.6	0.5	0.6	0.5

Table 4.3: Interclass correlation coefficient per topic in the HRI scenario.

Characteristic	Lottery	Food	School	Family	Pet
Valence	0.7	0.6	0.4	0.5	0.6
Arousal	0.6	0.6	0.6	0.4	0.5
Dominance	0.5	0.5	0.5	0.4	0.5
Emotion Concept	0.6	0.7	0.5	0.5	0.5

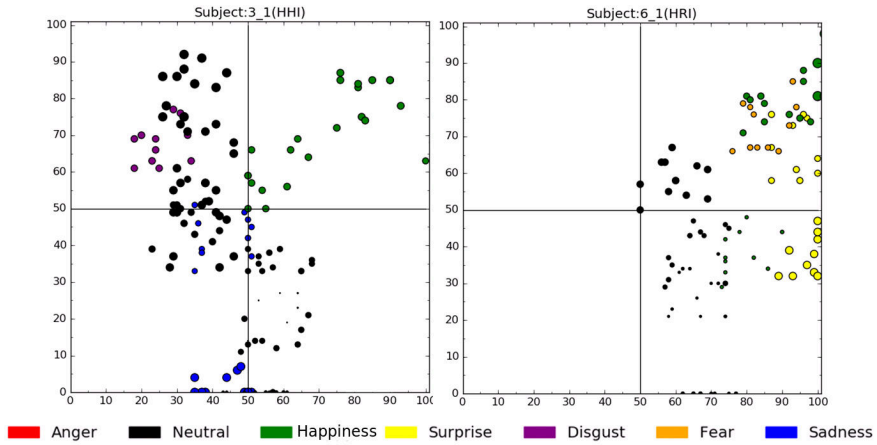

Figure 4.26: These plots show two examples of the spread of the annotations for the dataset separated per subjects. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

Analyzing the subjects, it is possible to see how they behave during the whole recording session. Figure 4.26 exhibits the behavior of two subjects per scenario. In the image it is possible to see that one subject from the HHI scenario presented mostly high arousal expressions, and highly more dominant ones. Also, the expressions were mostly with a negative valence, although annotated as neutral. This subject did not express any fear nor surprise expression during the five topics.

Subject 6_1 from the HRI scenario showed mostly positive expressions, with a high incidence of surprise and fear expressions. However, the dominance of this

subject was very weak mostly, although the arousal was more excited than calm. That probably means that this subject was nervous while talking to the robot, but could not present strong or intense expressions during most of the interaction.

The collected dataset contains a large amount of information, and many more interesting statistics could be provided. However, the ones that are necessary for the development and evaluation of the models presented in this thesis are the ones discussed in this section.

4.8 Summary

This chapter presented important concepts used by the models of this thesis and detailed the datasets used for the evaluation of such models. The neural network concepts described here are the basis to understand the proposed techniques and solutions, although what was discussed was only the fundamental concepts. Detailed views on network behavior, extensions, and possible limitations are discussed in the next chapters, which present our models.

The corpora used for our evaluations have different properties, which were necessary to evaluate the behavior of our model in different tasks: multimodal and spontaneous expression representation and recognition, and emotional concept learning. Although many corpora are available, none of them had an important characteristic: continuous scenarios representing different human-human and human-robot interactions. To address this problem, we proposed a new corpus. The acquisition, recording, labeling and analysis of this novel corpus was presented in this chapter and serves as the basis for comparison of our integrative emotional, behavioral analysis model, presented in Chapter 7.

Chapter 5

Emotion Perception with a Cross-channel Convolution Neural Network

5.1 Introduction

The first problem this thesis deals with is the multimodal emotion expression representation. As seen in Chapter 3, emotion expressions are very complex to be represented by computational models, in particular multimodal expressions. We propose here the use of a visual-cortex-inspired deep learning model to learn how to represent multimodal expressions.

Our model deals with multimodal stimuli, and takes into consideration visual, primary face expressions and body movements, and auditory information. It is implemented as a Cross-channel Convolutional Neural Network (CCCNN) and it extracts hierarchical features from the two modalities. The complex representation varies depending on the presented stimuli, and each hierarchical layer of the network learns a different level of abstraction. This means that the deeper layers will have a full representation of the input, while the first layers will have a local representation of some regions or parts of the input stimuli.

To be able to deal with sequences, cubic receptive fields are implemented [146] expanding the capability of the model into modeling dependencies between sequential information. Also, the use of shunting inhibitory fields [99], allows us to ensure a strong feature representation.

The proposed model is able to learn simple and complex features and to model the dependencies of these features in a sequence. Using a multichannel implementation, it is possible to learn different features of each stimulus. We use the concept of cross-channel learning to deal with differences within the modalities. This allows us to have regions of the network specified to learn features from face expressions and body movements, for example, but the final representation integrates both specific features.

To evaluate our model, we use different corpora of emotional expressions with

different modalities. We also introduce a study on the model's behavior, and a visualization analysis on the learned representation. Finally, we discuss the model behavior and architecture and its capability to learn multimodal spontaneous expressions.

5.2 Cross-channel Convolution Neural Network

To be able to deal with multimodal data, our network uses the concept of the Cross-channel Convolutional Neural Network (CCCNN). In the CCCNN architecture, several channels, each one composed of an independent sequence of convolution and pooling layers, are fully connected at the end to a Cross-channel layer, and trained as one single architecture. Our architecture is composed of two main streams: a visual and an auditory stream.

Our Cross-channel is inspired by the V4 area of the brain's visual pathway [91]. In this area, the representations obtained by the ventral and dorsal areas are integrated. In our case, we implement a Cross-channel layer which is composed of a layer of simple and complex cells. This layer receives the high-level representation of two different channels as input and integrates them.

Our model applies topological convolution, and because of this the size of the receptive field has an important impact on the learning process. The receptive fields in our Cross-channel should be large enough to be able to capture the whole concept of the stimulus, and not only part of it. With a small receptive field, our cross learning will not be able to capture the high-level features.

We apply our Cross-channel learning in two streams. Goodale and Milner [112] describe how the visual cortex is categorized into two streams, and how these streams are integrated into the V4 area. In their model, the ventral and dorsal streams extract different kinds of information from the input and are used as input to the V4 area. Hickok [131] describes a similar process occurring in the auditory pathway, where different kinds of information are processed by the ventral and dorsal stream, and integrated into the V4 area. Although we are not modeling the same pathways and information exactly as the ones present in the brain, the architecture of our model was developed in a way that resembles the brain's organizational structure. Also, we specify our model to deal with emotion expressions, and not general visual and auditory recognition.

5.2.1 Visual Representation

Inspired by the primate visual cortex model [91], the proposed model visual stream has two channels. The first channel is responsible for learning and extracting information about facial expressions, which comprises contour, shape and texture of a face, and mimics the encoding of information in the ventral area of the primate visual cortex. The second channel codes information about the orientation, direction and speed of changes within the torso of a person in a sequence of images, similar to the information coded by the dorsal area.

Although facial expressions are an important part of emotion expression determination, there is evidence that shows that in some cases facial expressions and body posture and movements are contradictory and carry a different meaning. This phenomenon was first observed by Darwin [60] and is referred to as micro expressions. Although micro expressions occur with other modalities as well, the face is the one in which this behavior is easily perceptible [243].

Ekman [83] demonstrates that facial micro expressions last from 40 to 300 milliseconds, and are composed of an involuntary pattern of the face, sometimes not directly related to the expression the person intended to perform. He also shows that micro expressions are too brief to convey an emotion, but usually are signs of concealed behaviors, giving the expression a different meaning. For example, facial micro expressions are usually the way to determine weather someone is angry while using a happy sarcastic expression. In this case, the addition of facial micro expressions as an observable modality can enhance the capability of the model to distinguish spontaneous expressions, but the observation of the facial micro expression alone does not carry any meaning [231].

Our architecture is tuned to deal with facial expressions and micro expressions. Our architecture is fed with frames comprising 1s, and our Face channel receives a smaller sequence representing 300 milliseconds. These intervals were found by experimenting with using different sequence lengths as input to the network, and the chosen values are congruent with evidence from [83]. That means that our network is able to recognize common face expressions, but also takes into consideration micro expressions.

To feed our visual stream, we must first find the faces on the images. To do so, we use the Viola-Jones face detection algorithm [291], which uses an Adaboost-based detection. Wang [297] discusses the Viola-Jones algorithm, and shows that it is robust and effective when applied to general face detection datasets. In our experiments, the Viola-Jones algorithm showed to be reliable in controlled environments. After finding the face, we create a bounding box to describe the torso movement. We extract face and torso from a sequence of frames corresponding to 1 second and use them as input to the network.

To define the input of the Movement channel, we use a motion representation. Feeding this stream with this representation, and not the whole image, allows us to specialize the channel into learning motion descriptors. This way, we can train the network with a smaller amount of data, and use a shallow network to obtain high-level descriptors.

To obtain the motion representation, an additional layer is used for pre-processing the input of the Movement channel. This layer receives 10 gray scale frames and creates a representation based on the difference between each pair of frames. This approach was used in previous works to learn gestures, and showed to be successful by [14]. The layer computes an absolute difference and sums up the resulting frame to a stack of frames. This operation generates one image representing the motion of the sequence, defined here as M:

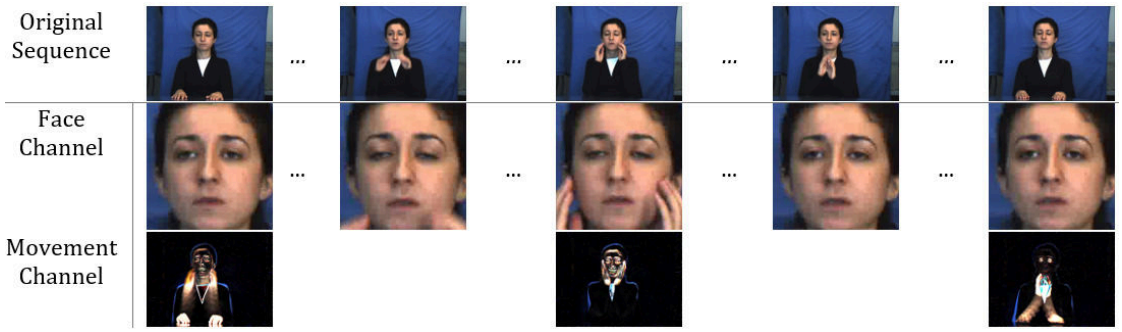

Figure 5.1: Example of input for the visual stream. We feed the network with 1s of expressions, which are processed into 3 movement frames and 9 facial expressions.

$$M = \sum_{i=1}^{N} |(F_{i-1} - F_i)|\,(i/t), \tag{5.1}$$

where N is the number of frames, F_i represents the current frame and (i/t) represents the weighted shadow. The weighted shadow is used to create different gray scale shadows in the final representation according to the time that each frame is presented. This means that every frame of the image will have a different gray tone in the final image. The weight t starts as 0 in the first frame and is increased over time, so each frame has a different weight. The absolute difference of each pair of frames removes non-changing parts of the image, being able to extract the background or any other detail in the image that is not important to the motion representation. By summing up all the absolute differences of each pair of images it is possible to create a shape representation of the motion and with the help of the weighted shadows, the information of when each single posture happened. Figure 5.1 displays a common input of our visual stream, containing examples of the Face and Movement channels.

The Face channel is composed of two convolution and pooling layers. The first convolution layer implements 5 filters with cubic receptive fields, each one with a dimension of 5x5x3. The second layer implements 5 filter maps, also with a dimension of 5x5, and a shunting inhibitory field. Both layers implement max-pooling operators with a receptive field of 2x2. In our experiments, we use a rate of 30 frames per second, which means that the 300 milliseconds are represented by 9 frames. Each frame is resized to 50x50 pixels.

The Movement channel implements three convolution and pooling layers. The first convolution layer implements 5 filters with cubic receptive fields, each one with a dimension of 5x5x3. The second and third channels implement 5 filters, each one with a dimension of 5x5 and all channels implement max-pooling with a receptive field of 2x2. We feed this channel with 1s of expressions, meaning that we feed the network with 30 frames. We compute the motion representation for every 10 frames, meaning that we feed the Movement channel with 3 motion representations. All the images are resized to 128x96 pixels.

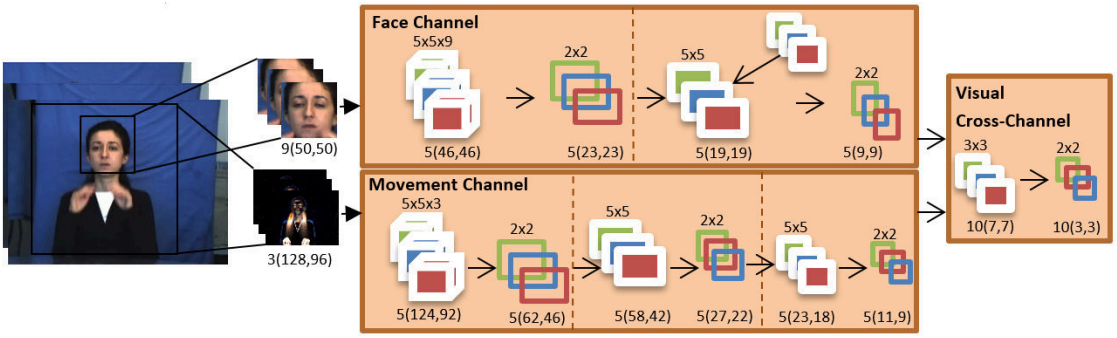

Figure 5.2: The Visual stream of our network is composed of two channels: the Face and the Movement channels. The Face channel implements two layers, each one with convolution and pooling, and applies inhibitory fields in the second layer, while the Movement channel implements three layers, with pooling and convolution. Both channels implement cubic receptive fields in the first layer. The final output of each channel is fed to a Cross-channel which implements convolution and pooling and produces a final visual representation.

We apply a Cross-channel to the visual stream. This Cross-channel receives as input the Face and Movement channels, and it is composed of one convolution channel with 10 filters, each one with a dimension of 3x3, and one max-pooling with a receptive field of 2x2. We have to ensure that the input of the Cross-channel has the same dimension, to do so we resize the output representation of the Movement channel to 9x9, the same as the Face channel. Figure 5.2 illustrates the visual stream of the network.

5.2.2 Auditory Representation

The dorsal and ventral streams of the brain process different auditory information [131]. While the ventral stream deals with speech information, the dorsal one maps auditory sensory representation. In earlier stages of the dorsal stream, the auditory information is decomposed into a series of representations, which are not connected to phonetic representations. We use this concept to separate the perception of auditory information in our network into two channels. One deals mostly with speech signals, and the other with general sounds, including music.

Evidence in the work of [264] shows that the use of Mel-Cepstral Coefficients (MFCC) is suited for speech representation, but does not provide much information when describing music. MFCCs are described as the coefficients derived from the cepstral representation of an audio sequence, which converts the power spectrum of an audio clip into the Mel-scale frequency. The Mel scale showed to be closer to human auditory system's response than the linear frequency [238].

When trying to describe general music information, spectral representations, such as power spectrograms, showed good results [104]. Power spectrograms are calculated in smaller sequences of audio clips, by applying a discrete Fourier transform in each clip. This operation describes the distribution of frequency compo-

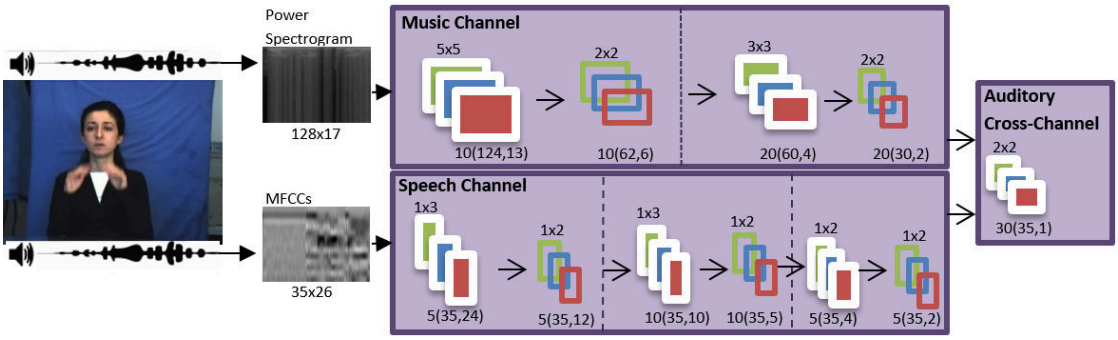

Figure 5.3: The Auditory stream of our network implements two channels: the Music channel and the Speech channel, which implements filters with one dimension. We feed the network with 1s audio clips, and calculate a power spectrogram as input for the Music channel and MFCCs as input for the Speech channel. The output of both channels is used as input for the auditory Cross-channel.

nents on each clip.

To use the auditory representation in CNNs, the MFCCs and power spectrograms are represented as images. But there is a fundamental difference when dealing with these representations. Usually, the input of CNNs is processed by a filter matrix which is applied in both, height and width axis. The filter is trained to capture local information of the region where it is applied. When this concept is applied to an auditory representation, to learn from a 2D region can generate a problem. In auditory input, each axis represents different kinds of information, where usually the X axis represents time and the Y axis the spectral representation. For the power spectrogram representations, the use of 2D filters showed to be ideal, because each filter captures the spectral representation in a certain region of the audio clip [128].

On the MFFCs representation, the use of 2D filters does not work. To Extract the MFCCs, a cosine transformation is applied and this projects each value of the Y axis into the Mel frequency space, which may not preserve locality. Because of the topological nature of 2D filters, the network will try to learn patterns in adjacent regions, which are not represented adjacently in the Mel frequency domain. Abdel-Hamid et al. [1] propose the use of 1D filters to solve this problem. The convolution process is the same, but the network applies 1D filters on each value of the Y axis of the image. This means that the filters will learn how to correlate the representation per axis and not within neighbors. Pooling is also applied in one dimension, always keeping the same topological structure.

We build our auditory stream based on the speech and music representation. We use two channels, which are connected to a Cross-channel. We use audio clips with 1s as input, and each clip is re-sampled to 16000 Hz. We compute the power spectrum and the MFCC of the audio clip and feed them to the two channels. The power spectrogram is the input of the Music channel, and it is computed over a window of 25ms with a slide of 10ms. The frequency resolution is 2048. This generates a spectrogram with 1024 bins, each one with 136 descriptors. We resize

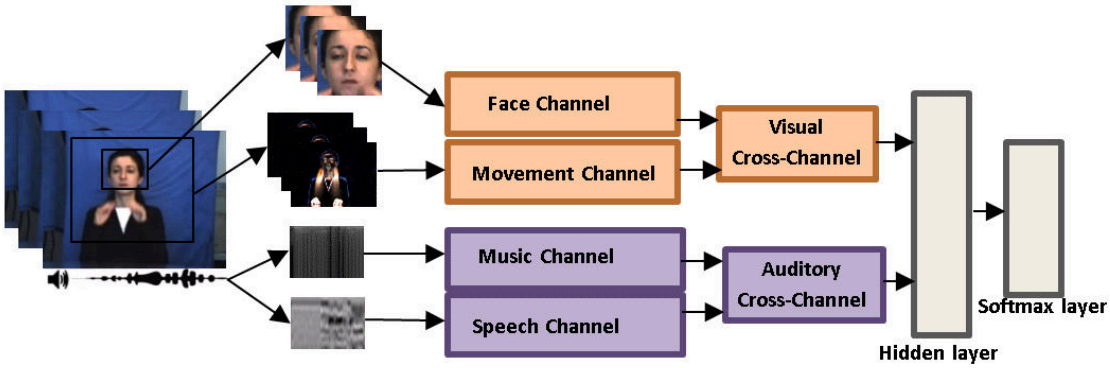

Figure 5.4: Final crossmodal architecture, which extracts features from the visual and the auditory input and classifies them in emotion expressions. We connect the outputs of each stream to a fully connected hidden layer and then to a softmax layer, which will give us a classification probability.

the spectrogram by a factor of 8, resulting in an input size of 128x7. The MFCC is used as input for the Speech channel, and it is calculated over the same window and slide as the power spectrogram. We change the frequency resolution to 1024, which generated a representation with 35 bins each one with 26 descriptors.

The Music channel is composed of two layers, the first one with 10 filters, and each one with a dimension of 5x5. The second layer has 20 filters, with a dimension of 3x3. Both layers implement pooling, with a receptive field of 2x2. The Speech channel is composed of three layers, each one with one-dimensional filters. The first has 5 filters, with a dimension of 1x3, the second one has 10 filters with a dimension of 1x3 and the third one 20 filters with a dimension of 1x2. All three layers apply pooling with a receptive field of 1x2.

The Cross-channel applied to our Auditory stream has one layer, with 30 filters, each one with a dimension of 2x2, without the application of pooling. To be able to use the Cross-channel, both channels must output data with the same dimensions and our results showed that resizing the Music channel output produced better performance. This can be explained by the fact that the Speech channel depends strongly on the non-locality of the features. Figure 5.3 illustrates our Auditory stream.

5.2.3 Crossmodal Representation

We integrate both streams into one Multichannel Convolutional Neural Network architecture. We connect each Cross-channel with a fully connected hidden layer, with 500 units, which is then connected to a softmax layer. This way, each modality, visual and auditory, has its own high abstraction level representation preserved. Figure 5.4 illustrates our final architecture.

It is possible to train the filters of a CNN using either a supervised [136] or an unsupervised approach [247]. Although the unsupervised approach does not depend on strongly labeled data, evidence [150, 296], showed that the use of su-

pervised training improved the effectiveness of CNNs. Also, the use of supervised training allows us to train the network with a smaller amount of data, which would not be possible when using unsupervised training.

Erhan et al. [88] show evidence that the use of pre-training steps improves the capability of the filters to be tuned faster in a specific domain. We follow this strategy and pre-train each channel of our network to learn specific representation from specific data. After the filters of each channel are trained, we then train our fully connected hidden layer and the softmax layer to classify the input. This strategy allows us to decrease the amount of time needed to train our network and increased the generalization property of our filters.

5.3 Methodology

To evaluate our model, we perform three sets of experiments. In the first set we perform a parameter evaluation, determining the impact of some of the parameters of the network. In the second set of experiments we evaluate some aspects of the architecture: the impact of the input length and the use of the inhibitory fields. The last set of experiments evaluates the capability of the CCCNN to learn specific and crossmodal representations, and use them to classify emotion expressions.

One extra corpus is used to train the Music stream of the auditory information. The GTZAN corpus [286] is not directly related to emotion recognition, but to music genre classification. The task of music genre classification is similar to music emotion classification [157] because the task is to cluster audio segments which are closely related based on auditory features. Music genres can also be used for emotion classification, since for example blues and soul music is more related to sadness or feelings of loneliness, and pop music more to happiness [157]. This database contains 1000 song snippets, each one with 30 seconds and a sampling rate of 22050 Hz at 16 bit, labeled into ten musical genres: Blues, Classical, Country, Disco, Hip hop, Jazz, Metal, Pop, Reggae, and Rock.

For all experiments, 30 experiment routines were performed and the mean of the accuracy was collected for each expression individually, which helps us to understand better our model.

5.3.1 Experiment 1: Parameter Evaluation

To evaluate how the parameters affect the network, a parameter exploration experiment is performed. Three parameters are chosen, with the range of values based on the evidence found by Simard et al. [279] and our previous experiments with CNNs. The three parameters were chosen because of their major influence on the network response. A total of three different values is chosen for each parameter, generating a total of 27 experiments. Table 5.1 shows the chosen parameters and the range of values.

The number of filter maps affects directly the number of features extracted, and what these features represent. A large number of feature maps introduces

Table 5.1: Parameter sets evaluated for each experiment. The combination of all values for all parameters was evaluated and discussed.

Parameter	Values		
Filter maps layer 1 and 2	10 and 20	20 and 40	30 and 60
Receptive field size layer 1	3x3x2	11x11x2	21x21x2
Receptive field size layer 2	3x3	7x7	11x11

redundancy, and a small number is not enough to extract a proper description of the emotion sequence. The minimum and maximum values of 10 and 30 filter maps were chosen based on preliminary experiments, where these values represented the limits where the network showed a big variation for the accuracy. The number of filter maps on the second layer [279], is selected as twice the number of filter maps on the first layer. This selection leads to more specialized features on the second layer to expand the representations on the first layer, which are mostly edge-like detectors. The size of the receptive fields determines which pixel structures are important for the model. On the first layer, the receptive fields are connected directly to the image, and they will enhance structures present in the original data. If the receptive fields are too small, they will not be able to enhance important pixel structures, and will generate redundancy for the next layers. If they are too large, they will consume more pixel structures than necessary, and they will not be able to determine or to react to these structures, aggregating more than one structure into one filter map. This can generate very specific filter maps for the data while training the network, which leads to an overfitting of the model. For our experiments, we chose a range between the smaller and maximum receptive field sizes which were able to extract meaningful information from our input.

5.3.2 Experiment 2: Aspects of the Architecture

An expression occurs between 300 milliseconds and 2 seconds [83]. To evaluate an optimal approximation of sequence length, we evaluated our Face channel with four different input lengths: 40ms, 300ms, 600ms and 1s. For this experiment we use the FABO corpus, explained in details in Chapter 4, Section 4.6, and as the sequences in this corpus were recorded with a frame rate of 30 frames per second, that means that we evaluate the use of 1, 9, 18 and 36 frames as input. We also evaluated the input length for the Movement channel. First, we evaluate the use of 2 frames to compose a movement representation, then 5 frames, 10 frames and lastly 15 frames. This leads to feeding the network with 15, 6, 3 and 2 movement images respectively.

We then evaluate the use of the inhibitory fields on the visual stream, by applying it in different layers. We show how the inhibitory fields affect each representation of each layer and why we only use them on our Face channel.

For the auditory representation, we follow indications in the work of [1] for

the Speech channel and [104] for the Music channel. Separating the same 1s of representation and using the window and earlier slide values indicated in this work produced the best results, so we kept them. Also, the use of inhibitory fields on the auditory channel did not produce any improvement on the results, causing exactly the opposite: an overfitting of the filters made the network lose completely the focus during training.

5.3.3 Experiment 3: Emotion Expression Recognition

Visual Emotion Expressions

Using the FABO corpus we evaluate the visual stream of the network. In this set of experiment, we evaluate the use of the Face and Movement channels individually and then both of them at the same time.

With this experiment we show in detail the impact that each modality has in different expressions. As the FABO corpus deals with secondary expressions, it is possible to see how our visual representation behaves for very different expressions, such as happiness and sadness, or very similar ones, as boredom and puzzlement.

Auditory Emotion Expressions

We use the SAVEE, GTZAN and EmotiW corpora, all detailed in Chapter 4, Section 4.6, to evaluate the auditory stream of the network. For all datasets we extracted 1s of each audio input to train our networks. To recognize each audio clip, we used a sliding window approach of 1 second as input to the network. So, if the original audio input has 30 seconds, we split the audio into 30 parts of 1 second and use them as input to the network. With 30 results, we identified the most frequently occurring ones, leading to a final classification result for the 30 seconds audio input.

We performed experiments using the two channels individually and the Cross-channel on the three datasets. This allows us to explain the behavior of the model when un-suitable information is used as input for each task. We also performed a Pre-training strategy, where the Music-specific architecture was trained exclusively with the GTZAN set, the Speech-specific architecture with the SAVEE set and the Cross-channel architecture uses the pre-trained features of both previous architectures and trains its own Cross-channel with the EmotiW corpus. This way we ensure that the Cross-channel architecture uses the specific representation learned through the specific architectures to construct a higher abstraction level of auditory features. The mean and standard deviation of the accuracy over 30 training runs are calculated for all the experiments.

Multimodal Emotion Experiments

The EmotiW corpus contains the most complex emotion expressions in our experiments. The video clips contain different subjects (sometimes at the same time),

music, speech, different lighting conditions in the same scene and various face positions. This makes the emotion classification in this dataset very difficult.

We evaluate the use of our channels trained with this dataset; first each channel individually and then the integration of visual only streams and auditory only streams. Finally, we evaluate the audio-visual representation. Each of these experiments is performed with two different training strategies: one with and one without the pre-training of the filters. We use the FABO corpus to pre-train the filters of the visual stream and the SAVEE and GTZAN corpus to pre-train the auditory stream.

All results are compared and we show for the six basic emotions [81], plus a neutral category, how each of the modalities behaves and the advantage of using the pre-training strategy.

As the SAVEE corpus also has visual information, with the recording of the faces of the subjects, we also evaluate only the Face channel and the crossmodal representation obtained with the use of the auditory channels and the Face channel.

5.4 Results

5.4.1 Experiment 1: Parameter Evaluation

After performing the parameter exploration experiments, the average of the accuracy was computed. Table 5.2 shows the results for all the parameter combinations. For the first set of experiments, we locked the number of filter maps to 10. The best result was achieved with a configuration of a receptive field size of 3x3 pixels in the first and second layer, with an accuracy of 91.3%, while the worst result found, with a configuration of kernel size in the first layer of 11x11 pixels and in the second layer of 21x21 pixels, was 87.89%. We found a trend: when the size of the receptive fields was increased, in both layers, the network produced the poorer results.

When locking the number of filter maps on the first layer at 20, the results obtained showed a similar trend: increasing the size of the receptive fields of the filter maps for the first and second layer decreases the accuracy. For this set of experiments, the best result was also with the smaller receptive field size, in both layers, achieving 91.25% of accuracy. The worst result can be observed when using the maximum value of the receptive field size. This configuration achieved an accuracy of 87.3%.

The trend can also be found when the number of filter maps on the first layer was locked at 30. The best result was also with the smaller receptive field sizes with an accuracy of 91.18%. The worst accuracy was 88.9%, when using the largest kernel size for both layers.

Evaluating the parameters, it is possible to find some trends in the network behavior. Figure 5.5 shows a box plot with the individual analysis of the parameters. The plot shows the variance of the accuracy of each parameter. The plot depicts that using a smaller receptive field in both layers the accuracy improves. Looking

Table 5.2: Average accuracy for each parameter combination computed during the parameter exploration experiments.

Receptive field size		Filter maps		
1st Layer	2nd Layer	10	20	30
3	3	**91.30%**	**91.25%**	**91.18 %**
3	7	90.93 %	89.83 %	91.00 %
3	11	89.93 %	87.92 %	90.47 %
11	3	90.77 %	90.83 %	90.01 %
11	7	90.04 %	89.75 %	91.02 %
11	11	87.34 %	88.65 %	90.43 %
21	3	90.08 %	89.82 %	89.90%
21	7	90.01 %	88.92 %	90.42 %
21	11	87.89 %	87.30 %	88.90 %

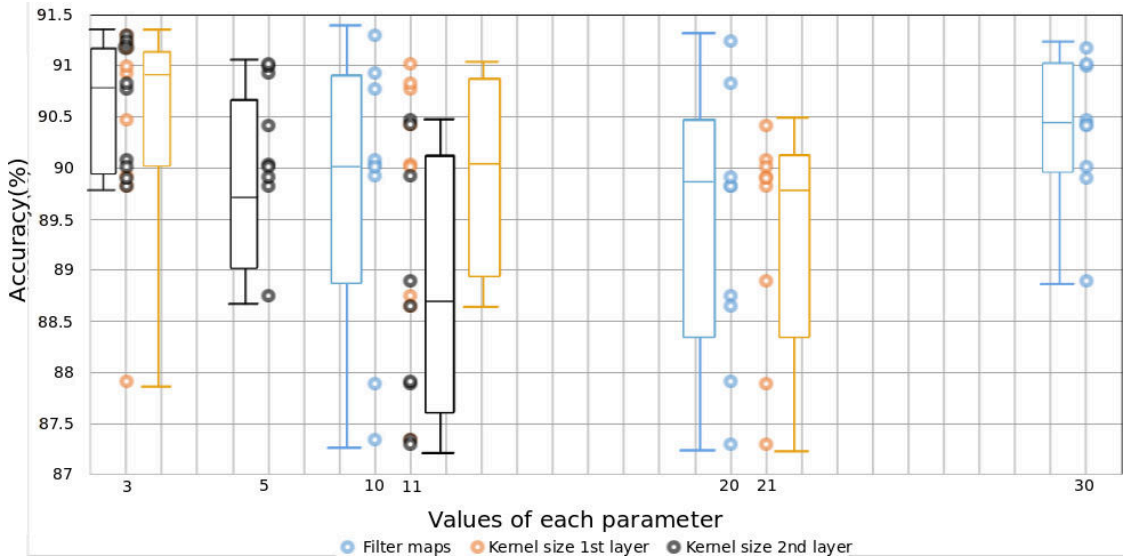

Figure 5.5: Individual analysis for the parameter exploration. It is possible to see a trend that when the kernel sizes are smaller the accuracy tends to be higher. When increasing the number of filter maps, the average of the results is also higher, despite the best accuracy being lower.

at the plot, it is possible to see a clear trend in the spread of the results. When using smaller receptive fields, the network produces results with a better accuracy median and smaller variance. Also, increasing the number of filter maps decreases the accuracy variance between the experiments. This shows that when we increase the number of filter maps, the results of our network tend to be more stable. Using a smaller receptive field allows the network to learn those general structures which occur more often in the images. Passing these general features through our layers generate higher level features in deeper layers. But, as our network is not very deep, we need to increase the number of filter maps in order to expand the variety of features extracted in the same layer.

When evaluating the combinations of parameters, it is possible to visualize how the number of filter maps influences the results. Figure 5.6 illustrates the plot with the combination of the parameters. It is possible to see that when we increased the size of filter maps, the variance in the accuracies values is small. Also, it is possible to see that increasing the number of filter maps in the second layer produces a lower accuracy. However, when using 30 filter maps and using a receptive field of 7x7 in the second layer, the network produces better results. This occurs because extending the number of filter maps will result in the network generating different feature representations at the same level and thus it generates redundant features which allow a better generalization. It is important to note that with too many redundant filters, the network loses the capability of generalizing and ends up overfitting.

5.4.2 Experiment 2: Aspects of the Architecture

The results obtained when training the Face channel with different sequence lengths showed us that the use of 9 frames produced the best results, as shown in Table 5.3. As the FABO corpus was recorded with 30 frames per second, the use of 9 frames means that the sequence has an approximate length of 300 milliseconds. A sequence with this length is congruent with the description of facial expressions and micro expressions, meaning that our model performed best when both expressions could be perceived. The use of longer expressions, with 1s, produced the weakest results.

The Movement channel receives a sequence of motion representations of 1s of the expression as input. This means that each representation of this sequence is composed of several frames. The results, exhibited in Table 5.3, show that the use of 3 movement representations gave the best performance, meaning that each movement representation is composed of 10 frames. This means that each motion representation captures 300ms. Our results showed that using a minimum number of frames to capture the movement, 2 frames per motion representation and 15 frames as the channel's input, produced the worst result.

In the following experiments, we evaluate the use of inhibitory neurons in our visual channels. We evaluate the use of the inhibitory fields on each of the layers, and in combination with all layers of each channel and Table 5.4 shows the results. The application of the inhibitory fields to the Movement channel did not

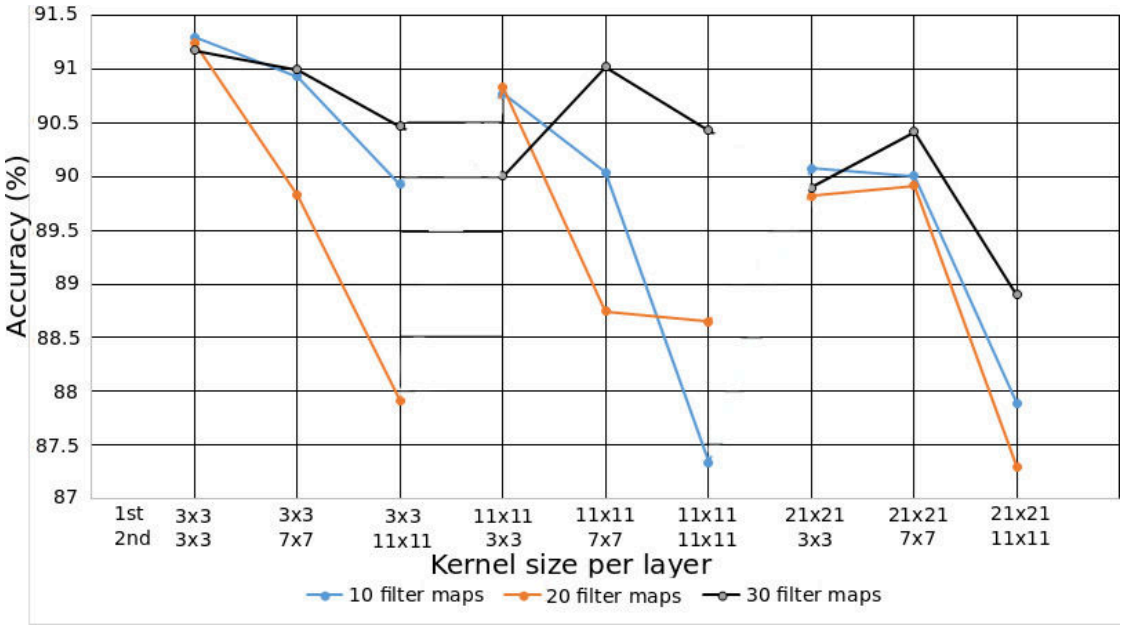

Figure 5.6: Combination analysis for the parameter exploration. When smaller receptive fields are used, the accuracy is higher. Also, when comparing the same size of receptive fields, but increasing the number of filter maps, it is possible to see that the average of the accuracy increases, although the best result was found with the smaller number of filter maps.

Table 5.3: Average accuracy, in percentage, for different lengths of the input sequence, in frames, for the Face channel, and in movement representations, for the Movement channel trained with the FABO corpus.

Face Channel				
Sequence Length	1	9	18	36
Accuracy(%)	64.8	*80.6*	73.6	49.4

Movement Channel				
Motion Representations	15	6	3	2
Accuracy(%)	48.3	67.9	*74.8*	66.2

produce better results, which is due to the fact that the movement representation is already a specified stimulus, and the filters alone were capable of coping with the complexity of the representation. The inhibitory fields produced better results when applied to the last layer of the Face channel, which confirms that the strong extra-specification on the last layer is beneficial for face expression recognition.

Table 5.4: Average accuracy, in percentage, for the use of inhibitory neurons in different layers of the Face and Movement channels trained with the FABO corpus.

Face Channel				
Layers	None	L1	L2	All
Accuracy(%)	80.6	59.9	*87.3*	64.4

Movement Channel					
Layers	None	L1	L2	L3	All
Accuracy(%)	*74.8*	41.3	47.8	48.8	45.8

5.4.3 Experiment 3: Emotion Expression Recognition

Visual Emotion Expressions

The combined results of the FABO experiments are presented in Table 5.5 and we can see that the overall mean accuracy of the integrated representation is the highest. Also, it is possible to see how some expressions behave with different modalities. For example, "Anxiety" and "Puzzlement" expressions had a performance similar to the Face and Movement channels individually, but increased when the integrated representation was used. Also, there was a great increase in the performance for "Disgust" and "Negative Surprise" expressions, showing that for these expressions the integrated representation provided more information than each modality individually.

Comparing our model to state-of-the-art approaches using the FABO corpus shows that our network performed similar, and in the Face representation better. Table 5.6 shows this comparison. The works of Chen et al. [45] and Gunes and Piccardi [118] extract several landmark features from the face, and diverse movement descriptors for the body movement. They create a huge feature descriptor for each modality, and use techniques as SVM and Random Forest, respectively, for classification. It is possible to see that the fusion of both modalities improved their results, but the performance is still lower than ours. In previous work, we used a Multichannel Convolution Neural Networks (MCCNN) [15], to extract facial and movement features. That network produces a joint representation, but our current CCCNN improved this representation through the use of separated channels per modality and the application of inhibitory fields. One can see a substantial improvement on the movement representation, mostly because we use a different movement representation in the CCCNN.

Auditory Emotion Expressions

Our Music-specific channel obtained the best accuracy, with a total value of 96.4%. The second best result appeared when using the pre-trained filters on the Cross-

Table 5.5: Accuracy, in percentage, for the visual stream channels trained with the FABO corpus. The results are for the Face channel (F), Movement channel (M) and the integrated Face and Movement channel, representing the visual stream (V).

Class	F	M	V
Anger	74.5	66.3	95.9
Anxiety	78.6	80.5	91.2
Uncertainty	82.3	75.8	86.4
Boredom	93.4	76.3	92.3
Disgust	78.3	65.9	93.2
Fear	96.3	80.0	94.7
Happiness	93.7	60.3	98.8
Negative Surprise	67.2	32.4	99.6
Positive Surprise	85.7	65.7	89.6
Puzzlement	85.4	84.8	88.7
Sadness	89.6	80.1	99.8
Mean	87.3	74.8	93.65

Table 5.6: Comparison of the accuracies, in percentage, of our model with state-of-the-art approaches reported with the FABO corpus for representations of the face, the movement, and both integrated.

Approach	Face	Movement	Both
MCCNN	72.7	57.8	91.3
Chen et al. [45]	66.5	66.7	75.0
Gunes and Piccardi [118]	32.49	76.0	82.5
CCCNN	87.3	74.8	93.65

channel architecture, with a total value of 90.5%, which still almost 6% less than using only the Music-specific architecture. Using the Speech-specific architecture, the accuracy was the lowest, reaching the minimum score of 62.5% when applying the pre-training strategy. Table 5.7 exhibits all the experimental results on the GTZAN dataset.

On the SAVEE dataset, the Speech-specific channel was the one which obtained the best mean accuracy (92.0%). It was followed closely by the pre-trained version of the Cross-channel architecture, with 87.3%. The trained version of the Cross-channel obtained a total of 82.9%. Here the Music-specific architecture obtained the worst results, with a minimum of 63.1% on the trained version. The pre-trained

Table 5.7: Average accuracy and standard deviation for all the experiments using the GTZAN dataset.

Experiment	Accuracy (STD)	
	Trained	Pre-Trained
Music-Specific	96.4% (+/- 3.4)	-
Speech-Specific	68.7% (+/- 3.2)	62.5%(+/- 1.6)
Cross-channel	83.9% (+/- 2.3)	90.5%(+/- 2.2)

Table 5.8: Average accuracy and standard deviation for all the experiments using the SAVEE dataset.

Experiment	Accuracy (STD)	
	Trained	Pre-Trained
Music-Specific	63.1% (+/- 2.7)	64.5% (+/- 2.3)
Speech-Specific	92.0% (+/- 3.9)	-
Cross-channel	82.9% (+/- 2.0)	87.3%(+/- 1.8)

Table 5.9: Average accuracy and standard deviation for all the experiments using the EmotiW dataset.

Experiment	Accuracy (STD)	
	Trained	Pre-Trained
Music-Specific	22.1% (+/- 1.4)	23.1% (+/- 2.2)
Speech-Specific	21.7% (+/- 2.3)	21.0%(+/- 1.2)
Cross-channel	22.4% (+/- 1.1)	30.0%(+/- 3.3)

version obtained slightly better results, reaching 64.5%. Table 5.8 shows all the experimental results on the SAVEE dataset.

For the EmotiW dataset the pre-trained version of the Cross-channel architecture gave the highest mean accuracy of 30.0%. All the other combinations, including the trained version of the Cross-channel, achieved accuracies around 20%. See table 5.9 for the results of our experiments with the EmotiW dataset.

The results achieved by our architectures are not far away from the state-of-the-art results in the literature. For the GTZAN dataset, our specific architecture performs close to the system proposed by Sarkar et al [267]. Their approach uses empiric mode decomposition to compute pitch-based features from the audio input. They classify their features using a multilayer perceptron. Following the same evaluation protocol we used, they could reach a total of 97.70% of accuracy,

Table 5.10: Performance of state-of-the-art approaches on the GTZAN dataset. All experiments use 10-fold cross validation and calculate the mean accuracy. The results obtained by Sgtia et al. [277] were using a different data split, using 50% of the data for training, 25% for validation and 25% for testing.

Methodology	Accuracy(%)
Arabi et al. [5]	90.79
Panagakis et al.[222]	93.70
Sgtia et al.[277]*	83.0
Huang et al. [138]	97.20
Sarkar et al.[267]	97.70
Music-specific	96.40
Cross-Channel	90.50

slightly more than our 96.4% with the Music-specific architecture.

Our Cross-channel architecture, when using the pre-training strategy, obtained a lower accuracy, but still competitive when compared to other results using different approaches. Table 5.10 exhibits the results obtained on the GTZAN dataset. All the proposed techniques use a combination of several features, and a generic classifier such as SVM or MLP. However, using such a large number of audio features, their approaches are not suitable for generalization, a property that our Music-specific architecture has. The approach of Sgtia et al. [277] is similar to ours. They evaluate the application of techniques such as dropout and Hessian Free training, but do not report the performance of the network neither for learning different features nor for generalization aspects.

For the SAVEE dataset, our approach is competitive. This dataset contains only speech signals, which are very different from the music signals. That explains the different accuracies obtained by the Music-specific architecture and the Speech-specific architecture. Once more the pre-trained Cross-channel architecture showed its generalization capabilities and was able to achieve a result which was comparable to the Music-specific architecture, having less than 3% of accuracy difference. When compared to state-of-the-art approaches, our Music-specific architecture obtained a result comparable with the work of Muthusamy et al. [215]. They use a particle swarm optimization technique to enhance the feature selection over a total of five different features, with many dimensions. They use an extreme learning machine technique to recognize the selected features. Their work showed an interesting degree of generalization, but still a huge effort is necessary, with the training step consuming enormous amounts of time and computational resources. The authors of the SAVEE dataset also did a study to examine the human performance for the same task. Using the same protocol, a 4-fold cross validation, they evaluated the performance of 10 subjects on the recognition of emotions on the audio data. The results showed that most approaches exceeded human per-

Table 5.11: Performance of state-of-the-art approaches on the SAVEE dataset. All the experiments use 4-fold cross validation and we have calculated the mean accuracy.

Methodology	Accuracy(%)
Banda et al. [9]	79.0
Fulmare et al.[101]	74.39
Haq et al.[124]	63.0
Muthusamy et al. [215]	94.01
Speech-specific	92.0
Cross-Channel	87.3
Human Performance [123]	66.5

Table 5.12: Performance of state-of-the-art approaches on the EmotiW dataset. All the results are the mean accuracy on the validation split of the dataset.

Methodology	Accuracy(%)
Liu et al. [196]	30.73
Kahou et al.[149]	29.3
Baseline results[70]	26.10
Cross-Channel	30.0

formance on this dataset. Table 5.11 exhibits the state-of-art results and human performance on the SAVEE dataset.

The EmotiW dataset proved to be a very difficult challenge. On this dataset, our specific models did not work so well, but as Table 5.12 shows this is also a much harder task. Due to the huge variability of the data, neither of them was able to learn strong and meaningful features by itself. When the Cross-channel architecture was used with the pre-training strategy, the network was able to learn to correlate the features of each channel, and use them to overcome the complexity of the dataset. Our Cross-channel architecture results are competitive with the state-of-the-art approaches, and performed better than the baseline values for the competition. The work of Liu et al. [196] and Kahou et al. [149] extract more than 100 auditory features each, and use classifiers such as SVM or multi-layer perceptrons to categorize them. Our Cross-channel architecture results showed that we can actually obtain similar generalization capability using a simple and direct pre-training strategy without the necessity of relying on several different feature representations. Table 5.12 exhibits the results on the EmotiW dataset.

Table 5.13: Average accuracy, in percentage, for the auditory and visual stream channels trained with the SAVEE corpus. The results are for the Face channel (F), Speech channel (S), Speech and pre-trained Music channel, representing the auditory stream (A) and the integrated audio-visual streams, with the Face, Speech and Music channels (AV).

Class	F	S	A	AV
Anger	95.4	95.0	92.6	100
Disgust	95.6	100	88.0	100
Fear	89.7	88.0	85.5	100
Happiness	100	81.1	86.1	95.0
Neutral	100	100	91.3	100
Sadness	90.0	93.5	87.4	96.5
Surprise	86.7	86.5	80.5	96.7
Mean	93.9	92.0	87.3	98.3

Multimodal Emotion Expressions

The results of the SAVEE experiments are exhibited in Table 5.13. It is possible to see that the auditory information yielded the lowest accuracy, and among them the pre-trained representation was the one with the lowest general accuracy. This happens because the data in the SAVEE corpus does not contain music, only speech, which reflects directly on the performance achieved by the network. Still, it is possible to see that the auditory channel composed of the Speech and Music does not decrease substantially the performance of the network, and makes it more robust to deal with speech and music data.

We also see that the face representation obtained a similar performance to the auditory one, but when combined, the performance tends to increase. This is due to the fact that when both, face and auditory information, are present, the network can distinguish better between the expressions. This is demonstrated by the performance of the expressions "Anger", "Sadness" and "Surprise", which have a similar performance in individual channels and a higher one in the integrated representation.

Our approach proved to be competitive when evaluated on the SAVEE corpus. When compared to state-of-the-art approaches, our representations showed a result comparable to the work of Banda et al. [9]. They use a decision-based fusion framework to infer emotion from audio-visual inputs. They process each modality differently, using linear binary patterns to represent the facial expressions and a series of audio features to represent speech. After that, a pairwise SVM strategy is used to train the representations. Our network has a similar performance for face representation, but a higher accuracy for audio. We improved 10% the accuracy more than the speech representation. For the multimodal integration, our network

Table 5.14: Performance of state-of-the-art approaches on the SAVEE dataset.

Methodology	Face	Audio	Both
[9]	95.0	79.0	98.0
[124]	95.4	56.3	97.5
CCCNN	93.9	92.0	98.31
Human Performance	88.0	66.5	91.8

has been shown to be competitive, and performed similarly, but with a much less costly feature representation process. The authors of the SAVEE dataset, [124], also did a study to examine the human performance for the same task. Using the same protocol, a 4-fold cross validation, they evaluated the performance of 10 subjects on the recognition of emotions on the audio and video data. The results indicate that most approaches exceeded human performance on this dataset. Table 5.14 exhibits state-of-art results and human performance on the SAVEE dataset.

The EmotiW corpus proved to be a very difficult challenge. Table 5.15 illustrates all results on the corpus. It is possible to see that the visual representations, represented by the columns F, M and V, reached better results than the auditory representation, presented in columns S, Mu and A.

The visual representations indicate a very interesting distribution of accuracies. It is possible to see that when the expressions were represented by the movement, Column M, "Happy" and "Sad" expressions were recognized better on our model than the others, showing that for these expressions the movements were more reliable than the face expression itself. When integrated, the visual representation improved the performance of most expressions, in particular surprised, "Anger" and "Happy" expressions, which indicates that these expressions are better recognized when movement and facial expressions are taken into consideration.

The auditory representation indicates that most of the expressions are not well recognized only with auditory information, exceptions are angry and happy emotions. This can be related to the nature of the dataset, because usually in movies happy and angry are expressed with similar song tracks or intonations. The integrated representation for the auditory stream performed better than the individual ones in all the expressions.

Finally, the multimodal representation was the one with the best performance. We see an improvement in classification of "Sad" and "Anger" expressions, but also in fear and surprised ones. This is due to the fact that the combination of different sound tracks, facial expressions and movement for these expressions represents them better than a single modality. In general, it is possible to see that surprised, disgusted and sad expressions were the ones with the lowest performance in all modalities.

Table 5.16 shows the results on the EmotiW dataset. On this dataset, the performance of our model dropped, but as Table 5.16 shows this is also a much

Table 5.15: Average accuracy, in percentage, for the auditory and visual stream channels trained with the validation set of the EmotiW corpus. The results are for the Face channel (F), Movement channel (M), Face and Movement channel together, representing the visual stream (V), Speech channel (S), Music channel (Mu), Speech and Music channel together, representing the auditory stream (A) and visual-auditory integration (AV).

Class	F	M	V	S	Mu	A	AV
Anger	70.2	50.8	77.8	56.4	50.7	70.1	80.3
Disgust	18.2	9.4	18.7	12.4	2.6	15.2	23.4
Fear	21.4	16.8	20.2	7.8	6.5	7.2	30.8
Happiness	67.2	75.6	77.8	59.1	65.4	72.0	81.2
Neutral	67.2	57.7	70.9	10.8	15.6	25.4	68.7
Sadness	22.4	21.2	23.2	8.3	9.8	16.2	24.5
Surprise	5.4	10.0	12.1	0.0	2.1	4.1	14.0
Mean	38.8	34.5	42.9	22.1	21.8	30.0	46.1

Table 5.16: Performance of state-of-the-art approaches on the EmotiW dataset. All the results calculate the mean accuracy on the validation split of the dataset.

Methodology	Video	Audio	Both
[196]	45.28	30.73	48.53
[149]	38.1	29.3	41.1
[70]	33.15	26.10	28.19
CCCNN	42.9	30.0	46.1

harder task. Due to the variability of the data, neither of the modalities provides an overall high accuracy. Our model results are competitive with the state-of-the-art approaches, and performed better than the baseline values of the competition. The works of Liu et al. [196] and Kahou et al. [149] extracts more than 100 auditory features each, and use several CNNs to extract facial features. A vector composed of the output of the CNN are used by several classifiers such as SVM and multi-layer perceptron to categorize the input into emotions. The results of our models showed that we can actually obtain similar generalization capability using a simple and direct pre-training strategy without the necessity of relying on several different feature representations.

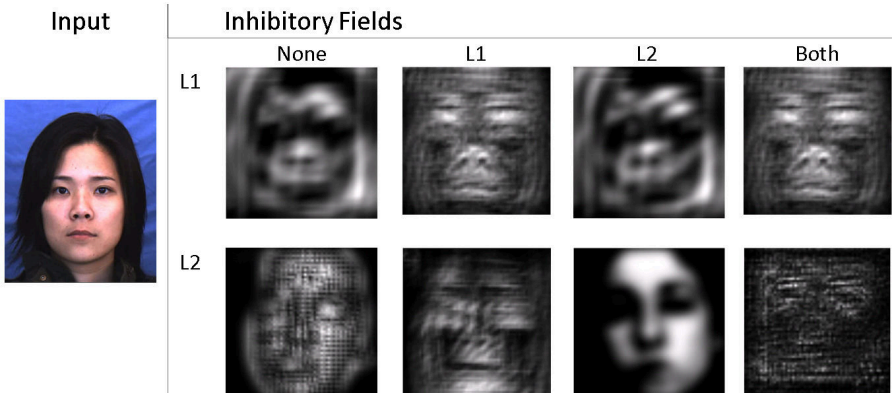

Figure 5.7: Implementing inhibitory fields in different layers of the network produces different features. Each visualization corresponds to one filter on a determined layer. It is possible to see how the inhibitory fields affect the feature extraction capabilities of each layer.

5.5 Discussion

In this section we discuss two concepts of our network. First we analyze the CCCNN architecture, and how the introduction of inhibitory fields and the Cross-channel contribute to the representation. Then we discuss how the model represents multimodal stimuli, how the expression is decomposed inside the model and what each layer represents.

5.5.1 Inhibitory Fields and Cross Channels

The application of the inhibitory fields has been shown to increase the performance of the network only when they were implemented in the last layer of the face channel. That was caused by the over specialization of that the inhibitory fields produced in the layer's filters, which turned to be beneficial for the model. When the inhibitory fields were applied to the first layer, the filters learned more complex patterns, which did not help in the feature generalization. That phenomenon is easily visible when we visualize the features that the network learned using the deconvolution process illustrated in Image 5.7, which shows the visualizations of the internal knowledge of one filter in different layers of the network.

When no inhibitory filter was implemented, the first layer the network learned some edge detectors, which could filter mostly the background and hair of the person. In the second layer, the network constructed a higher level of abstraction, mostly the shape of the face, and some regions such as eyes, mouth and nose are roughly highlighted. When we implement the inhibitory fields in the first layer only, we find that the more information is filtered. The filters detected more specific regions, filtering much more information which is relevant to represent the facial expression. This causes a problem in the second layer, which then tries to learn very specific concepts, and constructed a very limited representation. When the inhibitory fields are applied in the last layer, we found a very clear distinction in

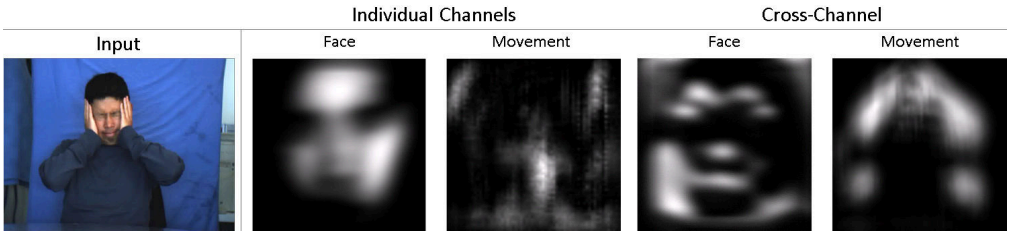

Figure 5.8: Applying the Cross-channel on the individual representations brings results on different features. Note that the face representation after the application of the Cross-channel changed to include the hands movement.

the representation. The shape of the face is very clear, but regions as eyes, nose and mouth are better represented than when no inhibitory fields are applied. Finally, when we applied the inhibitory fields in both layers, the final representation does not contain any reliable information with some very rough representation of the eyes and nose.

The cross channels also have an impact on the quality of the extracted filters. Our Cross-channels integrates two channels into one representation, which was shown to be more efficient and robust, but also reduced the dimensionality of the data. The application of the Cross-channels created a new representation of the input stimuli, which is different from the individual representation. Figure 5.8 illustrates the visualizations of the last layer of the individual channels and the Cross-channel. We can see that the Cross-channel features are different from the individual representation, and they changed to capture an important feature: the hands over the face. Furthermore, we see that the facial features changed drastically to incorporate the movement of the hands, which is now also highlighted in the movement channel.

5.5.2 Expression Representation

The application of the visualizations also helps us understand how the network represents an expression. It is possible to see how the expressions are decomposed inside the network, and have an insight on the role of each layer of the network to build the expression representation. By visualizing the same region of neurons for several images, it is possible to identify for which regions those neurons activate most strongly. This way, we can analyze which parts of the input stimuli activate each filter of the network. Figure 5.9 illustrates this concept, where it is possible to see what each filter codes for in each layer. To generate these visualizations, we created an average per filter in the Face channel for all the images in the FABO corpus.

The filters learn to represent different things, which are complementary for the emotion expression. In the first layer, mostly background and hair information is filtered. Filter 5 highlights the region of the mouth out of the image, while filter 2 keeps the eye information.

The most interesting representations occur in the second layer, where filters

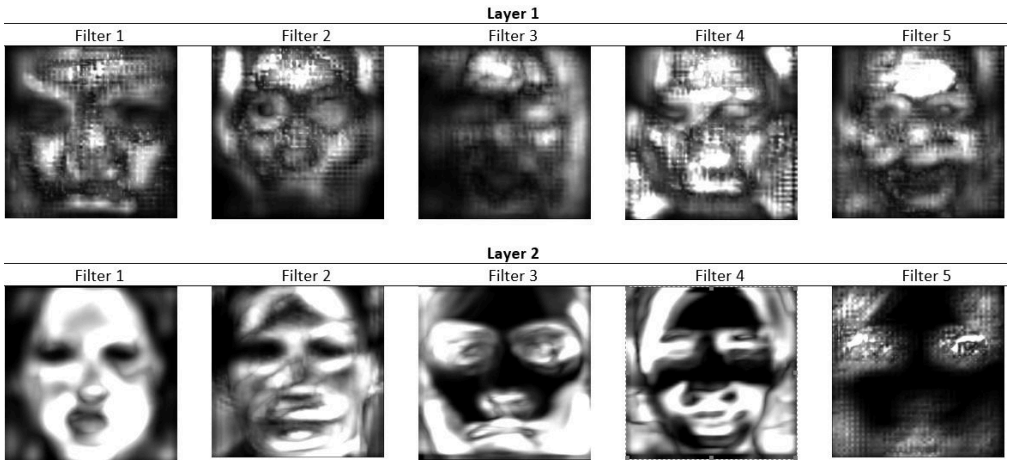

Figure 5.9: Mean visualization from all images in the FABO corpus per filter in all the layers of the Face channel. It is possible to specialized filters, which helps us to understand how the expression representation is created.

1 and 2 represent mostly the face shape and positions of eyes, nose and mouth. Filters 3 and 4 represent the eyes, nose and mouth shapes, where filter 3 activates mostly for the cheeks and closed mouths and filter 4 for open mouths. Different from the others, filter 5 specialized on eyebrows mainly.

The filters in our network show that our network actually builds the face expressions based on the changes of some particular regions, which is consistent with the Facial Action Coding System (FACS) [86]. The FACS is a coding scheme to represent facial expressions based on movement of facial muscles. The movements in FACS are mostly related to eyes, mouth, eyebrows and cheek regions, which are very similar to the regions detected by our network, showing that the features which emerge in the filters are actually related to human facial expression perception.

Our network filters react to very specific patterns in the input images, which are related to human facial expressions. We can see how these patterns are strong when we send to the network images which resemble human expressions, this is illustrated in Figure 5.10. The network highlighted regions which were closely related to human features. In the image with the dog, the position of the eyes and mouth were detected, and in the Don Quixote painting, the shape of the face was highlighted. In all images, it is possible to see that the filters of the network highlighted regions of interest that have a similar contrast as some facial features, as face, eyes and mouth shapes. On the other hand, the network is strongly domain restricted. It will always try to find human facial features in the images, even when they are not present. This can cause problems, especially in the EmotiW corpus, illustrated in the last column of Figure 5.10.

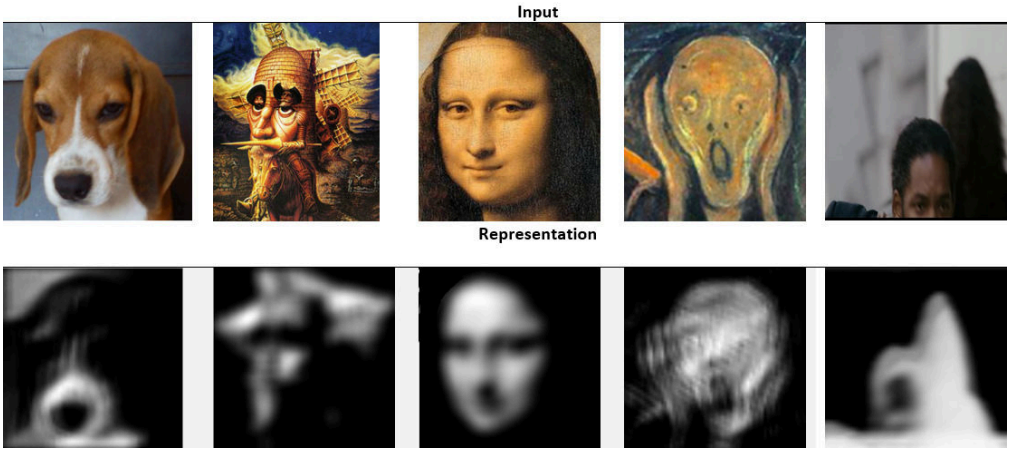

Figure 5.10: Visualization of the facial representation for different images. We see that the network tries to find human facial features, such as mouths, eyes and face shapes in the images.

5.6 Summary

In this chapter, we proposed a novel architecture for multimodal emotion expression representation. Our model introduces Cross-channel Convolution Neural Networks (CCCNN) to learn specific features of audio-visual stimuli. The network implements several channels, and each one learns different emotional features from each modality and applies a cross-convolution learning scheme to integrate auditory and visual representations of emotion expressions.

The network architecture is inspired by the ventral and dorsal stream models of the visual and auditory systems in the brain. Also, it implements shunting inhibitory neurons to specialize the deeper layers of the network, avoiding the necessity of a very deep neural network. Such mechanisms showed to help us to represent spontaneous and multimodal expressions from different subjects.

We also introduce visualization tools which allow us to understand and identify the knowledge of the network. By using the deconvolution process to visualize the internal representation of the CCCNN filters we showed how our model learns different expressions in a hierarchical manner.

To evaluate our model, we use three different corpora: the Bi-modal face and body benchmark database (FABO) with visual expressions, the Surrey Audio-Visual Expressed Emotion (SAVEE) Database with audio-visual expressions and the corpus for the Emotion-Recognition-In-the-Wild-Challenge (EmotiW), which contains audio-visual clips extracted from different movies. Each corpus contains different expression information, and we use them to fine tune the training to evaluate each modality. Our network showed to be competitive, and in the case of the FABO corpus better when compared to state-of-the-art approaches.

Chapter 6

Learning Emotional Concepts with Self-Organizing Networks

6.1 Introduction

To classify emotion expressions is a difficult task: First the observation of various different modalities is necessary. Second, the concept of emotion itself is not precise, and the idea of classifying what another person is expressing based on very strict concepts, makes the analysis of such models difficult.

Dealing with such set of restricted emotions is a serious constraint to HRI systems. Humans have the capability to learn emotion expressions and adapt their internal representation to a newly perceived emotion. This is explained by Hamlin [120] as a developmental learning process. Her work shows that human babies perceive interactions into two very clear directions: positive and negative. When the baby is growing, this perception is shaped based on the observation of human interaction. Eventually, concepts such as the five universal emotions are formed. After observing individual actions toward others, humans can learn how to categorize complex emotions and also concepts such as trust and empathy. The same process was also described by Harter et al. [125], Lewis et al. [188] and Pons et al. [242].

In the previous chapter we introduced the CCCNN for multimodal emotion expression representation, and in this chapter we extend the model by adapting it to learn different representations and cluster them into emotional concepts. Based on the findings of Hamlin et al. [120], and her emotional learning theory, we introduce the use of the CCCNN as innate perception mechanism, and make use of self-organizing layers to learn new emotional concepts. The model is evaluated in the recognition and learning of new expressions and we proceed with an analysis of the model while it learns multimodal expressions, and study the development of emotional learning from different subjects.

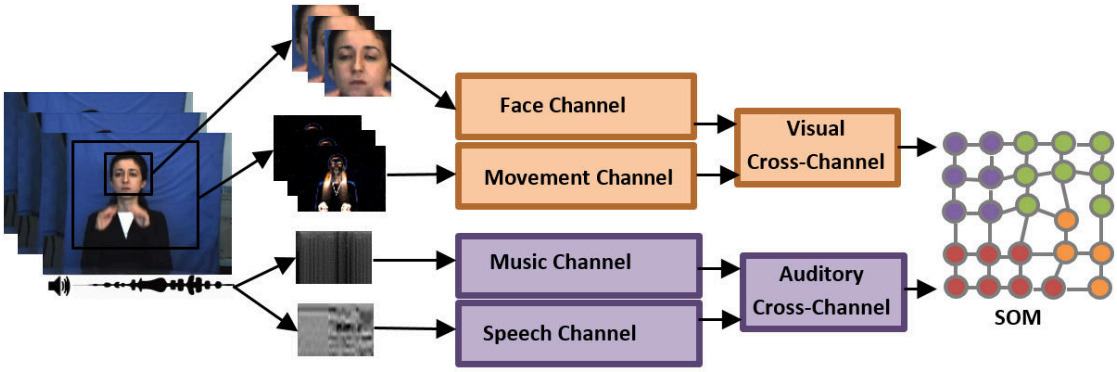

Figure 6.1: Crossmodal architecture used as input for the SOM. This architecture extracts multimodal features from audio-visual inputs and clusters the representation in different regions, which represent emotion expressions.

6.2 Emotion Expression Learning

To create a developmental emotion perception mechanism, we focus on the dimensional model representation. We follow the idea of Hamlin [120] of developmental learning, and we train our CCCNN to learn strong and reliable emotion expression representations in different modalities. We then replace the fully connected hidden and softmax layers of our network with a layer which implements Self-Organizing Maps (SOMs) [163]. The SOMs are neural models where the neurons are trained in an unsupervised fashion to create a topological grid that represents the input stimuli. In a SOM, each neuron is trained to be a prototype of the input stimuli, meaning that after training, each neuron will have a strong emotion representation and neurons which are neighbors are related to similar expressions.

In our architecture, we implement a SOM with 40 neurons in each dimension. Empirically this was shown to be enough to represent up to 11 emotions for the visual stream and up to 7 emotions using crossmodal representation. Figure 6.1 illustrates the updated version of our model.

6.2.1 Perception Representation

After training, a SOM will create a grid of neurons each one with the same dimensionality as the input stimuli. The neurons of a SOM organize a projection of a high-dimensional data space into a set of neurons spread in a grid. This means that the knowledge of a SOM is represented by its topology. One way to interpret the neurons in a SOM is to use the U-Matrix [287]. The U-Matrix creates a visual representation of the distances between the neurons. Basically, you calculate the distance between adjacent neurons. The U-Matrix gives us a very important representation of the structural behavior of the SOM, in which we can identify different clusters of neurons. The U-Matrix of a SOM is defined as:

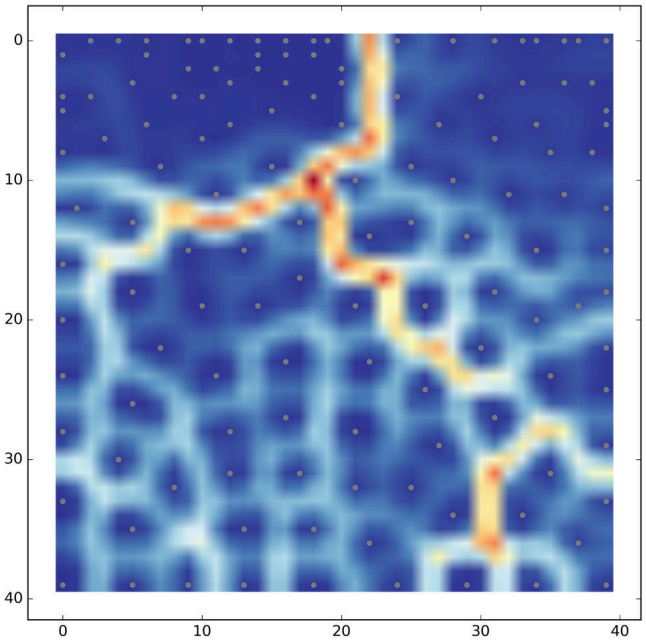

Figure 6.2: U-Matrix of a SOM with 40 neurons in each dimension and trained with happy, sad and neutral expressions. It is possible to see the neurons, represented by dots, in different regions, which represent the distances among the neurons.

$$U - Matrix = \sum M_{m=1} \sum_{i=1}^{k} d(w - w_m), \tag{6.1}$$

where M indexes the neighbor neurons, and w is the set of weights of each neuron. The distance calculation is given by $d(x, y)$, and is usually the Euclidean distance.

After training, our SOM has neurons which represent emotion expressions, and we can visualize them by calculating the U-Matrix. Our SOM is trained completely in an unsupervised fashion, which means that we do not identify the expressions we are showing to the network with any class and the U-Matrix shows the distribution of the neurons, or emotion expressions, over a grid. We use this grid to identify regions of neurons that have a similar representation, and find certain patterns of the neuron distribution. Figure 6.2 illustrates an example of a U-Matrix calculated of a SOM with 40 neurons in each dimension and trained with three different expressions: happy, sad and neutral. It is possible to see the neurons, marked as the dots, and different regions based on the distances between the neurons.

In a SOM we can calculate the distance of a certain input for all the neurons, and the neuron which has the smallest distance is selected as the best matching unit, which represents the neuron that mostly resembles the presented input. However, instead of using only one neuron to represent an expression, we can use the distances of each neuron to the input to create an activation map, showing which neurons of the SOM are more related to the input. This way, we can, for example,

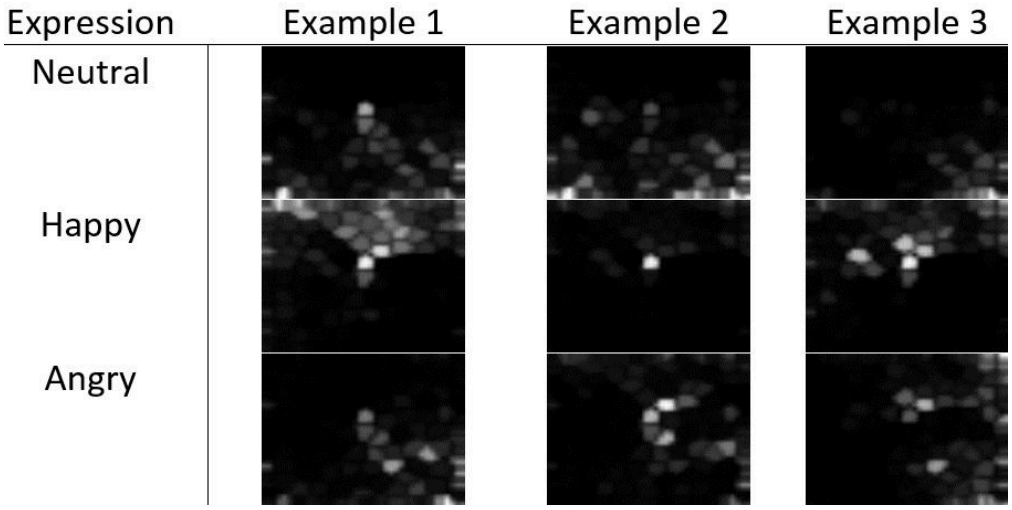

Figure 6.3: Examples of activation maps when three different expressions for each class are presented. It is possible to see that each class has an activation pattern different from the other classes.

identify which regions of the SOM activate mostly when happy or sad expressions are presented.

The neurons which are strongly related to a presented input, will activate most: for instance, a certain neuron that activates for a happy expression will have a lower activation when a sad expression is presented. This way, by visualizing several activation maps, we can have an emotion representation which is very close to a dimensional perception, but learned in an unsupervised way. Figure 6.3 illustrates different activation maps. It is possible to see that the activation pattern changes when different happy, angry or neutral expressions are presented to the network.

The visualization of the knowledge learned by the SOM is not easy, similar to the human perception of emotions. Emotion expressions are learned by humans in a continuous process of perceiving new expressions and adapting them to previous knowledge [120]. This process happens through childhood by assimilating similar emotions with known concepts, such as happiness, pain or depressive states. This means that each person has their own emotion perception mechanism, based on different features and different perceived emotions. We simulate this process by using a very strong feature representation, learned by the CCNN, and updating our SOM with perceived expressions. In this case, our SOM represents a unique perception representation, which could be related to a person's own perception.

This mechanism helps us to interpret emotion expressions in general. By creating the person's specific activation maps, we can identify how this person's expression behavior differs. Also, we can create specific SOM's for specific person's, or for a specific group of persons. This way we can create very personal representations of expressions, which are suited for different subjects and can be updated differently. This allows the model to have an individual knowledge about how a particular person expresses its own emotions. This way, a person that expresses

themselves in a more shy way will have a different neuron structure than one which is more excited, and both can be represented by different SOMs.

6.2.2 Expression Categorization

With the use of the U-Matrix and the activation maps, we can identify patterns in the SOM structure. We can assimilate concepts to similar patterns, finding which regions of the network fire most for known expressions. This means that we can identify network regions which fire for happy or sad expressions.

Using the same principle, we can create a categorical view of the network's representation. This helps us to use our model in emotion recognition tasks. The advantage of using our model is that we can create different categorical models without re-training the network. If we want to analyze simple separations as positive and negative emotions, we can easily identify which regions of the network fire for these categories. If we want to increase the number of categories, we just have to increase the number of clusters. So, instead of finding regions that fire only for negative or positive, we can find regions that fire for happy, sad, surprised and disgusted.

To find these clusters, we use the U-Matrix to create a topological representation of the neurons and the K-means algorithm [201] to cluster them. The K-means algorithm partitions a set of observations into N clusters, based on the distance of individual observations to each other. The goal of K-means is to minimize the within-cluster sum of squares, which is be defined as

$$K = argmin \sum_{i=1}^{k} ||(c - \mu_i)|| \,, \tag{6.2}$$

where K indicates a cluster, c is one observation and μ is the mean of each observation.

The limitation of our model is directly related to the SOM architecture limitation: we have to define the number of neurons before training them, which restricts the number of expressions that can be categorized. However, with an optimal number of neurons, we can create different categories of expressions without re-training the network.

Using the expression categorization, we can use our network to recognize different emotion categories. If, at first, we want to recognize only positive and negative emotions, we just have to define two clusters. Then, if we need to identify between a happy and an excited expression, we can apply the K-means algorithm only to the region of the network which has a bigger probability to activate for these concepts. In the same way, if we want to identify different kinds of happy expressions, we can create clusters only for this specific region. Figure 6.4 illustrates the application of K-means to the network illustrated in Figure 6.2. In this example, the network is clustered for three classes: happy, sad and neutral.

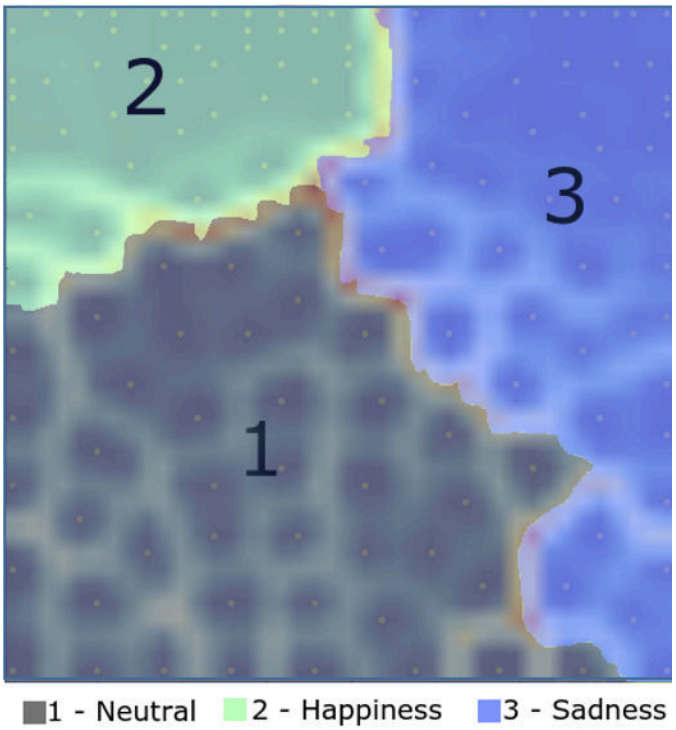

Figure 6.4: K-means algorithm applied to the SOM illustrated in Figure 6.2. We cluster the neurons into three expressions: happy, sad and neutral. We use the K-means clusters to classify expressions.

6.3 Methodology

To evaluate our expression learning architecture, we use the trained filters of the CCCNN to extract high-level expression representations and trained a series of SOMs with them. We perform three sets of experiments: the first one to evaluate the capability of the model to classify expressions and the second to evaluate the capability to learn new expressions. In the last set of experiments, we evaluate the use of the model in creating behavior analysis for independent subjects.

For all experiments, 30 experiment routines were performed and the mean of the accuracy was collected for each expression individually, which helps us to understand our model better.

6.3.1 Experiment 1: Emotion Categorization

After training the CCCNN with the multimodal expressions of the EmotiW corpus, we train a SOM and use it in a classification task using K-means to cluster the neurons in a number of specified classes. We compare the use of the SOM with the CCCNN performance for classifying crossmodal data. These experiments show us the capability of the SOM to generalize expressions.

6.3.2 Experiment 2: Learning New Expressions

In this set of experiments, we measure the capability of the SOM to learn new expressions. For this purpose, we train a SOM with a limited set of expressions, composed by only sad and happy expressions. Then, we systematically present new expressions to the SOM, such as angry, disgusted and surprised ones, and we also calculate the mean of the activation maps for each expression. This way we show the capability of the SOM to learn different expressions. For these experiments we use the FABO corpus, because it contains a controllable environment, which is not present in the EmotiW dataset.

6.3.3 Experiment 3: Individual Behavior

In the last round of experiments, we use of the SOM for analyzing the behavior of expressions. We perform experiments with the SAVEE corpus only, which contains data from four different subjects. We train one SOM for each subject and compare the differences of the expressions based on the clusters of each SOM.

6.4 Results

6.4.1 Experiment 1: Emotion Categorization

For these experiments, we trained our SOM with the emotion representation obtained by the CCCNN in of the previous chapter. We then cluster the neurons of the SOM in 7 regions with K-means algorithm, so each region represents one class of the EmotiW corpus. Figure 6.5 illustrates the clustered regions from 0 to 6, respectively: anger, disgust, fear, happiness, neutral, sadness and surprise. It is possible to see that the neutral expressions, represented by class number 5, have almost all the others expressions as their neighbor. Also, angry expressions, class number 1, are between happy, class number 4, and sad expressions, class number 6. And finally, it is possible to see that fear expressions, class number 3, are closely related to surprise expressions, class number 7. In this case, some of the fear expressions are between happy and surprise.

Using the clusters, we calculated the accuracy of the SOM in the validation set of the EmotiW corpus. Table 6.1 shows the results. It is possible to see that with the SOM clustering, such expressions as disgust and sadness show an increase of almost 7% in performance. As we see in the cluster, sad and disgusted expressions are neighboring regions, and the application of the SOM created a better separation border, which would explain the performance increase. In general we have an improvement of more than 3% in the accuracy when using the SOM.

6.4.2 Experiment 2: Learning New Expressions

In our next experiment, we trained the SOM with happy and sad expressions from the FABO corpus. We then proceed by feeding angry, disgusted and surprised

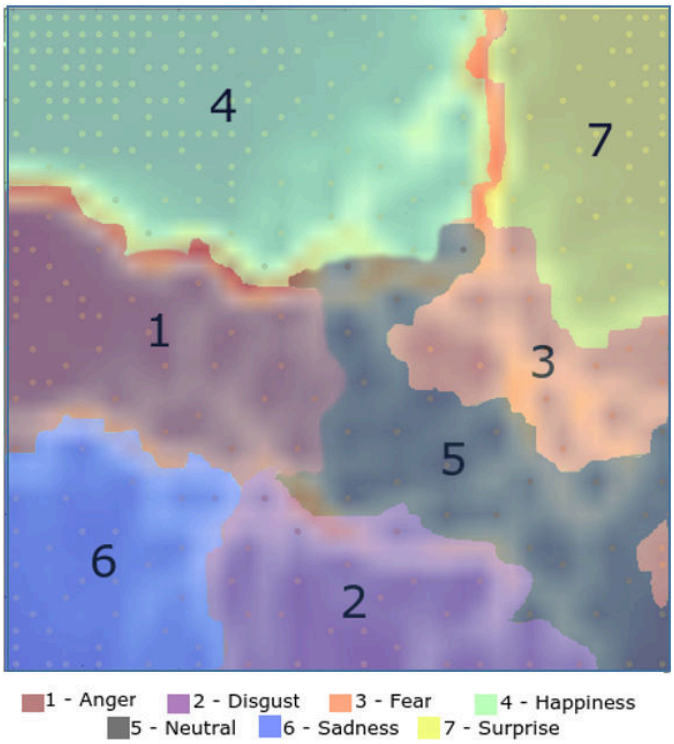

Figure 6.5: K-means algorithm applied to the SOM trained with the EmotiW multimodal representation. Six emotions were clustered: surprise, sadness, angry, happiness, fear, neutral and disgust.

Table 6.1: Mean accuracy, in percentage, for the multimodal representation in the validation set of the EmotiW corpus. The results are for the CCCNN and the SOM.

Class	CCCNN	SOM
Anger	80.3	85.3
Disgust	23.4	30.3
Fear	30.8	32.1
Happiness	81.2	82.3
Neutral	68.7	67.3
Sadness	24.5	31.7
Surprise	14.0	17.6
Mean	46.1	49.5

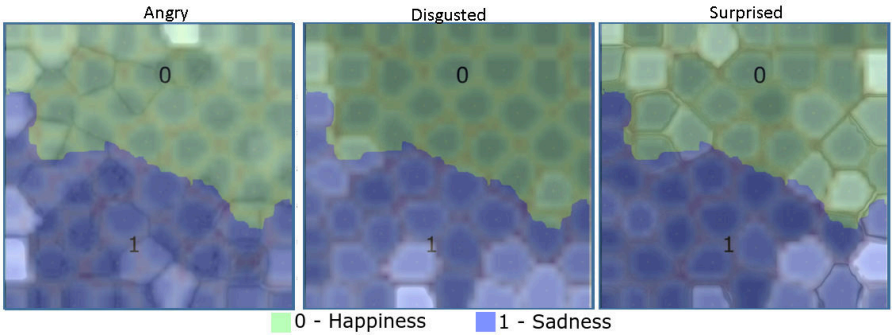

Figure 6.6: Activations plotted on top of a clustered SOM. The SOM was trained with sad and angry expressions and each activation shows the mean activation map when feeding the network with angry, disgusted and surprised expressions.

expressions to the network, and generate the mean of the activation maps for each set of expressions. Figure 6.6 illustrates the activations for each new set of expressions plotted on top of the clustered SOM. In this experiment, the network never saw angry, disgusted or surprised expressions and we can see how the neurons activate when these expressions are presented.

Angry expressions activated a mixed region of neurons, between the sad and happy regions. Two neurons had a higher activation, in both regions. This is congruent with the regions found when analyzing the EmotiW SOM, where angry expressions were represented between happy and sad. Expressions of "Disgust" were mostly activated by neurons on the sad region, which is also congruent with the cluster of the EmotiW SOM. And finally, the "Surprise" expressions were mostly activated in the "Happiness" regions, with some activation in the angry region.

We then proceeded to re-train the network on the new expression. We used the network trained with sad and happy expressions, and created four new networks, three of them trained with the addition of one new expression, and the fourth one with all five expressions. Figure 6.7 illustrates the clusters of each network. We can see that the disposition of the new clusters is similar to the activation maps of the network trained with only two expressions. That demonstrates how each emotion expression can be related to others, and our network is able to use this relation to learning new expressions.

6.4.3 Experiment 3: Individual Behavior

In the final set of experiments with the SOM, we train one SOM with expressions, represented by the Face and Speech channels, from each one of the four subjects from the SAVEE corpus, which are identified as DC, JE, JK and KL. We trained each SOM using a 4-fold cross validation strategy, only with the data of each individual subject. We then calculated the accuracy for each subject, which is shown in Table 6.2.

We separated the regions of each SOM into seven classes, and produced cluster

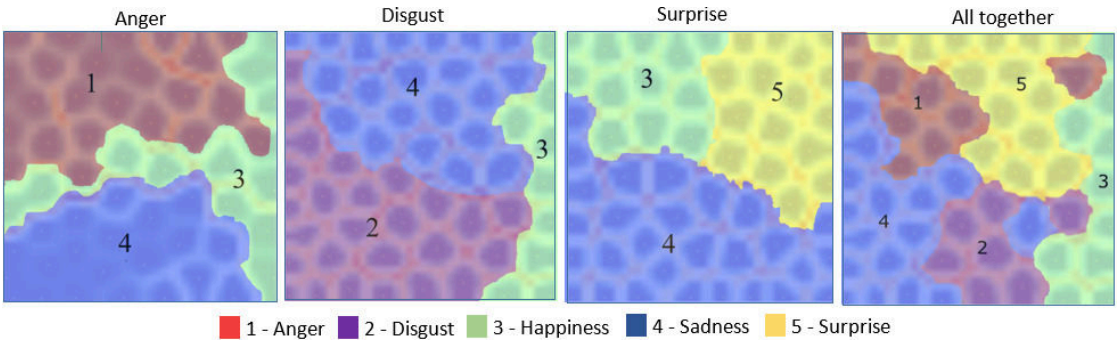

Figure 6.7: We train a network with two kinds of expressions: happy and sad. Systematically add one different expression and re-train the network. At the end, we train a network with the five expressions together.

Table 6.2: Mean accuracy, in percentage, for the auditory and visual stream channels trained with a SOM on the SAVEE corpus. The results are presented for the four different subjects: DC, JE, JK and KL.

Class	DC	JE	JK	KL
Anger	100.0	94.3	100.0	92.0
Disgust	100.0	100.0	100.0	90.9
Fear	100.0	100.0	96.7	100.
Happiness	99.4	99.1	100.	97.7
Neutral	98.3	100.0	100.0	96.7
Sadness	96.7	97.8	100.0	97.8
Surprise	100.0	100.0	97.9	98.2
Mean	99.1	98.7	99.2	98.3

images for each subject, which are illustrated in Figure 6.8. Analyzing each cluster, we can see that the same expressions have different regions for each subject. Analyzing these images, it is possible to obtain some information about how each subject expresses itself. For each subject, the same number of samples is recorded for each emotion category, so there is no bias to one expression in each subject.

Except for the network of subject JE, all others clustered expressions of "Surprise" in a neighboring region with "Happiness" expressions. On the other hand, all of them clustered "Surprise" in a neighbor region to "Angry" and "Fear" expressions. That indicates that JE "Surprise" expressions are less happy than the others. Also, the "Disgust" expression is different for each subject. Although all of them have "Disgust" expressions as a neighbor of "Sad" expressions, the other neighboring expressions are different. It is possible to see that for DC, disgusted expressions are closely related to "Angry", for JE with "Fear", JK with "Happy" and KL with "Surprise" expressions. Looking for the region that each expression takes part in, it is possible to see that JK's network clustered "Happy" expressions

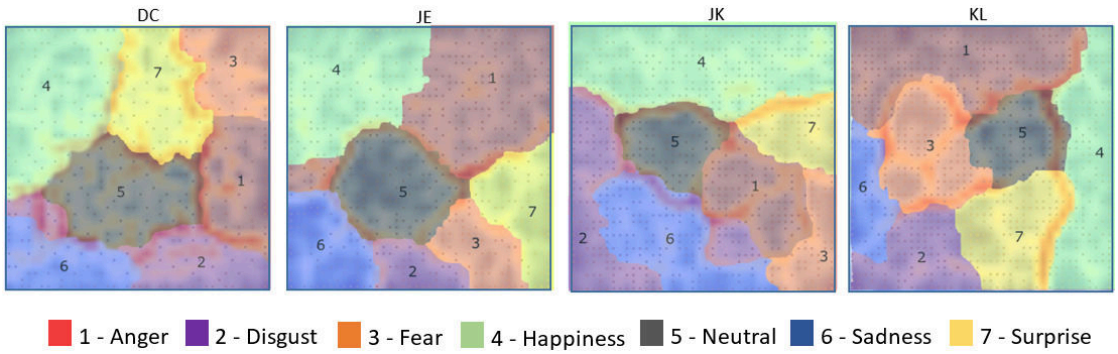

Figure 6.8: Trained networks with expressions of each subject of the SAVEE corpus: DC, JE, JK and KL. It is possible to visualize how different each subject expresses emotions by analyzing the network clusters.

with a larger region when compared to the other subjects, which could be an indication that JK expresses happiness different than the others. The same happens with JK's "Disgust" expression. On the other hand, his "Neutral" expressions have a smaller region than the others, indicating that most of his neutral expressions are very similar to one another.

6.5 Discussion

Some regions of the CCCNN code for specific features, such as face shape, eyes, mouth among others. But, once these features are related to an emotion expression, it is the role of the model, in this case of the fully connected hidden and the softmax layers to classify these features into emotions. These layers have some information about emotion expression, but they do not store any information about the expressions itself, only about the separation space. Replacing these neurons by a SOM gives the model a powerful tool to represent emotion expressions. Besides creating a more flexible separation region, the SOM allows the model itself to store information about the expressions.

6.5.1 The Prototype Expressions

Each neuron in the SOM represents a prototype of an expression, which is tuned to be similar to the data used to train the model. This means that each neuron alone codes for an expression, and neighbor neurons code similar expressions. In this way, we can simulate the spatial separation that the hidden and the softmax layers produce by clustering the neurons in different regions, giving the SOM the capability of classifying expressions. This means that a real expression has to be represented by one prototype expression in order to be classified. This showed how to improve the performance of classification tasks.

The prototype expressions also help our model to code the concept of the expression itself. While the filters on the CCCNN code for specific features from

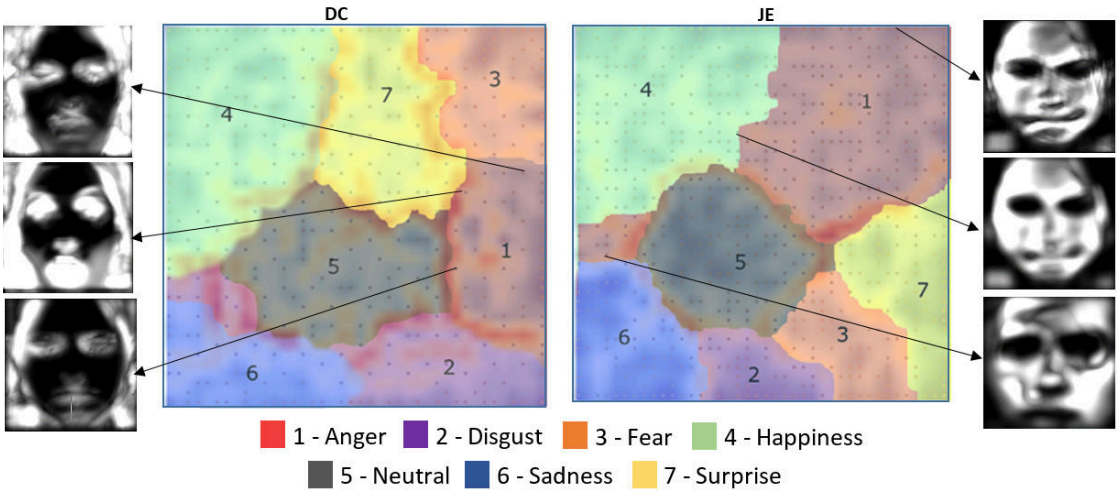

Figure 6.9: Visualization of the neural emotional representation for two subjects, DE and JE, of the SAVEE corpus. It is possible to see how neurons which are closer to different regions of the network, store different expressions.

the input stimulus, each group of neurons in the SOM code for similar expressions, giving our model a complete representation of the emotional expression, from the input stimuli to the expression representation itself.

6.5.2 Emotional Concept Representation

We can actually use the visualizations to gain insight into what expressions the model learns. When visualizing an input, we backpropagate the responses that the input produced by our filters, but by using the prototype neuron representation instead of the image representation, we can visualize which expression this neuron learned. By doing that for several images and several neurons, we can actually identify how these expressions change through the network, which helps us to understand the clusters of the SOM and the network representation itself.

Taking as an example the network trained for each subject of the SAVEE corpus, we can visualize the expressions learned by each neuron. Figure 6.9 illustrates some neurons of two subjects which are in the same region and correspond to angry expressions. It is possible to see that both networks have different representations for the angry expressions, depending on where the neurons are. For DC, it is possible to see that an expression closer to the fear region, produces a different mouth shape than the one closer to the surprise region. And for JE it is possible to see that all three representations have different eyes and mouth shapes.

6.6 Summary

In this chapter, we extended the Cross-channel Convolution Neural Network (CC-CNN) architecture by using a self-organizing layer to learn emotional concepts.

This structure gave the network the ability to cluster different expressions into similar concepts, in an unsupervised manner. This allows the model to learn how to identify different expressions into emotional clusters, introducing the capability to learn new expressions.

Using visualization techniques, we also showed how the model learned different clusters. It was possible to see the representation that each neuron learned, in the form of prototype expressions, and how they differ from each other, showing that in our model similar expressions are represented by neighboring neurons.

Our self-organizing layer presents two key features: First, it can learn new emotional concepts using expression cluster. These characteristics allows the model to adapt its knowledge to different domains and subjects, being able to generalize the representation, or specialize it, if necessary. Second, the model uses the prototype expressions to abstract emotional concepts, creating a robust emotional representation. In our experiments, we showed that this new emotional representation improved the recognition of spontaneous expressions.

We evaluated our new model in different tasks: cluster emotions, learn new emotions and behavioral analysis. The model was able to cluster different spontaneous expressions in similar regions and expand its learned representation with new emotional concepts. Used for behavioral analysis, the model was able to define patterns on how different subjects expressed themselves, and by visualizing the network topology, it was possible to represent these differences visually.

Chapter 7

Integration of Emotional Attention and Memory

7.1 Introduction

This chapter presents our integrated model for emotional attention and memory modulation, which contributes to our model in recognition and representation tasks. The models presented in Chapter 5 and Chapter 7 are able, respectively, to learn how to represent spontaneous expressions and to create emotional clusters which indicate affective concepts. However, the application of such models in real-world scenarios implies serious restrictions. First, attention is a very important part of emotional perception, and almost none of the proposed models implement any attention mechanism. Second, as explained in chapters 3 and 2, the understanding of emotional concepts are modulated by several memory mechanisms. By implementing memory modulators to our models, we intend to make it robust to emotion concepts determination, individual subjects behavioral learning and achieve a more detailed environment perception.

Cross-modal emotional learning happens in different brain regions [208, 92], and we aim to the develop our model as inspired by neural architectures. First, we introduce a visual emotional attention model, which is inspired by the two-stage hypothesis of emotional attention [32], which states that attention modulation happens in parallel processing between the amygdala complex and the visual cortex, but after a fast-forward only processing has happened. Second, we propose a memory model based on growing neural networks, which is applied to the learning of emotional concepts. Lastly, we integrate our model in an emotional learning architecture, which takes into consideration attention and memory modulation and apply it to different human-human and human-robot interaction tasks. We evaluate our model based on three characteristics: emotion expression learning, emotional attention, and individual affective maps estimation.

7.2 Emotional Attention

The first step towards an emotional attention mechanism is to introduce a selective attention system. Although many computational models of attention were introduced in the past decades [100], none of them take into consideration emotional modulation. That means, that none of these models implement any emotional aspect of the region of interest perception, which is an important mechanism found in humans for improving spatial attention [229, 233] and emotion recognition [288]. We introduce, here, an attention model which is closely related to emotional perception.

Our model combines the idea of hierarchical learning and selective emotional attention using convolutional neural networks (CNN). Our approach differs from traditional CNN-based approaches in two factors: first, the input stimuli are composed of the whole scene, which may or may not contain people expressing emotions. Second, the network is trained to a) localize spatially where the emotion expression is in the visual field and b) identify if the detected emotion expression is interesting enough to attract the attention of the model.

We use our model to detect emotional events conveyed by face expressions and bodily movements. In this scenario, each convolutional unit learns how to process facial and movement features from the entire image. We differ from simple classification tasks by not tuning the convolutional units to describe forms, but rather to identify where an expression is located in the image. Therefore, we implement a hierarchical localization representation where each layer deals with a sub-region of the image. The first layers will learn how to detect Regions of Interest (ROI) which will then be fine-tuned in the deeper layers. Because the pooling units increase the spatial invariance, we only apply them to our last layers, which means that our first layers are only composed of convolutional units stacked together. In this section 7.2, we describe our model, starting with common CNNs and our emotional attention model. To train our network as a localization model, we use a different learning strategy based on probability density functions, which will also be explained in this section.

Being a CNN-based model, our architecture learns implicit feature extractors and by investigating the learned representations, we identified regions of the network which coded for different visual modalities. While specific filters detected the facial expression, others focused on the body movement. That approximates our model from the first visual cortex areas, and we use this knowledge to integrate our attention model with our CCCNN architecture.

7.2.1 Attentional Salience Learning Strategy

To achieve the localization processing, we build our CNN architecture in a particular way: we only apply pooling operations in the last layers. Pooling usually introduces space invariance, which is in contrast with our concept of identifying a region of interest. Therefore, pooling is only applied in the last convolutional two layers, i.e. after we have identified a possible interest region by filtering information

with the convolutional units.

We observed that the use of the shunting neurons in the last layers of our network caused an interesting effect in the learned filters. As we feed the network with a full image and expect it to output a location, we are training the filters to detect unlabeled emotion expressions. Since the first two layers have no pooling, the filters are applied to the whole image to specify where possible expressions are occurring. By applying the shunting neurons on this layer, we increase the specification of the layer and assure that it can learn, indirectly, how to distinguish between different expressions.

CNNs are usually trained using strongly labeled classes, which provides an interpretation of the features learned by the filters. However, this also implicitly guides the learning process to shape the learned filters to detect features which are specific for a classification task. As an example, the concept of training the network to detect categorical expressions such as anger, happiness, and sadness, will enforce the filters to detect features which are important for such classes, ignoring possible other categorical emotions. The biggest problem, when using such models for localization, is that a CNN will reduce the input dimension and in the process introduces robustness against spatial translation. In a localization task, the spatial translation should be maintained, as we want to find where in the image a certain structure is located. To adapt the CNN to a localization task, we train our network to create salience regions on the input stimuli. This way, we do not have specific classes, such as the ones mentioned above. The model is trained using a probability distribution (instead of a class attribution) that indicates the location of interest.

Our model has two output layers, each one responsible for describing positions of the 2D visual field. One of the output layers gives the positions on the X axis and the other on the Y axis, which means that each output layer has different dimensions. Furthermore, using a probability distribution allows our model to converge faster since the concept of having a precise position would be hard to learn. The shape of the distribution changes, e.g. depending on how close the expression is being performed with respect to the camera. Using a probability distribution as output also allows us to identify other interesting regions in the images, allowing different expressions to be localized at the same time. Figure 2 illustrates an input sample to the network and the corresponding output distribution probabilities.

7.2.2 Attention Model

The model uses as input a sequence of 10 images, each one with a size of 400x300 pixels. That means that with a frame rate of 30 frames per second, our model localizes expressions at approximately 300ms, which is the interval of a duration of a common expression, between 200ms and 1s [83]. Our model is illustrated in Figure 7.2.

Our model implements 4 convolution layers, the first three with 4 filters and dimensions of 9x9 in the first layer and 5x5 in the second and third. A fourth convolution layer with 8 filters and a dimension of 3x3 is preceded and followed

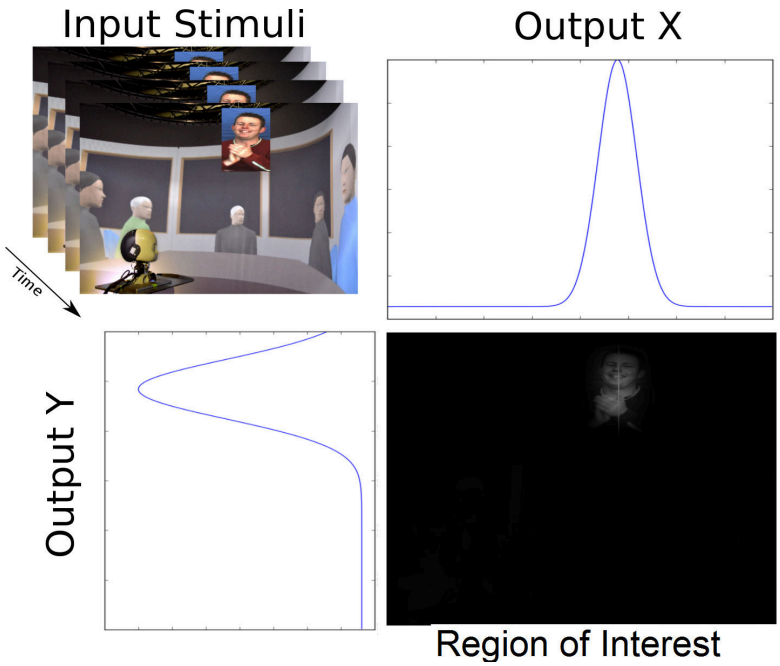

Figure 7.1: Example of output of the teaching signal. For each sequence of images, representing a scene with one or more expressions, two probability distributions are used as teaching signal, describing the region of interest by its position on the x- and y- axis, respectively.

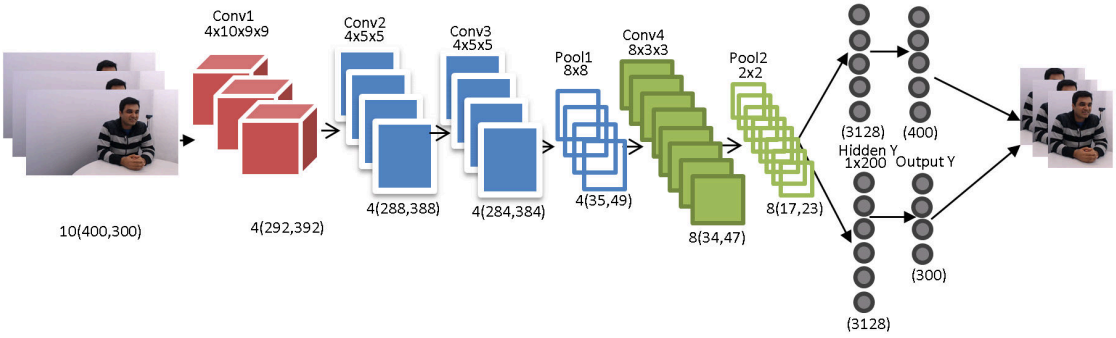

Figure 7.2: Proposed attention model. Our architecture has 4 convolution layers and only the two last ones are followed by a pooling layer. The convolution layers are fully connected with two separated hidden layers, which are connected with output units, computing the x and y location of the possible expression.

by pooling, with receptive field sizes of 8x8 and 2x2 respectively. These sizes were found to perform best based on empirical evaluations. The last convolutional layer, labeled Conv4 in Figure 7.2) implements shunting neurons. Here we use a pyramidal structure where the first layers implement four filters with large receptive field, and the last layers have 8 filters with smaller receptive fields. This is necessary because the role of the first layers is to detect, and thus filter, possible interest regions in the entire image. The larger receptive fields help filter large chunks of

the image where noise is easily filtered. The last layer applies a smaller filter, which looks for very specific information, such as facial expressions and body movements.

The convolution layers are connected to two separated fully connected hidden layers, which are then connected to two softmax output layers. We also normalize the teaching signal with a softmax function before using it for the network training. By separating the hidden layer, we assure that the same representation will be used for both outputs, but with independent connections. On the other hand, the network is trained with an error calculated from both hidden layers, meaning that each output influences the training of the convolution layers. Each hidden layer has 200 neurons and the output layers have 300 and 400 output units, one for each row and column of pixels, respectively, in the image input. The model applies L2 norm and dropout as regularization methods using momentum during the training with an adaptive learning rate. The input images are pre-processed with ZCA-whitening [49] before used as input to the network, which increases the contrast of the images and improved the final localization task.

7.2.3 Multicue Attention Stimuli

Our model is trained with a full image, without the necessity of any kind of segmentation. This means that the whole scene, including the facial expression and body movement, are processed by the convolutional filters. After analyzing the filters learned by the network, we could indicate how the network trained specific filters for specific visual stimuli: a set of filters were reacting to facial expressions and another set to body movements.

By visualizing the last layer of the network, using the deconvolution process, we are able to illustrate what the network has learned. Figure 7.3 illustrates the visualization of different filters of the pre-last layer of the network where one input stimulus was presented. It is possible to see that the network has learned how to detect different modalities. Filters 1, 2 and 3 learned how to filter features of the face, such as the eyes, mouth and face shape.

Filters 5 to 8, on the other hand, learned how to code movement. Furthermore, we can see that the network highlighted the hands, elbows, and hand movements in the representation, while other shapes were ignored. Another aspect of CNNs could be observed here: filters 5 to 8 detected partially the same movement shape, but with different activations, represented by the gray tone in the image. CNNs are known to use redundant filters in their classification process, and the movements were depicted by more filters than face expressions, mainly because the movement across the frames is the most salient feature. Filter 4, which did not pick any meaningful information, did not contribute to the localization of this expression.

7.2.4 Attention Modulation

The two-stage hypothesis of emotional attention [32] states that attention stimuli are first processed as a fast-forward signal by the amygdala complex, and then used as feedback for the visual cortex. This theory state that there are many feedback

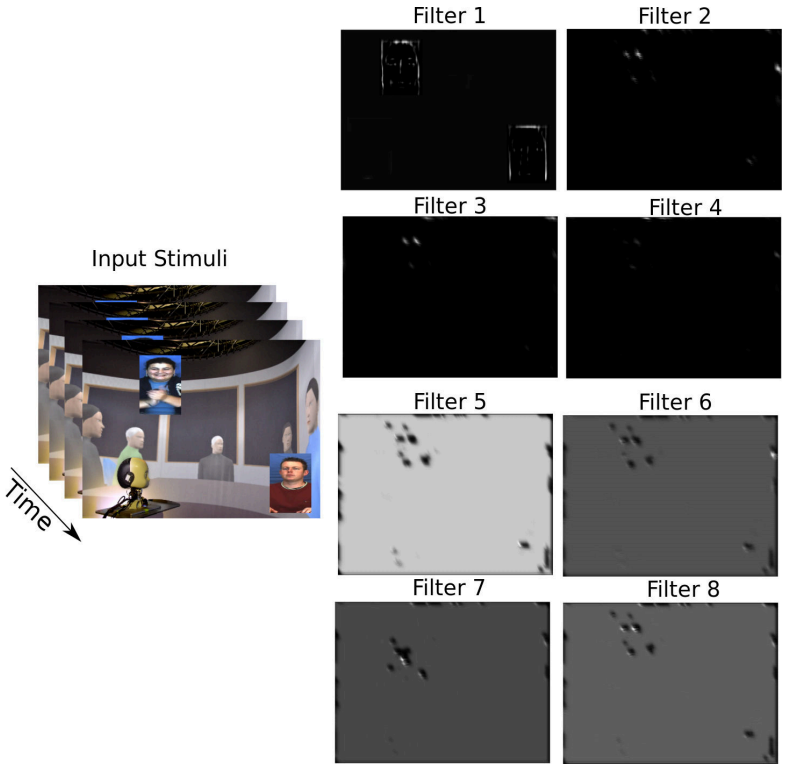

Figure 7.3: Examples of what each filter in the last layer of the network reacts to when presenting the stimuli shown on the left. It is possible to see that filters 1 to 3 focus on face features, filter 4 did not pick any meaningful information from this expression, and filters 5 to 8 focus on movement. In this figure, Filters 1 to 4 are inverted for better visualization.

connections between the visual cortex and the amygdala, however, to simplify our attention mechanism, we will use only one-side modulation: from the attention model to our perception model.

Based on this theory, our attention modulation starts as a fast-forward processing from the attention model, and then use the found region as input for the perception model. Finally, we use the use of specific features in the attention model as a modulator for the perception model. An image is fed to our attention CNN, and an interest region is obtained. However as shown in the previous section, our attention model has specific features which detect face expressions and body movements. To integrate both models, we create a connection between these filters and the second convolutional layer of the CCCNN. That means that our attention model feeds specific facial and movement features to the second layer of the CCNN. The second layer was chosen because, in the face channel, it already extracts features which are similar to the ones coming from the attention model, very related to final facial features. The movement channel still needs one more convolution layer to learn specific movement features. Figure 7.4 illustrates our final attention modulation model.

We choose the face-related features and add them to the input of the second

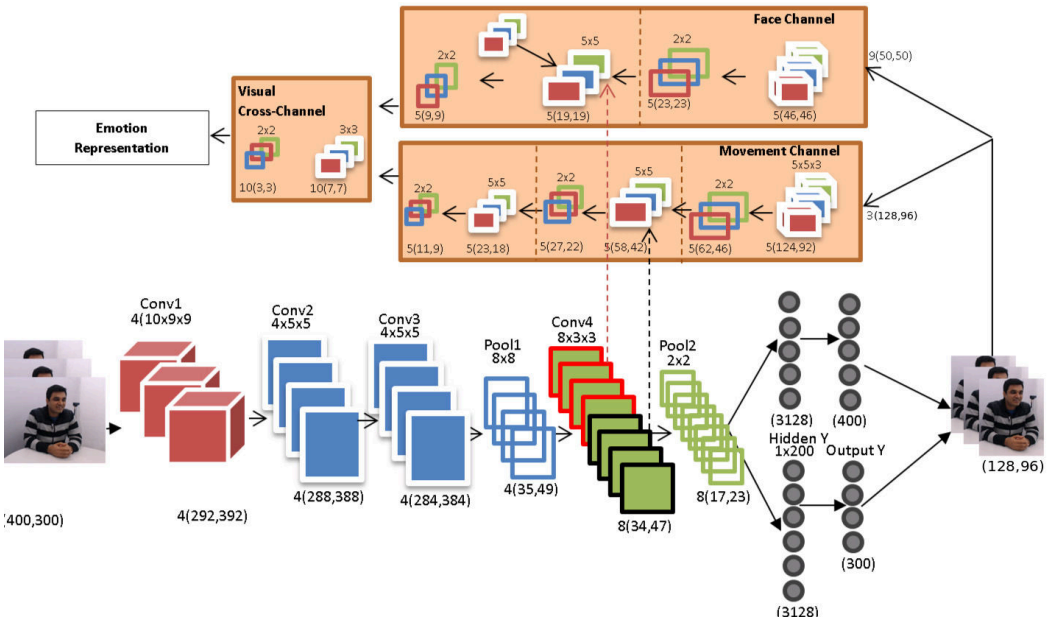

Figure 7.4: Our attention modulation model, which uses the specific features learned from the attention architecture as inputs for the specific channels of deeper layers of the CCCNN. In this picture the red dotted line represents the specific facial feature maps and the black dotted line the movement feature maps.

convolution layer of the face channel. This creates an extra set of inputs, which are correlated and processed by the convolution layer. By adding a new set of features to the input of the convolution layer, we are actually biasing the representation towards what was depicted on the attention model. The same occurs to the movement channel, and we connect only the movement related feature maps.

The new attention modulation model should be trained in parallel, with one teaching signal for each set of convolution filters. The final teaching signal is composed of the location of the face, to update the weights of the attention-related convolutions, and an emotion concept, to update the weights of the representation-related convolutions. Note that while training the CCCNN layers, we also update the specific facial and movement filters of the attention model on the fourth convolution layer. This is done as a way to integrate our representation, and assure that the attention modulation actually has a meaning on the learning from both models. We can also see this as a feedback connection, which ensures that the models can learn with each other.

7.3 Affective Memory

In a neural network, the idea of memory is usually related to the weights, and the knowledge attached to them [220]. With this concept in mind, the use of self-organizing architectures introduce a different type of memory to neural networks: instead of carrying representation about how to separate the input data, the neu-

rons in such models carry a knowledge about the data itself. In a self-organizing model, each neuron can be seen as a memory unit, which is trained to resemble the input data [165].

A common use of self-organizing neural networks is in associative memory tasks [162, 164]. In such tasks, the neurons in a self-organizing model will learn how to memorize the association between two concepts. We use a similar concept in the self-organizing layer of our CCCNN to associate auditory and visual modalities, and then generate a memory of what the network learned, grouping similarly learned concepts together. However, such model has a restrictive problem: the number of neurons affects directly what the network can learn. Also, restricting the topology of the neurons in a grid can create relations which are not present in the input data, in a way that neighboring regions may not be as closely related as the proximity indicates [135].

Emotion concepts are known to be very related to memory modulation, and thus have a strong participation on how memory is created, stored and processed. In this section, we introduce the use of growing self-organizing networks to simulate different memory stages, but also to learn and forget emotion concepts. To give our model the capability to use such concepts to improve the learning strategy, we introduce the use of a modulation system, which affects how the memory model stores and forget. Finally, we introduce the use of such a system in an emotional neural circuitry which encodes different stages of emotion perception and learning.

7.3.1 Growing Neural Memory

To address the problems one faces when using a Self-Organizing Map (SOM) we propose the update of our memory system by using a Growing-When-Required Neural Network (GWR) [206] to learn emotion concepts. Such networks have the ability to grow, by adding more neurons, in any direction. This means that the network is not restricted to a number of neurons, either by any topological structure. The GWR grows to adapt to the input data, meaning that the expression distribution which is shown to the network is actually better fitted, which produces a better-learned representation than in SOM.

The GWR gives our model three important new characteristics: it removes the limitation on the number and topological structure of the neurons, increases the capability of novelty detection, adapting to new expressions the moment they are presented to the network, and lastly, but most important, has the capability to learn and forget concepts. That means that we can use our GWR to learn how to associate different expression modalities, identify and learn never seen expressions and cluster them into new emotional concepts, and forget concepts which are not important anymore.

We first use a GWR model to learn general multimodal emotion expressions. This model represents the general knowledge of our perception architecture and is able to identify several different types of expression. We train this Perception GWR with different expressions coming from all our corpora, in a way that it produces the most general representation as possible. Figure 7.5 illustrates our

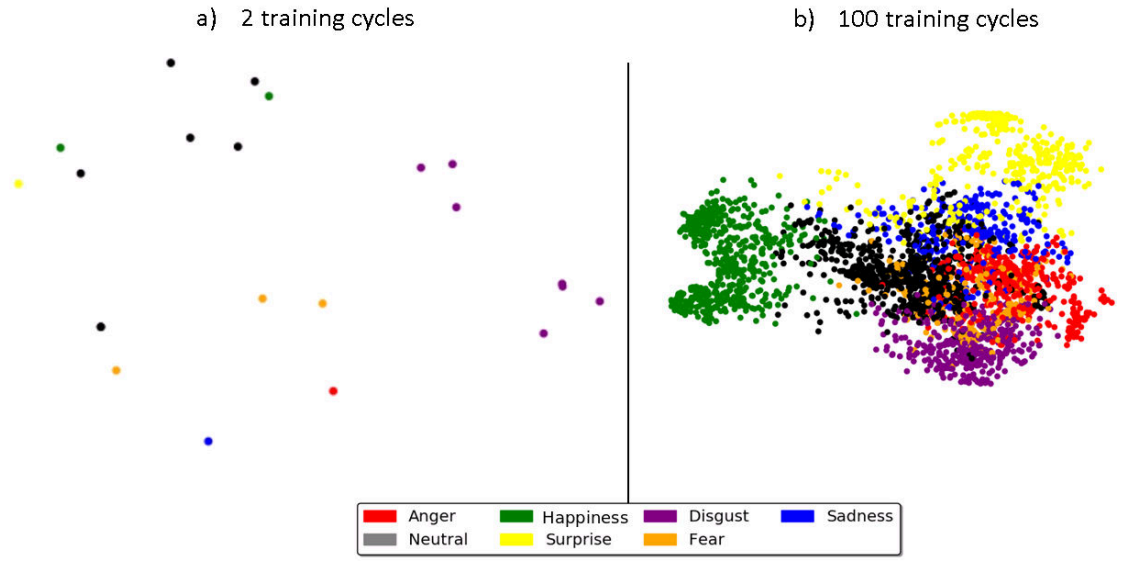

Figure 7.5: We proceed to train a Perception GWR, which will maintain our entire representation of multimodal emotion expression perception. The figure illustrates the general network trained with emotion expressions from all our corpora, in the first training cycle on the left, and after 100 ones on the right.

general network in the first interaction, on the left, and in the last interaction, on the right. It is possible to see that the network created clusters by itself, as we do not enforce any topological structure.

Training the GWR with different expressions gives us a very powerful associative tool which will adapt to the expressions which are presented to it. By adapting the learning and forgetting factors of the GWR we can determine how long the network will keep the learned information, simulating different stages of the human memory process. For example, training a GWR to forget quickly will make it associate and learn local expressions, in a similar way that the encoding stage works. By decreasing the forgetting factor of the network, it is possible to make it learn more expressions, meaning that it can adapt its own neurons topology to a set of expressions that was presented in a mid- to long-time span.

Figure 7.6 illustrates a GWR architecture used to represent an Affective Memory for a video sequence. We first proceed to use the Perception GWR to detect which expressions were been performed, and we feed this information to our Affective Memory GWR. In the beginning, represented by the topology on the left, it is possible to see that the network memorized mostly neutral concepts. However, at the end, different concepts were memorized. By changing the forgetting factor of this network, we can let it learn the expressions on the whole video, or just in one part of it.

Using the GWR we can create several kinds of emotional memory of what was perceived. By having other GWRs, with different learning and forgetting factors, we can simulate several types of emotional memory: short- and long-term memory, but also personal affective memory, related to a scene, person or object, and even

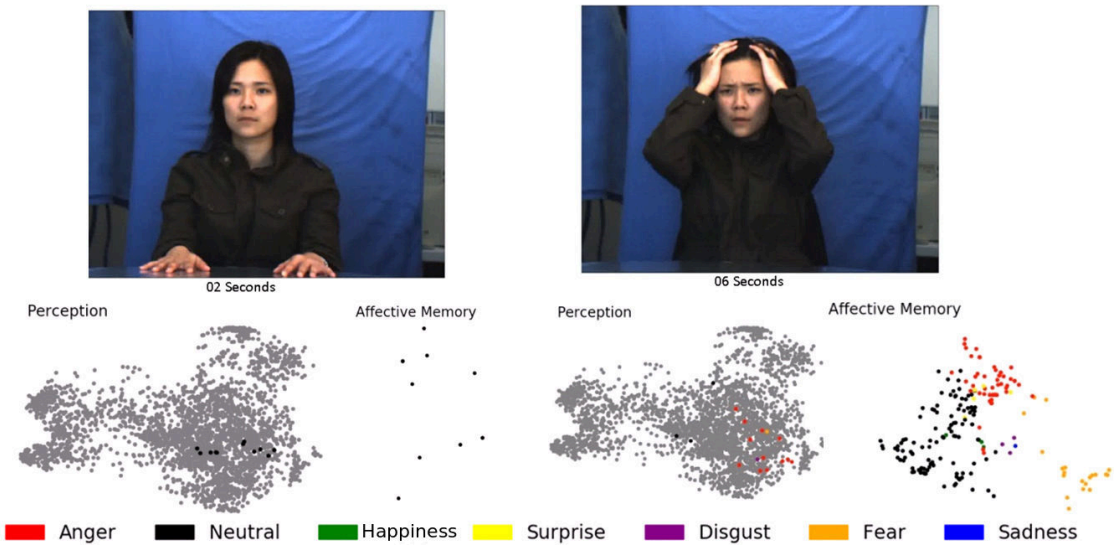

Figure 7.6: Using the expressions depicted on the Perception GWR, we proceed to train an Affective Memory GWR for a video. The network on the left illustrates the Affective Memory on the start of the video (02 seconds) and on the right at the end of the video (06 seconds). The colored dots in the Perception GWR indicate which neurons were activated when the expression is presented and the emotion concept associated with them. The colored neurons on the Affective Memory indicate which emotion concepts these neurons code.

mood. By feeding each of this memories with the Perception GWR, we can create an end-to-end memory model, which will learn and adapt itself based on what was perceived. The Perception GWR can learn new expressions if presented, and each of the specific memories will adapt to it in an unsupervised fashion.

7.3.2 Memory Modulation

Many researchers describe the mood as a representation of the internal correlation of different emotional processes [240], like hunger or fear, others as a complex behavior which is modulated by perception, attention, and memory [12]. We can also identify the mood in as part of the definition of the core affect [262] in cognitive emotions, as discussed in Chapter 2. In this sense, the mood would not affect how you perceive something, but also how you interpret the perceived expression, and how you store that as an emotional memory. In other words, the mood could be described as a medium-term memory modulator which affects and is affected by different sensory and behavioral mechanisms [218].

With the addition of such a modulator, humans have a sharper perception level in natural communication. Depending on a person's mood, he or she can show interest in different aspects of the communication, and identifying other people's mood can make you adapt the dialogue or interaction to avoid certain topics. Mood also reflects the way we perceive things, there is a consensus in the field that the

valence of our mood directly affects how we perceive certain expressions [26, 189]. This makes us more empathic towards each other, as we can adapt our perception to our general mood.

Creating a robot that integrates such modulator into its perception would make such an autonomous system capable of understanding and interpreting certain expressions better. A common problem is that automatic systems do not adapt their own representation of what was perceived, and this decreases the natural perception of the dialogue with humans, as was seen in our Human-Robot-Interaction scenario of the WTM Emotional Interaction Corpus.

We introduce, here, the use of a memory modulator, based on what was perceived to improve our models' adaptability. This modulator is implemented as a GWR network which is updated based on what the robot sees at the moment (short-term memory), and on a current mood (medium-term memory). The first updates on the current mood are basically copies from what the robot sees. However, after a certain amount of memory stored, the robot applies a modulation based on the mood's valence.

The modulation is applied as a function and calculates the amount of expressions necessary to update the memory. First, we have to identify the robots mood, based on the mean of the valences of all the neurons in its Mood Memory. Then, we calculate the modulator factor M:

$$M = \begin{cases} v_p > 0.5, & e + e.\left(\frac{1}{e^{-v_m}}\right) \\ v_p = 0.5, & e \\ v_p < 0.5, & e - e.\left(\frac{1}{e^{-v_m}}\right) \end{cases} \tag{7.1}$$

where v_p is the valence of the perceived expression, e is a constant indicating the modulator strength, and v_m is the mean valence of the memory. The modulator factor indicates the strength of the relation between the perceived expression and the Mood Memory. It will increase if the valences of the perceived expressions and memory are similar, and decrease if not.

We then proceed to update the memory using the perceived expression. To do that, we create M copies of the perceived expression and update the Mood memory with it. The forgetting factor of the Mood Memory is set to a mid-range term, meaning that as many expressions of the same type are presented, much stronger they will be remembered if an expression which is not strongly presented during the update, meaning a weak memory relation, it will generate fewer neurons and connections, and will be forgotten quickly.

The way the memory modulator is built increases the connection with expressions with the same valence but allows the memory to be updated with the opposite valence. That is an important mechanism because it allows the memory to change from a positive to a negative valence, completely based on the perceived expressions.

Applying the modulator factor to other memories could also create different modulations. For example, introducing the robot to a person which it associates

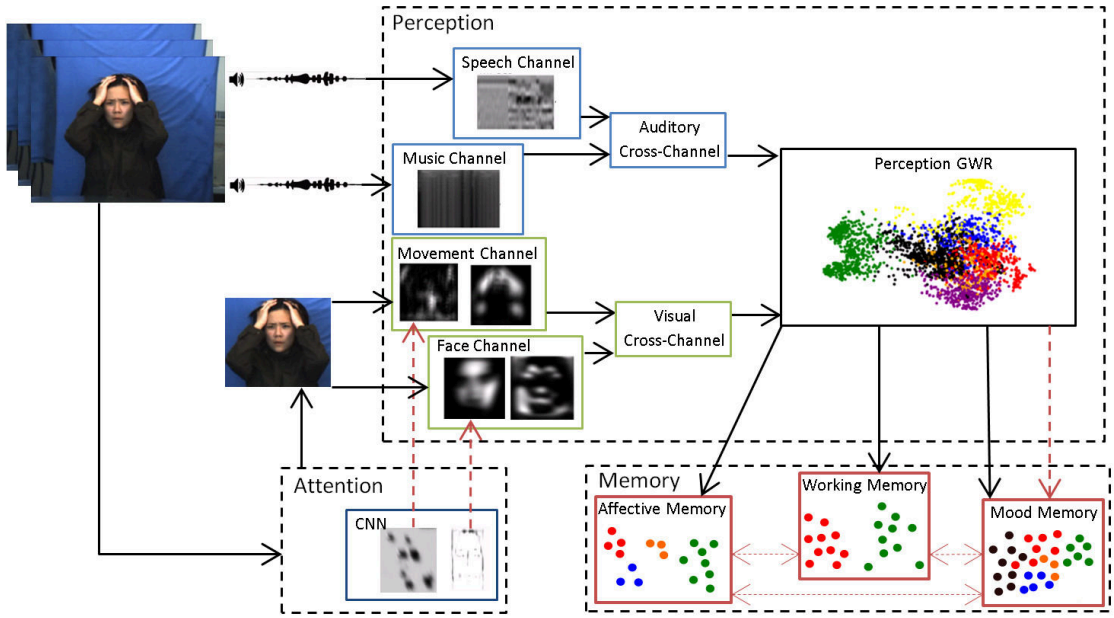

Figure 7.7: The Emotion Deep Neural Circuitry which integrates our attention, perception and memory models. The red dotted arrows indicate where modulation and feedback connection happens: mostly between attention and perception, and within memory mechanisms.

a strong positive valence memory with, could affect the Mood Memory of the robot. In the same way, if the robot has a very negative valence Mood Memory, it could affect perceive differently when a person communicates with it using negative expressions.

7.3.3 Emotional Deep Neural Circuitry

To integrate our the proposed systems and methods we propose an emotional deep neural circuitry, as illustrated in Figure 7.7. The model integrates our CCCNN with attention modulation, the perception GWR, and different memory mechanisms. With this model, it is possible to identify emotion expressions in a scene, by visual attention means, represent the expression using multimodal information, visual and auditory representation, and cluster the expression into different emotional concepts. The memory mechanisms introduce the Mood Memory, Affective Memory, connected directly to the individual subject in the scene, and a Working Memory, which can store emotion concepts from different interactions.

For each subject, a new Affective Memory model is created and only updated when that particular subject is present in a scene. This creates an individual measure for each subject, which gives us a tool to measure how that particular subject expressed themselves to the robot in a certain time-span. This memory can be a long- or mid-term memory, depending on the forgetting factor chosen for it. In this thesis, we adapted this memory to be updated during our experiments, in a way that it creates a measure within all the interactions. Such memory could

be related to the concept of affection and empathy, as it will store how a particular subject behaved while in an interaction with the robot.

Our Working Memory encodes the interaction of all subjects within which one certain type of interaction is performed. This means that it can be used to identify the robot's perception for an entire afternoon of work, or in a particular dialogue task. In the same way, could also be expanded to encode long-term memory, encoding expressions from a long time span, like days or weeks.

The Mood Memory is directly modulated by the Perception GWR and encodes the robot's own perception based on a certain range of past expressions. Our Mood Memory acts as the main modulator for all the other memories but is also modulated by Affective Memory and Working Memory. That means that if the robot is in a negative mood and interacts with a person which it relates positive expressions to in the past, the chances of the robot to change its mood towards a positive one is higher.

Our Emotional Deep Neural Circuitry is trained in different steps. First, the CCCNN and attention model are trained with pertinent data, to give us a very robust expression representation and attention mechanism. Without this strong pre-training, our model becomes weakly reliable, as all its representations are based on a robust expression representation. These two mechanisms are the ones which demand more time for training, as they implement deep neural networks and require a large amount of data to learn meaningful representations.

Our Perception GWR is pre-trained with the same data used to train the CCCNN and attention model. This gives our model very robust initial emotional concepts, however, this model can be trained online at any time. That means that our Perception GWR can learn new expressions and emotional concepts, which were not present during the CCCNN training. Lastly, our different memories are trained in an online fashion, while performing the experiments. This way, each of our memory mechanisms can learn different information from the interactions, and regulate each other using the memory modulators.

To estimate the valence, used for the memory modulators, we introduce the use of a Multi-Layer Perceptron (MLP). We use the trained CCCNN filters to represent different expressions and proceed to feed this general representation to an MLP with two hidden layers. The MLP outputs arousal and valence values. Similarly to the attention training strategy, we use a Gaussian distribution for each output dimension as a teaching signal. This gives this MLP the capability to identify a distribution-based information about the arousal and valence of each expression. In a similar way, we use the MLP used during the CCCNN training to classify categorical expressions into the six universal emotions concepts [81]. Both MLPs help us to identify what our network is depicting in a high-level abstraction, approximating a model of human knowledge via the internal representation of our networks.

7.4 Methodology

To evaluate our model we proceed with three different set of experiments: one to evaluate the emotion architecture and the use of attention modulation in our CCCNN model. The second set of experiments evaluates the use of our Affective and Working memory to map emotion expression scenarios. Finally, our third experiment evaluates the impact of the memory modulation on the perception of different emotion expression scenarios.

For all the models, the CCCNN architecture was pre-trained with a combination of all the corpora presented in this thesis, on Chapter 4, in order to learn robust representations for spontaneous expressions. We train our arousal and valence MLP with the KT Emotion Interaction Corpus and our categorical emotion MLP with data from all the other corpus presented in this thesis. This allows us to understand the knowledge of our model and to create a comparable and repeatable experiment.

7.4.1 Experiment 1: Emotional Attention

To evaluate our emotional attention model and modulator we use two training strategies: first, we evaluate the attention only model, and later we evaluate the effect of the attention modulation in the expression recognition task.

To evaluate the emotional attention by itself, we train our model in an emotion-attention scenario. That means that if no emotion is present, the output of the model should be no ROI at all, thus the probability distribution for the "no emotion" condition is always 0 for the whole output vector. If one expression is present, we want the model to trigger higher attention to displayed expressions, meaning that happy expressions should have a higher probability distribution with respect to neutral ones. For this purpose, when a neutral expression was present, we penalized the distribution by dividing the probabilities by 2. This still produced an attention factor to the neutral expression, but in a smaller intensity than the one present in the happy expression. Finally, when both expressions were present, the probability distribution was split into 1/3 to the neutral expression and 2/3 to the happy expression.

We perform our experiments on two different corpora: the Emotion Attention corpus and the WTM Emotional Interaction corpus. For both, we perform three different experiments: one with only the facial expressions, one with only the body motion and one with both. For the face expression experiments, we pre-processed the training data by using only the face of the expressions. For the body movement experiments, we remove the face of the expression and we maintain the whole expression for the experiments with both stimuli. We performed each experiment 10 times with a random selection of 70% for each category data for training and 30% for testing.

For the WTM Emotional Interaction corpus, we assume three different expressive states: no expression present, mainly no person in the scene, a neutral expression or an expressive one. An expressive expression could be any expression

which was not a neutral one.

We used a top-20 accuracy rate measurement. This error rate describes whether the top-20 activations in one feed-forward step match with the top-20 values of the teaching signal. We counted how many of the top-20 activations match and then averaged it. The value of 20 activations translates to 20 pixels of precision, which was the mean of the size of expression images. Finally, to help us understand how the model learns to detect expressions, we applied a de-convolution process to visualize the features of each expression and analyzed the output of the network for different inputs.

To evaluate the emotion-attention modulation impact, we perform a multimodal expression recognition experiment on the FABO corpus and on the KT Emotional Interaction Corpus, both described in Chapter 4. To evaluate the expression recognition, we follow the same protocol as presented in Chapter 5. For the FABO corpus, the expressions are recognized using the 11 emotional labels, while the WTM Emotional Interaction corpus is evaluated using the six universal expressions. We selected only the video clips which had more than 2 annotators agreeing with the same expression. We perform each experiment 30 times, each time chose the dataset for training (70%)and testing (30%) and take the mean of the accuracy.

7.4.2 Experiment 2: Affective Memory

To evaluate our memory models we perform experiments using our Emotional Deep Neural Circuitry without the use of memory modulation. This way, we can evaluate the use of the growing networks to code expressions with a direct relation to what was expressed. For this scenario we use the KT Emotion Interaction Corpus in two different tasks: evaluate the memory for individual subjects and for individual topics. Each task is evaluated with data from both Human-Human Interaction (HHI) and Human-Robot Interaction (HRI) scenarios.

The evaluation is done by presenting all the videos to the model, and letting each memory model learn without restriction. Each neuron on the memories will code one expression representation, and we proceed to use our MLPs for creating a valence/arousal and an emotion concept classification. At the end, we have one Affective Memory for each subject, and one Working Memory for each topic.

We proceed to calculate the intraclass correlation coefficient between the neurons in each memory and the annotators opinion on each of the subjects and each of the topics. We then calculate the mean of this correlation as a measure of how far the network memory was from what the annotators perceived.

Integrating the memory modulation in our architecture leads to a different perception and memory encoding. To show how the modulation affects each memory, we proceed with two analysis experiments: the first one with only mood modulation and the second one with mood modulating the Affective Memory.

We perform both analyses using the same schema as in the previous experiment: for two subjects and three of the topics in both HHI and HRI scenarios, we proceed to evaluate how each subject performed for each of the topics. This way

Table 7.1: Reported top-20 accuracy, in percentage, and standard deviation for different modalities and numbers of expressions presented to the model when trained with the Emotional Attention corpus.

-	Top-20 accuracy		
No Expression	75.5% (3.2)		
	Face	Body Movement	Both
One Expression	87.7% (1.5)	85.0% (2.3)	93.4% (3.3)
Two Expressions	76.4% (2.7)	68.2% (1.4)	84.4% (3.4)

we can evaluate how the development of the dialogue session affected the robot's perception.

7.5 Results

7.5.1 Experiment 1: Emotional Attention

Our results for the Emotional Attention corpus are shown in Table 7.1. It is possible to see that when no expressions were present, the model top-20 error was 75.5%. When no expression was present, the model should output a distribution filled with 0s, showing that no expression was presented.

After evaluating the model with one expression, we can see how each modality performs. If only the face was present, the top-20 accuracy was 87.7%. In this case, the presence of facial expression was effective in identifying a point of interest. In this scenario, the use of happy or neutral face expressions led to the strong assumptions of interest regions by the network. When evaluating only body movement, it is possible to see that there was a slight drop in the top-20 accuracy rate, which became 85.0%. This was likely caused by the presence of the movements in the image, without identification or clarification of which movement was performed (either happy or neutral). Finally, we can see that the best results were reached by the combination of both cues, reaching a top-20 accuracy of 93.4% and showing that the integration of facial expression and body movement was more accurate than processing the modalities alone.

Presenting two expressions of two persons being displayed at the same time to the network caused a drop of attention precision while making clearer the relevance of using both modalities. Using only facial expressions reached a top-20 accuracy of 76.4% while using only body movement had 68.2%. In this scenario, the identification of the type of expression became important. The distinction between happy and neutral expression represents a key point, and the facial expression became more dominant in this case. However, when presenting both modalities at the same time, the model obtained 84.4% of top-20 accuracy. This shows that both modalities were better in distinguishing between happy and neutral expressions

Table 7.2: Reported top-20 accuracy, in percentage, and standard deviation for different modalities for the WTM Emotion Interaction corpus.

-	Top-20 accuracy		
No Expression	63.2% (1.2)		
	Face	Body Movement	Both
One Expression	96.2% (2.3)	91.0% (1.4)	98.45% (1.2)

with respect to using one modality alone.

The results in the KT Emotional Interaction Corpus are showed in Table 7.2. It is possible to see a similar behavior, where the combination of body movement and face obtained a better result, reaching 98.45%. An interesting aspect is that the "no expression" experiment achieved a low accuracy of 63.2%, mostly due to the fact that this corpus has very little data with no expression being performed.

The second round of experiments deals with the use of the attention modulation for emotion recognition. Table 7.3 shows the results obtained while training the attention modulation recognition model with the FABO corpus in comparison of the common CCCNN architecture. It is possible to see that the mean general recognition rate increased from 93.65% to 95.13% through the use of the attention modulation. Although some of the expressions presented higher recognition, expressions such as "Boredom", "Fear" and "Happiness" presented slightly smaller accuracy. A cause for that to happen is that these expressions probably presented hand-over-face or very slight movements, which were depicted by the CCCNN but ruled out by the attention mechanism.

Evaluating the CCCNN with and without the attention modulation produced the results showed on Table 7.4. It is possible to see that the attention mechanism increased the accuracy of expressions as "Fear", "Happiness", and "Sadness" more than the others. Mostly this happens because this expression presents a high degree of variety in the dataset, which is easily perceived by the attention model, and sent to the CCCNN as an important representation.

7.5.2 Experiment 2: Affective Memory

Training our Emotional (deep) Neural Circuitry without memory modulation gave us a direct correlation between what was expressed and the stored memory. This means that the memory learned how to create neurons that will code for the presented emotional concepts. We fed all the videos of the HHI and HRI scenario to the model and proceed to calculate the interclass correlation coefficient per subject and per topic between the network representation and each of the annotator's labels.

The interclass correlation coefficients per topic for the HHI scenario are presented in Table 7.5. It is possible to see high correlations for both dimensions, valence, and arousal, for at least two scenarios: Lottery and Food. These two

Table 7.3: Reported accuracy, in percentage, for the visual stream channels of the CCCNN trained with the FABO corpus with and without attention modulation.

Class	Without Attention	With Attention
Anger	95.9	95.7
Anxiety	91.2	95.4
Uncertainty	86.4	92.1
Boredom	92.3	90.3
Disgust	93.2	93.3
Fear	94.7	94.5
Happiness	98.8	98.0
Negative Surprise	99.6	99.7
Positive Surprise	89.6	94.8
Puzzlement	88.7	93.2
Sadness	99.8	99.5
Mean	93.65	95.13

Table 7.4: Reported accuracy, in percentage, for the visual stream channels of the CCCNN trained with the KT Emotion Interaction Corpus corpus with and without attention modulation.

Class	Without Attention	With Attention
Anger	85.4	89.7
Disgust	91.3	91.2
Fear	79.0	81.2
Happiness	92.3	94.5
Neutral	80.5	84.3
Surprise	86.7	87.9
Sadness	87.1	90.2
Mean	86.0	88.4

scenarios were the ones with a stronger correlation also within the annotators, and possibly the ones where the expressions were most easily distinguishable for all the subjects. The emotion concept correlation also shows a strong agreement value for these two scenarios, while showing a slight disagreement on the School scenario.

The correlation coefficients for the HRI scenario are presented in Table 7.6. It is possible to see that, similarly to the HHI scenario, the topics with the highest correlation were Lottery and Food, while the lowest ones were Pet and Family. Here the correlation values are slightly smaller than in the HHI scenario, indicating

Table 7.5: Interclass correlation coefficient of our Emotional (deep) Neural Circuitry per topic in the HHI scenario.

Characteristic	Lottery	Food	School	Family	Pet
Valence	0.65	0.64	0.41	0.67	0.57
Arousal	0.67	0.72	0.42	0.56	0.49
Emotional Concept	0.84	0.71	0.47	0.52	0.53

Table 7.6: Interclass correlation coefficient per topic in the HRI scenario.

Characteristic	Lottery	Food	School	Family	Pet
Valence	0.78	0.67	0.31	0.58	0.47
Arousal	0.72	0.61	0.57	0.49	0.42
Emotion Concept	0.79	0.75	0.62	0.51	0.57

Table 7.7: Interclass correlation coefficient of our Emotional (deep) Neural Circuitry per subject in the HHI scenario.

Session	2		3		4		5	
Subject	S0	S1	S0	S1	S0	S1	S0	S1
Valence	0.63	0.54	0.67	0.59	0.69	0.67	0.54	0.59
Arousal	0.55	0.57	0.67	0.59	0.67	0.60	0.57	0.53
Emotional Concepts	0.79	0.67	0.74	0.79	0.61	0.74	0.67	0.59
Session	6		7		8			
Subject	S0	S1	S0	S1	S0	S1		
Valence	0.57	0.61	0.64	0.61	0.49	0.68		
Arousal	0.54	0.61	0.50	0.87	0.71	0.84		
Emotional Concepts	0.68	0.87	0.68	0.63	0.64	0.76		

that for these expressions were more difficult to annotate, which is reflected in our model's behavior.

The correlation coefficients calculated on the Affective Memory for the HHI scenario are showed in Table 7.7. Here, it is possible to see that for most of the subjects, the network presented a slightly good correlation, while only a few presented a very good one. Also, it is possible to see that the correlations obtained by the emotion concept were again the highest.

For the subjects in the HRI scenario, the correlation coefficients are presented in Table 7.8. It is possible to see that, different from the HHI scenario, the correlation between the model's results on recognizing emotion concepts and the annotators

Table 7.8: Interclass correlation coefficient of our Emotional (deep) Neural Circuitry per subject in the HRI scenario.

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9
Valence	0.58	0.42	0.67	0.59	0.42	0.80	0.45	0.61	0.78
Arousal	0.52	0.62	0.57	0.60	0.67	0.62	0.59	0.48	0.58
Emotional Concept	0.74	0.57	0.62	0.61	0.57	0.59	0.57	0.69	0.72

are rather small, showing that for this scenario the network could not represent the expressions as well as in the HHI scenario. The other dimensions show a similar behavior, showing that our Affective Memory can learn and represent subject behaviors in a similar manner as the annotators perceived them.

7.6 Discussion

In this chapter we presented two different emotional modulators: attention and memory. These modulators were combined with our perception and representation models, and we introduce our Emotional Deep Neural Circuitry. This model was used in different tasks, from emotion expression recognition to human behavioral description, and interaction analysis. Although we showed how the model behaved in this tasks, a deeper analysis on the model's representation helps us to understand better the behavior of our architecture.Thus, the sections below discuss two properties of our model: different emotional attention mechanisms, and the effect of memory modulation.

7.6.1 Emotional Attention Mechanisms

In our emotional attention model, we are using probability distributions based on image regions for the target output values. During the training, do not rely on a discrete teaching signal such as target labels that shape the filters into finding determined structures, as in traditional CNN-based approaches.

Instead, the model implicitly learns that a determined structure is in the region of interest, in agreement with the teaching signal. Because of that and the challenge of the localization task, our model takes a longer time to train, requiring more than a common classifier. On the other hand, the structures that the model learns to filter are not strongly labeled, and by visualizing this knowledge using the deconvolution process presented in Chapter 4, we are able to identify how the network learned to localize expressions and to identify different expressions.

Another interesting aspect of our architecture is how it behaves when different combinations of stimuli are presented. Our teaching signal determines the locations where expressions are displayed, but it also provides information on which expression to focus on. We showed this in a scenario with only one expression being

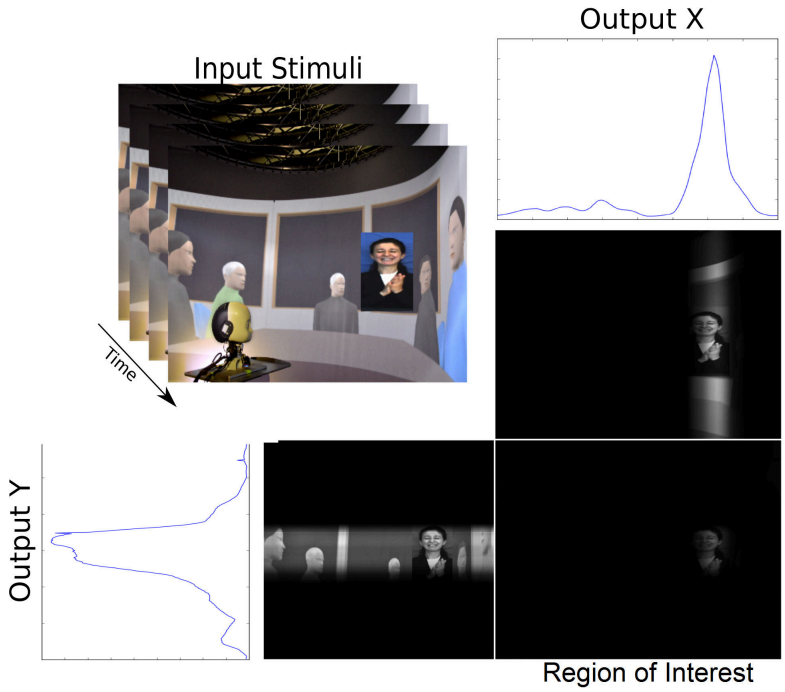

Figure 7.8: Example of the output of the model, when one "Happy" expression was presented. The input image values were multiplied by the network output values along the corresponding axes resulting in the illustrated region of interest.

displayed (Figure 7.8). We can see how the model detected the expression on the x- and y- axes, and how the shape of the output distribution changes depending on the size of the area displaying an expression.

In a second example, we used two expressions as input: one "Happy" and one "Neutral" expression. The output of the network changed in this scenario, and instead of a clear bump, it produced a wider distribution with a smaller bump showing where the neutral expression was located. Figure 7.9 illustrates this effect: the network correctly focuses on the happy expression.

Lastly, we evaluated the performance of the model in a scenario where two "Happy" expressions were presented, with one being more expressive than the other. In this case, the network chose to focus on the most expressive one. Figure 7.10 illustrates this process. It is possible to see on the x axis that the network has two clear bumps, but a bigger one for the most expressive expression. On the y axis, the network produced a larger bump, large enough to cover both expressions.

7.6.2 Memory Modulation

In our experiments with the Emotional Deep Neural Circuitry, we evaluated how the network behaves without the memory modulation. To evaluate how the modulation changes the learned representations, we proceed to show one full dialogue interaction to the model, containing one subject and one topic. We proceed with the evaluation on two subjects, 5_6 and 5_4, of the HHI scenario, with videos for

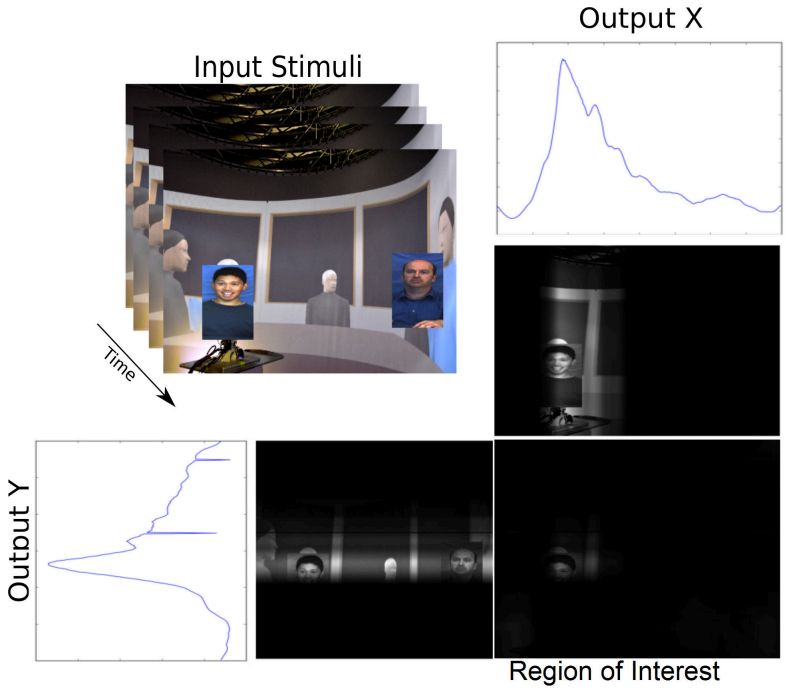

Figure 7.9: Example of the output of the model, when two expressions, "Happy" and neutral, were presented. It is possible to see that the model tends to detect both expressions, but has a stronger activation on the "Happy" one.

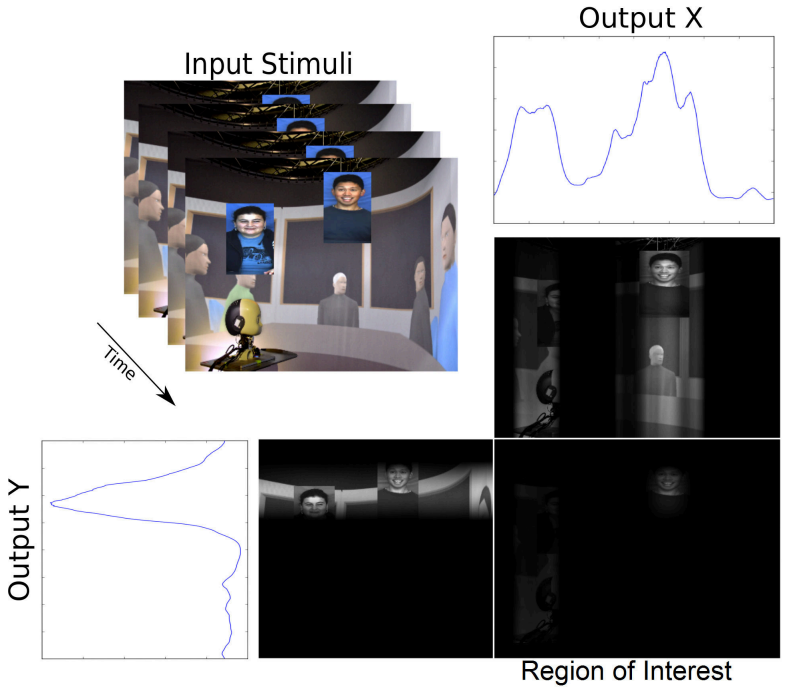

Figure 7.10: Example of the output of the model, when two "Happy" expressions were presented. The emotion which produced the higher activation peak gets primarily highlighted in the region of interest.

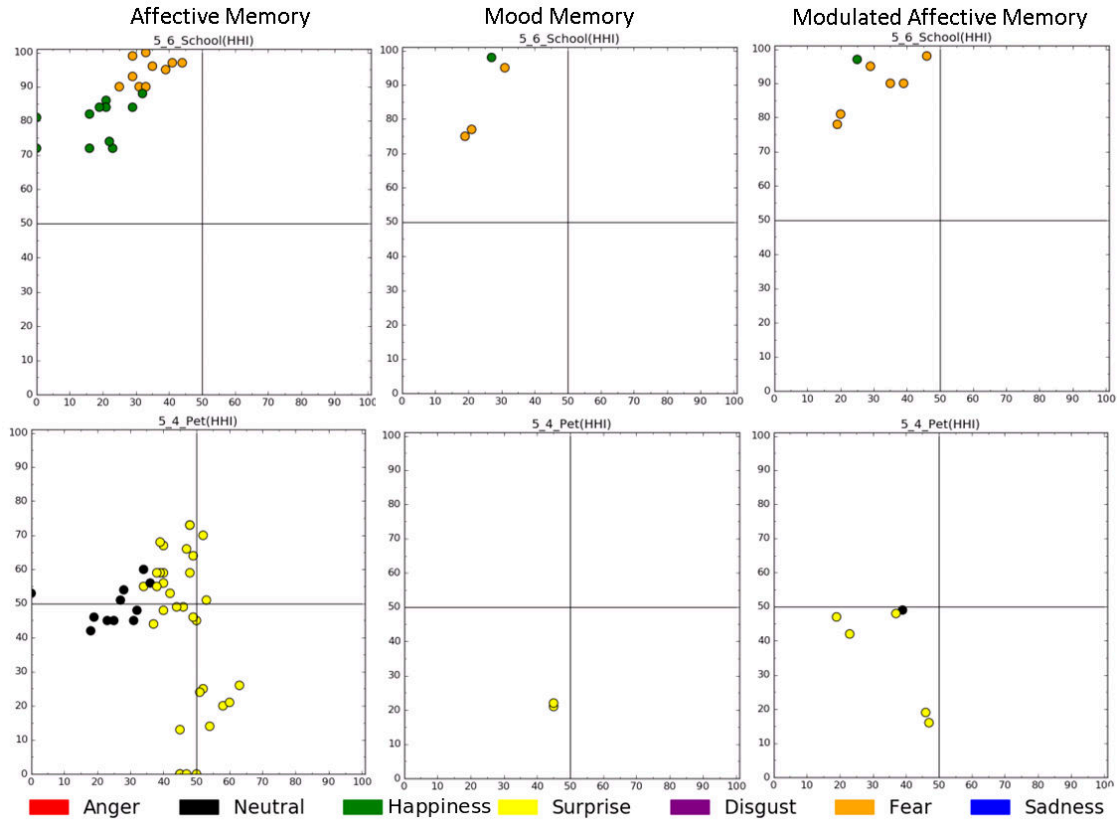

Figure 7.11: Illustration of the effects of memory modulation from the Mood Memory in the Affective Memory for two subjects in the Food and Pet scenarios.

the topics School and Pet. Figure 7.11 illustrates the arousal, valence and emotional concepts of the neurons for the Affective Memory, Mood Memory and an Affective Memory with mood modulation for this experiment.

It is possible to see that the Mood Memory contains much less information, however, code for the general behavior of the scene. On subject 5_6 it is possible to see that the Mood Memory codes information with very similar arousal, valence and emotional concepts. When used as a modulator, what happens is that the amount of information decreases drastically, however, the structure and amplitude of the three dimensions do not change much.

We proceed then to investigate only one subject, 2_7, but in two different topics in the HRI scenario: Food and Pet. It is possible to see again how the Mood Memory reduced the perceived information, while keeping the same topological structure. In the Pet topic, the Mood Memory had very little coded information, and thus had a stronger effect on the modulated Affective Memory. This happens due to the fact that this interaction was much shorter than the others, presenting to the network fewer expressions. When a larger amount of expressions are present, as in the Food topic, the network tends to behave better, as the Mood Memory has more information to update.

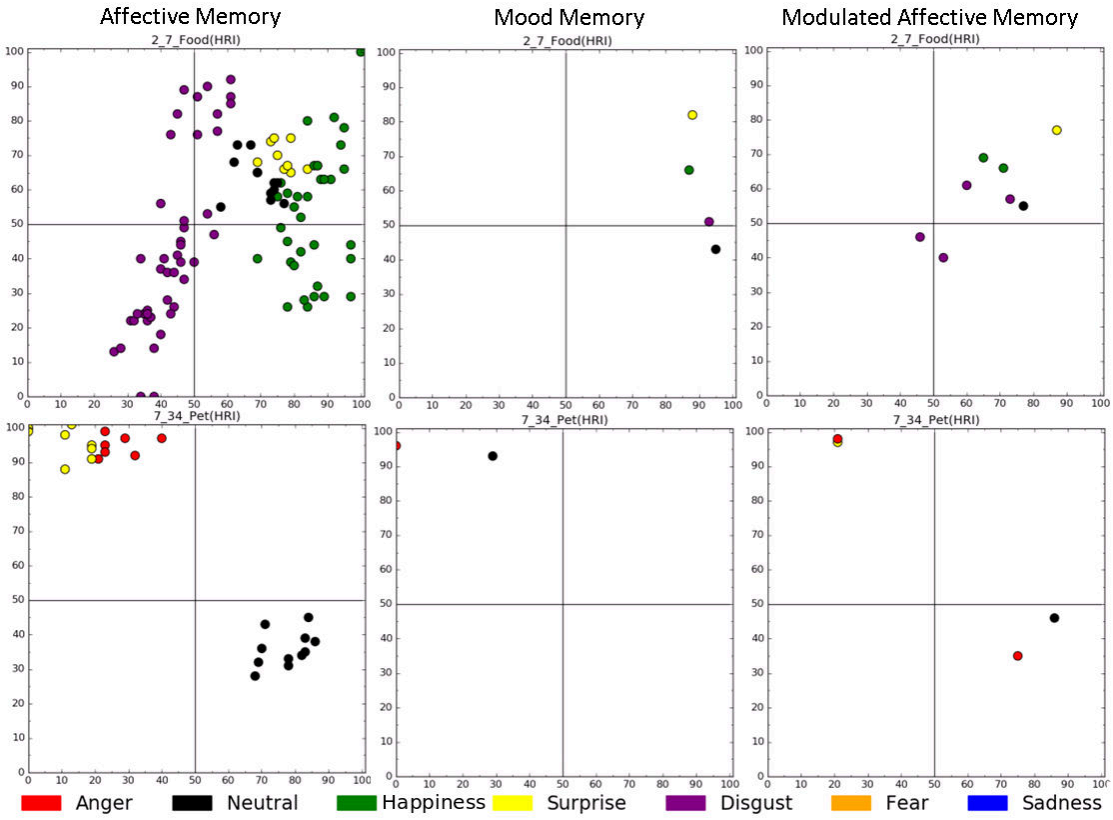

Figure 7.12: Illustration of the effects of memory modulation from the Mood Memory in the Affective Memory for two subjects in the School and Pet scenarios.

7.7 Summary

Our last model deals with the integration of different modulation mechanisms to improve the emotion concept learning and representation. The first of these mechanisms was an emotional attention model. This model introduces the use of Convolutional Neural Networks (CNNs) for localization tasks, using the convolutional filters to identify where, in an input image, an expression is being performed. By visualizing the model's internal representation, we showed that this network learned, implicitly, how to represent face expressions and body movements in different filters. We use these filters as modulators for our Cross-Channel Convolutional Neural Network (CCCNN) architecture, which improved the capability of the model to recognize visual expressions.

The second modulation mechanism proposed in this chapter was a memory model. Using Growing When Required Networks (GWR), it was possible to introduce a novel emotional memory to the network, which could be adjusted to learning and forgetting in different time steps. This allowed us to create individual memories which are related to particular persons, situations and periods of time, simulating different affective memory mechanisms.

Using the attention and memory modulations, we introduced the Emotional

Deep Neural Circuitry which integrates our multimodal expression representation model and the attention and memory modulation. Such model was used in different emotional behavior analysis task using the KT Emotional Interaction corpus, and the results showed that our model could represent complex emotional behavior from different users and situations. With this model, we were able to describe the interactions present in the corpus, and identify differences between subjects and different human-human and human-robot interactions.

Chapter 8

Discussions and Conclusions

Emotions are present in several different parts of our lives. They are present when we communicate with each other, when we learn and experience new things, and when we remember past events, for example. Our understanding of emotions is still in the beginning, and it is an interdisciplinary topic ranging from philosophy to neuroscience and robotics. In particular, the development of such concepts and ideas in autonomous intelligent systems is still an open field, with several different problems to be solved. This thesis aims to address some of these problems, by using different computational models inspired by behavioral and neural mechanisms.

This chapter shows a discussion on how each of the models presented in this thesis addresses each of the proposed research questions, limitations and future works, and the final conclusions.

8.1 Emotion Representation

The first question addressed by this thesis is: "Can a deep neural network represent multimodal spontaneous human expressions?". People express themselves differently, depending on many different factors. The same person can display happiness differently in the morning, when he sees a bird singing, or in the afternoon, when meeting with a close friend. Although the six universal emotions [81] are said to be understandable independent of cultural background, the expression itself can be very person dependent. Adding this to the fact that the perception of such expressions is a multimodal problem, involving, among others, auditory and visual systems, to represent these expressions in an artificial system is a challenging task.

Several different expression descriptor computational models were proposed and evaluated, as discussed in Chapter 3, however, most of them have severe limitations: not being able to deal with multimodal information [307], spontaneous expressions [186] or different subjects [1]. Such limitations occur mostly because the models heavily rely on very extensive feature pre-processing techniques [283] which post substantial restrictions: using the same lighting conditions, having a low number of subjects, noise-free data, among others.

To address our question, and minimize the limitations showed in the presented

research, our model, introduced in Chapter 5, implements a Cross-Channel Convolutional Neural Network (CCCNN), which can learn implicit features from multimodal data using different modality-specific channels. Our model is based on Convolutional Neural Network (CNN) [179], which were evaluated in different emotion expression recognition tasks [149, 204]. Such models showed an improvement in the capacity of dealing with spontaneous expressions because such networks learn how to represent the input stimuli, without relying on human knowledge [1]. However, these models usually need a large amount of data to learn general representations [302]. Another known problem is that such models rely on many neural connections to learn how to represent the input stimuli, and by using multimodal information the number of neurons necessary to obtain a general representation tends to increase, making it harder to train the network [274].

In the CCCNN, we designed specifics channel to process independent modality streams, in a structure that resembles the ventral and dorsal stream in the visual and auditory cortices of the brain [111]. This architecture helps to induce specific filters for each modality, reducing the number of neurons in the architecture. Each modality-specific channels use a Cross-channel layer to integrate the multimodal representation. This architecture uses the high-level feature abstraction obtained by each of the specific channels to create a multimodal representation without the necessity of using many convolutional layers. We also introduced the use of shunting inhibitory neurons in our model, in a way to force a high specialization on our face expression channel, which improved the expression representation.

In our evaluations, we showed that the CCCNN performed well in different emotion expression recognition tasks: individual visual and auditory modalities processing, multimodal streams representation, acted and spontaneous expression recognition. In each of these tasks, the CCCNN presented competitive, and in some cases even better results when compared to state-of-the-art techniques. Our network was shown to have a high degree of generalization, being able to use transfer learning principles to create robust expression representations.

By providing analyses on the CCCNN internal learned representation, we gain important insight into the networks' behavior. We could show that the network specialized specific regions of neurons which react to particular features in the input. We could demonstrate that the network had a hierarchical behavior on feature representation, showing very low-level features, such as lines and motion direction in the first layers and complex facial structures and movement patterns in deeper ones. The visualizations help us to understand why our network can generalize so well, and how it learns multimodal expressions.

Our CCCNN demonstrated that a deep neural network was able to learn how to represent multimodal expressions. The model introduced the use of modality-specific channels, shunting inhibitory neurons and Cross-channels, which made the learned features more robust than ordinary CNNs. These mechanisms made our model able to recognize multimodal spontaneous expressions with a high degree of generalization and introduced the use of implicit feature representation, extending the concept of automatic emotion recognition systems.

8.2 Emotional Concept Learning

Once we introduced our robust multimodal spontaneous expression descriptor model, we proceed to address our second question: "How to learn different emotional concepts from multimodal spontaneous expression representations?". Emotions can be represented by different concepts, as presented in Chapter 2: using categorical labels [240] or dimensional values [13], as arousal and valence. Although these concepts are used in various tasks, both can be used to represent emotion expressions, with their contextual information. This means that while the categorical labels give a universal meaning for the expression (Happy, Sad, or Surprised one), the dimensional representation tries to remove the cultural and personal context when representing it. This would make a smile be represented by different persons the same way using a dimensional approach, but interpreted as different emotional concepts based on the person's experience and background.

To approach this problem, we introduce the use of a self-organizing layer attached to our Cross-Channel Convolutional Neural Network (CCCNN), as detailed in Chapter 6. In Chapter 5, we showed that the CCCNN learned how to represent multimodal expressions in a hierarchical way, and we use this property as a basis for our emotional concepts learning model. Then, we extend the model by adopting a self-organizing layer, which creates neural clusters to represent different emotion concepts.

The CCCNN uses a feedforward connections to create a separation space and classifies different expressions into categorical emotions. Replacing these connections with a self-organizing layer [163], we give the model the capability to learn how to represent similar expressions into clusters. The self-organizing layer is trained unsupervised, which means that the network itself identifies how similar the expressions are, and represent them with neighboring neurons. These neural clusters are used as a topological representation of the network's knowledge, and they can be understood in different contextual tasks. This means that neighboring neurons can be used to represent an emotional category, such as Happy or Sad, or to identify similar expressions, such as a calm smile or an open mouth scared face.

We evaluate our model in three different tasks: emotion categorical classification, emotional concept learning, and behavioral analysis. The classification task showed that the neural clusters are more reliable in determining emotion concepts than categorical emotions-based systems, and showed better recognition and generalization capabilities. This happens because the clusters introduce neighboring neurons that represent similar input, which means that each neuron is trained to adapt its knowledge to similar expressions. As the network is trained unsupervised, the network learning process is not biased by any emotional concept and each cluster represents different expressions and not necessarily emotion categories. This gives the model the capability to learn expressions and not emotion concepts. However, the model can be used in tasks where emotion concepts are important, such as emotion recognition tasks, by selecting clusters which better represent the necessary concepts.

Our emotional concept learning experiments showed that the neural clusters

could be used to learn and represent humans interactions in different scenarios. Each neuron in the self-organizing layer acts as a prototype expression, meaning that each neuron indicates how the network represents what it was presented. We evaluate the capability of the network to learn new expressions by visualizing how different clusters were formed and which neurons were triggered. The network was able to represent different emotional concepts, and the topological structure showed to be consistent with both emotional categories representation [240] and dimensional ones [13].

We demonstrated how the network could be used on an interaction analysis task. By identifying different clusters and prototype neurons, we showed how different subjects behaved when expressing the same emotions. This visualization helped us to describe patterns of the human's behavior, identifying how different the subjects expressed differently while trying to express happiness or sadness, for example.

Our second model shows that we can learn and represent different emotional concepts, expand the idea of emotion recognition. The use of self-organizing unsupervised networks allows our model to learn from the CCCNN representations without the necessity of any label, and how to create similar emotional concepts. The evaluation of the network showed that it could learn new expressions, adapting its internal knowledge to never seen data, and it can be used in different emotional analysis tasks

8.3 Attention and Memory Modulation Integration

The last question addressed by this thesis is: "How to adapt attention and memory mechanisms as modulators for emotion perception and learning?". Attention is one of the most important modulators in human perception, and it is present in several other processing mechanisms, and emotional attention was shown to influence what and how we perceive the environment around us, as detailed in Chapter 3. Emotional memory also has an important role in modulating different processes, from perception to recognition and learning. The presence of these mechanisms enhance the way humans deal with emotional experiences, and introducing them to affective computing models could increase their robustness in representation, recognition, and generalization.

We introduce the use of an emotional attention model and different emotional memory architectures, all presented in Chapter 7. Besides their main tasks, as attention and memory models, we introduce them in a modulation architecture, which enhanced the capability of our Cross-Channel Convolution Neural Network (CCCNN) and self-organizing layer to learn, represent and recognize emotional expressions.

First, we introduced our emotional attention model, which is implemented based on Convolution Neural Networks (CNN). Usually, CNNs have expertise on

classification tasks, but not yet explored for localization and attention scenarios. We introduce the use of a CNN for a visual region of interest determination. However, we adapt it for emotional attention tasks. Our model learned, without any enforcement, how to differ face expressions from body movements, and how to use this information to determine happy and sad expressions from neutral ones.

Our attention model was able to learn how to identify different expressions, without the use of emotion categories. Also, the model could identify when more than one expression was presented, representing them with different activation intensities. That means that the model learned, implicitly, how to determine neutral expressions from happy or sad ones, and even detect when both happens together.

By visualizing our attention model, it was possible to see that it learned how to represent face expressions and body movements in different parts of the network. The hierarchical structure of the CNN allowed the model to specify different filters in describing high-level facial and movement features. We proceed to integrate this model to our CCCNN, using such filters as a modulator for our facial and movement specific channels. First, we use the region detected by the attention model as input for the CCCNN, and then we use feedback connections between the attention model's specific filters and the CCCNN channels. By doing this, we were able to improve the recognition capabilities of the CCCNN, and to present an integrated model which can detect and recognize emotion expressions in an entire visual scene.

Our memory model is based on unsupervised Growing-When-Required (GWR) networks [206]. Such networks are used to represent perceived expressions, and their ability to learn and forget learned concepts is the basis of our architecture. We proceed to create several GWRs, one representing a different type of memory with different forgetting factors. This gives our model a robustness to learn different expressions from several persons and scenarios or to create its internal representation based on specific events. This allows us to have specific emotional memories for a particular person, a place, or a period of time.

Integrating all these models, we propose an Emotion Deep Neural Circuitry. Such circuitry connects our CCCNN, with the attention modulator, with different types of memories. We introduce a memory modulator to be used to improve the model's representation by adapting it to different persons, scenarios or even the system's mood. This system was able to identify complex behavior in various human-human and human-robot interactions, and enhanced the perception capabilities of the model. Also, the use of the modulators improved the generalization and recognition capabilities of the CCCNN.

Our Emotional Deep Neural Circuitry was evaluated in several different tasks: from recognition to localization and memory storage. Analyses on how the network learned several emotional concepts from various persons and the role of the modulators showed that our model could be used in very complex emotional analyses tasks. The model integrates different aspects of emotions in an unsupervised architecture which can adapt itself in an online fashion, moving a step closer to real-world scenarios.

8.4 Limitations and Future Work

The models and ideas presented in this thesis were evaluated in several different tasks, however, to provide their use in a real-world scenario, the proposed techniques need to be refined in several directions. Currently, the model cannot be deployed in most of the robotic platforms as it needs high-end computational hardware to be trained. The amount of data necessary for a real-world generalization is enormous, and the data used during the evaluation of the model restricts its use in such cases.

The use of recurrent models for feature extraction and learning is encouraged and the study of how such models affect the learned representation would provide valuable information about the model. A deeper analysis of the auditory representations will be necessary, to understand how the model learns emotional auditory features. Adapting the model to learn from raw audio signals would be the ideal.

The self-organizing layers provide an associative relation between the visual and auditory modalities of the CCCNN. However, our current model does not deal with information conflict. Adding such mechanism would increase the robustness and the capability of the model to learn multimodal expressions.

Our attention system does not take into consideration auditory information, which could increase the model's accuracy. Similarly to the self-organizing layers, such a model could make use of conflict solving mechanisms to deal with the presence of conflicting attention information found in real world scenarios.

The analysis of our model could be expanded to several different levels. From the representation of memory, we encourage to study how the mechanisms proposed here behave when facing complex outdoor scenarios. The development of such model is not bounded by any constraint, and the update of different concepts proposed here, if beneficial to the model, is also encouraged.

8.5 Conclusions

In conclusion, this thesis contributes to the affective computing field demonstrating how different emotional concepts can be integrated. Dealing with spontaneous multimodal expression is hard, but the use of deep neural models which learn how to create better expression representations showed to be a good solution. The use of self-organizing networks to learn different emotional concepts provided the model with interesting properties, which are not common in the works on the field, such as learning of new expressions. The use of emotional attention and memory gave the proposed models a robustness on dealing with very complex behavioral scenarios, and enhanced their recognition and generalization capabilities, contributing to the affective computing field.

Appendix A

KT Emotional Interaction Corpus

This appendix shows the plots for the complete analysis on each of the topics of the KT Emotional Interaction Corpus. The first two plots, displayed in Figures A.1 and A.2 display how the annotators evaluated the behavior of the subjects while performing the dialogues of each of the topics. The analysis of how the annotators evaluated each subject behavior are displayed on the plots of Figures displayed in Figures A.3 and A.4 for HHI scenarios and Figures A.5 and A.6 for HRI scenarios.

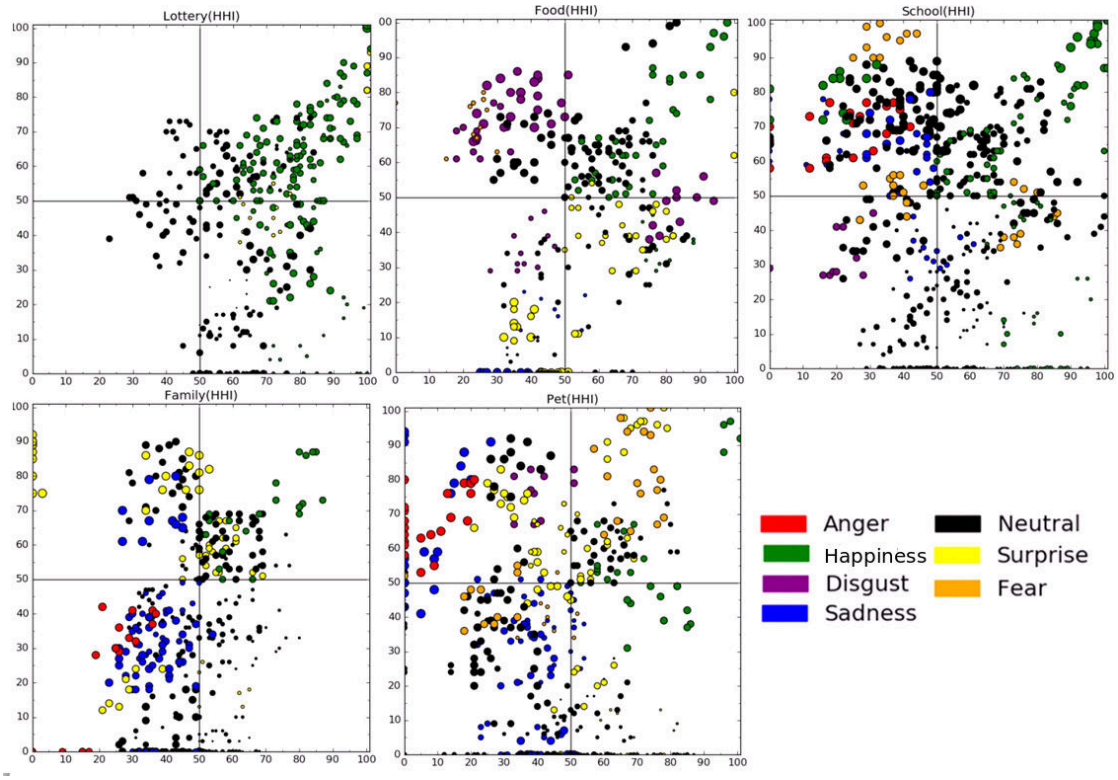

Figure A.1: Plots that shows the distribution of annotations for the HHI scenario, separated by topics. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

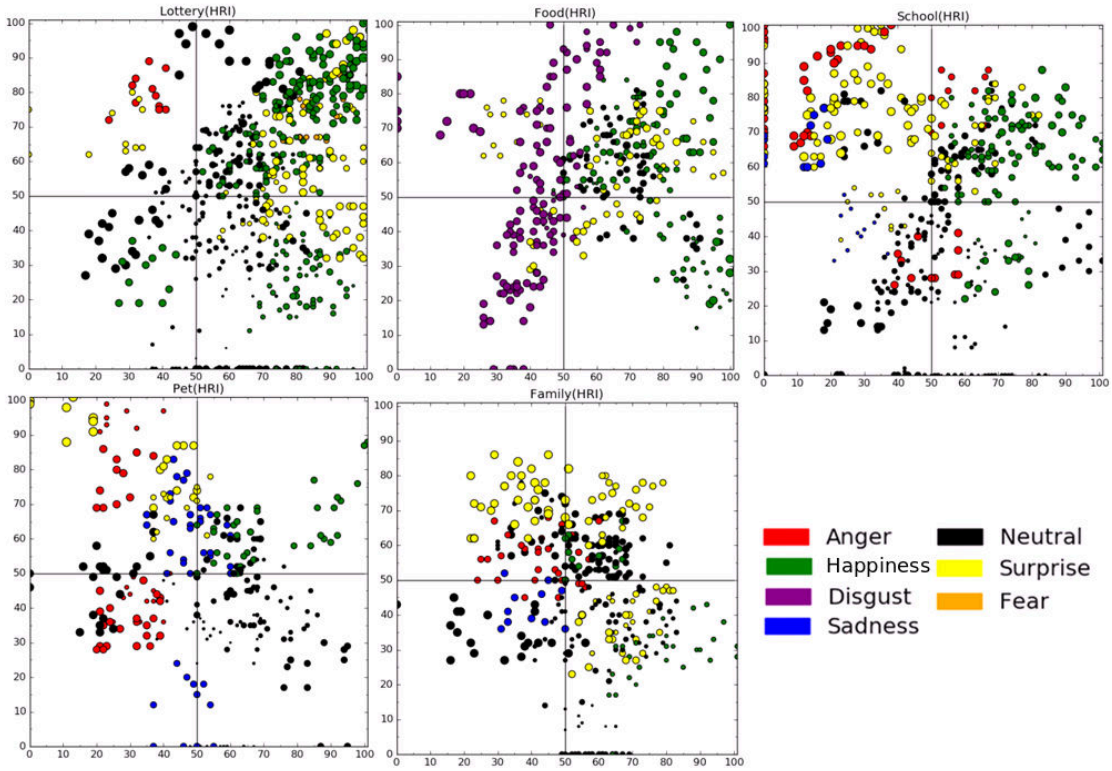

Figure A.2: Plots that shows the distribution of annotations for the HRI scenario, separated by topics. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

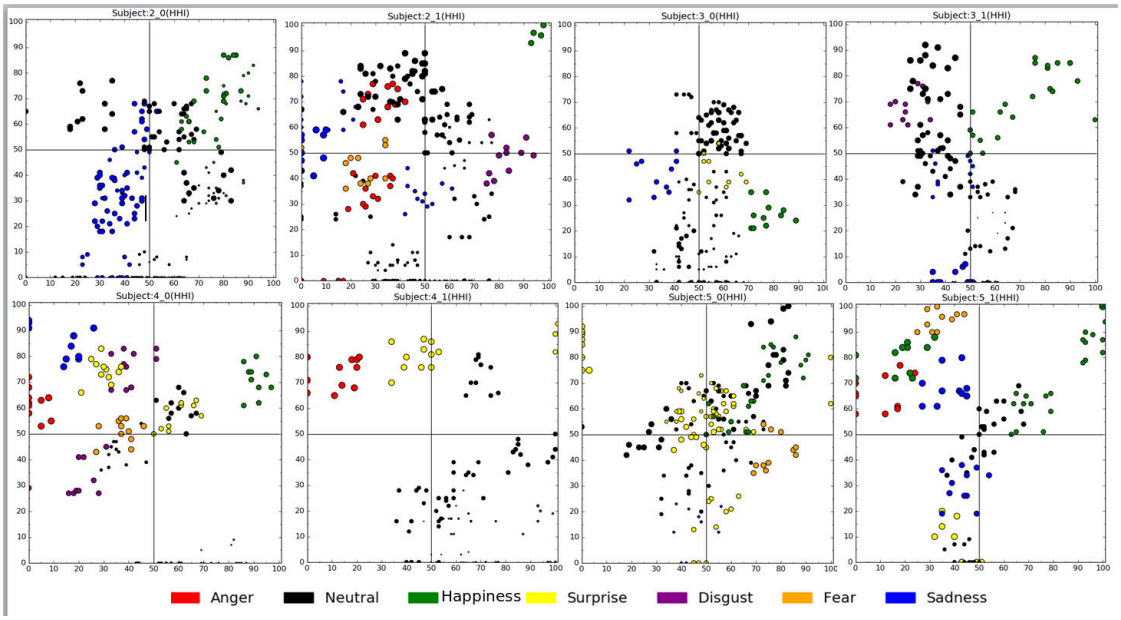

Figure A.3: Plots that shows the distribution of annotations for the HHI scenario, separated by Subjects. In this figure, the first 8 subjects are shown. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

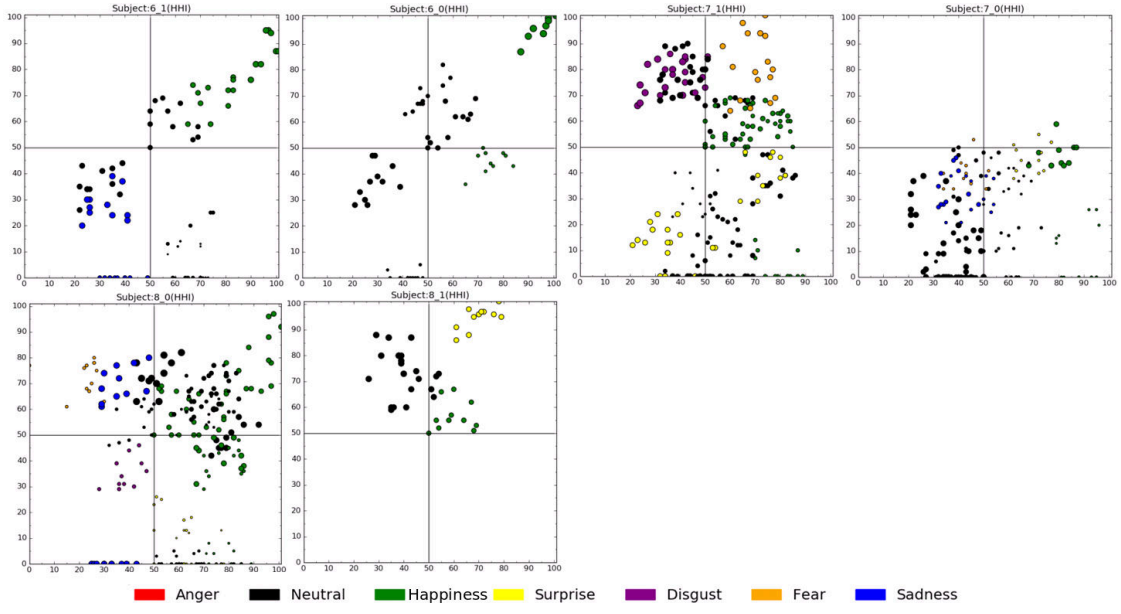

Figure A.4: Plots that shows the distribution of annotations for the HHI scenario, separated by Subjects. In this figure, the last 7 subjects are shown. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

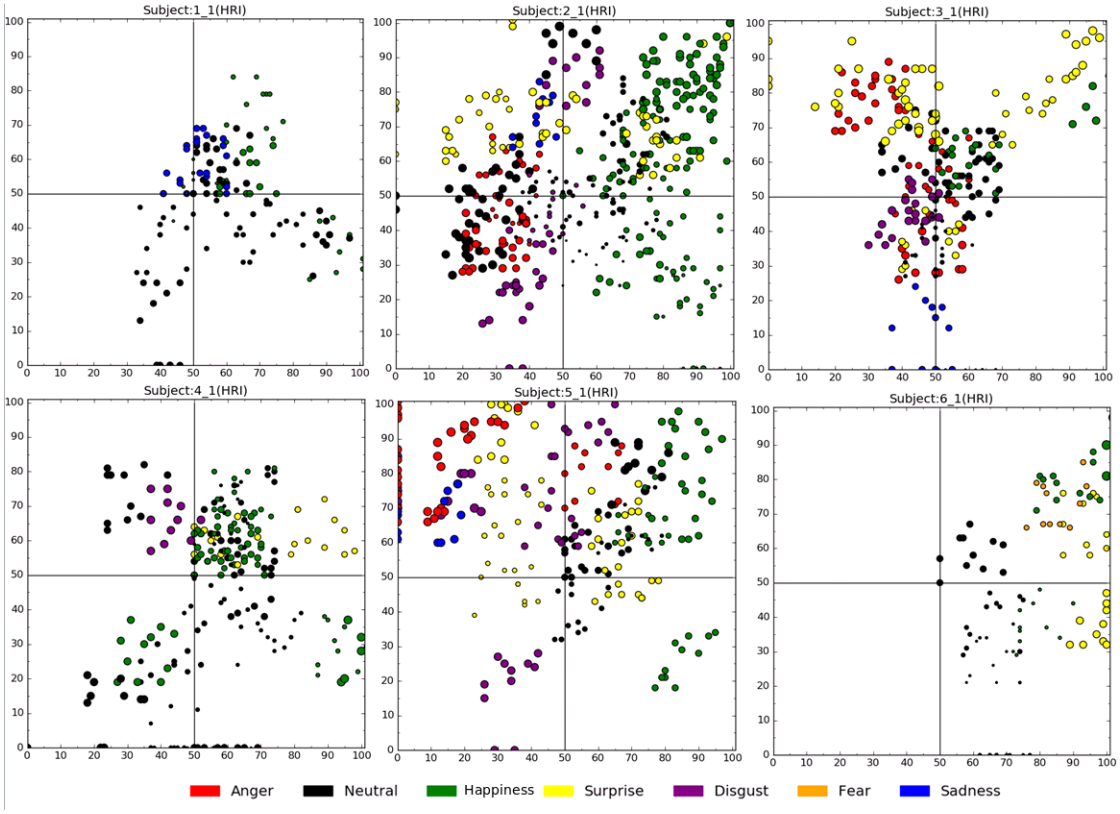

Figure A.5: Plots that shows the distribution of annotations for the HRI scenario, separated by Subjects. In this figure, the first 6 subjects are shown. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

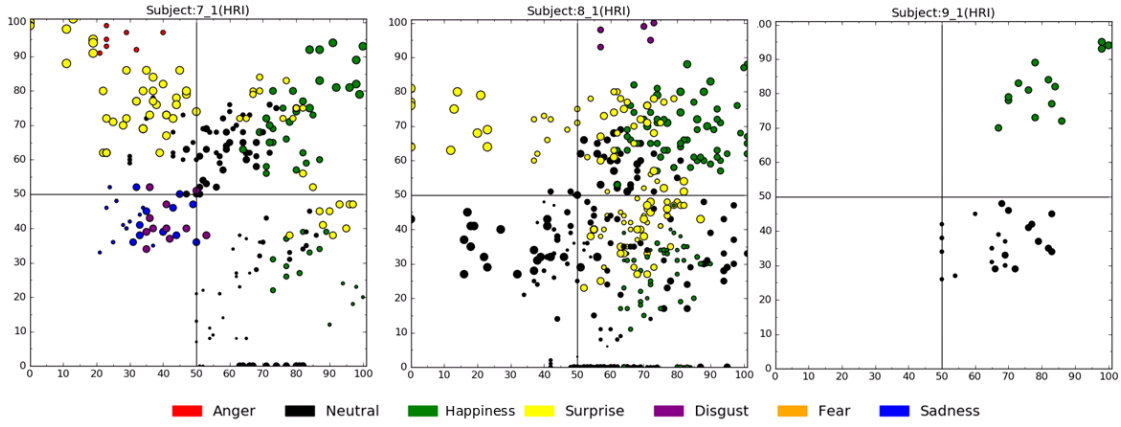

Figure A.6: Plots that shows the distribution of annotations for the HRI scenario, separated by Subjects. In this figure, the last 3 subjects are shown. The x axis represents valence, and the y axis represents arousal. The dot size represents dominance, where a small dot is a weak dominance and a large dot a strong dominance.

Appendix B

Publications Originating from this Thesis

Some of the concepts, models and experiments described in this thesis were published in different journals and conference proceedings.

- Barros, P., Wermter, S. Developing Crossmodal Expression Recognition based on a Deep Neural Model. Adaptive Behavior, Volume 24, Pages 373-396, 2016.
- Barros,P. ,Weber, C., Wermter, S. Learning Auditory Representations for Emotion Recognition. Proceedings of International Joint Conference on Neural Networks (IJCNN), pp. 921-928, Vancouver, Canada, 2016.
- Barros, P., Strahl, E., Wermter, S. The iCub Chronicles - Attention to Emotions! , Proceedings of the 10th AAAI Video Competition at the Conference on Artificial Intelligence (AAAI-16), Phoenix, USA, 2016.
- Barros, P., Jirak, D., Weber, C., Wermter, S. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. Neural Networks, Volume 72, Pages 140-151, December, 2015.
- Barros, P., Weber, C., Wermter, S. Emotional Expression Recognition with a Cross-Channel Convolutional Neural Network for Human-Robot Interaction. Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 582-587, Seoul, South Korea, 2015.
- Barros, P., Wermter, S. Recognizing Complex Mental States with Deep Hierarchical Features for Human-Robot Interaction. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4065-4070, Hamburg, Germany, 2015.
- Barros, P., Parisi, G. I., Jirak D. and Wermter, S. Real-time Gesture Recognition Using a Humanoid Robot with a Deep Neural Architecture. Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 83-88, Spain, 2014.

- Barros, P., Magg, S., Weber, C., Wermter, S. A Multichannel Convolutional Neural Network for Hand Posture Recognition. In Wermter, S., et al., editors. Proceedings of the 24th International Conference on Artificial Neural Networks (ICANN 2014), pp. 403-410, Hamburg, Germany, 2014.

The models and concepts presented in this thesis were also applied to different domains, and published in different journals and conference proceedings.

- Hinz, T., Barros, P., Wermter, S. The Effects of Regularization on Learning Facial Expressions with Convolutional Neural Networks. In Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), In Press, Barcelona, Spain, September 2016.

- Mousavi, N., Siqueira, H., Barros, P.,Fernandes, B., Wermter, S. Understanding How Deep Neural Networks Learn Face Expressions. Proceedings of International Joint Conference on Neural Networks (IJCNN), In press, Vancouver, Canada, 2016.

- Speck, D., Barros, P., Weber, C. and Wermter, S. Ball Localization for Robocup Soccer using Convolutional Neural Networks. RoboCup Symposium, Leipzig, Germany, 2016. - Best Paper Award

- Tsironi, E., Barros, P., and Wermter, S. Gesture Recognition with a Convolutional Long Short-Term Memory Recurrent Neural Network. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 213-218, Bruges, Belgium, 2016.

- Hinaut, X., Twiefel, J., Borghetti Soares, M., Barros, P., Mici, L., Wermter, S. Humanoidly Speaking How the Nao humanoid robot can learn the name of objects and interact with them through common speech. International Joint Conference on Artificial Intelligence (IJCAI), Video Competition, Buenos Aires, Argentina, 2015.

- Hamester, D., Barros, P., Wermter, S. Face Expression Recognition with a 2-Channel Convolutional Neural Network. Proceedings of International Joint Conference on Neural Networks (IJCNN), pp. 1787-1794, Killarney, Ireland, 2015.

- Jirak, D., Barros, P., Wermter, S. Dynamic Gesture Recognition Using Echo State Networks. Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN'15, pp. 475-480, Bruges, Belgium, 2015.

- Borghetti Soares, M., Barros, P., Parisi, G. I., Wermter, S. Learning objects from RGB-D sensors using point cloud-based neural networks. Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN'15, pp. 439-444, Bruges, Belgium, 2015.

- Borghetti Soares, M., Barros, P., Wermter, S. Learning Objects From RGB-D Sensors for Cleaning Tasks Using a Team of Cooperative Humanoid Robots. Proceedings of 15th Towards Autonomous Robots, TAROS 2014, LNAI 8717, pp. 273-274, Springer Heidelberg. Birmingham, UK, October 1-3, 2014.
- Parisi, G. I., Barros, P.,Wermter, S. FINGeR: Framework for Interactive Neural-based Gesture Recognition. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN '14), pp. 443-447, Bruges, Belgium, 2014.

Appendix C

Acknowledgements

The Ph.D. studies are very long and solitaire times. However, during the three years since I started this journey, I had so many persons in my life that helped me, supported me, guided me and, the most important, believed in me and on my work. I would like here to express some words of gratitude towards these persons.

I would like to, firstly, thank my supervisor, Prof. Stefan Wermter, for so many valuable advice, guidance, and support during my Ph.D. journey. I would like also to thank Dr. Cornelius Weber, Dr. Sascha Griffiths and Dr. Sven Magg for the helpful discussions and feedback. Would like to highlight here the importance of the opportunity you gave me to start my Ph.D. here, and how this changed my life.

Although the Ph.D. is a very lonely journey, I could rely on the Knowledge Technology staff to support me. I would like to thank, especially, Katja Kösters and Erik Strahl, for all the help and kindness they showed to me since I arrived in the group. Also would like to thank the support of the many colleagues from the group, the various discussions and talks we had, beers we drank and friendships we forged.

A big thank you to Prof. Bruno Fernandes and Prof. Byron Leite, and the entire RPPDI research group from the University of Pernambuco. I thank their support and guidance, and I am looking forward to proceeding with our collaborations. Yes, Vancouver was fantastic. Also would like to thank Prof. Jorge Correia, who help me since I started my graduation and gave me valuable feedback on this thesis and in my work.

Outside the academic environment, many different persons were part of the development of my studies. First, I would like to remember my fantastic NEX group, which proved that distance is just a number, and we are united wherever we are. Also important to mention my dear UAST/Chronos friends! This work started there, ten years ago, in Mineko's restaurant. I am so proud of what all of us reached in our lives, and I am looking forward to seeing where we go from here. Important to mention my around-the-world family, which are the best friends a person can have: Marcio "Winne" Alfredo, Ilzy Sousa, "Marcelo Mars Mauricio", Francisco "Galego" Junior, Sara "Saritz" and João Victor Cavalcanti.

In my life my family was always present, even if we do not see each other so

often. I have persons who inspired me, but none most than my parents. I hope I continue to learn with them as they still have so much to teach me. And family is not only the ones you share blood with, but the ones you find in your life, and here in Germany I found a very important part of my life. I would like to thank Doreen Jirak for all the support, help and inspiration that she so kindly offered me. I would like to say that you are an integral part of this work, and without you, for sure it would not have happened. I am looking forward to seeing how we will change this world together.

Finally, I would like to say that this work was partially supported by CAPES, the Brazilian Federal Agency for the Support and Evaluation of Graduate Education under the project number 5951-13-5, the German Research Foundation DFG under project CML (TRR 169), and the Hamburg Landesforschungsförderungsprojekt.

Muito Obrigado!

Bibliography

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(10):1533–1545, 2014.

[2] Ralph Adolphs. Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, 12(2):169 – 177, 2002.

[3] S. Afzal and P. Robinson. Natural affect data - collection and annotation in a learning context. In *3rd International Conference on Affective Computing and Intelligent Interaction.*, pages 1–7, Sept 2009.

[4] Alessandra Angelucci and Jean Bullier. Reaching beyond the classical receptive field of v1 neurons: horizontal or feedback axons? *Journal of Physiology-Paris*, 97(2):141–154, 2003.

[5] Arash Foroughmand Arabi and Guojun Lu. Enhanced polyphonic music genre classification using high level features. In *Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on*, pages 101–106. IEEE, 2009.

[6] Magda B Arnold. *Emotion and personality.* Columbia University Press, 1960.

[7] Minoru Asada. Development of artificial empathy. *Neuroscience research*, 90:41–50, 2015.

[8] Anthony P Atkinson and Ralph Adolphs. Visual emotion perception. *Emotion and consciousness*, page 150, 2005.

[9] Ntombikayise Banda and Peter Robinson. Noise analysis in audio-visual emotion recognition. In *Proceedings of the International Conference on Multimodal Interaction*, pages 1–4. Citeseer, 2011.

[10] Debarati Bandyopadhyay, V.S. Chandrasekhar Pammi, and Narayanan Srinivasan. Chapter 3 - role of affect in decision making. In V.S. Chandrasekhar Pammi and Narayanan Srinivasan, editors, *Decision Making Neural and Behavioural Approaches*, volume 202 of *Progress in Brain Research*, pages 37 – 53. Elsevier, 2013.

[11] Lauren Barghout-Stein. *How Global Perceptual Context Changes Local Contrast Processing*. University of California, Berkeley, 2003.

[12] Lisa Feldman Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46, 2006.

[13] Lisa Feldman Barrett and Moshe Bar. See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521):1325–1334, 2009.

[14] P. Barros, G.I. Parisi, D. Jirak, and S. Wermter. Real-time gesture recognition using a humanoid robot with a deep neural architecture. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 646–651, Nov 2014.

[15] Pablo Barros, Doreen Jirak, Cornelius Weber, and Stefan Wermter. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, 72:140–151, 2015.

[16] Marian Stewart Bartlett, Paul A Viola, Terrence J Sejnowski, Beatrice A Golomb, Jan Larsen, Joseph C Hager, and Paul Ekman. Classifying facial action. *Advances in neural information processing systems*, pages 823–829, 1996.

[17] Johannes Bauer, Jorge Dávila-Chacón, Erik Strahl, and Stefan Wermter. Smoke and mirrorsvirtual realities for sensor fusion experiments in biomimetic robotics. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, pages 114–119. IEEE, 2012.

[18] James Beament. *How we hear music: The relationship between music and the hearing mechanism*. Boydell Press, 2003.

[19] Michael S Beauchamp. The social mysteries of the superior temporal sulcus. *Trends in cognitive sciences*, 19(9):489–490, 2015.

[20] David James Beymer. Face recognition under varying pose. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 756–761. IEEE, 1994.

[21] Boris Birmaher, Suneeta Khetarpal, David Brent, Marlane Cully, Lisa Balach, Joan Kaufman, and Sandra McKenzie Neer. The screen for child anxiety related emotional disorders (scared): scale construction and psychometric characteristics. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(4):545–553, 1997.

[22] Hugh T Blair, Glenn E Schafe, Elizabeth P Bauer, Sarina M Rodrigues, and Joseph E LeDoux. Synaptic plasticity in the lateral amygdala: a cellular hypothesis of fear conditioning. *Learning & memory*, 8(5):229–242, 2001.

[23] Fredda Blanchard-Fields. Everyday problem solving and emotion an adult developmental perspective. *Current Directions in Psychological Science*, 16(1):26–31, 2007.

[24] Richard T Born and David C Bradley. Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28:157–189, 2005.

[25] Danny Oude Bos. Eeg-based emotion recognition. *The Influence of Visual and Auditory Stimuli*, pages 1–17, 2006.

[26] Gordon H Bower. Mood and memory. *American psychologist*, 36(2):129–148, 1981.

[27] Oliver J Braddick, Justin MD O'Brien, John Wattam-Bell, Janette Atkinson, Tom Hartley, and Robert Turner. Brain areas sensitive to coherent visual motion. *Perception*, 30(1):61–72, 2001.

[28] Clive R Bramham and Elhoucine Messaoudi. Bdnf function in adult synaptic plasticity: the synaptic consolidation hypothesis. *Progress in neurobiology*, 76(2):99–125, 2005.

[29] Margaret J Briggs-Gowan, Alice S Carter, Julia R Irwin, Karen Wachtel, and Domenic V Cicchetti. The brief infant-toddler social and emotional assessment: screening for social-emotional problems and delays in competence. *Journal of pediatric psychology*, 29(2):143–155, 2004.

[30] Charles Bruce, Robert Desimone, and Charles G Gross. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of neurophysiology*, 46(2):369–384, 1981.

[31] Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986.

[32] Jean Bullier. Integrated model of visual processing. *Brain Research Reviews*, 36(2):96–107, 2001.

[33] William E Bunney and David A Hamburg. Methods for reliable longitudinal observation of behavior: Development of a method for systematic observation of emotional behavior on psychiatric wards. *Archives of General Psychiatry*, 9(3):280–294, 1963.

[34] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[35] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan.

Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[36] Michel Cabanac. What is emotion? *Behavioural Processes*, 60(2):69 – 83, 2002.

[37] Larry Cahill and James L McGaugh. A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and cognition*, 4(4):410–421, 1995.

[38] Salvatore Campanella and Pascal Belin. Integrating face and voice in person perception. *Trends in cognitive sciences*, 11(12):535–543, 2007.

[39] R Campbell, Charles A Heywood, A Cowey, M Regard, and T Landis. Sensitivity to eye gaze in prosopagnosic patients and monkeys with superior temporal sulcus ablation. *Neuropsychologia*, 28(11):1123–1142, 1990.

[40] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 375–388. Springer, 2007.

[41] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, and Kostas Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 146–154. ACM, 2006.

[42] Joshua Michael Carlson, Tsafrir Greenberg, and Lilianne R Mujica-Parodi. Blind rage? heightened anger is associated with altered amygdala responses to masked and unmasked fearful faces. *Psychiatry Research: Neuroimaging*, 182(3):281–283, 2010.

[43] Ginevra Castellano, Loic Kessous, and George Caridakis. Emotion recognition through multiple modalities: Face, body gesture, speech. In Christian Peter and Russell Beale, editors, *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *Lecture Notes in Computer Science*, pages 92–103. Springer Berlin Heidelberg, 2008.

[44] Rama Chellappa, Charles L Wilson, and Saad Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–741, 1995.

[45] Shizhi Chen, YingLi Tian, Qingshan Liu, and Dimitris N. Metaxas. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, 31(2):175 – 185, 2013. Affect Analysis In Continuous Input.

[46] Sven-Ake Christianson. *The handbook of emotion and memory: Research and theory*. Psychology Press, 2014.

[47] Sven-Åke Christianson and Elizabeth F Loftus. Some characteristics of peoples traumatic memories. *Bulletin of the Psychonomic Society*, 28(3):195–198, 1990.

[48] Jason A Clark. Relations of homology between higher cognitive emotions and basic emotions. *Biology & Philosophy*, 25(1):75–94, 2010.

[49] Adam Coates and Andrew Y Ng. Selecting receptive fields in deep networks. In *Advances in Neural Information Processing Systems*, pages 2528–2536, 2011.

[50] Ira Cohen, Ashutosh Garg, Thomas S Huang, et al. Emotion recognition from facial expressions using multilevel hmm. In *Neural information processing systems*, volume 2. Citeseer, 2000.

[51] Jeffrey F Cohn, Adena J Zlochower, James Lien, and Takeo Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, 36(01):35–43, 1999.

[52] Randall Collins. Social movements and the focus of emotional attention. *Passionate politics: Emotions and social movements*, pages 27–44, 2001.

[53] Wikimedia Commons. Image showing dorsal stream (green) and ventral stream (purple) in the human brain visual system, 2007.

[54] Leda Cosmides and John Tooby. Evolutionary psychology and the emotions. *Handbook of emotions*, 2:91–115, 2000.

[55] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139, 2004.

[56] Roddy Cowie and Randolph R Cornelius. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1):5–32, 2003.

[57] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[58] Antonio R Damasio. *Descartes' error*. Random House, 2006.

[59] Justin d'Arms and Daniel Jacobson. The moralistic fallacy: on the'appropriateness' of emotions. *Philosophical and Phenomenological Research*, pages 65–90, 2000.

[60] Charles Darwin. *The expression of the emotions in man and animals*. John Murray, 1873.

[61] Beatrice De Gelder and Jean Vroomen. The perception of emotions by ear and by eye. *Cognition & Emotion*, 14(3):289–311, 2000.

[62] Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, volume 1, pages 397–401. IEEE, 1997.

[63] Marieke De Vries, Rob W Holland, and Cilia LM Witteman. Fitting decisions: Mood and intuitive versus deliberative decision strategies. *Cognition and Emotion*, 22(5):931–943, 2008.

[64] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

[65] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8599–8603. IEEE, 2013.

[66] Pieter MA Desmet. Design for mood: Twenty activity-based opportunities to design for mood regulation. *International Journal of Design*, 9(2):1–19, 2015.

[67] Diana Deutsch. Hearing music in ensembles. *Physics today*, 63(2):40–45, 2010.

[68] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[69] Abhinav Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[70] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka, and Tom Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 461–466. ACM, 2014.

[71] Thomas Dixon. *From passions to emotions: The creation of a secular psychological category*. Cambridge University Press, 2003.

[72] Raymond J Dolan. Emotion, cognition, and behavior. *science*, 298(5596):1191–1194, 2002.

[73] Raymond J Dolan, John S Morris, and Beatrice de Gelder. Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences*, 98(17):10006–10010, 2001.

[74] Jon Driver. A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1):53–78, 2001.

[75] R Dubner and SM Zeki. Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain research*, 35(2):528–532, 1971.

[76] Yadin Dudai. The neurobiology of consolidations, or, how stable is the engram? *Annu. Rev. Psychol.*, 55:51–86, 2004.

[77] Kristen Dunfield, Valerie A Kuhlmeier, Laura OConnell, and Elizabeth Kelley. Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy. *Infancy*, 16(3):227–247, 2011.

[78] John D Eastwood, Daniel Smilek, and Philip M Merikle. Differential attentional guidance by unattended faces expressing positive and negative emotion. *Perception & Psychophysics*, 63(6):1004–1013, 2001.

[79] Nancy Eisenberg. Emotion, regulation, and moral development. *Annual review of psychology*, 51(1):665–697, 2000.

[80] Paul Ekman. *The argument and evidence about universals in facial expressions of emotion.* John Wiley and Sons, 1989.

[81] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[82] Paul Ekman. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* Macmillan, 2007.

[83] Paul Ekman. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009.

[84] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[85] Paul Ekman and Wallace V Friesen. *Facial action coding system.* Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[86] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA, 1997.

[87] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[88] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660, March 2010.

[89] Irfan A Essa and Alex P Pentland. Facial expression recognition using a dynamic model and motion energy. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 360–367. IEEE, 1995.

[90] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[91] David C. Van Essen and Jack L. Gallant. Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1):1 – 10, 1994.

[92] Thomas Ethofer, Silke Anders, Michael Erb, Christina Droll, Lydia Royen, Ralf Saur, Susanne Reiterer, Wolfgang Grodd, and Dirk Wildgruber. Impact of voice on emotional judgment of faces: An event-related fmri study. *Human brain mapping*, 27(9):707–714, 2006.

[93] Michael W Eysenck. Arousal, learning, and memory. *Psychological Bulletin*, 83(3):389, 1976.

[94] B. Fasel. Head-pose invariant facial expression recognition using convolutional neural networks. In *Proceedings of the fourth IEEE International Conference on Multimodal Interfaces, 2002*, pages 529–534, 2002.

[95] Beverley Fehr and James A Russell. Concept of emotion viewed from a prototype perspective. *Journal of experimental psychology: General*, 113(3):464, 1984.

[96] Francesco Foroni and Gün R Semin. Language that puts you in touch with your bodily feelings the multimodal responsiveness of affective expressions. *Psychological Science*, 20(8):974–980, 2009.

[97] Elaine Fox. Processing emotional facial expressions: The role of anxiety and awareness. *Cognitive, Affective, & Behavioral Neuroscience*, 2(1):52–63, 2002.

[98] Elaine Fox. *Emotion Science: An Integration of Cognitive and Neuroscience Approaches*. Palgrave Macmillan, 2008.

[99] Yves Fregnac, Cyril Monier, Frederic Chavane, Pierre Baudot, and Lyle Graham. Shunting inhibition, a silent step in visual cortical computation. *Journal of Physiology*, pages 441–451, 2003.

[100] Simone Frintrop, Erich Rome, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6, 2010.

[101] Nilima Salankar Fulmare, Prasun Chakrabarti, and Divakar Yadav. Understanding and estimation of emotional expression using acoustic analysis of natural speech. *International Journal on Natural Language Computing (IJNLC)*, 2(4), 2013.

[102] Harn-M Gardiner, Ruth Clark Metcalf, and John G Beebe-Center. *Feeling and emotion: A history of theories.* American Book Publishing, 1937.

[103] Michael S Gazzaniga. *The cognitive neurosciences*. MIT press, 2004.

[104] Joe George and Lior Shamir. Unsupervised analysis of similarities between musicians and musical genres using spectrograms. *Artificial Intelligence Research*, 4(2):p61, 2015.

[105] Mark S George, Terence A Ketter, Debra S Gill, James V Haxby, Leslie G Ungerleider, Peter Herscovitch, and Robert M Post. Brain regions involved in recognizing facial emotion or identity: an oxygen-15 pet study. *The Journal of neuropsychiatry and clinical neurosciences*, 1993.

[106] Mark S George, Priti I Parekh, Ned Rosinsky, Terence A Ketter, Tim A Kimbrell, Kenneth M Heilman, Peter Herscovitch, and Robert M Post. Understanding emotional prosody activates right hemisphere regions. *Archives of neurology*, 53(7):665–670, 1996.

[107] Alessandra Geraci and Luca Surian. The developmental roots of fairness: Infants reactions to equal and unequal distributions of resources. *Developmental science*, 14(5):1012–1020, 2011.

[108] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey J. Gordon and David B. Dunson, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, volume 15, pages 315–323. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011.

[109] Erin Goddard, Damien J Mannion, J Scott McDonald, Samuel G Solomon, and Colin WG Clifford. Color responsiveness argues against a dorsal component of human v4. *Journal of vision*, 11(4):3–3, 2011.

[110] William V Good, James E Jan, Luis DeSa, A James Barkovich, Myryka Groenveld, et al. Cortical visual impairment in children. *Survey of ophthalmology*, 38(4):351–364, 1994.

[111] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[112] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[113] Stephen Grossberg. *Neural Networks and Natural Intelligence*. MIT Press, Cambridge, MA, USA, 1992.

[114] ED Grossman and R Blake. Brain activity evoked by inverted and imagined biological motion. *Vision research*, 41(10):1475–1482, 2001.

[115] Tobias Grossmann. The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, 28(2):219, 2010.

[116] Tobias Grossmann, Tricia Striano, and Angela D Friederici. Crossmodal integration of emotional information from face and voice in the infant brain. *Developmental Science*, 9(3):309–315, 2006.

[117] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition(ICPR), 2006*, volume 1, pages 1148–1153, 2006.

[118] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):64–84, Feb 2009.

[119] Jonathan Haidt, Craig Joseph, et al. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind*, 3:367–391, 2007.

[120] J Kiley Hamlin. Moral judgment and action in preverbal infants and toddlers evidence for an innate moral core. *Current Directions in Psychological Science*, 22(3):186–193, 2013.

[121] J Kiley Hamlin. The infantile origins of our moral brains. *The Moral Brain: A Multidisciplinary Perspective*, page 105, 2015.

[122] J Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450(7169):557–559, 2007.

[123] S. Haq and P.J.B. Jackson. *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA, Aug. 2010.

[124] Sanaul Haq, Philip JB Jackson, and J Edge. Speaker-dependent audio-visual emotion recognition. In *AVSP*, pages 53–58, 2009.

[125] Susan Harter and Bonnie J Buddin. Children's understanding of the simultaneity of two emotions: A five-stage developmental acquisition sequence. *Developmental psychology*, 23(3):388, 1987.

[126] ME Hasselmo, ET Rolls, GC Baylis, and V Nalwa. Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 75(2):417–429, 1989.

[127] Michael E Hasselmo, Edmund T Rolls, and Gordon C Baylis. The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural brain research*, 32(3):203–218, 1989.

[128] Darren Hau and Ke Chen. Exploring hierarchical speech representations with a deep convolutional neural network. *UKCI 2011 Accepted Papers*, page 37, 2011.

[129] Donald Olding Hebb. *The organization of behavior: A neuropsychological approach*. John Wiley & Sons, 1949.

[130] Gregory Hickok. The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of communication disorders*, 45(6):393–402, 2012.

[131] Gregory Hickok. The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of communication disorders*, 45(6):393–402, 2012.

[132] Gregory Hickok and David Poeppel. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1):67–99, 2004.

[133] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007.

[134] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[135] Geoffrey E Hinton and James A Anderson. *Parallel models of associative memory: updated edition*. Psychology press, 2014.

[136] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[137] Anett Hoppe and Marco Tabacchi. Towards a modelization of the elusive concept of wisdom using fuzzy techniques. In *Fuzzy Information Processing Society (NAFIPS), 2012 Annual Meeting of the North American*, pages 1–5. IEEE, 2012.

[138] Yin-Fu Huang, Sheng-Min Lin, Huan-Yu Wu, and Yu-Siou Li. Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. *Data & Knowledge Engineering*, 92:60–76, 2014.

[139] David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

[140] David Hume. Emotions and moods. *Organizational behavior*, pages 258–297, 2012.

[141] Spiros V Ioannou, Amaryllis T Raouzaiou, Vasilis A Tzouvaras, Theofilos P Mailis, Kostas C Karpouzis, and Stefanos D Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4):423–435, 2005.

[142] Carroll E Izard. *The psychology of emotions*. Springer Science & Business Media, 1991.

[143] Carroll E Izard. *Innate and universal facial expressions: evidence from developmental and cross-cultural research.* American Psychological Association, 1994.

[144] Carroll E Izard. *Human emotions*. Springer Science & Business Media, 2013.

[145] William James. What is an emotion? *Mind*, (34):188–205, 1884.

[146] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, Jan 2013.

[147] Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. Speech emotion recognition with acoustic and lexical features. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4749–4753, April 2015.

[148] Richard Joyce. *The evolution of morality*. MIT press, 2007.

[149] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, Mehdi Mirza, Sébastien Jean, Pierre-Luc Carrier, Yann Dauphin, Nicolas Boulanger-Lewandowski, Abhishek Aggarwal, Jeremie Zumer, Pascal Lamblin, Jean-Philippe Raymond, Guillaume Desjardins, Razvan Pascanu, David Warde-Farley, Atousa Torabi, Arjun Sharma, Emmanuel Bengio, Myriam Côté, Kishore Reddy Konda, and Zhenzhou Wu. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 543–550, New York, NY, USA, 2013. ACM.

[150] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732, June 2014.

[151] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[152] MASE Kenji. Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 74(10):3474–3483, 1991.

[153] Elizabeth A Kensinger. Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, 15(4):241–252, 2004.

[154] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[155] Hyung-O Kim, Soohwan Kim, and Sung-Kee Park. Pointing gesture-based unknown object extraction for learning objects with robot. In *International Conference on Control, Automation and Systems, 2008. ICCAS 2008*, pages 2156–2161, 2008.

[156] Kyung Hwan Kim, SW Bang, and SR Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.

[157] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics, 2010.

[158] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2013.

[159] Hiroshi Kobayashi and Fumio Hara. Recognition of six basic facial expression and their strength by neural network. In *Robot and Human Communication, 1992. Proceedings., IEEE International Workshop on*, pages 381–386. IEEE, 1992.

[160] Gary G Koch. Intraclass correlation coefficient. *Encyclopedia of statistical sciences*, 1982.

[161] Lawrence Kohlberg. Stages in the development of moral thought and action, 1969.

[162] Teuvo Kohonen. Adaptive, associative, and self-organizing functions in neural computing. *Applied Optics*, 26(23):4910–4918, 1987.

[163] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[164] Teuvo Kohonen. *Self-organization and associative memory*, volume 8. Springer Science & Business Media, 2012.

[165] Teuvo Kohonen and Panu Somervuo. Self-organizing maps of symbol strings. *Neurocomputing*, 21(1):19–30, 1998.

[166] David Konstan. *The emotions of the ancient Greeks: Studies in Aristotle and classical literature*, volume 5. University of Toronto Press, 2006.

[167] Irene Kotsia and Ioannis Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187, 2007.

[168] Robert Kraut. Love de re. *Midwest studies in philosophy*, 10(1):413–430, 1987.

[169] Richard J. Krauzlis, Lee P. Lovejoy, and Alexandre Zénon. Superior colliculus and visual spatial attention. *Annual review of neuroscience*, 36(1):165–182, July 2013.

[170] Mariska E Kret, Karin Roelofs, Jeroen J Stekelenburg, and Beatrice de Gelder. Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. *Frontiers in human neuroscience*, 7, 2013.

[171] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[172] Dharshan Kumaran. Short-term memory and the human hippocampus. *The Journal of Neuroscience*, 28(15):3837–3838, 2008.

[173] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *INTERSPEECH*. Citeseer, 2003.

[174] Kevin S LaBar and Elizabeth A Phelps. Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans. *Psychological Science*, 9(6):490–493, 1998.

[175] Peter J Lang. The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

[176] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[177] Richard S Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819, 1991.

[178] Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[179] Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[180] Joseph LeDoux. Rethinking the emotional brain. *Neuron*, 73(4):653–676, 2012.

[181] Joseph LeDoux and Jules R Bemporad. The emotional brain. *Journal of the American Academy of Psychoanalysis*, 25(3):525–528, 1997.

[182] Joseph E LeDoux. Emotion circuits in the brain. *Annu. Rev. Neurosci*, 23:155–184, 2000.

[183] Joseph E LeDoux. Emotion circuits in the brain. *Focus*, 7(2):274–274, 2009.

[184] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9):1162–1171, 2011.

[185] Geneviève Leuba and Rudolf Kraftsik. Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. *Anatomy and embryology*, 190(4):351–366, 1994.

[186] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 503–510. ACM, 2015.

[187] Linda J Levine and David A Pizarro. Emotion and memory research: A grumpy overview. *Social Cognition*, 22(5: Special issue):530–554, 2004.

[188] Michael Lewis. *Childrens emotions and moods: Developmental theory and measurement*. Springer Science & Business Media, 2012.

[189] Penelope A Lewis and Hugo D Critchley. Mood-dependent memory. *Trends in cognitive sciences*, 7(10):431–433, 2003.

[190] Tom LH Li, Antoni B Chan, and Andy HW Chun. Automatic musical pattern feature extraction using convolutional neural network. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1. Citeseer, 2010.

[191] James J Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. Automated facial expression recognition based on facs action units. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 390–395. IEEE, 1998.

[192] JJ-J Lien, T Kanade, JF Cohn, and Ching-Chung Li. Subtly different facial expression recognition and expression intensity estimation. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 853–859. IEEE, 1998.

[193] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *2005 International Conference on Machine Learning and Cybernetics*, volume 8, pages 4898–4901. IEEE, 2005.

[194] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

[195] Mengmeng Liu, Hui Chen, Yang Li, and Fengjun Zhang. Emotional tone-based audio continuous emotion recognition. In Xiangjian He, Suhuai Luo, Dacheng Tao, Changsheng Xu, Jie Yang, and MuhammadAbul Hasan, editors, *MultiMedia Modeling*, volume 8936 of *Lecture Notes in Computer Science*, pages 470–480. Springer International Publishing, 2015.

[196] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014.

[197] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based emotion recognition and its applications. In *Transactions on computational science XII*, pages 256–277. Springer, 2011.

[198] Damien Lockner and Nathalie Bonnardel. Towards the evaluation of emotional interfaces. In *Human-Computer Interaction: Design and Evaluation*, pages 500–511. Springer, 2015.

[199] R Lopez, Eva Balsa-Canto, and E Onate. Neural networks for variational problems in engineering. *International Journal for Numerical Methods in Engineering*, 75(11):1341–1360, 2008.

[200] Leo L Lui, James A Bourne, and Marcello GP Rosa. Functional response properties of neurons in the dorsomedial visual area of new world monkeys (callithrix jacchus). *Cerebral Cortex*, 16(2):162–177, 2006.

[201] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[202] Richard J Maddock, Amy S Garrett, and Michael H Buonocore. Posterior cingulate cortex activation by emotional words: fmri evidence from a valence decision task. *Human brain mapping*, 18(1):30–41, 2003.

[203] Jens Madsen, Bjørn Sand Jensen, and Jan Larsen. Learning combinations of multiple feature representations for music emotion prediction. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, ASM '15, pages 3–8, New York, NY, USA, 2015. ACM.

[204] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, 2014.

[205] Stacy Marsella and Jonathan Gratch. Modeling coping behavior in virtual humans: don't worry, be happy. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 313–320. ACM, 2003.

[206] Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. A self-organising network that grows when required. *Neural Networks*, 15(8):1041–1058, 2002.

[207] Grace B Martin and Russell D Clark. Distress crying in neonates: Species and peer specificity. *Developmental psychology*, 18(1):3, 1982.

[208] Dominic W Massaro and Peter B Egan. Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, 3(2):215–221, 1996.

[209] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5):555–559, 2003.

[210] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, pages 50–56. ACM, 2008.

[211] John Mikhail. *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press, 2011.

[212] Jeffrey Moran and Robert Desimone. Selective attention gates visual processing in the extrastriate cortex. *Frontiers in cognitive neuroscience*, 229:342–345, 1985.

[213] John S Morris, Beatrice DeGelder, Lawrence Weiskrantz, and Ray J Dolan. Differential extrageniculostriate and amygdala responses to presentation of emotional faces in a cortically blind field. *Brain*, 124(6):1241–1252, 2001.

[214] John S Morris, Arne Öhman, and Raymond J Dolan. Conscious and unconscious emotional learning in the human amygdala. *Nature*, 393(6684):467–470, 1998.

[215] Hariharan Muthusamy, Kemal Polat, and Sazali Yaacob. Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals. *PloS one*, 10(3):e0120344, 2015.

[216] Jin Narumoto, Tomohisa Okada, Norihiro Sadato, Kenji Fukui, and Yoshiharu Yonekura. Attention to emotion modulates fmri activity in human right superior temporal sulcus. *Cognitive Brain Research*, 12(2):225–231, 2001.

[217] Jerome Neu. *Emotion, Thought, & Therapy: A Study of Hume and Spinoza and the Relationship of Philosophical Theories of the Emotions to Psychological Theories of Therapy*. Univ of California Press, 1977.

[218] Kevin N Ochsner. Are affective events richly recollected or simply familiar? the experience and process of recognizing feelings past. *Journal of Experimental Psychology: General*, 129(2):242, 2000.

[219] Kevin N Ochsner and James J Gross. The cognitive control of emotion. *Trends in cognitive sciences*, 9(5):242–249, 2005.

[220] Randall C O'Reilly and Michael J Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2):283–328, 2006.

[221] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.

[222] Yannis Panagakis and Constantine Kotropoulos. Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Acoustics speech and signal processing (ICASSP), 2010 IEEE international conference on*, pages 249–252. IEEE, 2010.

[223] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 1998.

[224] Maja Pantic and Ioannis Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449, 2006.

[225] Maja Pantic and Leon JM Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1449–1461, 2004.

[226] Geoffrey JM Parker, Simona Luzzi, Daniel C Alexander, Claudia AM Wheeler-Kingshott, Olga Ciccarelli, and Matthew A Lambon Ralph. Lateralization of ventral and dorsal auditory-language pathways in the human brain. *Neuroimage*, 24(3):656–666, 2005.

[227] Joseph J Paton, Marina A Belova, Sara E Morrison, and C Daniel Salzman. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, 439(7078):865–870, 2006.

[228] Luiz Pessoa. On the relationship between emotion and cognition. *Nature reviews neuroscience*, 9(2):148–158, 2008.

[229] Luiz Pessoa, Sabine Kastner, and Leslie G Ungerleider. Attentional control of the processing of neutral and emotional stimuli. *Cognitive Brain Research*, 15(1):31–45, 2002.

[230] V Petrushin. Emotion in speech: Recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering*, volume 710, 1999.

[231] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *2011 International Conference on Computer Vision*, pages 1449–1456. IEEE, 2011.

[232] Elizabeth A Phelps. Human emotion and memory: interactions of the amygdala and hippocampal complex. *Current opinion in neurobiology*, 14(2):198–202, 2004.

[233] Elizabeth A. Phelps, Sam Ling, and Marisa Carrasco. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science*, 17(4):292–299, April 2006.

[234] Jean Piaget. *The moral judgement of the child*. Simon and Schuster, 1997.

[235] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[236] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[237] James O Pickles. *An introduction to the physiology of hearing*. Brill, 2012.

[238] Michael Pitz and Hermann Ney. Vocal tract normalization as linear transformation of mfcc. In *INTERSPEECH*, volume 3, pages 1445–1448, 2003.

[239] Plato and Sir Henry Desmond Pritchard Lee. *Plato, The Republic*. Penguin Books, 1955.

[240] Robert 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*, 89(4):344–350, 2001.

[241] Thomas S Polzin and Alex Waibel. Detecting emotions in speech. In *Proceedings of the CMC*, volume 16. Citeseer, 1998.

[242] Francisco Pons, Paul L Harris, and Marc de Rosnay. Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European journal of developmental psychology*, 1(2):127–152, 2004.

[243] Stephen Porter and Leanne Brinke. Reading between the lies identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19(5):508–514, 2008.

[244] Gilles Pourtois, Beatrice de Gelder, Anne Bol, and Marc Crommelinck. Perception of facial expressions and voices and of their combination in the human brain. *Cortex*, 41(1):49–59, 2005.

[245] Fangtu T Qiu and Rüdiger Von Der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, 47(1):155–166, 2005.

[246] P. Rani and N. Sarkar. Emotion-sensitive robots - a new paradigm for human-robot interaction. In *Humanoid Robots, 2004 4th IEEE/RAS International Conference on*, volume 1, pages 149–167 Vol. 1, Nov 2004.

[247] M. Ranzato, Fu Jie Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[248] Josef P Rauschecker. Ventral and dorsal streams in the evolution of speech and language. *Frontiers in evolutionary neuroscience*, 4:7, 2012.

[249] Anne Richards and Isabelle Blanchette. Independent manipulation of emotion in an emotional stroop task using classical conditioning. *Emotion*, 4(3):275, 2004.

[250] Jane M Richards and James J Gross. Emotion regulation and memory: the cognitive costs of keeping one's cool. *Journal of personality and social psychology*, 79(3):410, 2000.

[251] Fabien Ringeval, Shahin Amiriparian, Florian Eyben, Klaus Scherer, and Björn Schuller. Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 473–480, New York, NY, USA, 2014. ACM.

[252] Anna W Roe, Leonardo Chelazzi, Charles E Connor, Bevil R Conway, Ichiro Fujita, Jack L Gallant, Haidong Lu, and Wim Vanduffel. Toward a unified theory of visual area v4. *Neuron*, 74(1):12–29, 2012.

[253] EdmundT Rolls. *Emotion explained*. Oxford University Press, USA, 2005.

[254] Marcello GP Rosa and Rowan Tweedale. Visual areas in lateral and ventral extrastriate cortices of the marmoset monkey. *Journal of Comparative Neurology*, 422(4):621–651, 2000.

[255] Ira J Roseman. Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology*, 1984.

[256] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[257] Mark Rosenblum, Yaser Yacoob, and Larry S Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE transactions on neural networks*, 7(5):1121–1138, 1996.

[258] David C Rubin and Jennifer M Talarico. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8):802–808, 2009.

[259] David E Rumelhart, Richard Durbin, Richard Golden, and Yves Chauvin. Backpropagation: The basic theory. *Backpropagation: Theory, Architectures and Applications*, pages 1–34, 1995.

[260] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[261] James A Russell. 13-reading emotion from and into faces: resurrecting a dimensional-contextual perspective,. *The psychology of facial expression*, pages 295–320, 1997.

[262] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[263] Abraham Sagi and Martin L Hoffman. Empathic distress in the newborn. *Developmental Psychology*, 12(2):175, 1976.

[264] Tara N. Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel rahman Mohamed, George Dahl, and Bhuvana Ramabhadran. Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 64:39 – 48, 2015. Special Issue on Deep Learning of Representations.

[265] Hide-aki Saito, Masao Yukie, Keiji Tanaka, Kazuo Hikosaka, Yoshiro Fukada, and E Iwai. Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *The Journal of Neuroscience*, 6(1):145–157, 1986.

[266] Eliala Salvadori, Tatiana Blazsekova, Agnes Volein, Zsuzsanna Karap, Denis Tatone, Olivier Mascaro, and Gergely Csibra. Probing the strength of infants' preference for helpers over hinderers: two replication attempts of hamlin and wynn (2011). *PloS one*, 10(11):e0140570, 2015.

[267] Rajib Sarkar and Sanjoy Kumar Saha. Music genre classification using emd and pitch based feature. In *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*, pages 1–6. IEEE, 2015.

[268] Dorothee Saur, Björn W Kreher, Susanne Schnell, Dorothee Kümmerer, Philipp Kellmeyer, Magnus-Sebastian Vry, Roza Umarova, Mariacristina Musso, Volkmar Glauche, Stefanie Abel, et al. Ventral and dorsal pathways for language. *Proceedings of the national academy of Sciences*, 105(46):18035–18040, 2008.

[269] Daniel L Schacter. *Searching for Memory: The Brain, Th.* Basic Books, 2008.

[270] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[271] Stefan Scherer, Mohamed Oubbati, Friedhelm Schwenker, and Günther Palm. Real-time emotion recognition from speech using echo state networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 205–216. Springer, 2008.

[272] Jan Schluter and Sebastian Bock. Improved musical onset detection with convolutional neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6979–6983. IEEE, 2014.

[273] Allan N Schore. *Affect regulation and the origin of the self: The neurobiology of emotional development.* Routledge, 2015.

[274] Nicu Sebe, Ira Cohen, Thomas S Huang, et al. Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*, 4:387–419, 2005.

[275] Charles R Seger, Eliot R Smith, Zoe Kinias, and Diane M Mackie. Knowing how they feel: Perceiving emotions felt by outgroups. *Journal of Experimental Social Psychology*, 45(1):80–89, 2009.

[276] Daniel Senkowski, Till R Schneider, John J Foxe, and Andreas K Engel. Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in neurosciences*, 31(8):401–409, 2008.

[277] Siddharth Sigtia and Sam Dixon. Improved music feature learning with deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6959–6963. IEEE, 2014.

[278] Adam M Sillito, Javier Cudeiro, and Helen E Jones. Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in neurosciences*, 29(6):307–316, 2006.

[279] Patrice Y Simard, David Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, volume 3, pages 958–962, 2003.

[280] Craig A Smith and Phoebe C Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813, 1985.

[281] Craig A Smith and Richard S Lazarus. *Emotion and adaptation.* Guilford Press, 1990.

[282] Larry R Squire. *Memory and brain.* 1987.

[283] Bo Sun, Liandong Li, Guoyan Zhou, and Jun He. Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging*, 25(6):061407–061407, 2016.

[284] Raquel Tato, Rocio Santos, Ralf Kompe, and José M Pardo. Emotional space improves emotion recognition. In *INTERSPEECH*, 2002.

[285] Don M Tucker, Douglas Derryberry, Phan Luu, and KL Phan. Anatomy and physiology of human emotion: Vertical integration of brainstem, limbic, and cortical systems. *The neuropsychology of emotion*, pages 56–79, 2000.

[286] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[287] Alfred Ultsch. U*-matrix: a tool to visualize clusters in high dimensional data. Technical Report 10, Universtiy of Marburg, 2003.

[288] Mauro Ursino, Cristiano Cuppini, and Elisa Magosso. Neurocomputational approaches to modelling multisensory integration in the brain: A review. *Neural Networks*, 60:141 – 165, 2014.

[289] Diana Roupas Van Lancker and Gerald J Canter. Impairment of voice and face recognition in patients with hemispheric damage. *Brain and cognition*, 1(2):185–195, 1982.

[290] S. Velusamy, H. Kannan, B. Anand, A. Sharma, and B. Navathe. A method to infer emotions from facial action units. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2028–2031, May 2011.

[291] Paul Viola and MichaelJ. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[292] Barbara Von Eckardt. *What is cognitive science?* MIT press, 1995.

[293] Patrik Vuilleumier. How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12):585–594, 2005.

[294] Patrik Vuilleumier and Yang-Ming Huang. Emotional attention uncovering the mechanisms of affective biases in perception. *Current Directions in Psychological Science*, 18(3):148–152, 2009.

[295] Patrik Vuilleumier and Sophie Schwartz. Emotional facial expressions capture attention. *Neurology*, 56(2):153–158, 2001.

[296] Li Wang, Ting Liu, Gang Wang, Kap Luk Chan, and Qingxiong Yang. Video tracking using learned hierarchical features. *Image Processing, IEEE Transactions on*, 24(4):1424–1435, April 2015.

[297] Yi-Qing Wang. An Analysis of the Viola-Jones Face Detection Algorithm. *Image Processing On Line*, 4:128–148, 2014.

[298] Felix Warneken and Michael Tomasello. Varieties of altruism in children and chimpanzees. *Trends in cognitive sciences*, 13(9):397–402, 2009.

[299] Philip C Watkins, Karen Vache, Steven P Verney, and Andrew Mathews. Unconscious mood-congruent memory bias in depression. *Journal of Abnormal Psychology*, 105(1):34, 1996.

[300] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6):495–501, 2004.

[301] Sherif M Yacoub, Steven J Simske, Xiaofan Lin, and John Burns. Recognition of emotions in interactive voice response systems. In *INTERSPEECH*, 2003.

[302] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 451–458. ACM, 2015.

[303] Masaki Yuki, William W Maddux, and Takahiko Masuda. Are the windows to the soul the same in the east and west? cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology*, 43(2):303–311, 2007.

[304] Jonathan R Zadra and Gerald L Clore. Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science*, 2(6):676–685, 2011.

[305] Leslie A Zebrowitz. *Social perception.* Thomson Brooks/Cole Publishing Co, 1990.

[306] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer, 2014.

[307] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

Declaration of Oath
Eidesstattliche Versicherung

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, Tuesday 23rd August 2016
City and Date
Ort und Datum

Pablo Vinicius Alves de Barros
Signature
Unterschrift