Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Management, Production and Design (35^{th}cycle)

Facial Expression and Emotion Recognition using Deep Learning

3D FER to support Human-Computer Interaction

Francesca Nonis

Supervisor:
Prof. Sandro Moos

Doctoral Examination Committee:
Prof. Antonio Gloria, Università degli Studi di Napoli Federico II
Prof. Domenico Speranza, Università degli Studi di Cassino e del Lazio Meridionale
Prof. Barbara Motyl, Università degli Studi di Udine
Prof. Gabriele Baronio, Università degli Studi di Brescia
Prof. Bartolomeo Montrucchio, Politecnico di Torino

Politecnico di Torino
2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Francesca Nonis
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Ad Andrea, che prima come compagno di Università, poi come amico, confidente, fidanzato, e ora compagno di vita, mi è sempre stato vicino.

Dori

Abstract

Emotions play a crucial role in people's everyday lives because they have important functions (intrapersonal, interpersonal, and social and cultural), and it is impossible to imagine life without emotions. While recognizing and managing emotions is an innate human ability, providing emotional intelligence to computer systems is a current research challenge.

This study focuses on the recognition of facial expressions, one of the most common non-verbal communication cues, through deep learning, i.e., a machine learning technique that teaches computers to perform classification tasks learning directly from images, text, or sounds. Deep learning methods, often referred to deep neural networks, were first theorized in the 1980s, but have only recently become valuable because they require large amounts of labelled data and processing power. In particular, wishing to recognize not only canonical expressions obtained by asking people to act and reproduce them but also, above all, spontaneous expressions, *ecologically valid* facial expression databases are needed.

Emotions can be aroused in people through imagery techniques, sounds and music, or images and films, and many databases have been developed for this purpose. While belonging to visual stimuli, Virtual Reality (VR) is supposed to be more impactful than traditional methods because it allows individuals to enter the situation related to the emotion that should be stimulated, move, interact and experience the surrounding environment more deeply. Still, there are no affective virtual environments databases as it happens for pictures, videos, and sounds.

VR technology, favoured by cost reduction, availability, and acceptance by users, has also become common in industry and has gained cost competitiveness with applications in various sectors, e.g., aerospace, automotive, and entertainment. Both emotional design and virtual reality are involved in the conceptual design, the early phase of the design process, where collecting emotional feedback for early

identification of successful customer-oriented products is increasingly attractive because consumers tend to make buying decisions emotionally. Questioning methods, physiological measurements, and observational methods are the three main ways to capture emotional feedback and can be used individually or in a multi-modal approach for a more comprehensive response, getting closer to the *ground truth* of emotion. Therefore, it is necessary to develop increasingly complex and precise deep learning algorithms, with spontaneous 3D data and cutting-edge architectures, for the recognition of facial expressions to meet the requirements of real applications, trying to bridge the gap between humans and machines and enabling them to communicate naturally and emotionally.

The thesis retraces all these fundamental steps to deeply analyze emotion recognition through facial expressions. Some possible applications are then proposed, with the adoption of increasingly advanced neural networks analyzing still images and videos. The elicitation of emotions through images or designed ad hoc virtual environments combining dynamism, interaction, and semantic and sensory elements, has been addressed, resulting in a novel spontaneous 3D Facial Expression Database, the CALD3R, and a database of affective Virtual Environments.

Introduction

Is Artificial Intelligence intelligent? What is the ground truth of emotion? Are facial expressions a reliable tool for understanding and capturing human emotions?

This dissertation starts with some research questions and aims to retrace the steps taken in the PhD programme in Management, Production and Design, a three-year journey with ever-increasing awareness.

The object of study is emotions, i.e., feelings we consciously or unconsciously experience daily, which play an essential role in how we think and behave. Emotions can help people make decisions, help others understand you better, and allow you to understand others. Thus, the concept of *emotional intelligence*, a term first used in 1985 by Wayne Payne but later popularized by psychologist Daniel Goleman, should be introduced. It is an inborn human characteristic, a complex process involving perceiving, reasoning, understanding, and managing emotions.

A current research question is *Can Artificial Intelligence have Emotional Intelligence?* [101]. Wanting to make a comparison, it can be stated that emotional intelligence is about emotions, like artificial intelligence is about collecting, analyzing, and interpreting data. Going deeper, artificial intelligence provides computer systems with a lot of cognitive intelligence (ability to learn) but no emotional intelligence (ability to deal with emotions), posing a problem not only in Human-Computer Interaction (HCI) but also when people connect and communicate with each other through technology. Indeed, most communication is conveyed through non-verbal cues, such as facial expressions, tone of voice, and body language, that are likely to be lost. *Affective computing*, a concept pioneered by Rosalind Picard in 1995 and currently widespread throughout the world, aims to bridge the gap between humans and machines, enabling them to communicate naturally and emotionally.

Affective computing technology [106], also known as Emotion AI, has applications in various business functions and industries, from marketing (to analyze, for instance, what makes customers engaged and organize communication strategies accordingly), to education (to assess how satisfied or frustrated students are with the lesson or to help autistic children recognize other people's emotions in the school environment). Other affective computing applications and use cases regard job interviews, observing how stressed candidates are and how they communicate emotions to make better recruitment decisions, employee training, helping employers improve their empathy and customer service skills, and safety, tracking drivers' emotional states and providing alerts for unsafe driving. Nowadays, vehicles are provided with external cameras for safety and operational applications. Still, internal cameras endowed with AI could increase road safety and make the occupants more comfortable adapting music, lighting, and temperature to the observed state. Moreover, another excellent potential could be to identify the influence of negative emotional situations on employees' productivity in a working environment and improve physical workspace design and comfort through sentiment analysis. Therefore, we can say that a system capable of perceiving emotions would benefit many people in various application fields. HCI is more similar to human communication than a sterile connection based on interfaces, where machines are no longer straightforward tools but intelligent collaborating assistants.

Many techniques have been studied over the years to perceive and discern people's states and emotions, including Facial Expression Recognition (FER), i.e., the recognition of emotions through the study of facial expressions. The detection and processing can be achieved through traditional methods by extracting facial surface geometric information and feeding it to various classifiers or Deep Learning techniques for feature extraction and classification. Deep Learning is a subset of Machine Learning (ML) based on Artificial Neural Networks (ANNs), usually called neural networks (NNs), i.e., computing systems inspired by the biological neural networks consisting of one input, one output, and multiple fully-connected hidden layers in between. Both Machine Learning and Deep Learning are subsets of Artificial Intelligence, a larger set that includes any technique that enables computers to mimic human intelligence. Among the various deep learning architectures, Convolutional Neural Networks (CNNs) and, more recently, Vision Transformers (ViTs) emerged, yielding state-of-the-art results in computer vision applications and overcoming the difficulties of the traditional methods.

On the other hand, deep neural networks are considered black boxes because it is hard to gain a comprehensive understanding of their inner working after they have been trained, making them hard to debug and understand how they make decisions, and they require large amounts of data. Creating an emotion database is an essential task, but simultaneously tricky and time-consuming, to train a system to recognize human emotions. Most public databases include posed facial expressions only, where participants are asked to act rather than experience emotions and display natural expressions. *Ecologically valid* facial expression databases require more effort to select proper stimuli to elicit intended emotions and manually label them by trained individuals. However, having valid data to train machine learning algorithms is crucial to perform facial emotion and expression recognition in real-life situations. Visual stimuli are the most common and used method to elicit emotions, but Virtual Reality (VR) can be successfully employed for this purpose. Indeed, VR allows individuals to enter the situation related to the emotion that should be stimulated, move, interact and experience the surrounding environment more deeply, all favoured by cost reduction, availability, and acceptance by end-users of VR technology.

This dissertation starts with an introduction (Chapter 1) to how Emotional Design and Virtual Reality are involved in Product Lifecycle Management (PLM), especially in the early phase of the design process, investigating a relevant application of affective computing technology. It continues with a literature review (Chapter 2) highlighting the weaknesses and strengths of traditional and deep learning techniques to perform Facial Expression Recognition based on the six basic emotions, i.e., anger, disgust, fear, happiness, sadness, and surprise, theorized by Paul Ekman and accepted as universal. Thanks to technological development, 3D facial data have become increasingly used because naturally overcome challenging problems of illumination and pose variations, and deep learning architectures have been successfully employed in computer vision. Chapter 3 aims to open the black boxes, i.e., as neural networks are called, and understand abstractions in CNNs through linear correlation and symbolic regression between 3D facial descriptors and activations of convolutional layers. Chapter 4 deals with the other main problem of deep neural networks, which is the lack of *ecologically valid* data (where *ecological validity* refers to the fact that participants were not asked to act but to react spontaneously to stimuli) to train machine learning algorithms, and presents the CALD3R, a novel spontaneous facial expression database. A multimodal emotion recognition system has been trained and tested on it, demonstrating the advantage of employing three-dimensional data. In the

CALD3R database, emotions are elicited with images from the IAPS and GAPED databases (visual stimuli), while in Chapter 5, VR has been proposed as a successful method to arouse emotions. Ten affective interactive Virtual Environments (VEs) have been developed following a design methodology considering different aspects that influence users' emotional experience. The VEs were then tested and validated, confirming their effectiveness in eliciting specific emotions. A transformer-based architecture using ecologically valid data elicited through the VEs has been used to recognize emotions in spontaneous videos, which represents a further improvement compared to the recognition of still images. The path ends with two case studies (Chapter 6), where education and driving assistance affective computing applications have been investigated. The first one aims to evaluate user engagement by combining the User Engagement Scale (UES) questionnaire and a deep learning-based Facial Expression Recognition system, while the second aims to monitor drivers' attention during a driving simulation.