COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

RECOGNIZING AND IMITATING EMOTIONS IN A ROBOTIC SYSTEM

Diploma Thesis

Bratislava, 2020 Bc. Oswaldo Macedo

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

RECOGNIZING AND IMITATING EMOTIONS IN A ROBOTIC SYSTEM

Diploma Thesis

Study programme:	Cognitive Science
Field of study:	Computer Science
Supervising department:	Department of Applied Informatics
Supervisor:	prof. Ing. Igor Farkaš, Dr.

Bratislava, 2020 **Bc. Oswaldo Macedo**

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Oswaldo Macedo
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Recognizing and Imitating Emotions in a Robotic System

Annotation: Emotions are an important component in human-robot interaction. Human companion robots need to be sociable and responsive towards emotions to better interact with the human environment they are expected to operate in. Therefore, designing such systems capable of learning is a challenge that requires interdisciplinary effort within cognitive robotics.

Aim:
1. Implement and test a neural network able to recognize basic human emotions.
2. Associate recognized human emotions with corresponding emotional state of the simulated humanoid robot NICO.
3. Test the ability of the robot to imitate human emotions.

Literature: Churamani M. et al. (2017). Teaching emotion expressions to a human companion robot using deep neural architectures. International Joint Conference on Neural Networks, IEEE.
Kerzel M. et al. (2017). NICO – Neuro-Inspired COmpanion: A Developmental Humanoid Robot Platform for Multimodal Interaction. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).
Sunderhauf N. et al. (2018). The limits and potentials of deep learning for robotics. Adaptive Behavior, The International Journal of Robotics Research, 37(4–5) 405–420.

Supervisor:	prof. Ing. Igor Farkaš, Dr.
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	prof. Ing. Igor Farkaš, Dr.
Assigned:	25.02.2019

Approved: 25.02.2019 prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

....

Student Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Oswaldo Macedo
Študijný program:	kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor:	informatika
Typ záverečnej práce:	diplomová
Jazyk záverečnej práce:	anglický
Sekundárny jazyk:	slovenský

Názov: Recognizing and Imitating Emotions in a Robotic System
Rozpoznávanie a Imitácia emócií v robotickom systéme

Anotácia: Emócie sú dôležitou súčasťou interakcie človek-robot. Robotickí spoločníci človeka by mali byť spoločenskí a vnímaví na emócie, aby mohli lepšie interagovať s prostredím človeka, v ktorom by mali fungovať. Preto, dizajn takýchto učiacich sa systémov je výzvou, ktorá vyžaduje interdisciplinárnu snahu v rámci kognitívnej robotiky.

Cieľ: 1. Implementujte a otestujte neurónovú sieť schopnú rozpoznať základné ľudské emócie.
2. Vytvorte asociáciu rozpoznaných emócií s korešpondujúcim emocionálnym stavom simulovaného humanoidného robota NICO.
3. Otestujte schopnosť robota imitovať ľudské emócie.

Literatúra: Churamani M. et al. (2017). Teaching emotion expressions to a human companion robot using deep neural architectures. International Joint Conference on Neural Networks, IEEE.
Kerzel M. et al. (2017). NICO – Neuro-Inspired COmpanion: A Developmental Humanoid Robot Platform for Multimodal Interaction. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).
Sunderhauf N. et al. (2018). The limits and potentials of deep learning for robotics. Adaptive Behavior, The International Journal of Robotics Research, 37(4–5) 405–420.

Vedúci:	prof. Ing. Igor Farkaš, Dr.
Katedra:	FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry:	prof. Ing. Igor Farkaš, Dr.
Dátum zadania:	25.02.2019

Dátum schválenia: 25.02.2019

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.... študent

.... vedúci práce

Declaration

I hereby declare that this diploma thesis is the product of my own work, and information taken from other authors is cited as such.

Acknowledgments

I would like to thank prof. Ing. Igor Farkaš Dr. for his valuable help, knowledge and patience throughout all the process of elaborating this thesis. I would also like to thank my parents for always allowing me to choose the path that I wanted to follow.

Abstract

For decades, interest in emotion recognition by artificial intelligence has been growing steadily. Lately, especially because of improvements in the processing capabilities of computers and the creation of new machine learning algorithms, this path of research grew even faster, substantially improving the accuracy of the task. Currently, emotion recognition through facial expressions can achieve close to 73% accuracy with state-of-the-art implementations tested with images in natural settings (Pramerdorfer & Kampel, 2016). Comparably, human accuracy tested on the same images reaches 65±5% (Goodfellow et al., 2013). This thesis consists of two main parts: first, a neural network based emotion recognition system trained with static images of faces with a neutral expression and six basic emotions (sadness, happiness, anger, fear, disgust, surprise); and second, an emotion imitation system trained on outputs of the emotion recognition system, learning to imitate the facial expressions of new images. The thesis combines state-of-the-art level facial expression recognition implementations with an emotion imitation system to produce a human–robot interactive system capable of perceiving and imitating human emotions through facial expressions. For this purpose, state-of-the-art neural network models are modified and trained with static images of faces from three commonly used datasets of facial expressions. Also, the imitation system is trained with associative learning, using the resultant information from the previously trained neural network. The emotion recognition system achieves an average accuracy of 62% on the most challenging, natural setting dataset used in this research; while achieving an average accuracy of 83% on laboratory setting datasets. The emotion imitation system correctly associates all images with an average accuracy of 96% on the laboratory setting dataset, which results in correctly imitating emotion with a 78% accuracy in combination with the emotion recognition system. Regarding the natural setting dataset, association accuracy is 91% on average, which combined with the recognition system results in 56% average imitation accuracy. The results show that the robotic system would not be very accurate at imitating emotions in a natural setting with current emotion recognition capabilities, nevertheless, on a laboratory setting, it could be viable.

Keywords: associative learning, computer vision, convolutional neural networks, emotion, facial expression recognition

Abstrakt

Už niekoľko desaťročí sa záujem o rozpoznávanie emócií umelou inteligenciou neustále zvyšuje. V poslednom čase, najmä z dôvodu zlepšenia spracovateľských schopností počítačov a vytvorenia nových algoritmov, výskum v tejto oblasti rástol ešte rýchlejšie, čím sa podstatne zvýšila presnosť rozpoznávania emócií. V súčasnosti môže rozpoznávanie emócií prostredníctvom výrazov tváre dosiahnuť takmer 73%-nú percentnú presnosť pomocou najmodernejších implementácií testovaných s obrázkami v prirodzenom prostredí (Pramerdorfer & Kampel, 2016). Táto práca sa skladá z dvoch hlavných častí: po prvé, systém rozpoznávania emócií, ktorý má umelú neurónovú sieť, ktorá je vycvičená pomocou statických obrazov tvárí s neutrálnym výrazom a vyjadrením 6 základných emócií (smútok, radosť, hnev, strach, znechutenie, prekvapenie); a po druhé, systém imitácie emócií, ktorý je trénovaný s výsledkami systému rozpoznávania emócií a napodobňuje výrazy tváre v nových obrazoch. Diplomová práca kombinuje najmodernejšie implementácie rozpoznávania výrazov tváre so systémom imitácie emócií a vytvára interaktívny systém človek-robot schopný vnímať a napodobňovať ľudské emócie prostredníctvom výrazov tváre. Za týmto účelom sú najmodernejšie modely neurónových sietí modifikované a trénované pomocou statických obrazov tvárí z troch bežne používaných súborov dát výrazov tváre. Imitačný systém je tiež trénovaný s asociatívnym učením, využívajúc výsledné informácie z predtým trénovaných neurónových sietí. Systém rozpoznávania emócií dosahuje presnosť 62% v prípade najnáročnejšieho dátového súboru s prírodnými podmienkami použitom v tomto výskume, pričom dosahuje presnosť 83% na dátovom súbore s laboratórnymi podmienkami. Imitačný systém správne asociuje všetky obrázky s priemernou presnosťou 96% na dátovom súbore s laboratórnymi podmienkami, čo vedie k správnej imitácii emócií so 78% presnosťou v kombinácii so systémom rozpoznávania emócií. Pokiaľ ide o dátový súbor s prírodnými podmienkami, presnosť asociácie je v priemere 91%, čo v spojení so systémom rozpoznávania vedie k priemernej presnosti imitácie 56%. Výsledky ukazujú, že robotický systém by nebol veľmi presný pri imitácii emócií v prirodzenom prostredí so súčasnými schopnosťami rozpoznávania emócií; napriek tomu, v laboratórnom prostredí by mohol byť realizovateľný.

Kľúčové slová: asociatívne učenie, počítačové videnie, konvolučné neurónové siete, emócie, rozpoznávanie výrazov tváre

Contents

List of Tables

List of Figures

Introduction

Affective computing has the potential to become part of our everyday life. Advances in research and technology allowed emotion recognition systems to be used in several fields of work like customer experience and fraud detection (Gartner, 2017). Also, in the coming years emotion recognition systems could be used more extensively in fields such as healthcare, education, transportation, and security (Dzedzickis, Kaklauskas, & Bucinskas, 2020; Lucey et al., 2010). Due to the importance and growing interest in the field of affective computing, this research is focused on emotion recognition and imitation through facial expressions, an interdisciplinary research which is tightly related to concepts from psychology, artificial intelligence, and to a lesser extent, neuroscience. Our objective was to use an interdisciplinary approach by evaluating artificial neural network models in conditions similar to human evaluations, with facial expressions datasets with minimal preprocessing, and evaluate a biologically inspired association model for the task of imitation of emotions. The main question to answer with this research is whether a computational system is capable of recognizing and imitating emotions accurately. Two hypotheses were formulated from this question; 1. Will the emotion recognition system be as performant as state-of-the-art implementations; 2. Will the imitation system be able to imitate emotions through associative learning?

The thesis is organized as follows. Chapter 1 deals with the many definitions of emotion and universality of emotions, it also gives a theoretical background for affective computing and convolutional neural networks. Chapter 2 reviews current state-of-the-art models in emotion recognition and imitation and compares human accuracy with current emotion recognition computational solutions. Chapters 3 and 4 present the emotion recognition and imitation systems, their methodology and results, respectively. Finally, in Chapter 5 we discuss the results, limitations of this thesis and ethical concerns. With the help of this research, it is expected that we will be able to better understand what architectures fit best emotion recognition and if it is plausible to teach robots to imitate emotions through associative learning.

1 Theoretical background

1.1 Definitions of emotion

It is important to mention first what is meant by emotion, and for that, several definitions of emotion are provided next to indicate that there is not one precise definition of emotion, and that authors of these definitions can, over time, modify them adding new findings or clarifications.

According to the American Psychological Association (n.d.) emotion is “a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event.”

Scherer defines emotion as “an episode of massive, synchronous recruitment of mental and somatic resources to adapt to, or cope with, a stimulus event that is subjectively appraised as being highly pertinent to the needs, goals, and values of the individual.” (Barrett, Niedenthal, & Winkielman, 2007, p. 314)

Sander and Scherer (2009, p. 106) define emotions “as transient, bio-psychosocial reactions designed to aid individuals in adapting to and coping with events that have implications for survival and well-being.”

We can see that the previous definitions do have common elements, emotions influence our behavior, emotions play a role in dealing with events that affect us, and they also seem to be of short duration, differentiating them from moods, attitudes or traits (Sander & Scherer, 2009, p. 202). Apart from these definitions from a psychological perspective, they are, neuroscientific perspectives such as “emotion is a fundamental property of the brain and is instantiated in distributed circuitry that enables emotion to interact with other major mental functions such as attention and memory” (Sander & Scherer, 2009, p. 199), and early philosophical perspectives as “an attempt to bundle together states that were supposedly marked by a degree of 'emotion', a metaphorical extension of the original sense of the word, namely agitated motion, or turbulence.” (Sander & Scherer, 2009, p. 200).

When we consider all the perspectives of what emotions are, in general, they seem to refer to the same phenomenon. However, it is in the details that differences arise, such as between Barrett, Niedenthal, and Winkielman (2007, p. 314) where emotion affects behavior to achieve individual needs, goals, and values, and Sander and Scherer (2009, p. 106) where emotion is a reaction of an individual in order to survive and avoid danger.

It is because of these details and other differences that are not even mentioned in general definitions, that Scherer created a framework that would allow several disciplines (neuroscience, psychology, philosophy, anthropology) and different perspectives inside those disciplines to continue their research having a common ground that would allow for more compatibility between these perspectives.

The framework is as follows:

> Emotions (1) are focused on specific events, (2) involve the appraisal of intrinsic features of objects or events as well as of their motive consistency and conduciveness to specific motives, (3) affect most or all bodily subsystems which may become to some extent synchronized, (4) are subject to rapid change due to the unfolding of events and reappraisals, and (5) have a strong impact on behaviour due to the generation of action readiness and control precedence (Sander & Scherer, 2009, p. 202).

The above-mentioned framework is used when referring to emotions throughout this thesis.

1.2 Universality of emotions

This is a long-debated topic – are emotions the same across all human cultures, or do they differ by culture – which can be attributed to the nature-vs-nurture debate. Nurture alone means that a person is a blank slate (*tabula rasa*) who learns everything from the culture they live in, while nature alone means that humans come ready when born having everything they need to survive, including emotions. Behaviorists from mid-20th century like B. F. Skinner were inclined to nurture albeit they would not disregard nature. By the end of the 20th century there was already more agreement that nature and nurture are entangled, and humans have something from both concepts (Sander & Scherer, 2009, p. 202).

Paul Ekman dedicated plenty of his research to find emotion universals, he proposed the theory of basic emotions, which includes the emotions anger, joy, surprise, fear, sadness, and disgust (Figure 1.1); and later he proposed to add other emotions that could be universal, such as, amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, and shame (Ekman, 2005; Reisenzein, 2015). Several experiments with participants from different countries, focused on the initial six emotions, led to the conclusion that basic emotions are universal, regardless of cultural differences (Ekman, 2005; Ekman & Friesen, 1971; Reisenzein, 2015). Nevertheless, social norms can influence the expression of any given emotion as seen in Ekman & Friesen (2003) where they compared cultural differences between American and Japanese participants when they express emotions: while Americans would express the emotions in group settings in the same way as when alone, Japanese would not display negative emotions in group settings, however, they would express them when alone (not being directly observed).

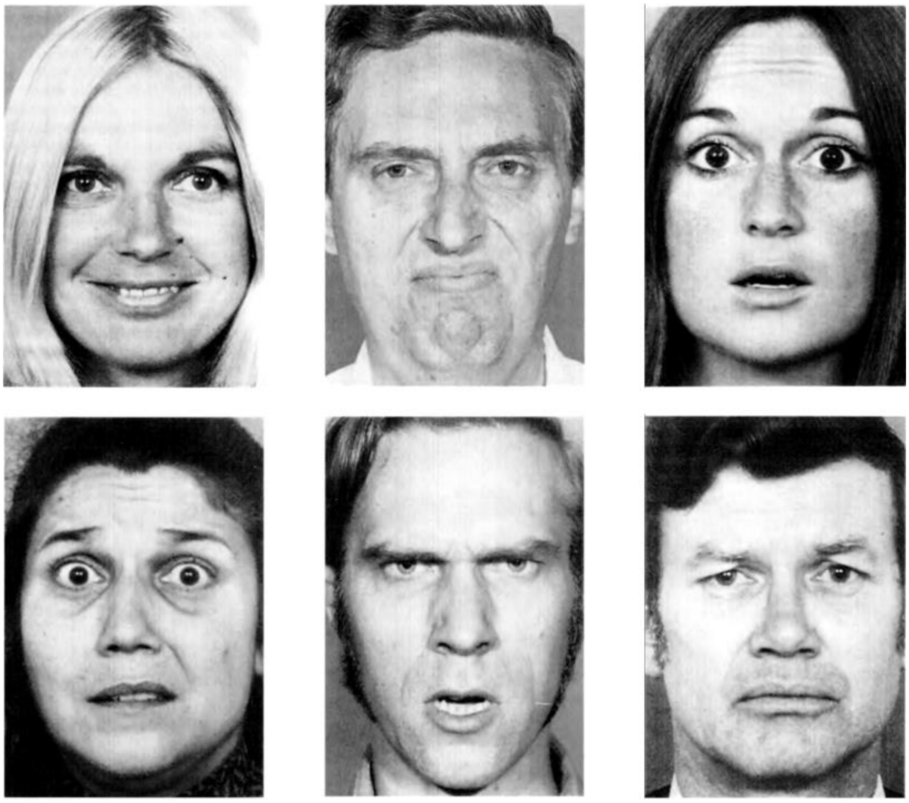


Figure 1.1. Examples of facial expressions of 6 basic emotions as defined by Paul Ekman. From top left to bottom right: happiness, disgust, surprise, fear, anger, sadness (Ekman & Friesen, 2003, pp. 104, 181, 175, 179, 185, 195).

1.3 Affective computing

Affective computing is an interdisciplinary field of research that uses computational systems to detect, analyze, simulate, and convey emotions and related affective states (Sander & Scherer, 2009). It is very important for human-robot interaction, were it aids communication by displaying empathy and emotional intelligence. However, it is important to notice that the display of the emotions by the robot, or another computational system does not mean that the robot feels them or understands them as we do. Human emotions may be simulated at different levels of complexity in a robot by adding modules that are inspired by the nervous and other systems of the human body.

In affective computing different sensors can be used to recognize emotions through several modalities, from which the most common are visual (video or static images) of facial expressions, body posture, and gestures; and auditive. Additionally, physiological modalities like electro-encephalography (EEG), electrocardiography (ECG), electrodermal activity (EDA) also called galvanic skin response (GSR), electromyography (EMG), heart rate variability (HRV), respiratory rate analysis (RR), skin temperature measurements (SKT), and electrooculography (EOG) can be used; usually in a multimodal manner where two or more modalities complement each other in order to recognize the desired emotions (Dzedzickis et al., 2020; Janssen et al., 2013; Liu, Sourina, & Nguyen, 2010).

1.4 Neural networks for emotion recognition

Neural networks are commonly used in computer vision problems such as object detection, classification, and emotion recognition; specifically, convolutional neural networks (CNN) have seen a great increase in popularity in the last decade because of the accuracy they achieve compared to other solutions (Kim, Roh, Dong, & Lee, 2016; Ko, 2018). They are part of the connectionist approach which is an interdisciplinary approach that tries to explain the inner workings of the brain with the premise that biologically inspired structures that resemble neurons and their connections can be created to simulate brain activity, in the case of CNNs they model areas of the human visual system (Kriegeskorte, 2015; Lindsay, 2020).

In the case of emotion recognition through facial expressions, the input of a CNN is an image with a human face and the output is a label that the CNN assigned to that image. They are several algorithms that are organized into preprocessing, training methods, and algorithms in the network architectures (Kim et al., 2016).

Training algorithms can be gradient based, which have as an objective to minimize the loss function to obtain a local minimum (Bushaev, 2017). The loss function shows how good a neural networks is for a task, while lower the value of a loss function is, the better the network is for the specified task (Bushaev, 2017).

In order to better understand how CNNs work, the general structure composed of layers is presented next:

- Convolutional layers split the information from the image into feature maps that store the information of where the feature occurs in the image, and have as an output an activation map (Can, 2017).
- Rectified linear unit (ReLU), which is a linear function with a threshold at zero, that propagates the gradient through the network avoiding the common problem of vanishing gradient that prevents weights from updating (Can, 2017). Also, because the function has a threshold at zero, negative values are not taken into account which avoids the cancellation problem were positive values and negative values would cancel out each other, thus improving robustness when the inputs are noisy (Can, 2017). ReLUs also present simple computations (comparisons between values) which makes them more efficient in CNNs compared to other functions (Can, 2017).
- Pooling layers, which reduce the spatial size of the activation map reducing computational requirements and minimizing overfitting of the data (Can, 2017). The pooling layer can discard more than 75% of the data with the smallest sized filter which can limit depth and performance, however, they are still used, especially in image recognition tasks, because the detected features are more important than their location in the image (Can, 2017).
- Fully connected layers, which in CNNs are multi-layer perceptrons that map the outputs from the combination of the previous layers and have the goal of tuning the weights of the network and producing the final output of the CNN (Can, 2017).

The previously mentioned layers are ordered in the architecture in an interspersed fashion of convolutional layers and pooling layers, with ReLU usually being part of the convolutional layer, and the fully connected layers at the end of the CNN.

2 State of the art in emotion recognition and imitation

There are not many studies in emotion recognition that compare human accuracy with machine learning accuracy, and the state-of-the-art machine learning solutions are often taken as the benchmark. However, these studies are necessary to assess machine learning performance on emotion recognition. Next, some of these comparative studies are presented, and the results show that by now machine learning can be as good or even better at recognizing emotions than a layperson (Janssen et al., 2013).

Currently, there already are companies on the market that offer emotion recognition software for different uses such as marketing, product placement, gaming, driving safety, education, and healthcare; however, specifically for facial expression recognition, the accuracy of their products did not achieve human accuracy levels when tested on two facial expressions datasets (BU-4DFE, UT-Dallas) in a study made by Dupré, Krumhuber, Küster, and McKeown (2019); human accuracy of 14 participants was 73%, while the emotion recognition software of 8 companies ranged from 49% to 62% accuracy. As Dupré et al. mention, the software had problems with spontaneous emotions from UT-Dallas dataset, and it was most accurate in recognizing a happy expression even surpassing human accuracy for this facial expression. On the other hand, one of the software companies claims an above 80% accuracy in emotion recognition of seven basic emotions (anger, contempt, disgust, fear, joy, sadness, and surprise), and mentions that their software is trained on millions of images of faces from 86 countries in real-world environments (Affectiva, 2017), which implies that the conditions of the environment, or the input data differ from the study by Dupré et al.

In another study, Janssen et al. (2013) set as the benchmark the human accuracy in recognizing emotions. The researchers recorded physiological signals and videos with audio of Dutch participants that elicited emotions by describing emotional events that they had experienced; later, they presented the videos in three modalities, audio only, video only, and audio-video to two other groups of participants and the emotion recognition software, and compared the accuracies between all the groups. One group was composed of Dutch native speakers, and the other of American participants that could not understand Dutch. The results show that for humans, semantic understanding and context are important for emotion recognition, and the accuracy of the recognition software surpasses that of a layperson considerably if the person does not have a semantic understanding of the elicited emotions. The accuracy of the recognition software was 65% with audio-video data, while for humans it was 47.9% for Dutch participants and 31% for Americans. They also mention that acted emotion elicitation is more intense and easier to detect for humans, in this case the emotion elicitation might have been more subtle because of the artificial setting the participants were in, and because the participants were not acting according to the tests they took to ascertain they were expressing the emotions they felt.

In a study by Esparza, Scherer, Brechmann, and Schwenker (2012), emotion recognition through audio was tested in humans and two machine learning algorithms, hidden Markov models and support vector machines. The participants analyzed audio recordings from two German datasets (WaSeP and emoDB) of acted emotional speech. The accuracy of humans for the datasets WaSeP and emoDB were 84% and 84.7% respectively, while the accuracy of the algorithms was 84% for WaSeP and 77% for emoDB.

In addition, it is important to mention the study by Cao, Cooper, Keutmann, and Gur (2014) who created the crowd-sourced emotional multimodal actors dataset (CREMA-D)(Cao et al., 2014). The researchers contacted actors to elicit five emotions (anger, disgust, fear, happiness, sadness) with varied intensities, and generated visual, audio, and audio-visual data. The purpose of the research was to determine how accurate can humans be in recognizing emotions of different intensities through different modalities. As with other research in human emotion recognition the results are similar, 63.6% of accuracy for audio-visual data, 58.2% for visual data, and 40.9% for audio-only data.

2.1 Datasets of facial expressions

Before mentioning state-of-the-art models of emotion recognition, it is important to describe some of the most common datasets used for facial expression recognition to have perspective on how varied they can be, and what challenges they can bring. They are different publicly available datasets that vary in the number of labeled emotions they contain, commonly from only two emotions up to eight. Further, datasets about facial expression recognition are divided into laboratory setting datasets, which have images of the faces taken frontally, centered, with specific lighting conditions, and oriented horizontally; and natural setting datasets (also called in the wild), which present different lighting conditions, they can be rotated, with the face uncentered, occluded, and have images taken in different environments (Dhall, Goecke, Gedeon, & Sebe, 2016). State-of-the-art solutions are tested on one or more of the following datasets.

FER-2013 dataset, which is a natural setting dataset created by Pierre Luc Carrier and Aaron Courville and used for the 2013 emotion recognition in the wild contest (Goodfellow et al., 2013); this is the biggest publicly available dataset to date of writing, with 35887 labeled images distributed into 6 emotions (anger, fear, sadness, surprise, happiness, disgust) and a neutral expression as seen in Figure 2.1.

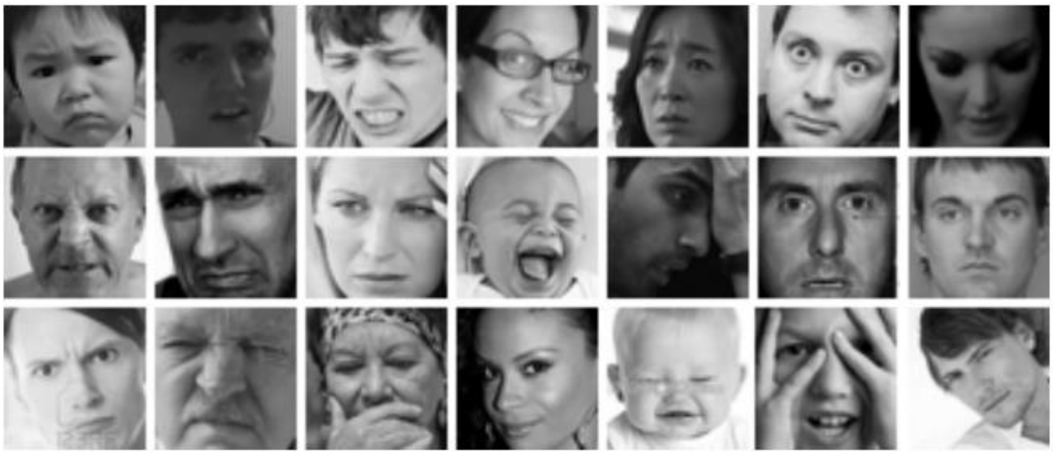


Figure 2.1. FER-2013 dataset sample images; from left to right: angry, disgust, fear, happy, sad, surprise, and neutral expressions (Pramerdorfer & Kampel, 2016).

GENKI-4K dataset (Http://mplab.ucsd.edu, 2009), which is a natural setting public dataset of 4000 smile and non-smile images (Figure 2.2), taken from the bigger 20 000 images dataset GENKI created to train algorithms for the use in cameras to take pictures of people when smiling (Whitehill, Littlewort, Fasel, Bartlett, & Movellan, 2009).

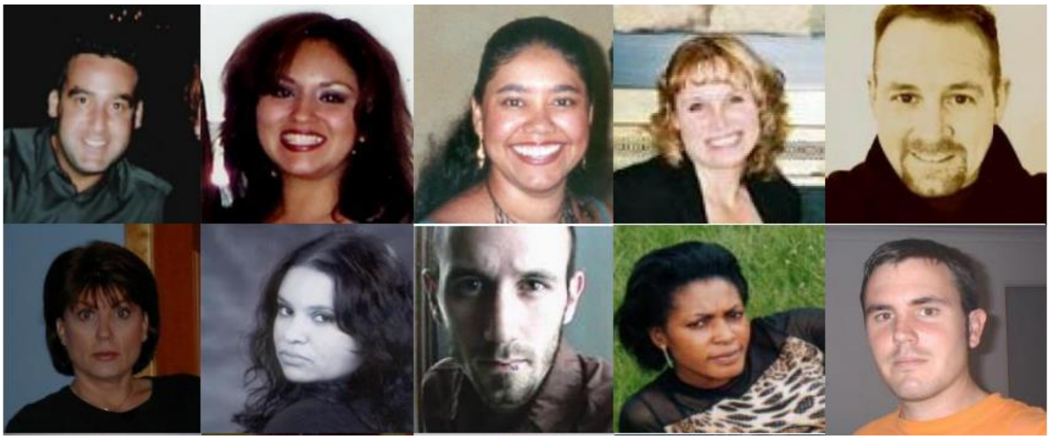


Figure 2.2. Smile (top) vs non-smile (bottom) images from GENKI-4K dataset (Sang, Bao Cuong, & Thuan, 2017).

SWEF (Static Facial Expressions in the Wild) database (Dhall, Goecke, Lucey, & Gedeon, 2011), which has 700 natural setting images for emotion recognition classified into 6 emotions (anger, fear, sadness, surprise, happiness, disgust) and a neutral expression (Figure 2.3), created from the Acted Facial Expression in the Wild (AFEW) dataset, which is a collection of labeled movie clips created for the purpose of having data that would be comparable to a real-world scenario; these datasets are used in the EmotiW challenges (Dhall et al., 2016).

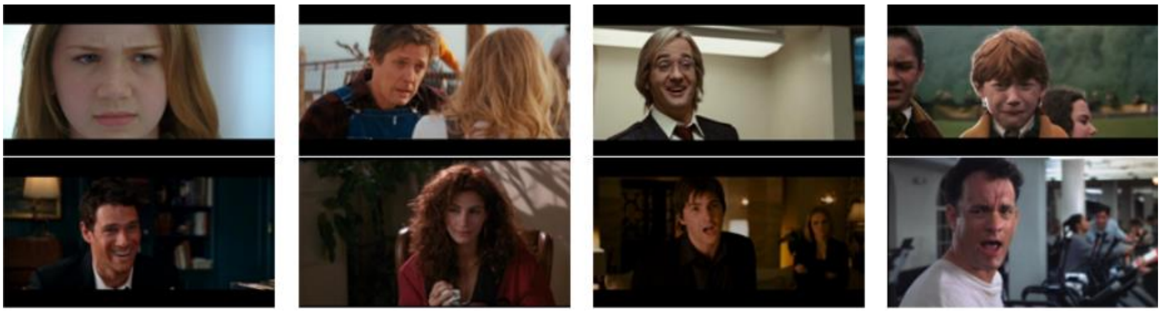


Figure 2.3. SFEW database sample images (Dhall et al., 2011).

JAFFE (Japanese Female Facial Expression) database (Lyons, Akamatsu, Kamachi, & Gyoba, 1998) which has 213 images of 10 female participants expressing anger, fear, sadness, surprise, happiness, disgust, and a neutral expression in a more subtle way than most laboratory setting datasets as seen in Figure 2.4; as well, the amount of data is considerably smaller compared to other datasets; regardless, this dataset is used commonly in research.

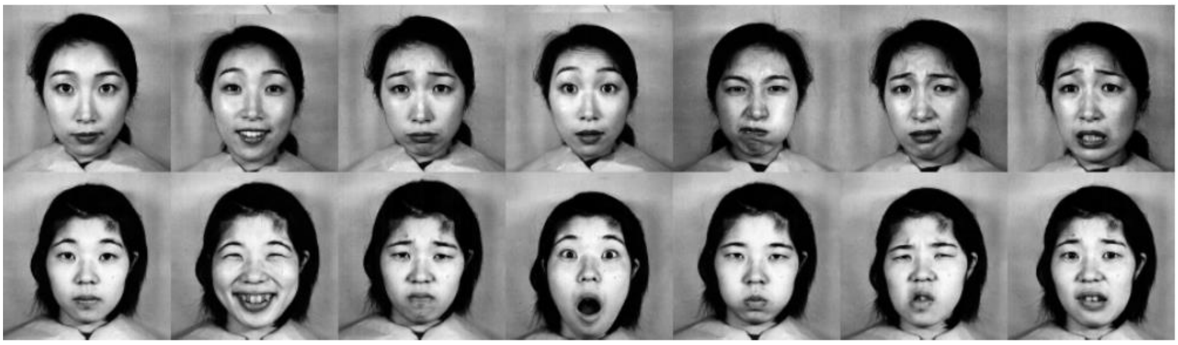


Figure 2.4. Sample images from JAFFE database of two out of ten participants expressing, from left to right, neutral, happy, sad, surprise, angry, disgust, and fear expressions (Lyons et al., 1998).

At last, CK+ (the extended Cohn-Kanade dataset) which has 593 sequences of emotions going from neutral to peak emotion in a laboratory setting from 123 participants, with contempt present as an additional emotion to the other six previously mentioned emotions (Figure 2.5) (Lucey et al., 2010).

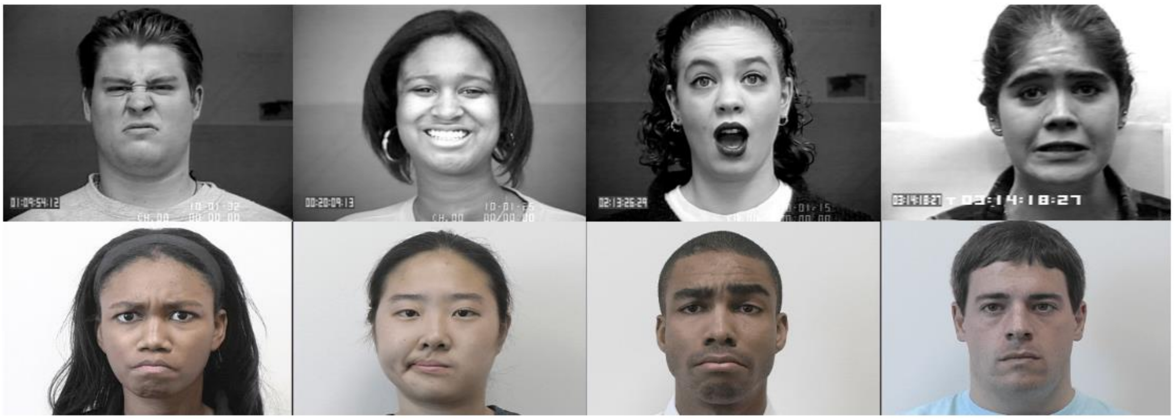


Figure 2.5. Sample images of all expressions from CK+ dataset; from top left to bottom right: disgust, happy, surprise, fear, angry, contempt, sad, and neutral (Lucey et al., 2010).

2.2 Models of emotion recognition

Currently, artificial neural networks are the most used computational models in emotion recognition, specifically convolutional neural networks (CNNs); next, two studies are presented that managed to achieve top accuracy in facial expression recognition.

2.2.1 Hierarchical committee of deep convolutional neural networks

This study by Kim, Roh, Dong, and Lee (2016) was the winning submission for the EmotiW2015 challenge (Dhall et al., 2016), it used a group of CNNs with different parameters and architectures, which allows the entire model to generalize better and improve accuracy. The group of CNNs was organized into committees where each output would get averaged in a hierarchical structure called exponentially-weighted average (EWA) (Kim et al., 2016) as seen in Figure 2.6.

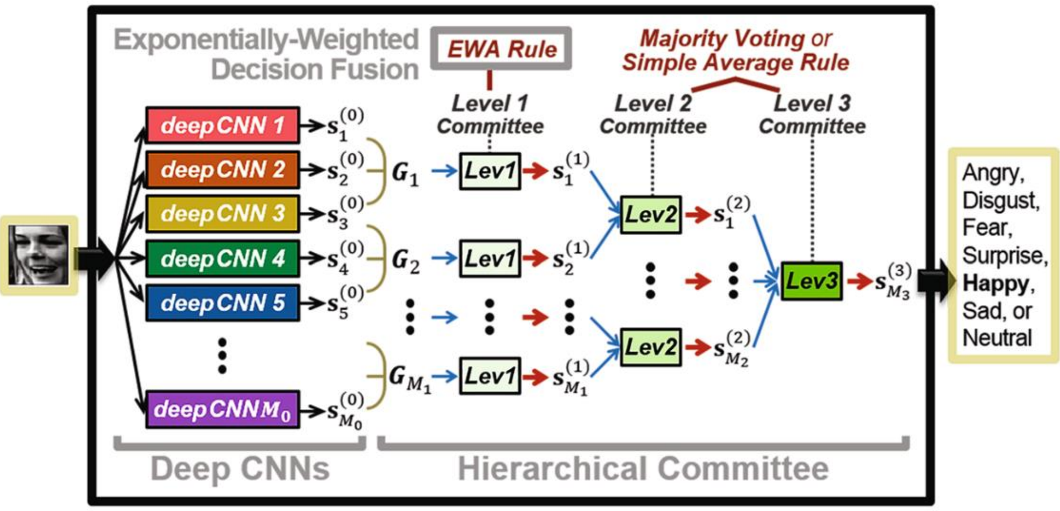


Figure 2.6. Overall architecture of a hierarchical committee of deep CNNs (Kim et al. , 2016).

With the best hierarchical architecture of this kind with 240 deep CNNs, Kim et al. managed to achieve an accuracy of 61.6% on the EmotiW2015 challenge SFEW database, while their best single deep CNN managed to achieve an accuracy of 57.3%, both significantly above the baseline of 39.1% accuracy on the SFEW database.

2.2.2 The use of deep neural networks for emotion recognition

Pramerdorfer and Kampel (2016) reviewed the state of the art of CNN solutions for facial expression recognition and searched for possible bottlenecks in those solutions; they found two main problems; the data available in the field of facial expression recognition is too small compared to other fields of research and CNNs that are used are too shallow in comparison with state of the art CNNs in other fields, such as object classification. When Pramerdorfer and Kampel tested three state-of-the-art CNNs from object classification on the FER-2013 dataset, they all performed better with 72.7%, 71.6%, and 72.4% accuracy in comparison with the best shallow model by Kim et al. (2016) with 70.58% accuracy on the same dataset. Additionally, when they created an ensemble of 8 of these CNNs, in a similar fashion as in Figure 2.6., they achieved an accuracy of 75.2%, compared to 72.72% from Kim et al. (2016).

2.3 Models of emotion imitation

In the field of human–robot interaction it is not only important for the robot to recognize human emotions, but also to imitate the human emotions, to allow a more meaningful communication between both. Churamani, Kerzel, Strahl, Barros, and Wermter (2017) created a model with 4 neural networks that recognizes facial expressions and learns to imitate them by association with rewards and penalties. They use the Neuro Inspired Companion robot (NICO) which is a child-sized modular robot with multi-modal sensors, with tactile, auditory, visual, and proprioception capabilities (Figure 2.7).

Churamani et al. (2017) model uses 4 emotions (anger, happiness, sadness, surprise) and a neutral expression for the recognition and imitation tasks because of the difficulty for participants to identify NICO’s disgust and fear expressions (Figure 2.8). The model was pre-trained using the CK+ dataset, and then trained with 5 participants that expressed one of the five emotions when prompted by the researchers. For the association learning to occur, the participants waited for NICO to express an emotion, and, if the expression was correct, the participant would express a happy expression as a reward, and an angry expression as a penalty. The learning by association was effective, and to improve accuracy, Churamani et al. simulated 100 additional interactions using images of the participants facial expressions. Their results show high accuracy in imitation of the emotions, with 100% accuracy for anger, happiness, and neutral, 78% for sadness, and 86% for surprise.

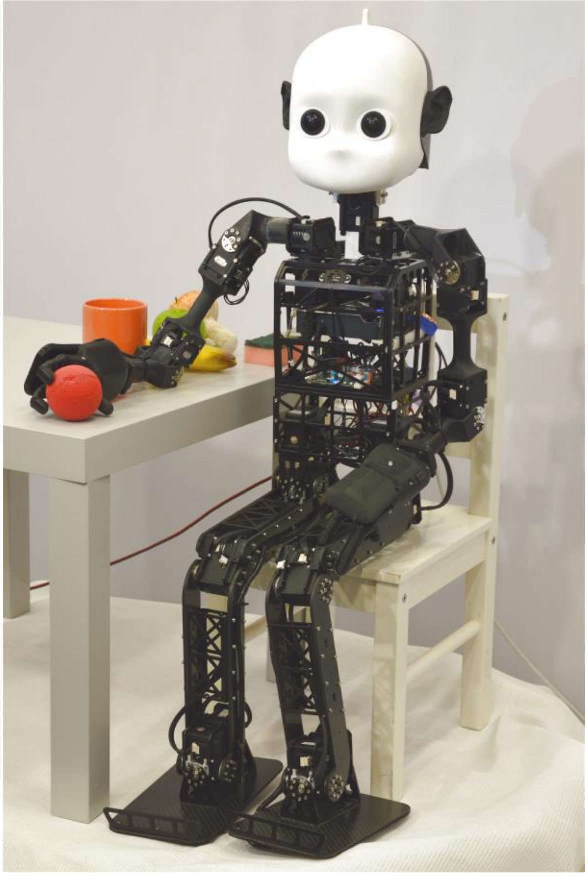


Figure 2.7. NICO (Neuro Inspired COmpanion) multimodal robot (Kerzel et al., 2017).

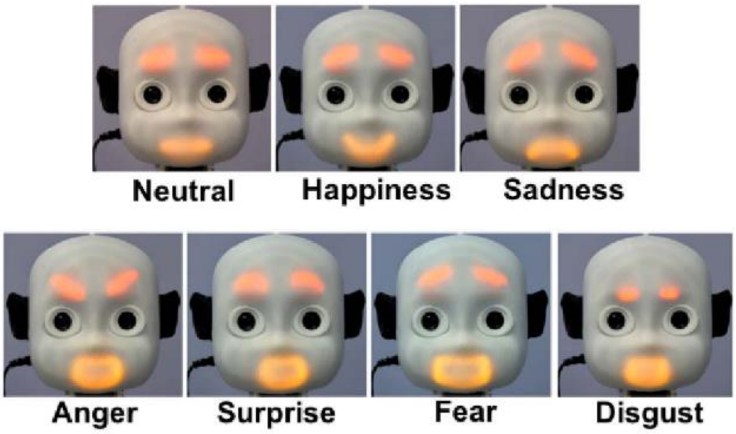


Figure 2.8. NICO expressing seven basic emotions (Churamani et al., 2017).

3 Emotion recognition system

The emotion recognition system has as objective accurately recognize emotions from facial expressions. The approach was to use supervised learning with CNNs that use as input labeled static images of human faces expressing anger, happiness, fear, sadness, surprise, disgust, and a neutral expression.

3.1 Methodology

CNNs were selected because they are inspired in the visual system (Kriegeskorte, 2015; Lindsay, 2020), and because they show improved accuracy over other methods in the task of emotion recognition through facial expressions (Goodfellow et al., 2013; Kim et al., 2016; Pons & Masip, 2018; Pramerdorfer & Kampel, 2016). The recognition system was based on an implementation of a facial expression recognition project on the JAFFE dataset by Patel (2018), and adapted to the version 3.5.6 of Python programming language.

The machine learning platform used was TensorFlow and the application programming interface (API) used to create and evaluate the neural networks was Keras by François Chollet (2015). Keras allows to create deep neural networks easily with modules for every type of layer, activation function, regularizers; as well as, preprocessing, training, and analysis of the data.

For the experiments, a DELL G5 5587 computer with Intel Core i7-8750H (2.2 GHz) CPU, 16 GB RAM, and a 6GB NVIDIA GeForce GTX 1060 with max-Q design graphics card (GPU) with compute capability of 6.1 was used. Additionally, for the more challenging computations, a GPU computing server with AMD Ryzen 7 2700 (3.2 GHz) CPU, 48 GB RAM, and a 12 GB NVIDIA Titan V GPU with compute capability of 7.0 was used. The minimum requirements to run TensorFlow on a GPU require to have a compute capability of at least 3.5, which both GPUs meet. Both CNN models were run on DELL G5 5587 when tested on the merged dataset because of the smaller dataset size, and both models were run on the GPU server when tested on the FER-2013 dataset because of the bigger dataset and consequently longer training times.

3.1.1 Datasets

Three publicly available datasets were used in this research, which are described in detail below:

- JAFFE (Japanese Female Facial Expression) database by Lyons et al. (1998) is composed of 213 monochromatic images with size 256×256 pixels of 10 Japanese female participants in TIFF file format. Each participant expressed each emotion from three up to four times providing varied expressions for every emotion. 6 basic emotions are provided: happiness, fear, anger, disgust, surprise, and sadness; additionally, a neutral expression is also provided (see Figure 2.4). The images were rated by 92 students.

- CK+ (the extended Cohn-Kanade dataset) by Lucey et al. (2010) is a dataset that contains 593 sequences of facial expressions from 123 participants in 8-bit grayscale images of size 640×480 or 640×490 pixels in PNG file format; 69% of participants were female, 81% Euro-American, 13% Afro-American, and 6% belonged to unspecified groups. The sequences have an initial neutral frame and each sequence ends with a peak expression frame; from the total of sequences, only 327 have expressions of emotions. Eight emotions are expressed: happiness, fear, anger, disgust, surprise, sadness, and contempt (see Figure 2.5). Contempt is excluded from this research because of the small amount of data and the incompatibility with the generation of emotions in the imitation system (see Figure 2.8). All frames of peak expressions were also coded in FACS (Facial Action Coding System) by one FACS certified expert; additionally, 15% of the images were compared between two certified FACS experts to check the validity of the coding of expressions.

- FER-2013 (Facial Expression Recognition 2013) dataset by Goodfellow et al. (2013) contains 35887 grayscale images with size 48×48 pixels obtained by using the google image search API; emotional keywords were used to find images of faces expressing emotions that were verified by human labelers; subsequently, the images were cropped and classified into six basic emotions and a neutral expression (see Figure 2.1). The images are codified into a CSV file format arranged in three columns: a label from 0 to 6 for the expressed emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral respectively); the pixels of the images in space-separated values in row-major order; and the usage, divided into training, public test, and private test with 28709, 3589, and 3589 images respectively. The images were organized in this manner for a machine learning contest on facial expression recognition (Goodfellow et al., 2013), and for this research the organization is kept the same using training images which correspond to 80% of the entire dataset for training the neural networks, using public test images as a validation set which corresponds to 10% of the dataset, and using private test images as the testing set which corresponds to 10% of the dataset as well.

Since JAFFE and CK+ datasets have few images, they are merged into one dataset with a total of 639 images. All 213 JAFFE images are used, and all CK+ peak emotional frames and neutral (initial) frames are used totaling 308 emotional expressions images and 118 neutral expression images; the neutral images are only 118 to avoid having a neutral expression repeated by the same participant several times. To achieve an optimal merge of both datasets and avoid images of different sizes, CK+ images where cropped to fit the face of the participants in a 256×256 pixel image which corresponds to the size of JAFFE dataset images. The faces where detected using a Haar-Cascade filter and the images from both datasets merged with an algorithm created by Duncan, Shine, and English (2016). The merged dataset is then split into training set with 461 images, validation set with 82 images, and testing set with 96 images.

For FER-2013 dataset, the CSV file which contained the information of all 35887 images was converted into individual JPG files for each instance of facial expression. The images were assigned a number from 0 to 6 corresponding to each emotion continued by the position of the image in the list (e.g. 0-28709.jpg) and stored in 3 folders for training, validation, and testing. For the conversion, an algorithm provided by Iftekharanam (2017) was used. The images were further classified into the folders angry, disgust, fear, happy, sad, surprise, and neutral in this order.

3.1.2 Kim et al. convolutional neural network

The most successful model from Kim et al. (2016) tested on FER-2013 dataset was adapted and used as the base model. The detailed architecture is shown in Table 3.1.

Table 3.1. Adapted CNNM from Kim et al. (2016)

Layer	Shape[a]	Kernel[b]
Input	48×48, 1	-
Convolution 1	48×48, 32	5×5 (1)
Max-pooling 1	23×23, 32	3×3 (2)
Convolution 2	23×23, 32	4×4 (1)
Max-pooling 2	11×11, 32	3×3 (2)
Convolution 3	11×11, 64	5×5 (1)
Max-pooling 3	5×5, 64	3×3 (2)
	Neurons[c]	
Fully connected hidden	512, or 3072	
Fully connected output	7	

The model uses ReLU activation functions after each convolutional layer. After each pooling layer and the fully connected hidden layer dropout regularization is used, which discards a percentage of units from the previous layer reducing overfitting and improving accuracy (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The fully connected output layer has a softmax activation function and 7 neurons to classify all 6 emotions and the neutral expression.

[a] Dimensions of each feature map and the number of feature maps.

[b] Dimensions of kernels and in parenthesis, stride, which is the number of positions the kernel moves at a time on the feature map.

[c] Number of neurons in the fully connected layers.

3.1.3 VGG convolutional neural network

The most successful deep model from Pramerdorfer and Kampel (2016) was used, which is a modified version of VGG-B model by Simonyan and Zisserman (2014). The detailed architecture is shown in Table 3.2.

Table 3.2. Adapted VGG-B model from Pramerdorfer and Kampel (2016)

Layer	Shape[a]	Kernel[b]
Input	48×48, 1	-
Convolution 1	48×48, 64	3×3 (1)
Convolution 2	48×48, 64	3×3 (1)
Max-pooling 1	24×24, 64	2×2 (2)
Convolution 3	24×24, 128	3×3 (1)
Convolution 4	24×24, 128	3×3 (1)
Max-pooling 2	12×12, 128	2×2 (2)
Convolution 5	12×12, 256	3×3 (1)
Convolution 6	12×12, 256	3×3 (1)
Max-pooling 3	6×6, 256	2×2 (2)
Convolution 7	6×6, 512	3×3 (1)
Convolution 8	6×6, 512	3×3 (1)
Max-pooling 4	3×3, 512	2×2 (2)
	Neurons[c]	
Fully connected hidden	1024	
Fully connected output	7	

[a] Dimensions of each feature map and the number of feature maps.

[b] Dimensions of kernels and in parenthesis, stride, which is the number of positions the kernel moves at a time on the feature map.

[c] Number of neurons in the fully connected layers.

The model uses ReLU activation functions after each convolutional layer, and batch normalization after each convolutional layer and the fully connected hidden layer. Batch normalization was implemented in the model by Pramerdorfer and Kampel (2016) to reduce the effect of suboptimal value initialization of weights in the network. Dropout regularization was also performed after each pooling layer and the fully connected hidden layer. The output layer has a softmax activation function to classify the output into one of the 6 emotions or the neutral expression.

3.1.4 Analysis

From the two previously mentioned models, the solution with higher accuracy in facial expression recognition was used as the base for the emotion imitation system. Both models were analyzed and evaluated on both FER-2013 and the merger of JAFFE and CK+ datasets separately. The models input varied between FER-2013 with 48×48 pixels images and the merged dataset with 256×256 pixels images.

Image preprocessing was minimal to keep data similar to what humans would perceive if they would have to evaluate the images. Only rescaling of images was performed from the original range of 0-255 to 0-1 pixel values to avoid too high values of the weights, which could difficult training (Brownlee, 2019; Omid, 2015). Data augmentation was not performed, to evaluate the models in a more real-world type of scenario, similar to how humans would not be aided by augmented data.

Also, the data was one-hot encoded so that categorical cross-entropy loss function can be applied at the moment of model compilation. Categorical cross-entropy outputs the probability of each predicted image belonging to a specific ground truth emotion class and is commonly used in classification problems with several categories (Chollet, 2015). Both models use Adam optimizer, a stochastic gradient descent-based optimizer also commonly used in CNNs which adapts the learning rate over time (Kingma & Ba, 2015).

Initial testing for finding the best parameters was performed on the merged dataset with Kim et al. (2016) model. Each run was analyzed with version 0.20.0 of scikit-learn machine learning library which shows probability values for precision, recall, and f1-score for each emotion, and computes micro average, macro average, and weighted average. The different model configurations were run 5 times each and the accuracies averaged for each configuration. After testing the accuracy of the base model, regularization methods were implemented in the convolutional layers and the fully connected hidden layer to reduce overfitting; a grid search of possible values was performed which are shown in Table 3.3.

Table 3.3. Regularization parameters for Kim et al. (2016) model

Regularization method	Values
Dropout[a]	0.25, 0.50, 0.80
L1[b], L2[c], and combined L1 and L2[d]	10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}

[a] Dropout regularization in the fully connected hidden layer is tested only with the values 0.50 and 0.80.
[b] L1 regularization also known as LASSO regularization.
[c] L2 regularization also known as Tikhonov regularization.
[d] L1 and L2 regularization applied with the same value for both regularizations.

After testing the regularization methods, the learning rate was analyzed. Initially, the learning rate was not modified and was left with the default value of 10^{-3} from Adam optimizer in Keras. After the tests with regularization methods, the learning rate schedule was modified, implementing a new method of varying learning rates called cyclical learning rates (CLR) created by Smith (2017), which reduces the amount of iterations needed for a model to converge and can slightly improve accuracy by 1% depending on the optimization method CLR is used with. The author also created a method for finding optimal learning rates (Figure 3.1) which linearly increases the value of the learning rate in a specified range for a partial amount of training epochs. An implementation of CLR and learning rate finder by Rosebrock (2019) was adapted to work with the emotion recognition system.

CLR has an oscillating learning rate which starts at a minimum learning rate and increases linearly until it reaches a maximum value, which is called a step; after the learning rate reaches its maximum, it decreases until reaching the minimum, completing a cycle. There can be many cycles when training the network defined by the step size, the batch size, and the number of epochs; Smith (2017) recommends to go through at least 4 cycles to see improvements in training. CLR has 3 policies of varying learning rate schedules: triangular policy (Figure 3.2), which keeps the same minimum and maximum learning rates throughout the training; triangular2 policy, which keeps the minimum learning rate and halves the maximum learning rate each cycle; and exp_range policy, which keeps the minimum learning rate and exponentially decreases the maximum learning rate each cycle.

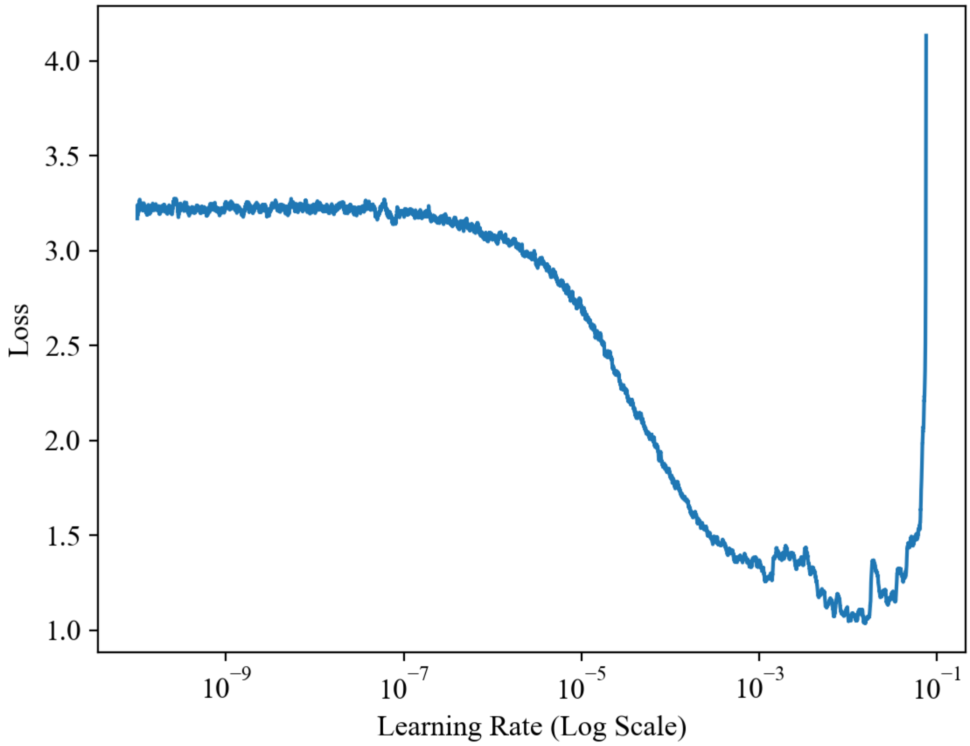


Figure 3.1. Example of learning rate estimation by the learning rate finder.

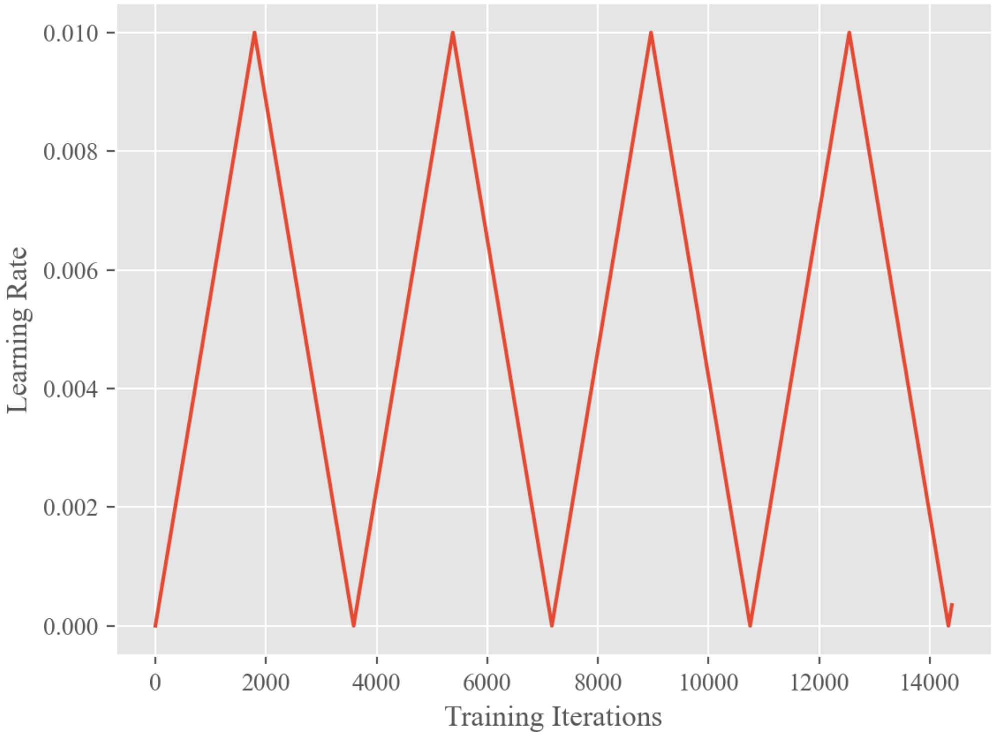


Figure 3.2. Example of CLR triangular policy learning rate oscillation.

In Figure 3.1 we can see the change in loss across values of the learning rate generated by the learning rate finder that starts at 10^{-10} and ends at 10^{-1}. The optimal learning rate boundaries are between 10^{-5} where the slope starts to be steeper, and 10^{-2} before the loss starts to increase.

The final training parameters for both Kim et al. (2016), and Pramerdorfer and Kampel (2016) models are shown in Table 3.4. These parameters were used when evaluating the accuracy of the models after the initial tests with regularization methods and learning rate estimations were performed.

Table 3.4. Training parameters for each model and dataset.

	Adapted CNNM[a]		Adapted VGG-B[b]	
Training parameters	Merged[c]	FER-2013	Merged	FER-2013
Epochs	256	128	128	64
Batch size	64	256	64	128
CLR policy	Triangular	Triangular	Triangular	Triangular
Step size[d]	8	8	8	8
Minimum LR[e]	10^{-5}	10^{-4}	10^{-5}	10^{-5}
Maximum LR[f]	10^{-3}	10^{-3}	10^{-2}	10^{-2}
Image size	256×256	48×48	128×128	48×48

[a] Model based on Kim et al. (2016).

[b] Model based on Pramerdorfer and Kampel (2016).

[c] JAFFE and CK+ merged dataset.

[d] The number of epochs the learning rate takes to reach its maximum value from its minimum value and vice versa.

[e] Minimum learning rate as estimated with the learning rate finder.

[f] Maximum learning rate as estimated with the learning rate finder.

3.2 Results

Kim et al. (2016) base model mean accuracies of all dropout configurations are shown in Table 3.5.

Table 3.5. Mean training and testing accuracy of CNN*M* model with dropout parameters on merged dataset

Configuration[a]	Training accuracy (*M*, *SD*)	Testing accuracy (*M*, *SD*)
No dropout	0.96, 0.01	0.76, 0.03
Conv: 0.25, FC: 0.50	0.95, 0.00	0.74, 0.02
Conv: 0.50, FC: 0.50	0. 92, 0.02	0. 69, 0.04
Conv: 0.80, FC: 0.50	0.26, 0.06	0.22, 0.04
Conv: 0.25, FC: 0.80	0.95, 0.00	0.76, 0.03
Conv: 0.50, FC: 0.80	0.77, 0.09	0.58, 0.09
Conv: 0.80, FC: 0.80	0.26, 0.04	0.23, 0.02

[a] Conv and FC: convolutional layers and fully connected hidden layer respectively; the values stand for the percentage of units that are randomly dropped from the network temporarily.

The dropout values from the fully connected hidden layer were specified in the research by Kim et al. (2016) as 0.50 and 0.80, therefore those were tested; additionally, Pramerdorfer and Kampel (2016) did a grid search to find optimal dropout values after each pooling layer; however, they did not specify the values they used; therefore, the values 0.25, 0.50, and 0.80 were tested based on the findings made by Srivastava et al. (2014) about dropout regularization. Figure 3.3 shows the mean testing accuracy with the different dropout parameters.

To test whether the mean testing accuracy between dropout configurations had a statistically significant difference, a one-way ANOVA test was performed. The results show there was a statistically significant difference between means with $F(6,28) = 132.04$, $p < .001$. Furthermore, post-hoc analysis with multiple comparisons with the best (MCB; Hsu, 1992) was performed, which indicates that dropout parameters Conv: 0.25, FC: 0.80 (see Table 3.5 for reference) with the highest mean accuracy (*M* = 0.76 *SD* = 0.03) were not significantly different from no dropout (*M* = 0.76, *SD* = 0.03); Conv: 0.25, FC: 0.50 (*M* = 0.74, *SD* = 0.02); and Conv: 0.50, FC: 0.50 (*M* = 0. 69, *SD* = 0.04). However, they were significantly different from Conv: 0.80, FC: 0.50 (*M* = 0.22, *SD* = 0.04); Conv: 0.50, FC: 0.80 (*M* = 0.58, *SD* = 0.09); and Conv: 0.80, FC: 0.80 (*M* = 0.23, *SD* = 0.02).

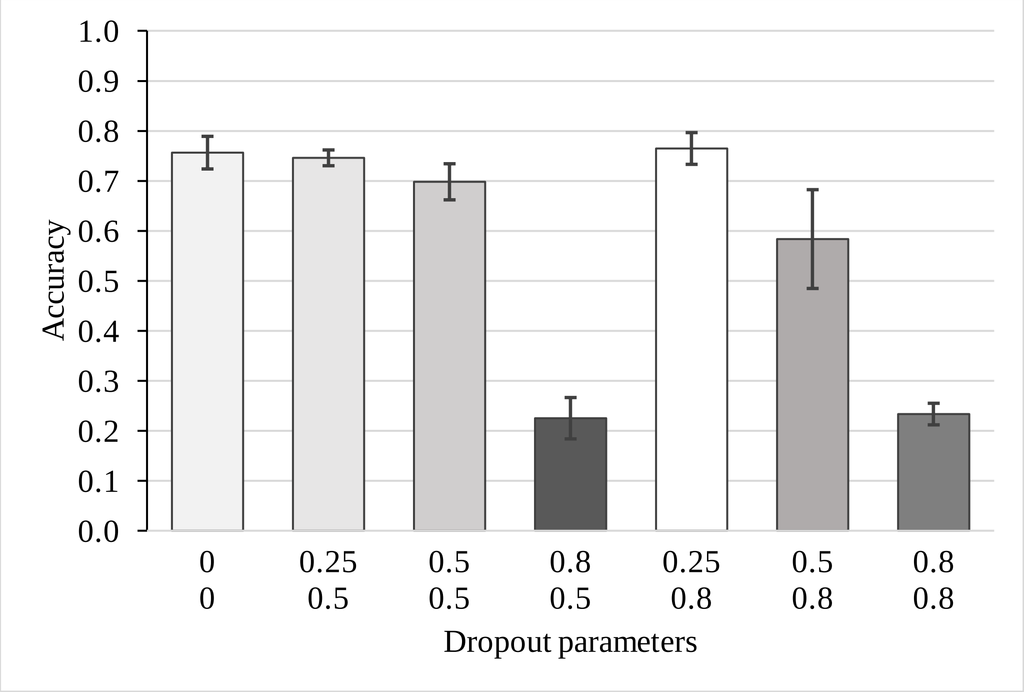


Figure 3.3. Mean accuracy of testing data with the use of dropout with standard deviations; first row of dropout parameters is from the convolutional layers and the second row is from the fully connected hidden layer.

Two other commonly used regularization methods, L1 and L2, were tested and combined with dropout. As suggested by Srivastava et al. (2014), combined with dropout they can improve the accuracy of the model even more. The base values of dropout were 0.25 in the convolutional layer, and 0.5 in the fully connected hidden layer for all tests with L1, L2, and combination of both regularizations. The values used for L1, L2, and L1 and L2 combination are shown in Table 3.3. The initial training and testing accuracies for L1 are shown in Table 3.6, for L2 in Table 3.7, and for L1 and L2 combination in Table 3.8.

Table 3.6. Initial parameters of L1 regularization and training and testing accuracies.

Parameters	Training accuracy	Testing accuracy
10^{-1}	0.24	0.22
10^{-2}	0.24	0.22
10^{-3}	0.35	0.38
10^{-4}	0.94	0.73
10^{-5}	0.96	0.73
10^{-6}	0.96	0.77
10^{-7}	0.96	0.78

L1 parameter values 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7} have the highest accuracy as seen in Table 3.6. These values were further analyzed performing 5 additional runs, and then means of the runs used for further analysis with one-way ANOVA to test if there is a statistically significant difference between means among these values. For training accuracy means, the results show there was not a statistically significant difference between means with $F(4,20) = 0.64$, $p > .05$. For testing accuracy means, there was as well not a statistically significant difference between means with $F(4,20) = 0.42$, $p > .05$. Figure 3.4 shows mean training and testing accuracies.

For L2 regularization, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6} values were further analyzed (Figure 3.5) because of their high accuracy (Table 3.7). A one-way ANOVA test was performed which showed that for training accuracy, there was not a statistically significant difference between means with $F(4,45) = 0.55$, $p > .05$. And for testing accuracy, there was not a statistically significant difference between means either, with $F(4,45) = 0.99$, $p > .05$.

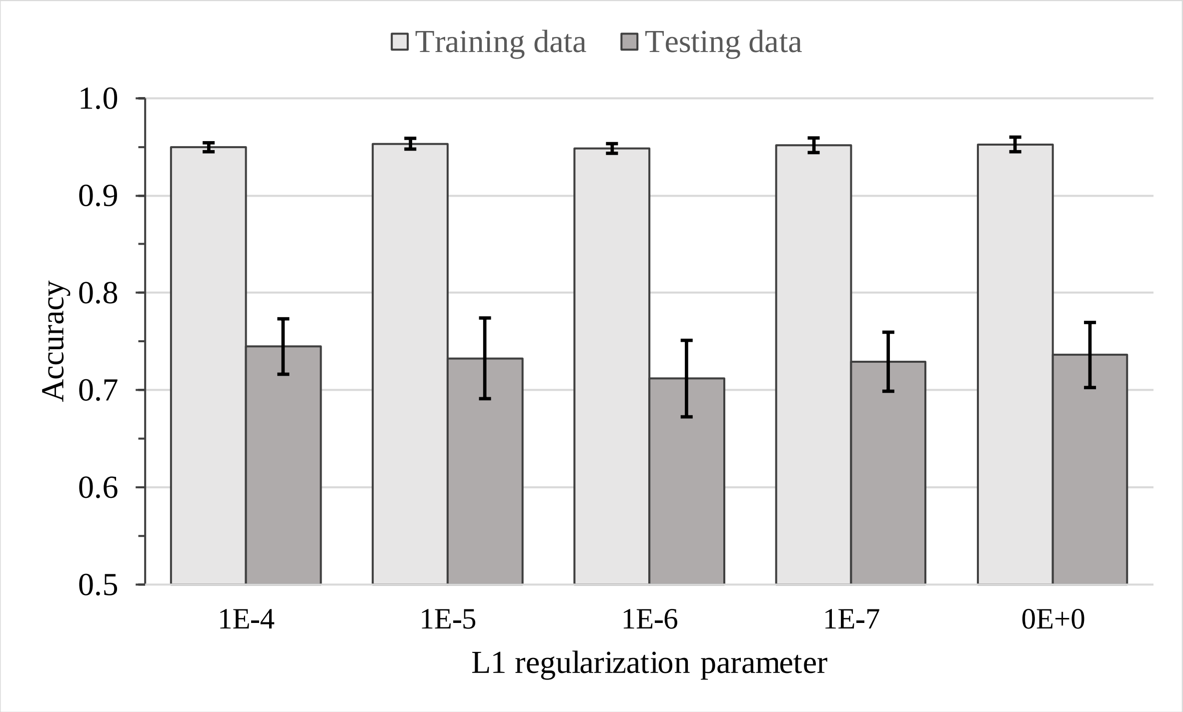


Figure 3.4. Mean accuracy of training versus testing with the use of dropout and L1 regularization for highest accuracy regularization parameters and the baseline with only dropout.

Table 3.7. Initial parameters of L2 regularization, and training and testing accuracies.

Parameters	Training accuracy	Testing accuracy
10^{-1}	0.24	0.23
10^{-2}	0.93	0.65
10^{-3}	0.95	0.72
10^{-4}	0.96	0.80
10^{-5}	0.95	0.69
10^{-6}	0.95	0.69

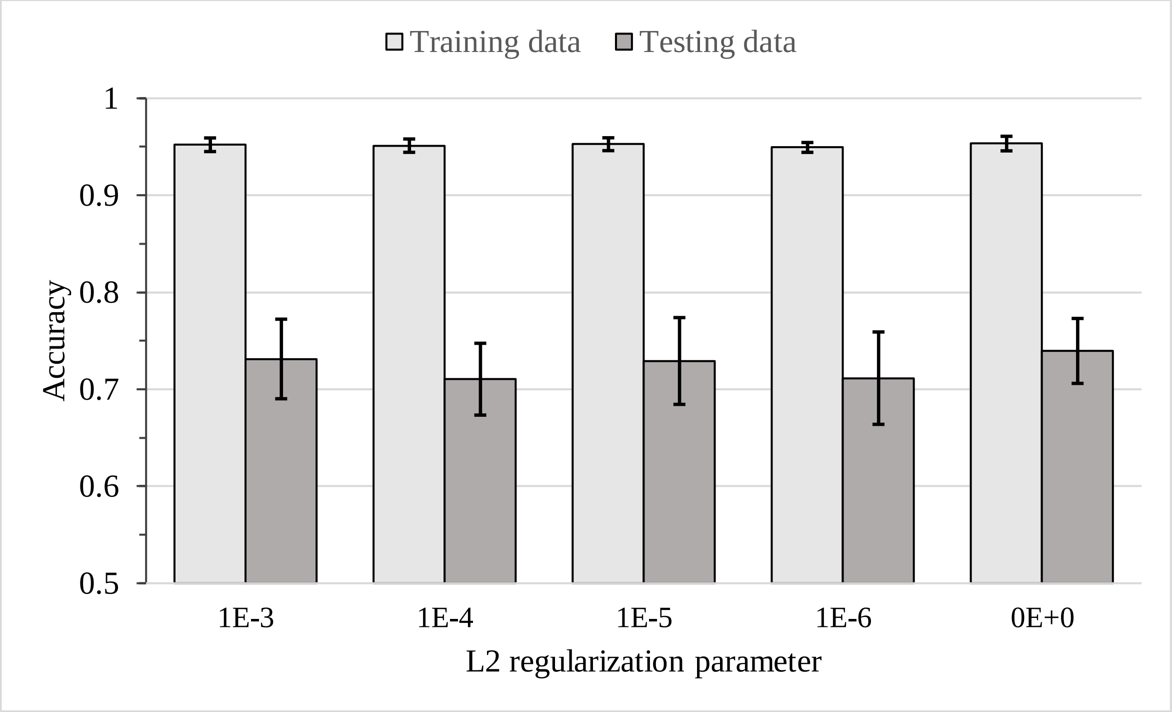


Figure 3.5. Mean accuracy of training versus testing with the use of dropout and L2 regularization for highest accuracy regularization parameters and the baseline with only dropout.

Table 3.8. Initial parameters for L1 and L2 combination and training and testing accuracies

Parameters	Training accuracy	Testing accuracy
10^{-1}	0.24	0.22
10^{-2}	0.24	0.22
10^{-3}	0.39	0.39
10^{-4}	0.95	0.76
10^{-5}	0.95	0.77
10^{-6}	0.96	0.79
10^{-7}	0.96	0.77

For L1 and L2 combination, 10^{-4}, 10^{-5}, 10^{-6}, and 10^{-7} values were analyzed (Figure 3.6) because of their high accuracy (Table 3.8). A one-way ANOVA was performed with results showing statistically significant difference between means for the training accuracy with $F(4,45) = 4.71$, $p < .05$; however, for the testing accuracy, the means were not statistically significantly different with $F(4,45) = 0.45$, $p > .05$. Since the means of the testing accuracy were not significantly different, further analysis was not performed.

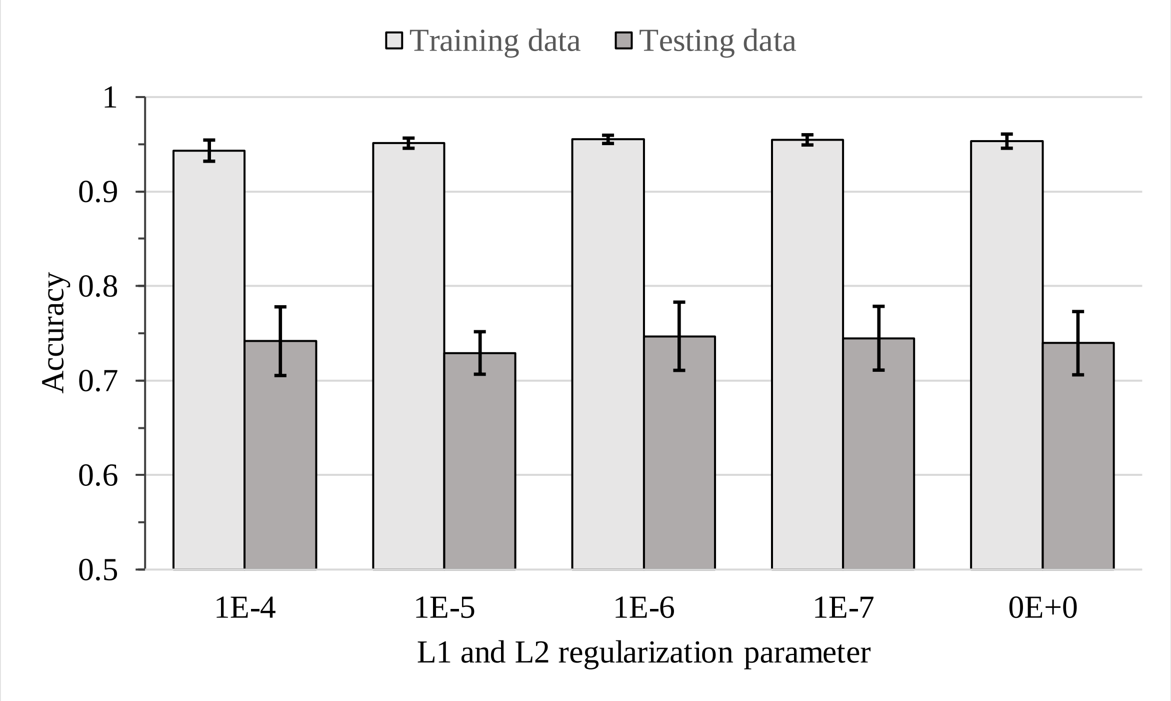


Figure 3.6. Mean accuracy of training versus testing data with the use of dropout and L1 and L2 regularization combination for highest accuracy regularization parameters and the baseline with only dropout.

After the experiments with regularization, dropout was kept in Kim et al. (2016) model with the values 0.25 for convolutional layers and 0.80 for the fully connected hidden layer. In Pramerdorfer and Kampel (2016) model, the value 0.25 for dropout was used after each pooling layer and the fully connected hidden layer was set to a dropout value of 0.50. No other regularization methods were implemented in the models.

Cyclical learning rates (CLR) were used to optimize the learning rate of both models; learning rate estimation was performed with the learning rate finder. The calculated values are shown in Table 3.4 for both models on both datasets; 4 graphs were generated and analyzed similar to the one in Figure 3.1.

The results of Kim et al. (2016) model on the merged dataset applying CLR are shown in Figure 3.7 and Figure 3.8. We can see losses and accuracies of training and validation sets in Figure 3.7 of one run of the model, and we can see mean accuracy and standard deviation of training and validation sets of all 5 runs in Figure 3.8.

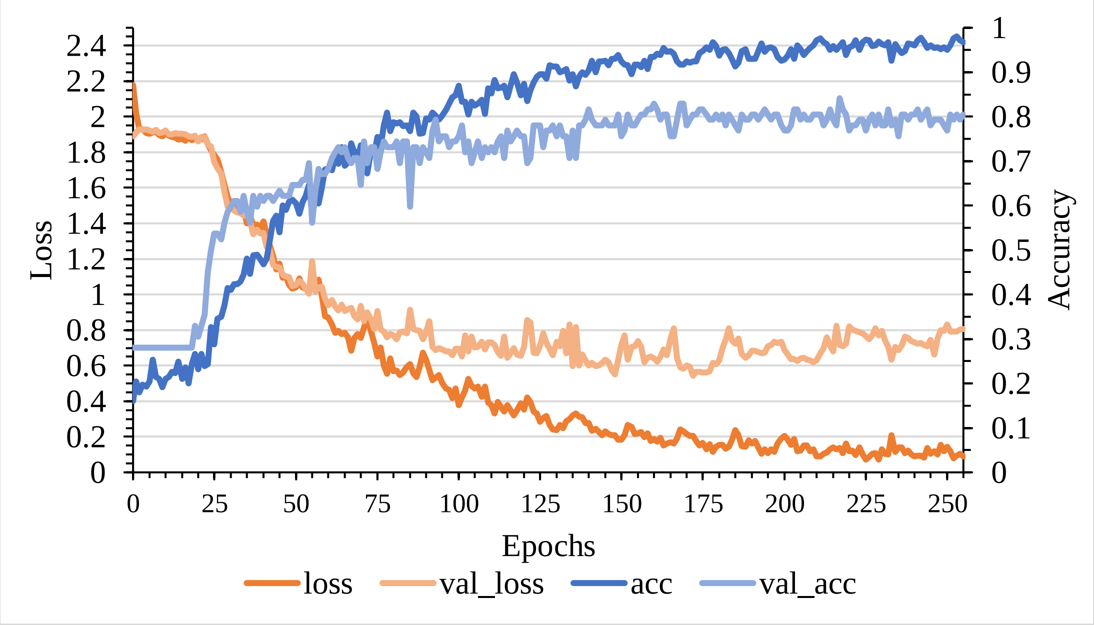


Figure 3.7. Loss and accuracy on training/validation sets with Kim et al. (2016) model tested on the merged dataset.

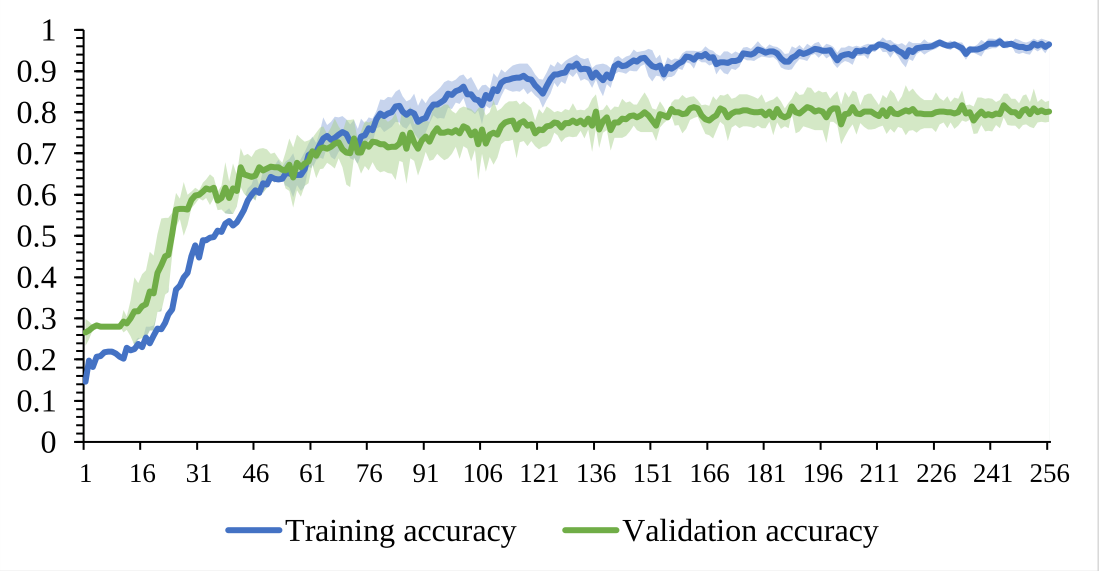
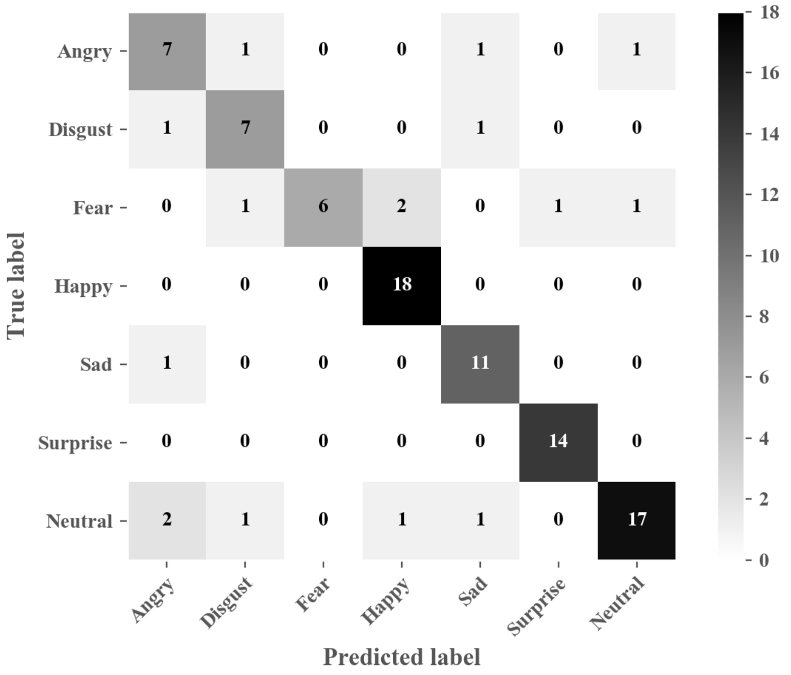


Figure 3.8. Mean and standard deviation of training and validation accuracies across all training epochs on merged dataset tested with Kim et al. (2016) model.

In Figure 3.9, we can see the confusion matrix of one run with Kim et al. (2016) model on the merged dataset, which shows all the 96 testing images classified into 7 facial expressions. The total of each row (true label) shows the ground truth of each emotion, which is the total of labeled images per class. On the other hand, the total of each column (predicted label) shows the number of images classified by the model into each class. In Figure 3.10, we can see the normalized confusion matrix of one run with Kim et al. (2016) model on the merged dataset, which shows the percentage of correctly classified and misclassified images for each class.



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	7	1	0	0	1	0	1
Disgust	1	7	0	0	1	0	0
Fear	0	1	6	2	0	1	1
Happy	0	0	0	18	0	0	0
Sad	1	0	0	0	11	0	0
Surprise	0	0	0	0	0	14	0
Neutral	2	1	0	1	1	0	17

Figure 3.9. Confusion matrix of the merged dataset results using Kim et al. (2016) model.

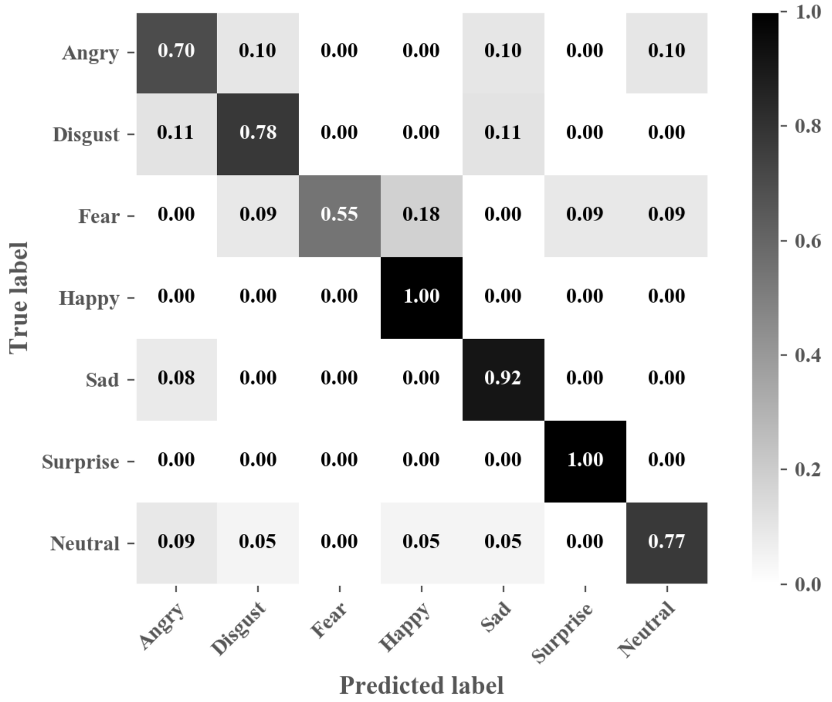


Figure 3.10. Normalized confusion matrix of the merged dataset results using Kim et al. (2016) model.

The results of Pramerdorfer and Kampel (2016) VGG model tested on the merged dataset with CLR are shown in Figure 3.11 and Figure 3.12., which show the losses and accuracies for one run, and mean accuracy and standard deviation of training and validation sets of all 5 runs respectively. The confusion matrix of one run with Pramerdorfer and Kampel (2016) model on the merged dataset is in Figure 3.13, and the normalized confusion matrix of one run with Pramerdorfer and Kampel (2016) model on the merged dataset is in Figure 3.14.

To compare the accuracy between Kim et al. (2016) and Pramerdorfer and Kampel (2016) models on the testing set of the merged dataset, a paired sample *t*-test was performed. The results show there was no significant difference in the testing set accuracies of Kim et al. (2016) model (*M* = 0.83, *SD* = 0.02) and Pramerdorfer and Kampel (2016) model (*M* = 0.82, *SD* = 0.02); *t*(4) = 0.63, *p* > .05.

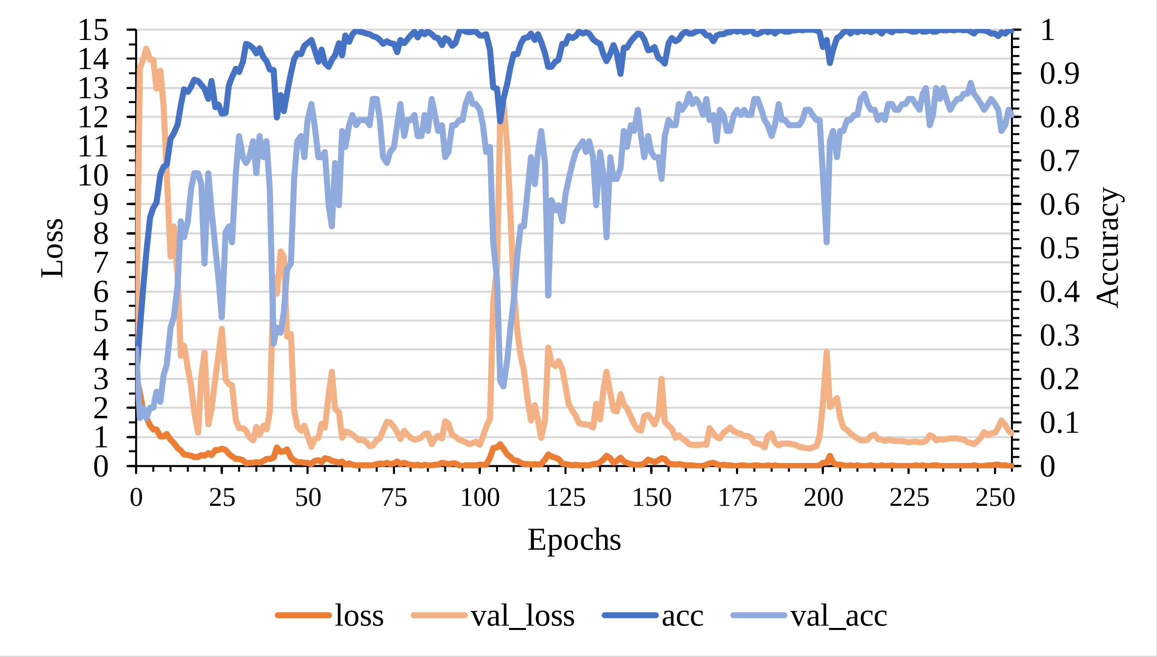


Figure 3.11. Loss and accuracy on training/validation sets with Pramerdorfer and Kampel (2016) VGG model tested on the merged dataset.

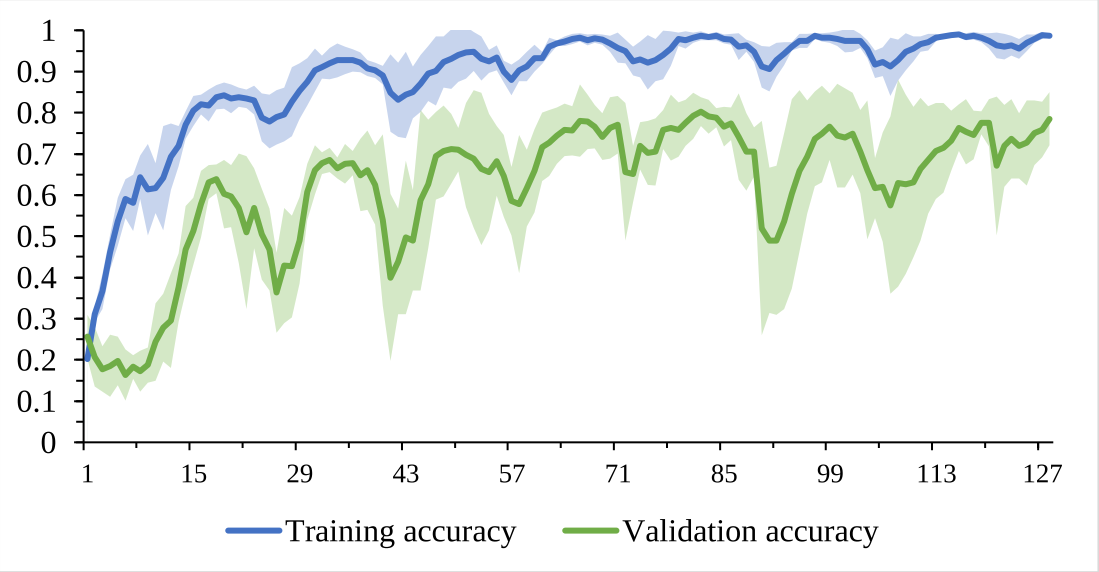


Figure 3.12. Mean and standard deviation of training and validation accuracies across all training epochs on merged dataset tested with Pramerdorfer and Kampel (2016) model.

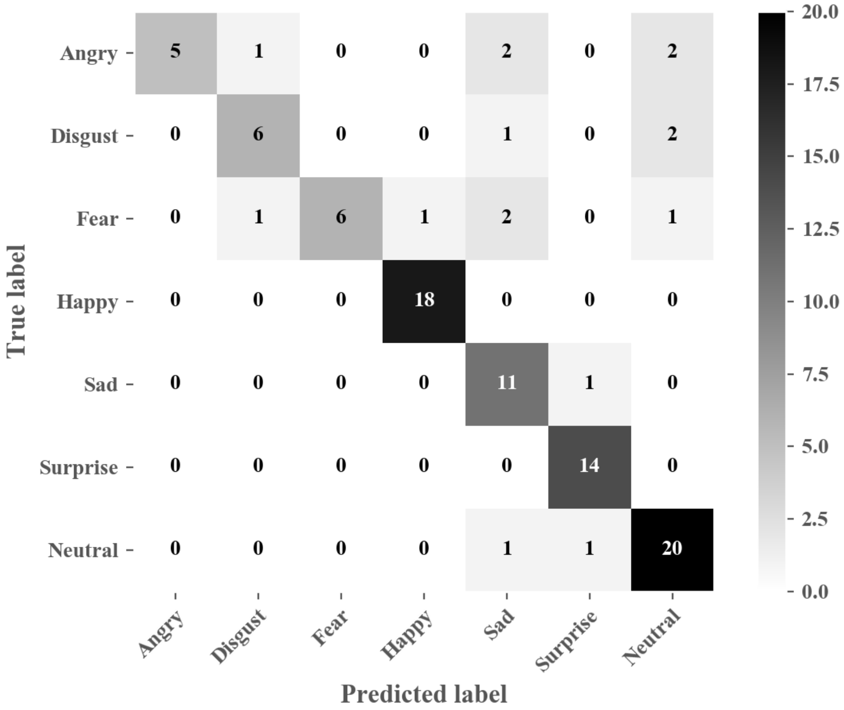


Figure 3.13. Confusion matrix of the merged dataset results using Pramerdorfer and Kampel (2016) VGG model.

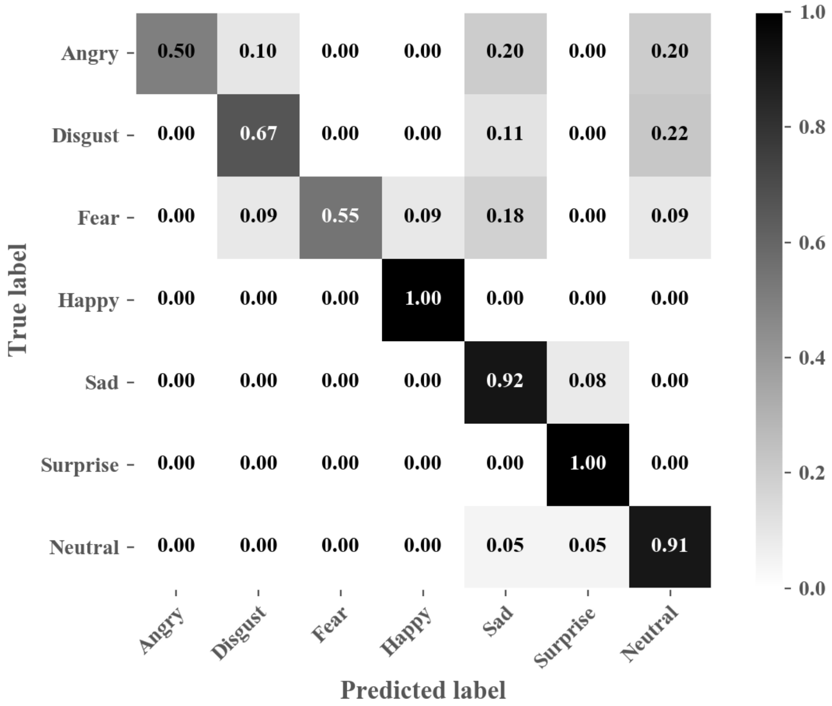


Figure 3.14. Normalized confusion matrix of the merged dataset results using Pramerdorfer and Kampel (2016) VGG model.

On the FER-2013 dataset the initial model tested was Kim et al. (2016) with CLR implemented. The losses and accuracies of one run of the model are in Figure 3.15, and the mean accuracies and standard deviation of the training and validation sets are in Figure 3.16. The confusion matrix in Figure 3.17 shows the classification of the 3589 images from FER-2013 testing set, and the normalized confusion matrix shows the percentage of correct classification and misclassification of images in Figure 3.18.

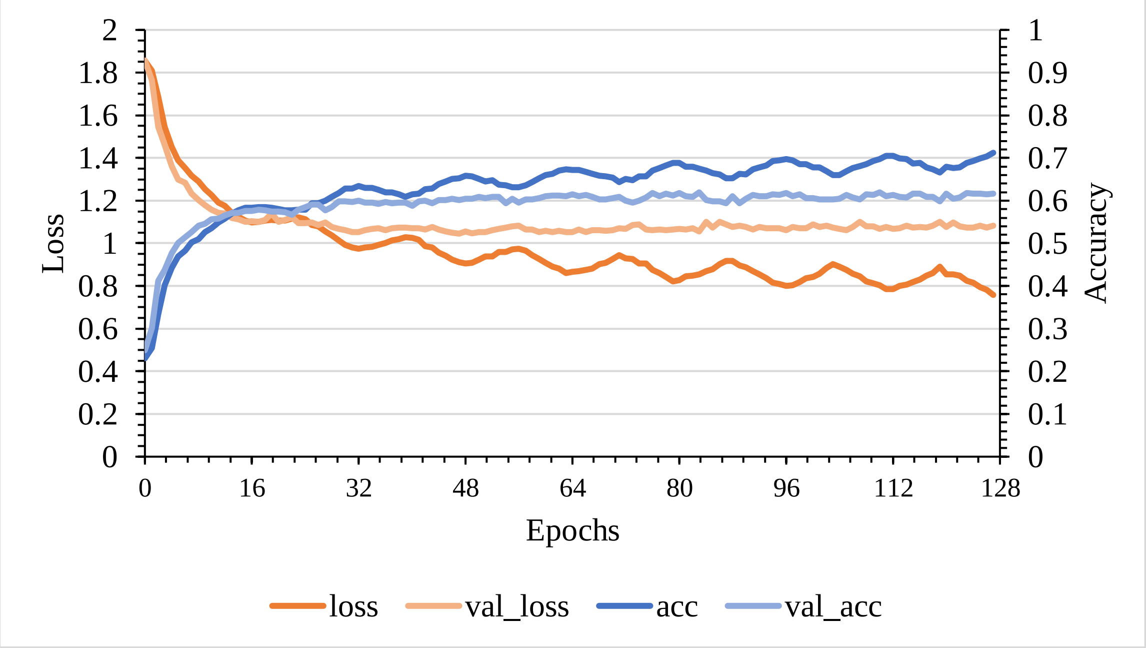


Figure 3.15. Loss and accuracy on training/validation sets with Kim et al. (2016) model tested on FER-2013 dataset.

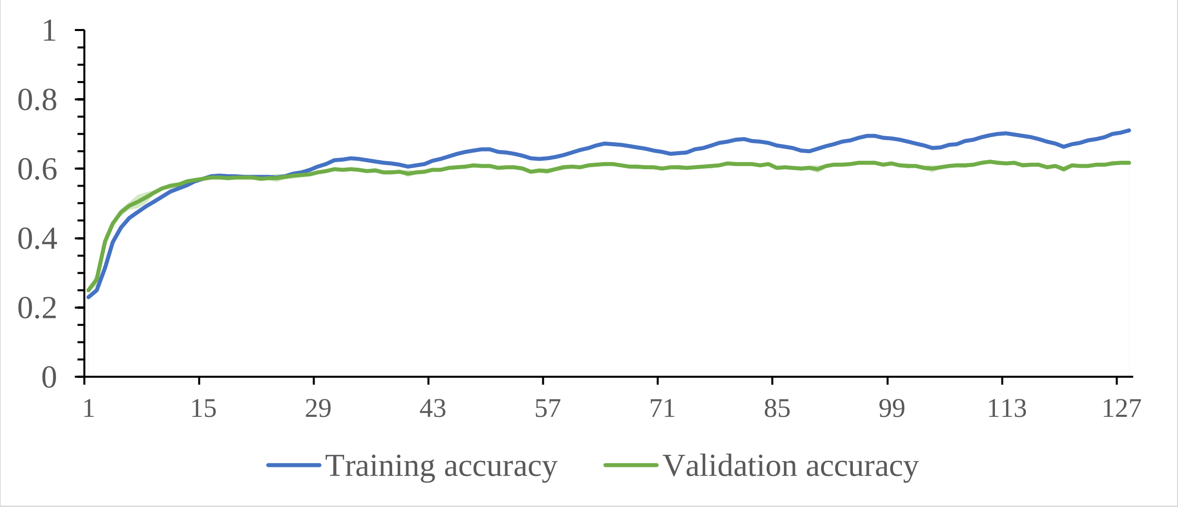


Figure 3.16. Mean and standard deviation of training and validation accuracies across all training epochs on FER-2013 dataset tested with Kim et al. (2016) model.

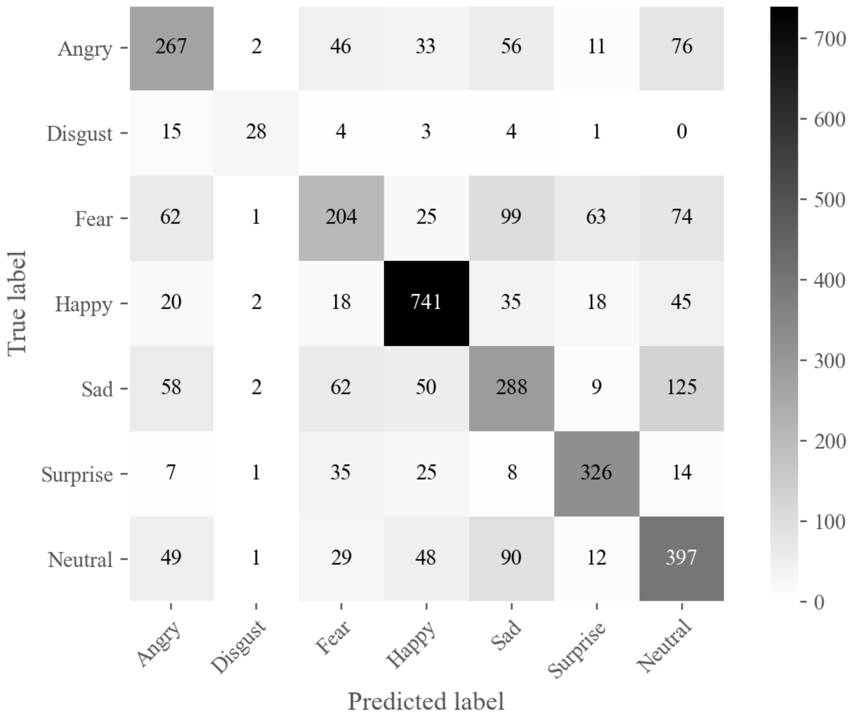


Figure 3.17. Confusion matrix of FER-2013 dataset results using Kim et al. (2016) model.

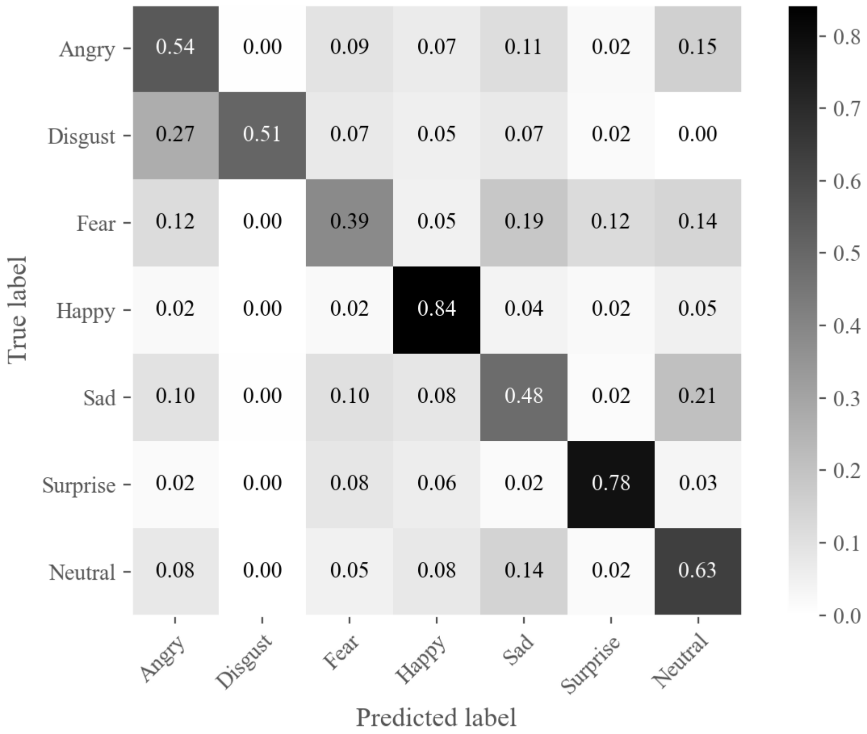


Figure 3.18. Normalized confusion matrix of FER-2013 dataset results using Kim et al. (2016) model.

The results of Pramerdorfer and Kampel (2016) VGG model tested on FER-2013 dataset with CLR are shown in Figure 3.19 for the losses and accuracies of one run of the model, and in Figure 3.20 for the mean accuracies and standard deviation of the training and validation sets. The confusion matrix of one run with Pramerdorfer and Kampel (2016) model on FER-2013 testing set is in Figure 3.21, and the normalized confusion matrix is in Figure 3.22.

To compare the accuracy of both models now on FER-2013 dataset, another paired sample *t*-test was performed. The results show there was no significant difference in the testing set accuracies of Kim et al. (2016) model (*M* = 0.62, *SD* = 0.01) and Pramerdorfer and Kampel (2016) model (*M* = 0.63, *SD* = 0.01); *t*(4) = –2.06, *p* > .05. Therefore, because of no significant differences between models on both datasets, Kim et al. (2016) model was chosen for a more stable architecture and overall lowest loss values (see Figure 3.15 and Figure 3.19 for comparison).

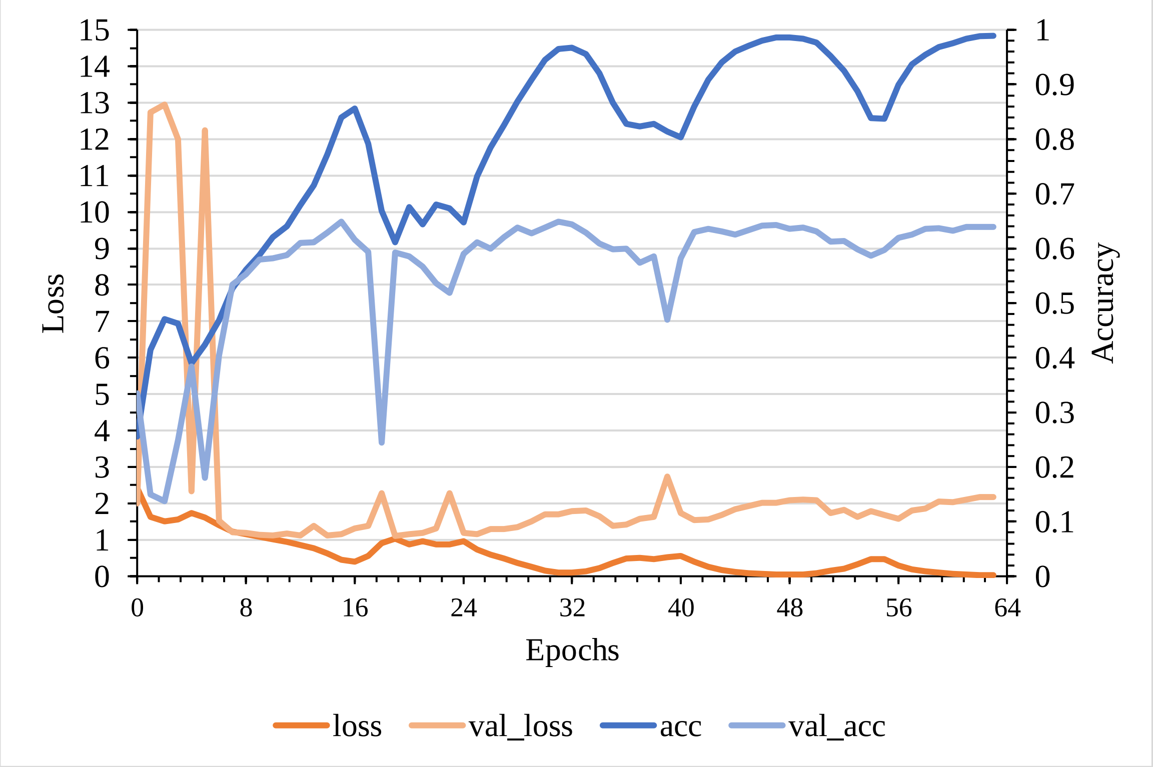


Figure 3.19. Loss and accuracy on training/validation sets with Pramerdorfer and Kampel (2016) VGG model tested on FER-2013 dataset.

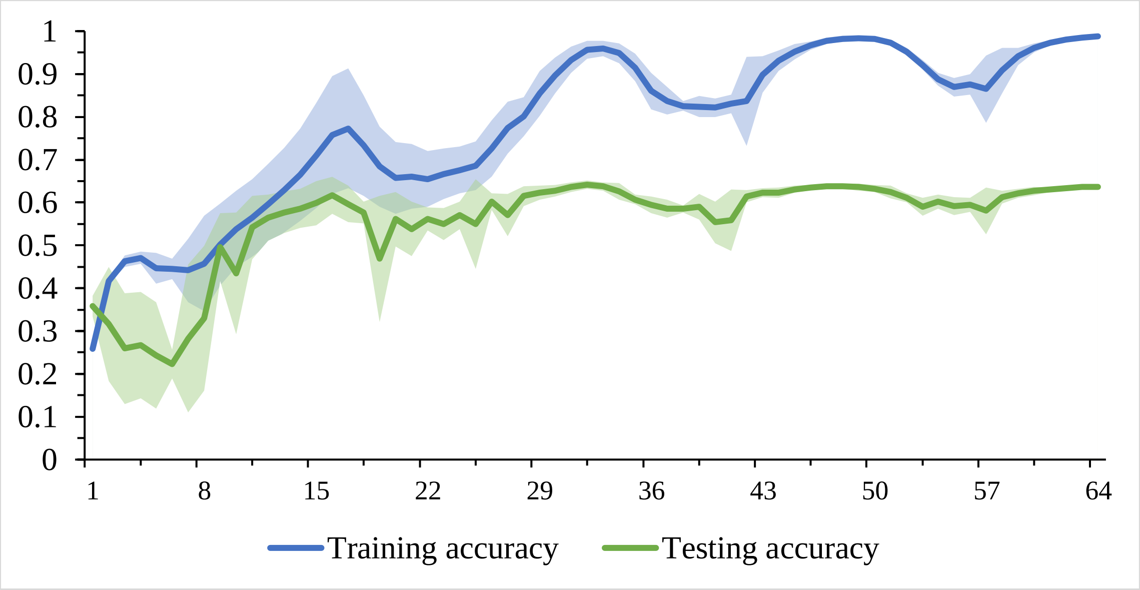
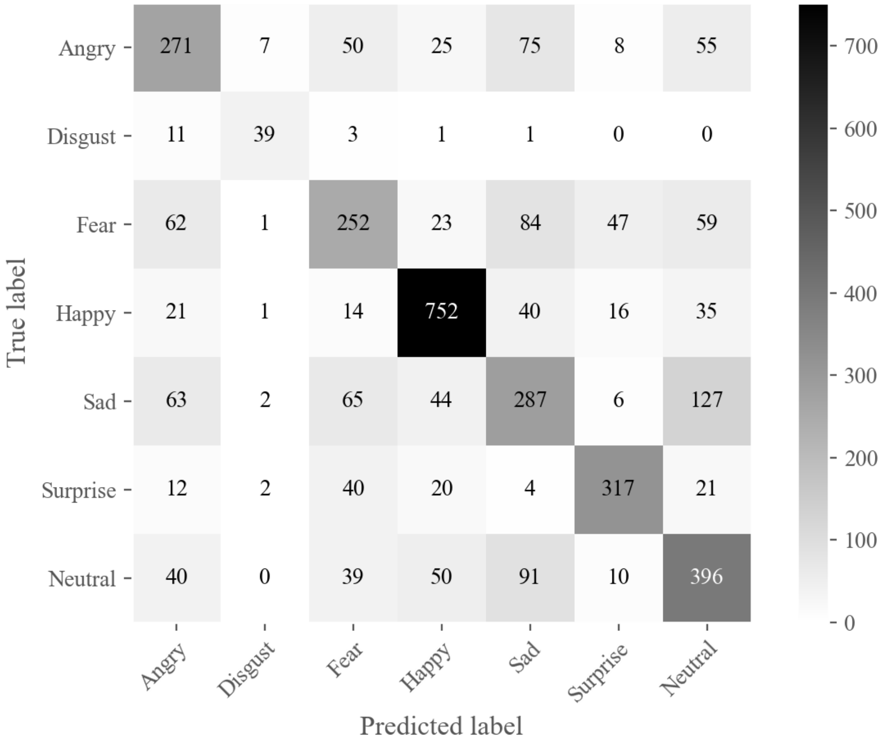


Figure 3.20. Mean and standard deviation of training and validation accuracies across all training epochs on FER-2013 dataset tested with Pramerdorfer and Kampel (2016) VGG model.



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	271	7	50	25	75	8	55
Disgust	11	39	3	1	1	0	0
Fear	62	1	252	23	84	47	59
Happy	21	1	14	752	40	16	35
Sad	63	2	65	44	287	6	127
Surprise	12	2	40	20	4	317	21
Neutral	40	0	39	50	91	10	396

Figure 3.21. Confusion matrix of FER-2013 dataset results using Pramerdorfer and Kampel (2016) VGG model.

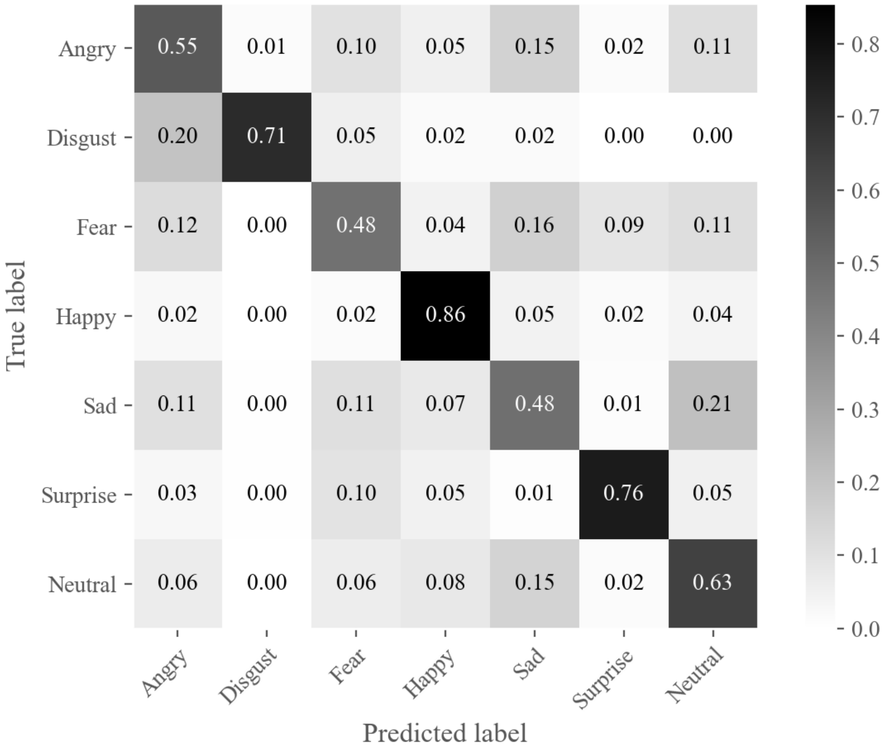


Figure 3.22. Normalized confusion matrix of FER-2013 dataset results using Pramerdorfer and Kampel (2016) VGG model.

In summary, dropout regularization did not improve significantly the accuracy of both models, neither L1 nor L2 regularization. For this reason, these parameters were set to the defaults used in Kim et al. (2016) and Pramerdorfer and Kampel (2016) when mentioned; otherwise, they were set based on the research by Srivastava et al. (2014) on dropout.

Regarding CLR (cyclical learning rates), the method reduced the training time considerably compared to the mentioned in Pramerdorfer and Kampel (2016). In their research they trained CNNs for 300 epochs on FER-2013 dataset; comparably, the adapted model used in this research was trained only for 64 epochs, which was enough to reach a plateau in accuracy. The same model was trained with 128 and 256 epochs; however, accuracy did not improve, and loss started to increase. For Kim et al. (2016) model, training epochs were higher (up to 256) because improvement in accuracy and loss did occur even passed 128 epochs. The amount of data trained with the network also has a role in the amount of epochs necessary to reach a plateau in accuracy; for both models on the merged dataset the training epochs were double of the training epochs in FER-2013 dataset, which is considerably bigger.

Ultimately, Kim et al. (2016) model was chosen to be used with the imitation system because it was more stable with lower losses, and accuracy (83% on merged, 62% on FER-2013 datasets) not significantly different from Pramerdorfer and Kampel (2016) VGG model.

4 Emotion imitation system

The emotion imitation system receives the output of the model from the recognition system of both datasets. When the imitation system receives an input, it associates it internally with an emotion; if it associates the emotion correctly, the system is rewarded; otherwise, the system is penalized.

4.1 Methodology

The imitation system is trained with the testing sets of merged and FER-2013 datasets individually. Kim et al. (2016) model training results are in the form of a matrix which contains the probability values of the emotions for each image, which is loaded into the imitation system to begin the association process.

The internal representation of emotions in the imitation system is stored in the form of a 7×7 weights matrix. To achieve association of perceived emotions from the recognition system with the internal representation of emotions from the imitation system, Oja's (1982) learning rule (1) was implemented. Oja’s rule is a biologically inspired model, which updates the network weights like the Hebbian rule with an additional component that controls the growth of the weights.

Oja’s learning rule which modifies the Hebbian rule to include weight normalization is:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha y_i(t)\left[x_j(t) - y_i(t) w_{ij}(t)\right], \tag{1}$$

where $w_{ij}(t+1)$ is the updated weight, $w_{ij}(t)$ is the current weight, α is the learning rate, and $y_i(t)$ is the current output at argmax position of the dot product between the weights matrix and the input vector x.

To begin the association learning, weights are initialized first. For weight initialization, two variants of He, Zhang, Ren, and Sun (2015) initialization were tested: the first one initializes weights in the form $W = rnd_{ij}\sqrt{1/classes}$, where W is the weights matrix, rnd_{ij} are the elements of the weights matrix randomly generated with values within the standard normal distribution, and *classes* is the number of layer inputs, in this case 7, for each emotion class; the second variant initializes the elements of the weights matrix W with random values within a uniform distribution of values between 0 and 1. He et al. (2015) weight initialization was implemented to avoid large initial weights, which could disrupt the association learning of the imitation system.

After weight initialization, reward learning rate α and penalty learning rate β are set. Then, argmax of the input vector x is selected, which would correspond to the emotion class with the highest probability value calculated by the recognition system. Afterwards, the output vector y is calculated by applying dot product between the weight matrix W and the input vector x as seen in (2).

$$y=W \cdot x. \tag{2}$$

Subsequently, the weights element w_{ij} on the argmax position of the output vector y_i is ready to be updated with a reward or a penalty. If argmax is equal for both the input vector and the output vector, the related weight is fortified with Oja's rule (1); otherwise, the learning rate α is multiplied with the input vector argmax x_j and the output y_i at argmax position of the output vector y_i and subtracted from the current weights element w_{ij}. The algorithm repeats this learning process for a given number of epochs as shown in Figure 4.1.

The emotion imitation system is first tested on the merged dataset, with two variants of He et al. (2015) weight initialization; and 3 different learning rates: 0.05, 0.1, and 0.2; the learning rates are individually set for the rewards and penalties. Each test is run 30 times with 30 epochs for each test. After concluding the tests on the merged dataset, the best association learning configuration is used and evaluated on FER-2013 dataset.

```
begin
set $\alpha$, $\beta$, $Epochs$
set $W = rnd_{ij}\sqrt{1/classes}$
for each $epoch$ in $Epochs$ do
    select $x$ from images data set
    $max\_x$ = argmax($x$)
    compute eq. 2
    $max\_y$ = argmax($y$)
    select $w_{ij}$ at $max\_y$ from $W$
    select $y_i$ from $y$ at $max\_y$
    if $max\_x == max\_y$
        compute eq. 1
        $W[max\_y, max\ x] = w_{ij}$
    else
        $w_{ij} = w_{ij} - \beta\, x_j\, y_i$
        $W[max\_y, max\ x] = w_{ij}$
end
```

Figure 4.1. Association learning algorithm for the imitation system.

4.2 Results

The imitation system learns to imitate emotions on both datasets. On the merged dataset, to compare the accuracy of both variants of He et al. (2015) weight initialization, a paired sample *t*-test was performed. The results show there was a significant difference between the first variant (*M* = 0.47, *SD* = 0.17) and the second variant (*M* = 0.96, *SD* = 0.07); *t*(29) = –15.29, p < .001. Further, choosing the second variant (*M* = 0.96, *SD* = 0.07) for its significantly higher accuracy, tests with learning rates were performed. 4 configurations of learning rates were tested as seen in Table 4.1.

Table 4.1. Learning rate configurations for the imitation system.

Learning rates	Accuracy *(M, SD)*
α:0.05, β:0.05	0.92, 0.08
α:0.1, β:0.1	0.96, 0.66
α:0.2, β:0.2	0.94, 0.07
α:0.2, β:0.1	0.92, 0.11

A one-way ANOVA was performed to test whether there is a statistically significant difference between means of the learning rates. The results show there was no statistically significant difference between the means with $F(3, 116) = 1.65$, $p > .05$. Since the means were not significantly different, the configuration with learning rate 0.1 was chosen because of its higher accuracy ($M = 0.96$, $SD = 0.66$).

With the results gathered on the merged dataset, the same configuration, with the second variant of He et al. (2015) weight initialization, and learning rate 0.1 was used to test how well can the imitation system associate the emotions in FER-2013 dataset. The results show that the imitation system achieves a mean accuracy of 92% ($M = 0.92$, $SD = 0.08$).

When the imitation system was tested, on both datasets they occurred instances when a weight was initialized with a too small value, and the model would fail to associate an emotion even after many epochs. In Figure 4.1 and Figure 4.2 we can see the mean accuracy and the range of values of accuracy in 30 epochs for 30 runs, which clearly shows that even after 13 epochs, they were instances where the imitation system would not be able to associate all emotions. Nevertheless, there were very few of these instances as shown by the mean accuracy (Figure 4.1, Figure 4.2) of the imitation system.

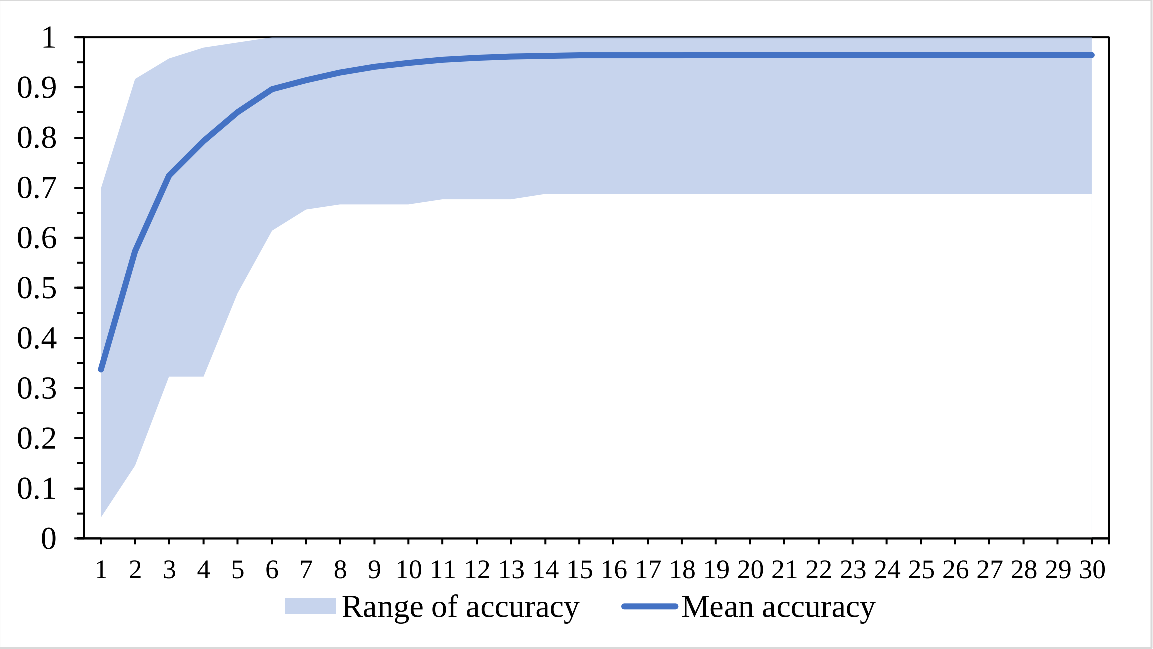


Figure 4.2. Mean accuracy of the association of emotions on the merged dataset with the range between minimum and maximum values per epoch.

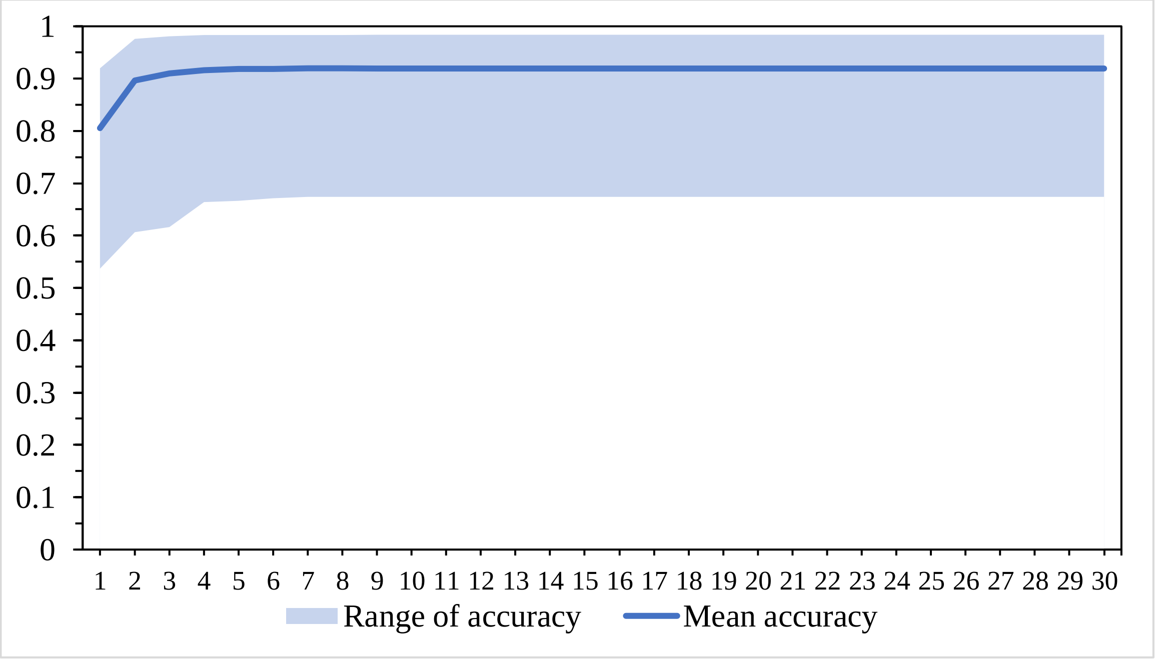


Figure 4.3. Mean accuracy of the association of emotions on FER-2013 dataset with the range between minimum and maximum values per epoch.

4.3 Integrated system

The integrated system composed of the emotion recognition and imitation systems, could be used in a robotic system; in Figure 4.4 a mockup of the interface is shown. The system was intended to be used with NICO robot (Kerzel et al., 2017). NICO is able to produce 7 facial expressions (neutral, happiness, sadness, anger, surprise, fear, and disgust) as seen in Figure 2.8, through a set of three LED arrays corresponding to the mouth and the eyebrows. In Figure 4.5 the scheme of the entire system is shown.

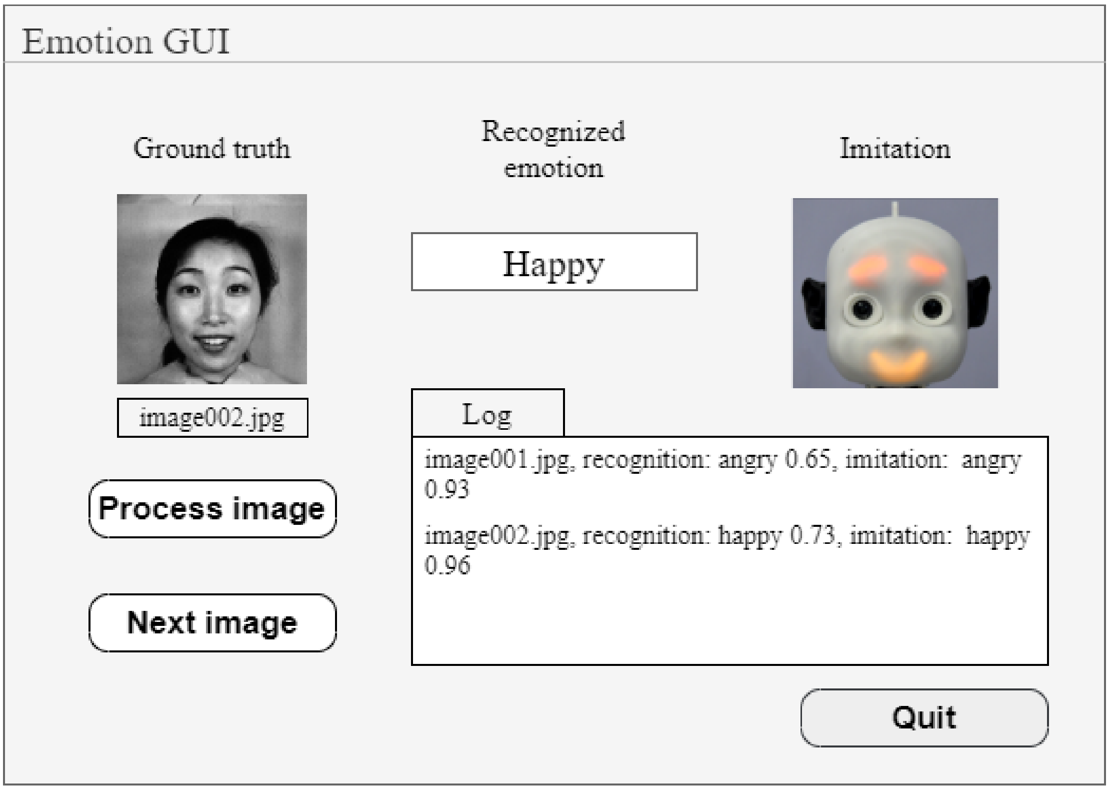


image002.jpg

Happy

```
image001.jpg, recognition: angry 0.65, imitation:  angry 0.93
image002.jpg, recognition: happy 0.73, imitation:  happy 0.96
```

Figure 4.4. Mockup of the interface, which would test the recognition and imitation capabilities of the system. The user would cycle through the images located in a folder with next image, and would receive the results from the integrated system with the process image button. The results would be printed in the log area, the recognized emotion would be printed in the text box in the recognized emotion area and the imitated expression would appear at the imitation area.

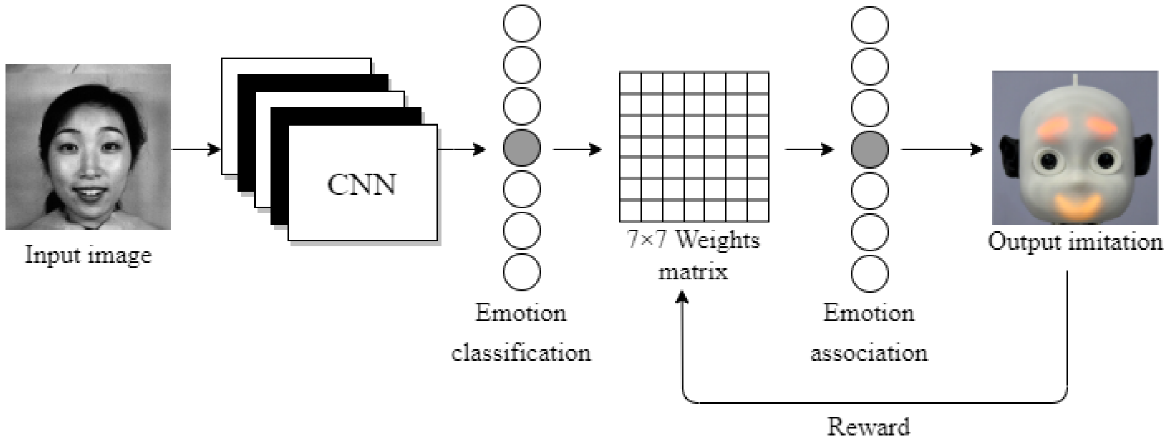


Figure 4.5. Integrated system scheme. The input image gets processed by the CNN (Kim et al., 2016) model, which outputs a probability value for emotion classification; then, the probability value gets processed with the weights matrix to output an emotion association; if the emotion imitated is correct a reward is sent to update the weights; otherwise, a penalty is sent to update the weights.

The overall emotion imitation accuracy of the system was, on average, 80% combining the imitation system average accuracy of 96%, with the recognition system average accuracy of 83%, based on the ground truth of the testing set from the images of the merged dataset. And, on FER-2013, the imitation accuracy of the system was, on average, 57% combining the imitation system average accuracy of 91%, with the recognition system average accuracy of 63%.

5 Discussion

The emotion imitation system associates emotions from the emotion recognition system with high accuracy on the merged dataset (96%) and on FER-2013 dataset (91%); however, imitation accuracy is highly dependent on the capabilities of the CNN model implemented in the recognition system; from the results on emotion recognition of 83% on the merged dataset, and 62% on FER-2013 dataset, the final accuracy of the integrated system is 80% and 57% respectively.

The big difference in accuracy between datasets can be attributed to the complexity of the images in FER-2013 compared to the merged dataset. Comparing Figure 3.7 with Figure 3.15 we can see that FER-2013 dataset is more challenging for Kim et al. (2016) model, as well as for Pramerdorfer and Kampel (2016) model (Figure 3.11 and Figure 3.19). This is understandable even for the large amount of data (28709 training images) in FER-2013, since the dataset has many pose variations, different illumination conditions, and occlusion across the data (Figure 2.1). On the other hand, the merged dataset (JAFFE and CK+) with less data (461 training images) only differs slightly in illumination conditions between images across data, and the faces are centered without any occlusion (Figures 2.4, 2.5).

Comparing the adapted models to the original implementations by Kim et al. (2016) and Pramerdorfer and Kampel (2016), they did not achieve the original models accuracy of 70% and 72% respectively on FER-2013 dataset. There are two possible factors that generated lower accuracy on the recognition of emotions: some parameters were not described in Kim et al. (2016) and Pramerdorfer and Kampel (2016), like learning rate initialization, weight initialization, and dropout regularization, which made more difficult to replicate the results; and data augmentation was not performed in this research, to represent a more natural-like input of data.

Initial tests with regularization methods on the merged dataset with Kim et al. (2016) model had approximately 3 mislabeled images within each group of emotions. Even with these labeling mistakes on a relatively small dataset (639 images), both models were able to achieve accuracies of 77% compared to the final tests with 83% accuracy, which implies that the networks are robust to some amount of mislabeled data. In the case of FER-2013 dataset, Goodfellow et al. (2013) mention that labelling mistakes were present; however, since the amount of data in the dataset is greater, they considered that these labeling mistakes should not be a problem, also because human accuracy did not seemed hindered by the labeling mistakes. On the final tests with regularization methods, we would expect that the regularization methods would improve accuracies as mentioned by Srivastava et al. (2014) from 1% to 2%. However, the differences were not significant, probably because individual experiments were run only 5 times and more tests would be necessary to increase the power of the test. For this reason, we used dropout regularization which yielded the best results.

A significant difference between Pramerdorfer and Kampel (2016) implementation and this research was present in the number of training epochs needed to reach a plateau in training accuracy in the recognition system. The use of cyclical learning rates (CLR) in the models did reduce the training of the model from 300 epochs in Pramerdorfer and Kampel (2016) to 64 epochs (see Figure 3.19). Kim et al. (2016). did not specify the number of epochs trained, as they use validation data as a stopping mechanism; therefore, training epochs cannot be compared with this research. Nevertheless, comparing both Kim et al. (2016) and Pramerdorfer and Kampel (2016) model, Kim et al. (2016) model needed the double of training epochs to reach a plateau compared to Pramerdorfer and Kampel (2016) model (see Table 3.4); this could be because of the architecture, which has less layers and parameters. The main objective for reducing training epochs was to allow shorter training times, and possibly make more evaluations in less time, which was very convenient. According to Smith (2017), it even can improve the accuracy of the model by 1%, aside from reducing the training time considerably. In this research statistical tests were not performed to examine the difference in accuracy between CLR and common methods like learning rate decay, because the initial results with CLR were similar to previous tests with other methods. We ultimately chose CLR for the reduced training times of the models.

On a different note, about the accuracy of the imitation recognition system, when the weights were randomly initialized with a standard normal distribution with He et al. (2015) weight initialization, the imitation system would fail to associate most of the emotions, which occurred because some weights would grow negatively when applying Oja's (1982) rule. Because of this issue, a uniform distribution in a range of 0 to 1 was applied to He et al. (2015) initialization, and this improved the mean accuracy up to 96%; they still were instances when the system would fail to associate an emotion, which occurred approximately 33% of the time. Analyzing the trained weights of the imitation system, we found that a problem would arise if the weight initialization was too small (less than 0.05) for a specific emotion. This low initialization value of the weight results in failure at learning association of that emotion. Additional tests with an initialization method that generates a different range of values, preferably higher than 0.05, would be necessary to evaluate if the problem lays on the close to zero value of the weight, or if it is more related to the classification accuracy from the emotion recognition system.

Also, the emotion imitation system learns what emotions to express based on association through rewards and penalties; however, humans, depending on the culture are heavily influenced by the context in which emotions are perceived to react to them in a specific manner, the so-called display rules (Ekman & Friesen, 2003); for future work, it would be interesting to expand the system and add a contextual system that defines contexts that require the expression of a different emotion from the one being recognized, interacting with the imitation recognition system. An example would be a human-robot interactive system in the healthcare services, where if fear or anger is perceived from a patient or staff, the system takes into account contextual information to keep a neutral expression or to show a happy expression; naturally, this would be part of a complete multimodal system, that would also react verbally, trying to show empathy or appease the affected person.

5.1 Limitations

To evaluate the accuracy of the emotion recognition system, we performed only 5 runs for each parameter configuration. For the experiment to have a higher statistical validity, many more tests would be necessary, which would be intensely time consuming. With more tests, it would probably be possible to find differences between the different regularization methods, and learning rates optimizations like CLR; nevertheless, the extensive research that was done on these parameters by other researchers, aided in the decision making when choosing the final parameters for the recognition system.

Also, the recognition system was trained on static images because of the greater availability of this kind of datasets. Currently, there are datasets that provide dynamic data with several sequences from neutral to peak emotional expression, which provide more information about the produced facial expression; however, these datasets are usually small, under 30 participants. If datasets of this kind were bigger and publicly available, there would be a possibility to create a more robust system that could include time as another dimension in its analysis of emotions.

Inferior performance on emotion recognition compared to humans was expected; however, according to research on human accuracy on FER-2013 dataset (Goodfellow et al., 2013) the accuracy of the recognition system did get close with 62% compared to human 65% accuracy. Still, the recognition system did not achieve the accuracy (72%) of state-of-the-art solutions (Kim et al., 2016; Pramerdorfer & Kampel, 2016).

Finally, NICO robot expressions are limited by the LED arrays which correspond to its mouth (16×16 pixels arrays) and eyebrows (two 8×8 pixels arrays). In research by Churamani et al. (2017), they found that people would not recognize disgust and fear easily, so they excluded those two emotions from their experiments. In this research all 6 emotions and a neutral expression were used, and the integrated system manages to recognize and imitate them to a certain extent; however, the integrated system was not tested on NICO directly and experiments were humans would recognize emotions imitated by the system were not performed.

5.2 Ethical considerations

We used publicly available datasets referencing indicated research as solicited by the authors. Since the datasets contain faces of the participants, it was important to show only images from participants who agreed for their information to be public. In the case of FER-2013 dataset, this was not possible because the dataset was generated from the internet; however, the images were in public domain.

In addition to ethical issues related to data acquisition, there should be considerations related to the operation of the robotic system; humans could wrongly perceive that the robot has emotions as humans do while this is not true, the expression of emotion aids in the communication with the robot and enriches the human-robot interaction (Cowie, 2015). It is important to familiarize people who meet the robot about general notions of how it operates and how it is expected to behave.

Conclusion

We implemented facial expression recognition with the use of convolutional neural networks and emotion imitation with associative learning to produce a computational system that imitates facial expressions that can be embedded into a robotic system.

The emotion recognition system achieves an accuracy of 83% on laboratory settings datasets JAFFE and CK+ and 62% on the more challenging natural setting FER-2013 dataset. Albeit not enough to reach the results achieved by Pramerdorfer and Kampel (2016) and Kim et al. (2016) with 72% and 70% accuracy respectively, tested on FER-2013 dataset, the emotion imitation system associates the emotions from merged and FER-2013 datasets with its internal representation of emotions. The final accuracy of the integrated system is 80% on JAFFE and CK+ datasets, and 57% on FER-2013 dataset. We can see that the accuracy of the integrated system is highly dependent on the accuracy of the emotion recognition system, which means that the focus of research should be in improving emotion recognition systems. However, it is important to choose between two main paths for emotion recognition research: pursue recognition and processing of emotions in robotic systems as close to how humans do it to understand and implement human behavior in robots, or choose to implement systems that top emotion recognition accuracy regardless of how they are implemented, fully focusing on performance and providing certainty in their results. Both paths have ethical issues that should be considered when creating systems that recognize emotions and interact with humans.

Bibliography

Affectiva. (2017). Emotion AI 101: All About Emotion Detection and Affectiva's Emotion Metrics. Retrieved May 21, 2020, from https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics

American Psychological Association. (n.d.). emotion. Retrieved May 21, 2020, from https://dictionary.apa.org/emotion

Barrett, L. F., Niedenthal, P. M., & Winkielman, P. (2007). *Emotion and Consciousness*. Guilford Publications. Retrieved from https://books.google.sk/books?id=eNsJy2sE7e8C

Brownlee, J. (2019). How to Manually Scale Image Pixel Data for Deep Learning. Retrieved July 24, 2020, from https://machinelearningmastery.com/how-to-manually-scale-image-pixel-data-for-deep-learning/

Bushaev, V. (2017). How do we 'train' neural networks ? Retrieved February 21, 2019, from https://towardsdatascience.com/how-do-we-train-neural-networks-edd985562b73

Can, A. (2017). Layers of a Convolutional Neural Network. Retrieved from https://wiki.tum.de/display/lfdv/Layers+of+a+Convolutional+Neural+Network

Cao, H., Cooper, D. G., Keutmann, M. K., & Gur, R. C. (2014). CREMA-D : Crowd-Sourced Emotional Multimodal Actors Dataset, *5*(4), 377–390.

Chollet, F. (2015). Keras. Retrieved from https://keras.io

Churamani, N., Kerzel, M., Strahl, E., Barros, P., & Wermter, S. (2017). Teaching emotion expressions to a human companion robot using deep neural architectures. *Proceedings of the International Joint Conference on Neural Networks, 2017-May*, 627–634.

Cowie, R. (2015). Ethical Issues in Affective Computing. In R. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.) (pp. 334–348). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.006

Dhall, A., Goecke, R., Gedeon, T., & Sebe, N. (2016). Emotion recognition in the wild. *Journal on Multimodal User Interfaces*, *10*(2), 95–97.

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. *Proceedings of the IEEE International Conference on Computer Vision*, 2106–2112.

Duncan, D., Shine, G., & English, C. (2016). Facial Emotion Recognition in Real Time.

Dupré, D., Krumhuber, E. G., Küster, D., & McKeown, G. (2019). Emotion recognition in humans and machine using posed and spontaneous facial expression, (June), 2–3. https://doi.org/10.31234/osf.io/kzhds

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors (Switzerland)*, *20*(3).

Ekman, P. (2005). Basic Emotions. In *Handbook of Cognition and Emotion* (pp. 45–60). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch3

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Malor Books. Retrieved from https://books.google.sk/books?id=TukNoJDgMTUC

Esparza, J., Scherer, S., Brechmann, A., & Schwenker, F. (2012). Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark. *2012 11th International Conference on Information Science, Signal Processing and Their Applications, ISSPA 2012*, (August 2014), 1253–1258.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., … Bengio, Y. (2013). Challenges in Representation Learning: A report on three machine learning contests. Retrieved from http://arxiv.org/abs/1307.0414

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1026–1034.

Hsu, J. C. (1992). Stepwise multiple comparisons with the best. *Journal of Statistical Planning and Inference*, *33*(2), 197–204.

Http://mplab.ucsd.edu. (2009). The MPLab GENKI Database, GENKI-4K Subset. Retrieved May 21, 2020, from https://inc.ucsd.edu/mplab/wordpress/index.html%3Fp=398.html

Iftekharanam. (2017). Challenges in Representation Learning: Facial Expression Recognition Challenge [Discussion post]. Retrieved July 23, 2020, from https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/discussion/29428

Janssen, J. H., Tacken, P., De Vries, J. J., Van Den Broek, E. L., Westerink, J. H. D. M., Haselager, P., & Ijsselsteijn, W. A. (2013). Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. *Human-Computer Interaction*, *28*(6), 479–517.

Jellinger, K. A. (2010). Oxford Companion to Emotion and the Affective Sciences. *European Journal of Neurology*, *17*(1), e7–e7.

Kerzel, M., Strahl, E., Magg, S., Navarro-Guerrero, N., Heinrich, S., & Wermter, S. (2017). NICO-Neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction. *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, *2017-Janua*, 113–120.

Kim, B.-K. K., Roh, J., Dong, S.-Y. Y., & Lee, S.-Y. Y. (2016). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, *10*(2), 173–189.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Ko, B. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, *18*(2), 20.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Lindsay, G. (2020). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 1–15.

Liu, Y., Sourina, O., & Nguyen, M. K. (2010). Real-time EEG-based human emotion recognition and visualization. *Proceedings - 2010 International Conference on Cyberworlds, CW 2010*, 262–269.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (Vol. 10, pp. 94–101). IEEE.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 200–205). IEEE Comput. Soc.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Omid, M. (2015). How can I normalize input and output data in training neural networks? Retrieved from https://www.researchgate.net/post/How_can_I_normalize_input_and_output_data_in_training_neural_networks

Patel, A. (2018). Facial Expression Recognization using JAFFE. Retrieved from https://github.com/ashishpatel26/Facial-Expression-Recognization-using-JAFFE

Pons, G., & Masip, D. (2018). Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. *IEEE Transactions on Affective Computing*, *9*(3), 343–350.

Pramerdorfer, C., & Kampel, M. (2016). Facial Expression Recognition using Convolutional Neural Networks: State of the Art. Retrieved from http://arxiv.org/abs/1612.02903

Reisenzein, R. (2015). A Short History of Psychological Perspectives on Emotion. In R. Calvo, S. D’Mello, J. Gratch, & A. Kappas (Eds.) (pp. 21–37). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.014

Rosebrock, A. (2019). Keras Learning Rate Finder. Retrieved July 27, 2020, from https://www.pyimagesearch.com/2019/08/05/keras-learning-rate-finder/

Sander, D., & Scherer, K. R. (Eds.). (2009). *The Oxford Companion to Emotion and the Affective Sciences*. *The Oxford Companion to Emotion and the Affective Sciences*. Oxford, England and New York, NY, USA: Oxford University Press.

Sang, D. V., Bao Cuong, L. T., & Thuan, D. P. (2017). Facial smile detection using convolutional neural networks. *Proceedings - 2017 9th International Conference on Knowledge and Systems Engineering, KSE 2017*, *2017-Janua*(June), 136–141.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *American Journal of Health-System Pharmacy*, *75*(6), 398–406.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, (April), 464–472.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., & Movellan, J. (2009). Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 2106–2111.