University of Tartu

Faculty of Science and Technology

Institute of Technology

Abdallah Hussein Sham

Facial Expression Recognition using Neural Network for Dyadic Interaction

Master's Thesis (30 ECTS)

Robotics and Computer Engineering

Supervisors:

Dr. Cagri Ozcinar

Prof. Pia Tikka

Prof. Gholamreza Anbarjafari

Tartu 2020

Abstract

Facial Expression Recognition using Neural Network for Dyadic Interaction

Computers are machines that don't share emotions as humans do. With the help of Machine Learning (ML) and Artificial Intelligence (AI), social robots can become a reality. These robots are currently capable of interacting with people at a certain level, but not exactly as a person would do. For them to reach that level, they would need to understand more about how people interact daily and to learn from the dyadic interaction of two people would be a good option. Participants' facial expressions are the main features that can be retrieved from dyadic interaction and this can be done using a trained Deep Neural Network (DNN) model. The DNN model, known as the Mini-Xception, is trained in this thesis using a dataset that has been pre-processed and can then be tested on images. Using a face detector algorithm, the model will be able to detect a person's facial expression on the image. After successful image results, the model can be tested using a different medium. First, the tests are carried out using a webcam, then videos with more than one participant. Since people react to expressions, their reactions can also be caused by a context in which, for example, sad news would be the reason for sad emotion. The results of the tests will, therefore, be used for analysis where a correlation can be constructed between facial expressions and context.

CERCS: T120 Systems engineering, computer technology; T125 Automation, robotics, control engineering; T111 Imaging, image processing; T121 Signal processing

Keywords: Machine Learning, Artificial Intelligence, Deep Neural Network, Social Robot, Mini-Xception, face detector algorithm, facial expression recognition

Kokkuvõte

Näoilmetuvastus kaksiksuhtluses kasutades tehisnärvivõrku

Arvutid on masinad, mis ei näita emotsioone selliselt nagu inimesed. Tänu masinõppele (ML) ja tehisintellektile (AI), võivad sotsiaalsed robotid saada reaalsuseks. Need robotid on hetkel võimelised inimestega teatud määral suhtlema, kuid mitte täpselt samal viisil kui inimene suhtleks. Selleks, et need robotid võiksid saavutada sama taseme, peaksid nad paremini mõistma, kuidas inimesed igapäevaselt suhtlevad. Hea moodus selleks oleks jälgida kahe inimese omavahelist suhtlust. Osapoolte näoilmed on peamised tunnusjooned, mida kahepoolsest suhtlusest on võimalik välja lugeda, ning seda saab teha treenitud sügava närvivõrgu (DNN) mudeliga. See DNN mudel, mida tuntakse Mini-Xception nime all, on selle töö raames treenitud kasutades eeltöödeldud andmestikku, seejärel saab seda mudelit testida piltide peal. Kasutades näotuvastus algoritmi, on see mudel võimeline tuvastama inimese näoilmeid piltide peal. Pärast edukaid pilditulemusi, on mudelit võimalik testida teise meediumi peal. Esiteks teostatakse katsed veebikaameraga, seejärel videote peal, kus on rohkem kui üks osaline. Kuna inimesed reageerivad näoilmetele, võib nende reaktsioon tuleneda kontekstist, näiteks võivad kurvad uudised olla põhjuseks kurva emotsiooni taga. Selle testi tulemusi kasutatakse seega analüüsimiseks juhul kui loodud on korrelatsioon näoilmete ja konteksti vahel.

CERCS: T120 Süsteemitehnoloogia, arvutitehnoloogia; T125 Automatiseerimine, robootika, juhtimistehnika; T111 Pilditehnika; T121 Signaalide töötlemine

Võtmesõnad: Masinõpe, Tehisintellekt, Sügav närvivõrk, Sotsiaalne robit, Mini-Xception, Näotuvastus algoritm, Näoilme tuvastus

Contents

List of Figures

List of Tables

Abbreviations, constants, definitions

AI - Artificial Intelligence

AFEC - Automated Facial Expression Coding

AR - Augmented Reality

BECV - Behavioral Ecology View

BET - Basic Emotions Theory

DNN - Deep Neural Network

csv - Comma Separated Values

FACS - Facial Action Coding System

FER - Facial Expression Recognition

FFSFP - Face to Face Still-Face Paradigm

FN - False Negative

FP - False Positive

GAN - Generative Adversarial Network

GSR - Galvanic Skin Response

HCI - Human Computer Interaction

HOG - Histogram of Oriented Gradient

HMD - Head Mounted Display

HR - Heart Rate

JAFFE - Japanese Female Facial Expression

ML - Machine Learning

NASA - National Aeronautics and Space Administration

NP - Normal Play

QoE - Quality of Experience

RGB - Red Green Blue

RP - Reunion Period

SVM - Support Vector Machine

TN - True Negative

TP - True Positive

VR - Virtual Reality

3D - 3 Dimension

1 Introduction

Human face is important for social interaction between two or more people, as it provides non-verbal indicators of gender, ethnicity, emotion and overall health [35]. Emotions and their physiological expressions are assumed to be the most significant of these indicators, as they reflect social roles, convey knowledge of people's thoughts, expectations and social experiences [46]. That being said, technological devices cannot feel emotions like humans or animals do. The field of affective computing [68], [14] combined with social robotics [12], machine learning, and big data have opened up new technological opportunities for a machine to learn to detect human face in real time. Some of the major applications include the rise of social humanoid robots like Sophia, Pepper, Nao, Sam, Lynx and many more [69]. Another application is the use of digital humans, including Soul Machines [40], Auxuman and Hatsune Miku [6], which are becoming more common especially in the Virtual Reality (VR) environment [8].

The study of human facial expression is not recent, but it dates back to Darwin C. (1872) from his publication on *The Expression of Emotions in Man and Animals*; he examined the facial expressions of animals to support his belief in evolution to better understand the evolution of human facial behavior [31]. Relating to Darwin's work, Ekman, P. (1973) published *Darwin and Facial Expression* where he explained about the facial expression of emotion in nonhuman primates and humans [24]. Later, Ekman, P. and Friesen, W. (1978) proposed a technique known as the Facial Action Coding System (FACS) that is used to measure facial expressions in the context of basic facial muscle activity [9]. This method has shown promising results in the field of psychology and cognitive neuroscience, in studying the relation of emotions and muscle activity of human faces. Rosenberg, E. And Ekman, P. (1997) published *What the Face Reveals* in which FACS succeeded in differentiating the different types of facial expressions based on universal and cultural differences for different age groups [25].

The use of FACS has also been used in studying two people interacting with one another, i.e.

dyadic interaction by implementing Machine Learning (ML). From [76], Won et al. described how non-verbal behavior predicts learning performance in teacher-student dyads using computer vision hardware and Machine Learning (ML) algorithms to analyse the participants' body gestures. Body gestures are part of non-verbal communication, which has a major impact on human social behavior.

Jacques et al. (2016), investigated how an Artificial Intelligent (AI) agent could predict whether it can be connected more naturally with its user and be able to express relevant facial expression and body language through response [36]. The agent used Machine Learning (ML) classifiers which have been trained for facial expression (using FACS) and body language (head, neck, thumbs, tips of the finger, four positions on each hand, and three positions on the spine) to predict whether the interaction between the user and AI agent is highly or lowly bonded.

On a different note, Cohn, J. (2006), pointed out that facial expressions and emotions are often inaccurately perceived as same phenomenon. To address this he proposed a new approach to FACS, which takes into account emotions and subjective feelings are not one and this statement is outdated [18]. The author also argued that the computing system should be capable of detecting context, representing individual differences, and formulating hypotheses about human emotions.

1.1 Problem Statement

It can be comprehended from the above-mentioned works that the results are satisfactory in the identification of body-gestures and facial expressions using ML algorithms. Today, FACS is most frequently used as it has been proved to be a good method for classification of face expressions.

From [33], Haines et al. (2019) used Automated Facial Expression Coding (AFEC) to classify facial expressions as positive or negative intensity or arousal. AFEC is an automated coding system which combines image recognition with Machine Learning (ML) while FACS is a manual coding system. However, whatever facial recognition technology one uses, none of the current systems can claim that they capture the full range of individual human emotions.

Thus, it can be understood that these techniques require continuous development. Specifically, a new approach is needed that can be used to allow a Deep Neural Network (DNN) model to

learn first individual human facial expressions so that this understanding can later be applied to dyadic interaction settings.

To tackle these problems stated above, this thesis will study following research questions (RQ):
RQ1: Can Deep Neural Network (DNN) model be efficient enough to detect facial expressions?
RQ2: Can the same model be applied in dyadic interactions?
RQ3: Does facial expressions depend on Context?

1.2 Objectives and Roadmap

1.2.1 Objectives

The thesis objectives (Obj) are:
Obj1: To train a Deep Neural Network (DNN) model in order to detect human facial expressions.
Obj2: To test the appropriate trained model in dyadic interactions.
Obj3: To gather information from the tests and analyze them, so that a relation can be found between facial expressions and context.

1.2.2 Roadmap

To meet the objectives of the thesis, following research activities are planned to be applied.

This work is divided into two parts which can be summarized in Figure 1.1 and Figure 1.2. From Figure 1.1, the first part of the research is to pre-process the chosen dataset. The Deep Neural Network (DNN) model is then defined along with some Fine-tuning parameters. Fine-tuning parameters are known to be a set of conditions which would restrict the Deep Neural Network (DNN) model from being either over-trained or under-trained. Once all parameters have been set, the DNN model can start training after the proper dataset is picked.

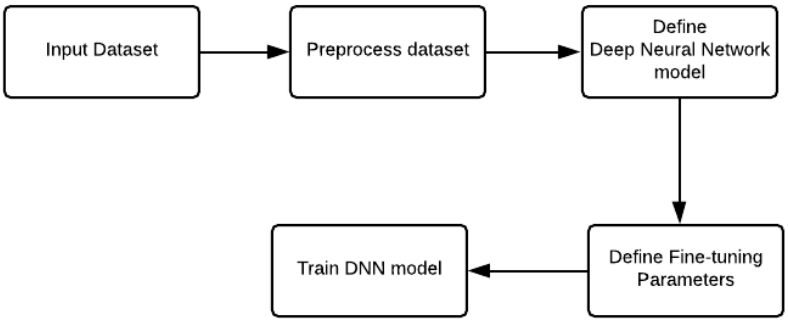


Figure 1.1: The Roadmap to train the Deep Neural Network (DNN) Model.

The second part of the research is to implement the objectives of the dissertation as shown in Figure 1.2. After the Deep Neural Network (DNN) model has been successfully trained, it must be able to detect human faces from a video or a computer webcam. Expressions are detected from identified human faces using the trained DNN model and stored for further processing.

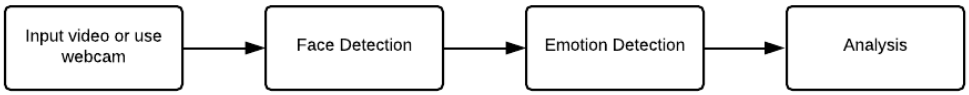


Figure 1.2: Objectives of the dissertation.

In layman's terms, this work is divided into two main parts: training and testing. Identifying and selecting an appropriate dataset is very important in the first part. Subsequently, an algorithm is written with respect to the chosen dataset, which would allow the computer program to gather meaningful information and is known as the dataset pre-processing. The processed information is then fed into the Deep Neural Network (DNN) training model. After training, the model is analyzed for accuracy. In the second part, the trained model is then ready for Facial Expression Recognition (FER) with satisfactory accuracy, which can be performed either by video or web-cam. The algorithm should then be able to detect human faces from the provided medium and classify human facial expression, which is then saved for further analysis.

A major part of this thesis will be dedicated on the Deep Neural Network (DNN) as it is the most important component. Therefore, this work is organized in the following manner. Section 2 covers the Literature Review. The methodology will be covered in section 3. The Experimental Results and Discussion will be presented in section 4. Conclusion and Future Work will be presented in section 5.

2 Literature Review

2.1 Virtual Reality

The affective computing plays an important role in the psychological studies of Virtual Reality. The word 'virtual' means physically non-existent and the word 'reality' refers to an environment that appears real, together referring to world perceived as a synthetic reality [64]. Bricken, W. (1990) [13] described several ways in which Virtual Reality (VR) can be implemented in different fields, such as education, gaming, aviation, military, entertainment and its physics is psychology. The emotion recognition and other affective computing implementations help to make the Virtual Reality (VR) environment more immersive and interactive. The implementation of the research carried out in the field of human psychology can thus help to develop the VR field.

Head Mounted Display (HMD) is assumed to be the main VR hardware but for creating immersive and interactive VR environments, the system requires other hardware devices. From [62], the authors used two different wearable hardware capable of measuring Galvanic Skin Response (GSR), Heart Rate (HR), Electroencephalogram (EEG) and eye-tracking for physiological metrics. The psycho-physiological measurements can help to analyse how to improve the Quality of Experience (QoE).

Besides the use of physiological or implicit metrics to understand the quality of experience (QoE), experts have also carried out surveys using explicit metrics. In other words, human parts such as the hand, the face, the head and even the whole body are considered explicit metrics. From [73], the authors conducted a survey by comparing emotional states induced 360°via two different ways, via the Head Mounted Display (HMD) and via computer screen. They concluded that the intended emotional states could be evoked in both ways but participants felt more immersed in using HMD. It is assumed that if the content that the person is watching is

emotionally engaging, both the computer screen and the Head Mounted Display (HMD) can allow equally emotional experiences.

2.2 Facial Expression Recognition Datasets

There are several datasets that are currently available for training Facial Expression Recognition models. From the following review paper [44], the author summarized details of datasets that are available in this field and some are illustrated in Table 2.1.

Table 2.1: Overview of Facial Expression Databases. Adapted from [44]

Database	Samples	Expressions	Conditions
CK+ [47]	593 images	7 + 1	Lab
MMI [71]	740 images and 2,900 videos	7	Lab
JAFFE [50]	213 images	7	Lab
TFD [66]	112,234 images	7	Lab
FER-2013 [29]	35,887 images	7	Internet
AFEW 7.0 [21]	1,809 videos	7	Movie
SFEW 2.0 [22]	1,766 images	7	Movie
BU-3DFE [78]	2,500 3D images	7	Lab
BU-4DFE [77]	606 3D sequences	7	Lab
RaFD [41]	1,608 images	7	Lab
KDEF [48]	4,900 images	7	Lab
EmotioNet [26]	1,000,000 images	23	Internet
RAF-DB [45]	29,672 images	7 and 12	Internet
AffectNet [53]	450,000 images	7	Internet
ExpW [79]	91,793 images	7	Internet
4DFAB [16]	1.8 million faces	7	Lab

Table 2.1 depicts an overview of some databases on Facial Expressions. It can be understood that most of these databases make use of the seven basic facial expressions such Anger, Disgust,

Scared, Happy, Sad, Surprised and Neutral. As for CK+ [47], the additional expression is contempt and for the RAF-DB [45] consists of the seven basic facial expressions and 12 compound expressions. EmotioNet contains 23 facials expressions which are both the basic and compound expressions. Compound expressions are normally a combination of other expressions, for instance, it can be happily-surprised.

The situation of the person is important in order to have the expected facial expression. The databases that are assembled in a laboratory condition are ignored, which is 9 databases in this case. This is because the participants are all facing the camera at the same angle, which can cause problems in the future. First, if the person's face were in a different position from the way the Deep Neural Network (DNN) model was trained, the outcome would not be as expected. In addition, the size and quality of some of the databases is high. If the Deep Neural Network (DNN) model is trained using these high-quality images, then the model will only work with high-quality images. It is thus clear that making use of such databases can cause problems in the future.

Out of the remaining 7 databases, only FER-2013 [29] can be easily found as the rest of the databases are overloaded with access requests and have to wait for them to be accepted. FER-2013 [29] appears to be a good choice as it contains images taken in the wild with a low image size and quality. In other words, there are images of people taken from different angles. The size and quality of the images are fair enough, because the faces are more or less centered and occupy roughly the same amount of space in each image.

2.3 Related Works

Facial Expression Recognition (FER) has previously been mentioned in a variety of fields. The use of FER in dyadic interaction with Deep Learning is currently being exploited. From [34], the authors proposed a new approach to generating contextually relevant facial expressions in dyadic interaction between the interviewer and the interviewee using the Generative Adversarial Network (GAN). The authors created their own database, which was recorded through video-conferencing during the university's consented admission interviews. In all videos, the interviewer was the same person. The result was to generate a facial expression of the interviewer based on the interviewee's response behaviour.

From [43], the authors performed an experiment to investigate the relational proximity of human dyadic interactions. The main objective was to understand the human to human behavior and try to map the behavioral codes onto the next generation of androids. The study of human dyadic interaction is therefore used to make human computer interaction more fluid.

Most of the studies that have been carried out in relation with dyadic interaction are mainly performed under specific situations in experimental laboratory settings, yet trying to introduce more ecologically valid situations. One example of dyadic interaction experiments is the parent-infant scenario where father-infant or mother-infant scenarios are recorded. The research applied the Face-to-Face Still-face Paradigm (FFSFP) [54], a validated procedure to estimate the socio-emotional regulation between two people. To study the facial interaction between parent and infant, two scenarios were set: the Normal Play (NP), in which the parents were instructed to be unresponsive and then to re-engage with the child referred to as the Reunion Period (RP) [55]. From this study, it can be understood that there are given scenarios which can affect the interaction between people.

From [37], the authors implemented the FFSFP in an interactive virtual learning environment and used the virtual agent as a pedagogical character, using the accumulated knowledge of human psychology with FFSFP as non-verbal feedback.

Yuon et al. (2016) [57] ran an experiment to see if improving one's smile in a virtual environment (through avatar) could result to a more proper communication experience. They concluded that behavioral measures in virtual environments can be used as a discreet measure that can estimate people's attitudes.

The applications of Virtual Agent are implemented in different various fields apart from gaming and learning. For instance, Figure 2.1 shows the implementation of virtual agent as doppelganger for self-awareness in health risk. This work showed the effect of drinking soft-drink on the human body. The virtual agent/doppelganger is shown in a room drinking from a bottle of fizzy drinks. The results are shown after 2 months, 1 year and 2 years where the latter has gained two, 10 and 20 pounds respectively from the scale. From this example, showing the virtual effect one self can help cause an impact on people.

Figure 2.1: From (A), a virtual doppelganger standing in the room holding a bottle of soft drink. As the virtual agent starts drinking from the bottle, the calendar and clock show the rapid passage of time, and fat starts to fall on the digital scale (B). The virtual agent has gained 10 pounds (C) after a year consuming soft drinks. The virtual agent has gained 20 pounds at the end of two years, as shown on scale (D). Adapted from [2].

In addition, VR is believed to be a specific narrative medium in the storytelling disciplines and the intelligent characters [7]. The idea behind storytelling is to make the viewer fully immerse oneself in the content and structure of the story [59]. Virtual characters play a key role in virtual environments in which they need to be socially intelligent, capable of responding to the viewer's non - verbal cues so that the characters can become more credible [72]. Some of the virtual characters can either be nonexistent humans such as Ross [15] or live actor based humans as created in [58]. In the latter work the concept of *Volumetric Intelligence* is proposed to describe production pipeline for a dynamically active human-like character.

From [67], [42] and [10], the authors conducted similar avatar studies that could lead to a better communication experience between humans and computers. The dyadic interaction experiments help to improve Human Computer Interaction (HCI).

3 Methodology

This section discusses the various methods that exist which can be used as one of the facial expression recognition methods. The overall system will ideally be able to detect human face from an image or video using one of the techniques used to detect any objects. The extracted feature is then translated into a Deep Neural Network (DNN) model, in this case, human face. In order for the model to work, it must be trained first. The overview of the Deep Neural Network (DNN) model is explained along with other architectures. For this work, one model is chosen and trained.

As mentioned earlier, this work is divided in two parts, training and testing. The training part is explained in this section and testing part is described in Section 4. Figure 3.1 shows the flow of events in this section.

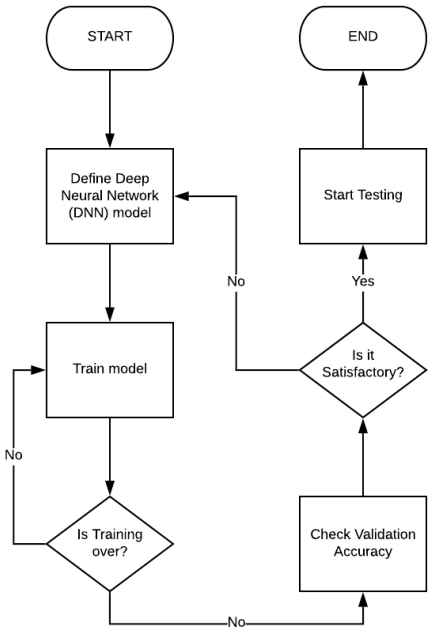


Figure 3.1: Overview of Training part.

Figure 3.1 describes a flowchart of the tasks that need to be done in this section. Each layer of

the Deep Neural Network (DNN) architecture is defined first. Once all the layers have been defined, the DNN model can start training. While training the DNN model, two types of accuracy are most often considered: training and validation. The training accuracy shows how accurate the model was taught and was able to predict the emotions from the images used during training while validation accuracy shows how exact can the model predict from a separate set of images which is completely unknown.

Before training, it is important to preprocess the dataset. This means that appropriate information is described for the computer so that it can recognize the features that are expected to be retrieved from images. In this case, these features relate to human face and emotional expression. The idea is that the model will take the human face as input and the emotion as output. In order to provide the human face information to the model, the concept of face detection needs to be understood.

3.1 Face Detection

Computers perceive an image of human face as a matrix of numbers with three different channels. Each channel will be representing a different color namely; Red, Green and Blue (RGB). One common practise for image processing is to use to convert the image to grayscale. After converting to grayscale, there is only one channel with values varying between 0 and 255. Zero means that it is a black pixel and 255 for a white pixel. In so doing, the computer system has a matrix of numbers which can be registered for a single image.

Imagine a person is given a matrix of numbers and is asked to find a human face from it. The person will not be able to do so, unless if the knowledge is acquired and the same applies to a computer system. The content analysis of faces and facial expressions from an image is done using object detection techniques, such as Haar-like cascade, Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) which are used as classifiers [19].

Deniz et al. [20] explains how Histogram of Oriented Gradients (HOG) is used to detect faces. The algorithm counts occurrences of edge orientations in a local neighborhood of an image. In doing so, the image is divided into smaller regions which are connected and known as cells. These cells are computed in a histogram of edges which are compared so that each feature of a human face can be extracted. The same method is commonly used from Dlib, a toolkit written

in C++ programming language that helps to solve real life problems [23]. This toolkit provides the face detector algorithm as well as facial landmarks. Kopp et al. [39] compared different types of face detector and alignment methods for face landmark detection and in most cases, Dlib performed better than other methods. In this thesis, the Dlib 68 2D dataset is chosen as it outputs 68 x-y coordinates of a facial landmarks from a two dimension image.

3.2 Deep Neural Network (DNN)

Technologies have developed tremendously over the years and biomimicry helped to make it possible. Biomimicry means learning from nature, and imitating it [60]. This practice has been implemented in various fields namely, transportation, aeronautic, aerospace and even architecture [49]. For instance, the biomimicry of the human brain functions may be understood to correspond to a type of Artificial Intelligence (AI) [52]. Similarly, human brain neuron functions can be imitated to some extent in the functions of artificial neurons.

Thousands or millions or any combinations of artificial neurons connected to each other constitutes a Deep Neural Network (DNN). DNN is comprised of three main layers: Input, Hidden and Output layers. Figure 3.2 shows an illustration of the main constituents of a Deep Neural Network (DNN) model. The number of hidden layers (H) is normally more than one. The input layer (I) is made up from the user input based on a database. The output layer (O) is used to provide the expected result whereby the one with the highest probability is considered to be the final result.

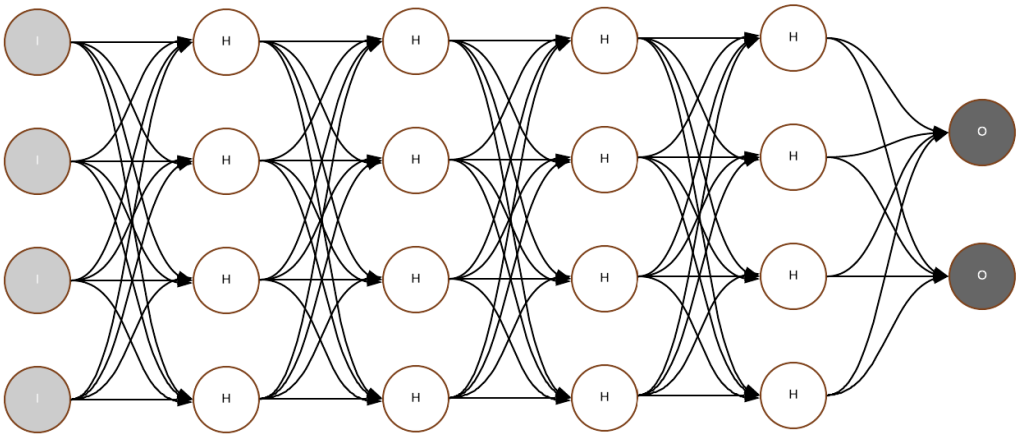


Figure 3.2: Sample Overview of an Deep Neural Network (DNN). Adapted from [11].

Between the hidden layers, each connection can be represented as shown in Figure 3.3 whereby

H is the value of input for a specific neuron in the current layer and W is the weight value of the current layer with respect to the preceding one. When both of them are multiplied, the result is the output of a neuron.

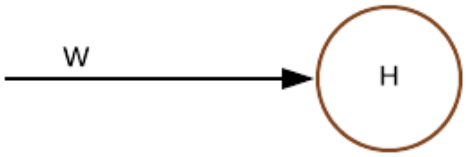


Figure 3.3: Sample Overview of a hidden connector. Adapted from [65].

3.2.1 Convolutional Neural Network

Fukushima [28] proposed the concept of *Neocognitron* - a neural network model for stronger visual pattern recognition. Later on, the same concept is evolved into Convolutional Neural Network which is more effective and commonly used in the field of Computer Vision. Figure 3.4 shows the general overview of how this method works. CNN consists of three main layers which are Convolutional, Pooling and Fully connected layers.

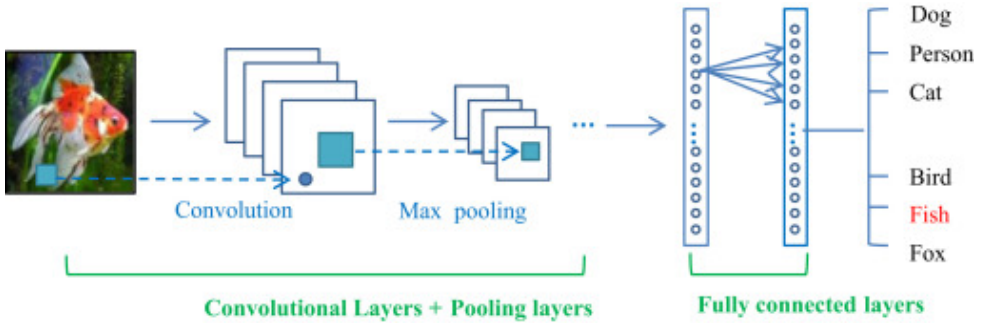


Figure 3.4: General Overview of Convolutional Neural Network Architecture. Adapted from [32].

Convolutional layers are layers whereby CNN uses a number of different kernels to convolve the image and also the intermediate feature maps. In doing so, the Neural Network is able to minimize the number of parameters, the local connections learn from the neighbouring pixels which results in faster learning.

Pooling layers are normally followed after convolutional layers. The purpose of this layer is to minimize the size of the next convolutional layer. This process reduces the number of pixels

as well as information. The loss information will be beneficial for the Neural Network against Overfitting. The most common techniques for pooling are Maximum or Average Pooling. Maximum pooling is considered to be the fastest.

Fully-connected layers always come after pooling layer. As the name suggests, neurons are connected to all activated neurons to the previous layer. It can be seen from Figure 3.4, Fully-connected layers are in one dimension from two dimension feature maps. Hence, it makes use of feed-forward the Neural Network to a vector which can be further processed [74].

3.2.2 Xception model

Xception architecture is inspired from its predecessor the Inception architecture. The term 'Xception' comes from Extreme Inception and makes use a new technique of convolution called depthwise separable convolution. This technique allows a separate layer to be working independently from every channel. Hence, depthwise separable convolution resulted in faster learning with better accuracy proven in [38]. Figure 3.5 shows the architecture of the Xpection model.

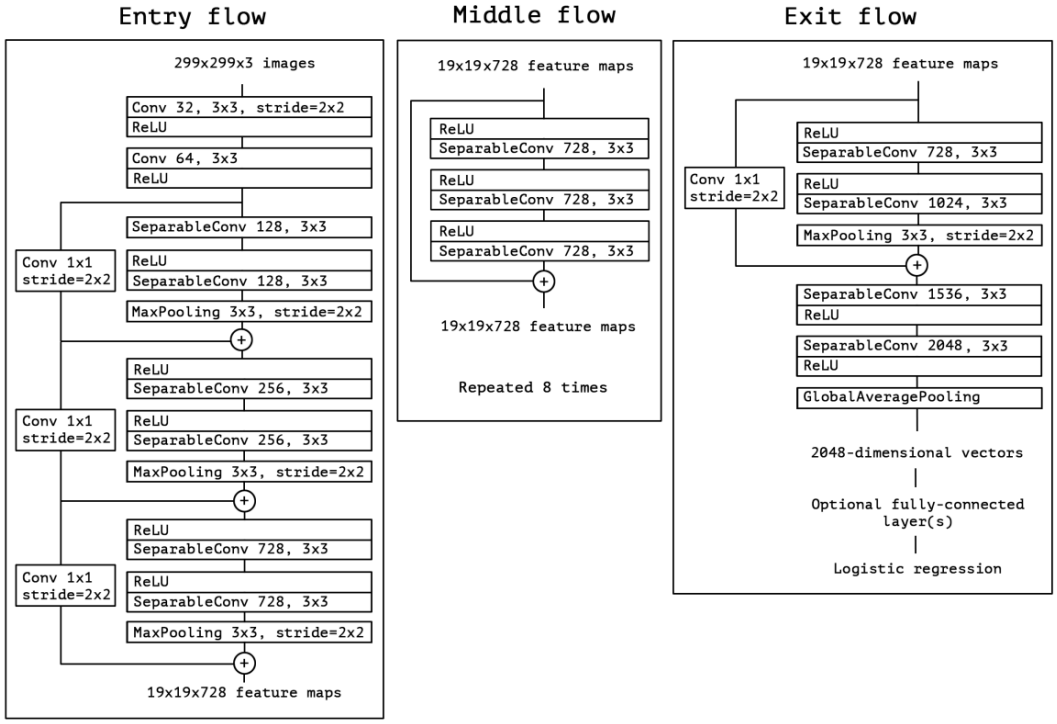


Figure 3.5: General Overview of Xception Model architecture. Adapted from [17].

From Figure 3.5, the architecture is divided into three main flows: Entry, Middle and Exit. In the entry flow, colored images of size 299 by 299 are inserted into the model. Afterwards, the

data are processed with three layers of separable convolution. In the middle flow, computations are performed into a three layered separable convolution which is repeated 8 times. Then the exit flow is used to minimize the number of parameters so that the output can be prepared to be inserted into a fully connected layer for logistic regression. It is important to mention that a batch normalisation is performed after every convolution and separable convolution process. The flows are used so that the coarse features are extracted in the entry, more complex features are extracted in the middle and features with more details are extracted before the global average pooling in the exit flow [61].

3.2.2.1 Mini-Xception model

The Xception model is a big and complex architecture which can cause the training of the model to use huge computational resources. As the chosen dataset (FER2013 dataset) makes use of grayscale images and Xception model takes normal coloured images, some layers can be removed to minimize time and resources. The mini-Xception model is considered, therefore, because of its advantages over its predecessor. Figure 3.6 shows the architecture of Mini-Xception model.

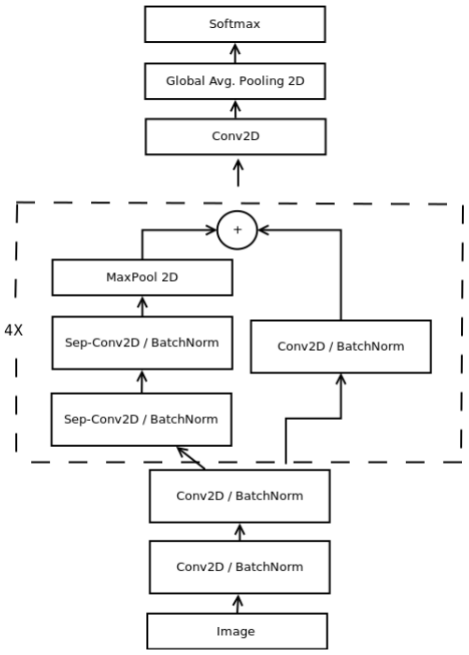


Figure 3.6: General Overview of Mini-Xception Model architecture. Adapted from [5].

From Figure 3.6, the model architecture is much smaller as compared to Xception model. The whole has been compiled into one flow and the fully-connected layers have been removed. In

doing so, the number of parameters is much less and this favors time as well as resources. The image is inserted, then is processed by two convolutional layers before getting the residual module and separable convolution layers which are both repeated four times. Finally, the output goes into one last convolution layer and global average pooling which will be sent to a softmax layer- a layer that is used for final output of a neural network through multi-class classification.

3.2.2.2 Training model

Once an appropriate model has been chosen, it does not mean that training can be started. It is important to make sure that the model will neither be over-fitted nor under-fitted. This means that the trained model should be able to learn from the dataset and generalize well from the learnt dataset. If it is not able to learn properly, it is said to be under-fitting. Else if, it is able to learn too well but unable to generalize outside the dataset, then it is said to be over-fitting.

In order to get a good fit model, the model needs to be fine tuned during training. The fine-tuning is shown in the table 3.1. The batch number is set to 32 and the input shape is set to 48. The number of Epochs is set to 1,000 but the model will stop training if the validation accuracy does not change after 50 epochs/ iterations. In this model, the learning rate is the speed at which the model learns and is initially set by default on the computer. The latter will reduce if there is no change to the validation accuracy after 12 consecutive epochs. In doing so, the system is not destined for worst-case scenarios mentioned above.

Table 3.1: Overview of fine-tuning parameters.

Parameters	Values
Batch Size	32
Number of Epochs	1,000
Input Shape	(48,48,1)
Patience	50

4 Experimental Results and Discussion

The system implementation can be carried out using the methods described above. Firstly, the DNN model is defined and trained using the dataset chosen (FER-2013 dataset). Some trials will be done to choose a model, and the one with the best results would be chosen. Once a model is chosen, the accuracy is checked and evaluated. Afterwards, the model is made to perform tests using images, then webcam and videos. All the results are saved so that they can be analyzed which can later be reviewed, evaluated and addressed in the Discussion section at the end of this chapter.

4.1 Convolution Neural Network (CNN)

CNN is widely used in Deep Neural Networks (DNN). In this subsection, all the tests that have conducted using CNN. It is further demonstrated how they were trained, followed with their results. The fine-tuning parameters were set to early stopping and reducing learning rate after 10 epochs while training. More details are provided respectively below:

4.1.1 CNN model using 3 Convolutional block layers and 1 fully-connected block layer

This model is the first attempt of training the model for this thesis. Table 4.1 depicts all the parameters, layer and output shape for the whole model. It can be summarized as the first 8 rows to be one block and hence, there are 3 blocks of convolutional layers and 1 block of fully-connected layer. The total number of parameters is 6,127,525 and 896 are non-trainable parameters. The latter was proposed by Serebgil, S. (2018) [63] which according to him, was supposed 66% accurate.

Table 4.1: Convolutional Neural Network(CNN) model architecture using 3 Convolutional layers and 1 fully-connected layer.

Layer/type	Output Shape	Number of Parameters
Convolution 2D	(None, 48, 48, 32)	320
ReLU	(None, 48, 48, 32)	0
Batch Normalization	(None, 48, 48, 32)	128
Convolution 2D	(None, 48, 48, 32)	9248
ReLU	(None, 48, 48, 32)	0
Batch Normalization	(None, 48, 48, 32)	128
Max Pooling	(None, 24, 24, 32)	0
Dropout	(None, 24, 24, 32)	0
Convolution 2D	(None, 24, 24, 64)	18496
ReLU	(None, 24, 24, 64)	0
Batch Normalization	(None, 24, 24, 64)	256
Convolution 2D	(None, 24, 24, 64)	36928
ReLU	(None, 24, 24, 64)	0
Batch Normalization	(None, 24, 24, 64)	256
MaxPooling	(None, 12, 12, 64)	0
Dropout	(None, 12, 12, 64)	0
Convolution 2D	(None, 12, 12, 128)	73856
ReLU	(None, 12, 12, 128)	0
Batch Normalization	(None, 12, 12, 128)	512
Convolution 2D	(None, 12, 12, 128)	147584
ReLU	(None, 12, 12, 128)	0
Batch Normalization	(None, 12, 12, 128)	512
MaxPooling	(None, 6, 6, 128)	0
Dropout	(None, 6, 6, 128)	0
Flatten	(None, 4608)	0
Dense	(None, 1024)	4719616
Dropout	(None, 1024)	0

Dense	(None, 1024)	1049600
Dropout	(None, 1024)	0
Dense	(None, 64)	65600
Dropout	(None, 64)	0
Dense	(None, 64)	4160
Dropout	(None, 64)	0
Dense	(None, 5)	325

After training the model, the validation accuracy was 29.79% and the overall of accuracy against epochs can be represented in Figure 4.1. The x and y axes are the epoch and the percentage of accuracy, respectively. The Red line represents the training accuracy which varied from 20% to 29.9% and the blue line represents the validation accuracy which alternated around 28.8% to 29.5%.

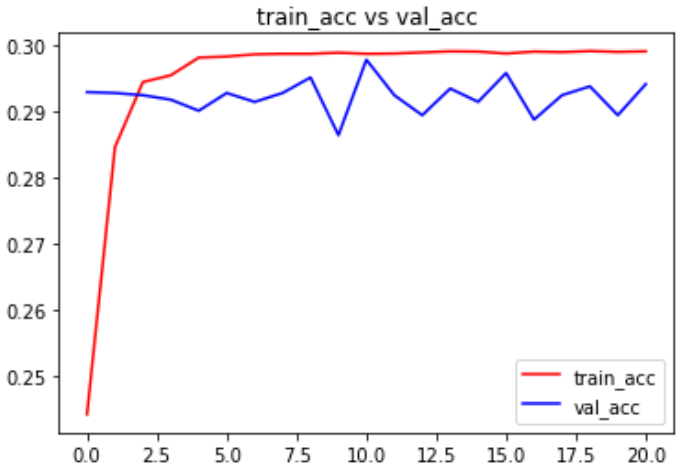


Figure 4.1: Overall accuracy versus Epoch for first attempt.

It can be understood from Figure 4.1 that the model is not accurate enough. In this case, a better model should be chosen.

4.1.2 CNN model using 2 Convolutional block layers and 1 fully-connected block layer

This model architecture is the second attempt of training the model for this thesis. Table 4.2 depicts all the parameters, layer and output shape for the whole model. It can be summarized

as the first 3 rows to be one block and hence, there are 2 blocks of convolutional layers and 1 block of fully-connected layer. The total number of parameters is 7,954,391 and all of them are trainable parameters. This model architecture was proposed by Maurya, R. (2018) [51] whereby he showed that this model has an accuracy of 56% and above.

Table 4.2: Convolutional Neural Network(CNN) model architecture using 2 Convolutional layers and 1 fully-connected layer.

Layer/type	Output Shape	Number of Parameters
Convolution 2D	(None, 128, 128, 6)	456
Activation	(None, 128, 128, 6)	0
Maximum Pooling	(None, 64, 64, 6)	0
Convolution 2D	(None, 64, 64, 16)	2416
Activation	(None, 64, 64, 16)	0
Maximum Pooling	(None, 32, 32, 16)	0
Convolution 2D	(None, 28, 28, 120)	48120
Activation	(None, 28, 28, 120)	0
Dropout	(None, 28, 28, 120)	0
Flatten	(None, 94080)	0
Dense	(None, 84)	7902804
Activation	(None, 84)	0
Dropout	(None, 84)	0
Dense	(None, 7)	595
Activation	(None, 7)	0

After training the model, the validation accuracy was 56.25% and the overall of accuracy against epochs can be represented in Figure 4.1. The x and y axes are the epoch and the percentage of accuracy, respectively. The Red line represents the training accuracy which varied from 78% to 90% and the blue line represents the validation accuracy which varied between 40.6% to 56.25%. Hence, the accuracy matched what the author initially proposed.

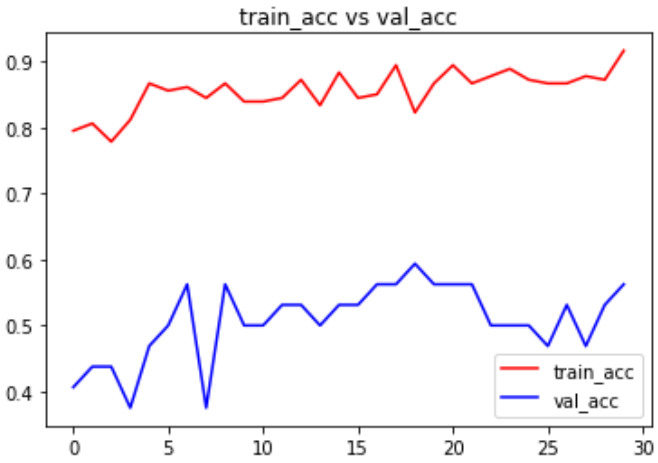


Figure 4.2: Overall accuracy versus Epoch for second attempt

From this validation accuracy, the model was tested and the result can be seen in Figure 4.3. The model predicted that the participant is sad and the same expression did not change throughout the whole testing. Since the model kept predicting 'sad' expression, this does not show to be an effective model.

Figure 4.3: Sample result from the model with 56%.

The difference between the two attempts is that the convolution layers were reduced and the fully connected was increased. Doing such thing may increase the accuracy but the computational time and resources would increase. Therefore, the mini-Xception model is considered to be worthy.

4.1.3 Proposed Method

As described earlier, the mini-Xception architecture was described in Figure 3.5. The script is implemented as explained in the paper from its code repository [27]. The layers are summarized and tabulated below in Table 4.3. The total number of parameters sums up to 58,423 out of which 1,472 are non-trainable parameters.

Table 4.3: Mini-Xception Neural Network Model.

Layer/type	Output Shape	Number of Parameters
Input Layer	(None, 64, 64, 1)	0
Convolution 2D	(None, 46, 46, 8)	72
Batch Normalization	(None, 46, 46, 8)	32
Activation	(None, 46, 46, 8)	0
Convolution 2D	(None, 44, 44, 8)	576
Batch Normalization	(None, 44, 44, 8)	32
Activation	(None, 44, 44, 8)	0
Separable Convolution	(None, 44, 44, 16)	200
Batch Normalization	(None, 44, 44, 16)	64
Activation	(None, 44, 44, 16)	0
Separable Convolution	(None, 44, 44, 16)	400
Batch Normalization	(None, 44, 44, 16)	64
Convolution 2D	(None, 22, 22, 16)	128
Maximum Pooling 2D	(None, 22, 22, 16)	0
Batch Normalization	(None, 22, 22, 16)	64
Addition	(None, 22, 22, 16)	0
Separable Convolution	(None, 22, 22, 32)	656
Batch Normalization	(None, 22, 22, 32)	128
Activation	(None, 22, 22, 32)	0
Separable Convolution	(None, 22, 22, 32)	1312
Batch Normalization	(None, 22, 22, 32)	128
Convolution 2D	(None, 11, 11, 32)	512
Maximum Pooling 2D	(None, 11, 11, 32)	0

Batch Normalization	(None, 11, 11, 32)	128
Addition	(None, 11, 11, 32)	0
Separable Convolution	(None, 11, 11, 64)	2336
Batch Normalization	(None, 11, 11, 64)	256
Activation	(None, 11, 11, 64)	0
Separable Convolution	(None, 11, 11, 64)	4672
Batch Normalization	(None, 11, 11, 64)	256
Convolution 2D	(None, 6, 6, 64)	2048
Maximum Pooling 2D	(None, 6, 6, 64)	0
Batch Normalization	(None, 6, 6, 64)	256
Addition	(None, 6, 6, 64)	0
Separable Convolution	(None, 6, 6, 128)	8768
Batch Normalization	(None, 6, 6, 128)	512
Activation	(None, 6, 6, 128)	0
Separable Convolution	(None, 6, 6, 128)	17536
Batch Normalization	(None, 6, 6, 128)	512
Convolution 2D	(None, 3, 3, 128)	8192
Maximum Pooling 2D	(None, 3, 3, 128)	0
Batch Normalization	(None, 3, 3, 128)	512
Addition	(None, 3, 3, 128)	0
Convolution 2D	(None, 3, 3, 7)	8071
Global Average Pooling 2D	(None, 7)	0
predictions (Activation)	(None, 7)	0

After the model has been described, the model is set to training along with the fine tuning parameters. As mentioned in Methodology section, the fine tuning parameter such as batch size, number of Epochs, Input shape, and patience are set accordingly. New set of parameters are implemented such the use of Image generator along with image augmentation from varying from -10°to 10°, rotating the images to a 10% range and allowing to flip them horizontally. The Input shape was changed from (48,48,1) to (68,68,1). It is believed that increasing the input form would improve the accuracy of the validation by 2%. The model is set to training and the

result was summarized in a confusion matrix depicted in Table 4.5. The Validation accuracy obtained was 65% and 67% after evaluation.

Table 4.4: Confusion Matrix for training the model on first attempt.

	Angry	Disgust	Scared	Happy	Sad	Surprised	Neutral
Angry	575	13	76	38	152	16	117
Disgust	24	64	10	0	10	1	6
Scared	128	7	393	43	121	91	91
Happy	29	3	23	1565	33	31	110
Sad	115	2	108	46	742	12	211
Surprised	27	2	89	46	18	624	21
Neutral	62	4	37	81	183	13	865

The model is trained again with new set of parameters are implemented such the use of Image generator along with image augmentation from varying from -25°to 25°, rotating the images to a 25% range and allowing to flip them horizontally. The model is set to training and the result was summarized in a confusion matrix depicted in Table 4.4. The Validation accuracy obtained was 65% and 69% after evaluation.

Table 4.5: Confusion Matrix for training the model on a second attempt.

	Angry	Disgust	Scared	Happy	Sad	Surprised	Neutral
Angry	626	7	64	35	115	21	134
Disgust	27	68	9	2	4	1	6
Scared	125	7	468	37	205	80	112
Happy	31	0	13	1587	32	23	102
Sad	111	1	92	43	788	7	234
Surprised	25	3	74	42	16	630	24
Neutral	60	1	44	76	138	10	818

The use of seven basic facial expressions in this thesis is not important. For example, the results of Haines using AFEC were categorized as Positive or Negative Intensity, as mentioned previously in Problem Statement. On another point, suppose a person has been presented with 7 different car make and asked to choose one. Although 3 different makes of cars are introduced to another person, it is presumed that an individual with more alternatives would take more time so that his / her option will result in satisfaction as suggested by Gray, L. & Tallman, I. [30].

Using three facial expressions instead of seven can help to avoid confusion and ease the decision-making process. Hence, the seven basics facial expressions are altered as follows: Angry and Sad are labeled as Negative, Happy & Surprised as Positive, Neutral remained the same but Disgust & Scared are removed. Once these changes are done, the same model and parameters are

used to train a new model and the confusion matrix is summarized in Table 4.6. The validation accuracy obtained was 80% and accuracy after evaluation was 82%.

Table 4.6: Confusion Matrix for training the model on third attempt.

	Negative	Positive	Neutral
Negative	1846	117	246
Positive	146	2280	100
Neutral	337	124	848

The process for calculating accuracy is based on Table 4.6 and the evaluation can be performed using the materials provided in [56] and [1] as reference. Columns represent the estimated values of the Trained model while rows indicate the actual values in the dataset. In order to check the model, a section of the dataset was kept aside so the model could predict the outcome based on what it learned during training. From the table, the second row and column show the number of times the model predicted Negative just as the actual outcome was Negative and this is known as True Positive (TP). Besides TP on the same row, these values are known as False Negative (FN) and in the same column, all other values besides TP are known as False Positive (FP). Using the defined terms, the 3 basic calculations can be done using the equations below and the same terms can be applied on the other classified outcomes:

$$Recall = \frac{TP}{(TP + FN)} \tag{4.1}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{4.2}$$

$$F1 - Score = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \tag{4.3}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{4.4}$$

After using equations 4.1, 4.2 and 4.3 for each classification. The mean-average value known as Macro-average of Recall, Precision and F1-score can then be calculated by the following equation:

$$Macro - F1 = \frac{(F1 - score_{Negative} + F1 - score_{Positive} + F1 - score_{Neutral})}{3} \tag{4.5}$$

Equation 4.5 can also be used to calculate the mean values for both Precision and Recall. Then, weighted-average value of classification can be calculated using the Equation 4.6.

$$Weighted - F1 = \frac{\sum_{i=Negative}^{Positive,Neutral}((TP + FN)_i * (F1 - Score)_i)}{(TP + FP + FN + TN)} \tag{4.6}$$

Using the above-mentioned equations, the results are as follows:

1. Negative:
   (a) Recall: 1846 / 2209 = 0.836
   (b) Precision: 1846 / 2329 = 0.793
   (c) F1-score: 2 x (0.793 x 0.836) / (0.793 + 0.836) = 0.814
2. Positive:
   (a) Recall: 2280 / 2526 = 0.903
   (b) Precision: 2280 / 2521 = 0.904
   (c) F1-score: 2 x (0.903 x 0.904) / (0.903 + 0.904) = 0.903
3. Neutral:
   (a) Recall: 848 / 1309 = 0.648
   (b) Precision: 848 / 1194 = 0.710
   (c) F1-score: 2 x (0.648 x 0.710) / (0.648 + 0.710) = 0.678

4. Macro-F1: (0.814 + 0.903 + 0.678) / 3 = 2.395 / 3 = 0.798

5. Macro-Precision: (0.793 + 0.904 + 0.710) / 3 = 2.407 / 3 = 0.802

6. Macro-Recall: (0.836 + 0.903 + 0.648) / 3 = 2.387 / 3 = 0.796

7. Accuracy: 4974 / (4974 + 1070) = 0.823

8. Weighted-F1: (2209 x 0.814 + 2526 x 0.903 + 1309 x 0.678) / 6044 = 4966.606 / 6044 = 0.822

9. Weighted-Precision: (2209 x 0.793 + 2526 x 0.904 + 1309 x 0.710) / 6044 = 4964.631 / 6044 = 0.821

10. Weighted-Recall: (2209 x 0.836 + 2526 x 0.903 + 1309 x 0.648) / 6044 = 4975.934 / 6044 = 0.823

The values of Accuracy, weighted-F1, weighted-Precision and weighted-Recall are compared. They are multiplied by 100% to get the percentages and they are all approximated to 82%. Hence, this is how the model accuracy is evaluated.

4.1.4 Analysis

Once the model has been trained successfully, testing can be carried out. Figure 4.4 and 4.5 are two examples that show the model works appropriately using images. The next step is to test the model over a video or webcam and this is where flaws can be spotted.

Figure 4.4: Tested image of an angry woman [4].

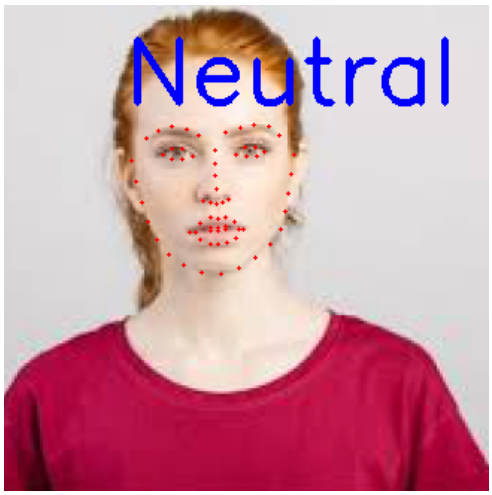


Figure 4.5: Tested image of a woman with neutral emotion [75].

Figure 4.7 shows the different tests carried out by a participant. In general, the model works perfectly for still face and emotion but some unexpected results were obtained. The participant deliberately obscured side of his face from part (a) and (b), and thus the device could not identify his face.

Nonetheless, Figure 4.6 shows that the device will operate if the face is partially blocked. It can be inferred that the system will operate if two-thirds of the face is visible. In part (c) and (e), the system predicts the exact facial expression, but not when the face is titled as shown in part (d). It was also found that the system was able to recognize the participant's face when the participant was positioned in that location. On the other hand, the device operates well with the user wearing eyeglasses.

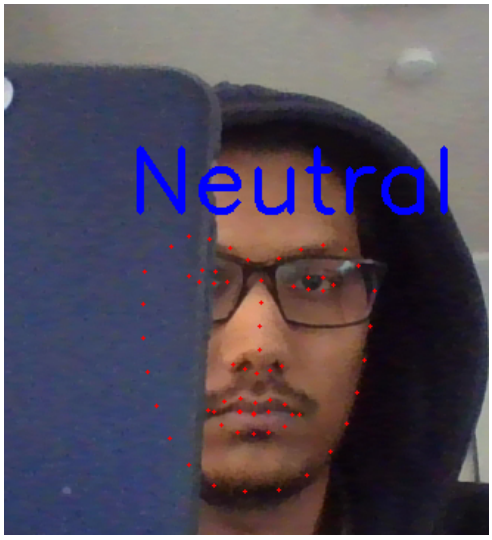


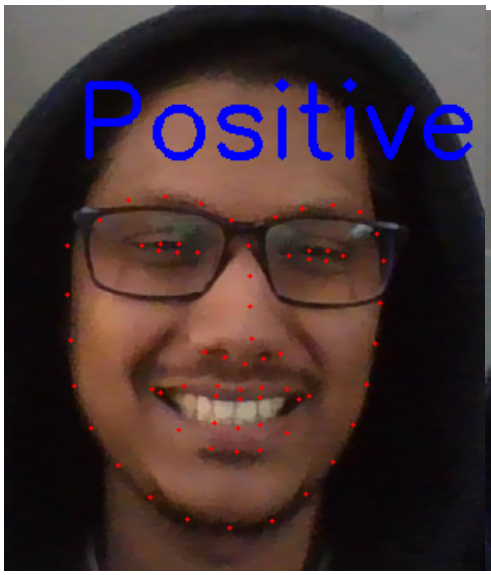
Figure 4.6: Part of the participant's face blocked and still detect the face.

((a)) Intentionally blocking the webcam vertically.

((b)) Intentionally blocking the webcam horizontally.

((c)) Captured Negative Facial Expression.

((d)) Captured Negative Facial Expression on Titled face



((e)) Captured Postive Facial Expression.

((f)) Face detection could not capture the face.

Figure 4.7: Tests conducted using webcam to check performance of both model and face detection

By considering all the limitations of the system, it then pushed to test the dyadic interaction from an online video about 'Therapist Guide to Coronavirus' from [70]. Figure 4.8 and 4.9 show the results of the system while the video is running. These two figures shows the trained model works well but the problem arises on Figure 4.10 and 4.11. The facial expressions changed out of a sudden for one frame or two and then they stabilize afterward.

Figure 4.8: Tested the model in a videocall.

Figure 4.9: Tested the model in a videocall.

Figure 4.10: Tested the model in a videocall.

Figure 4.11: Tested the model in a videocall.

After part of the video was tested, the model automatically generated a Comma Separated Values (csv) file as output. Table 4.7 depicts an overview of the sample output file generated by the model in csv file format. First two columns contain the x and y coordinates, respectively, of each facial landmark on both faces. Each cell in the xy coordinates column consists of 68 points of facial landmarks due to limited space, only the first coordinates are shown for clarity. The third and fourth columns show the recorded emotions or facial expressions of each participant that are set to Negative, Positive and Neutral.

Table 4.7: Sample Output file generated by model.

Face # 1	Face # 2	Emotion # 1	Emotion # 2
[[232, 351], ..]]	[[906, 356],,..]]	Positive	Positive
[[232, 351], ..]]	[[906, 356],..]]	Positive	Positive
[[233, 351],..]]	[[906, 358], ..]]	Positive	Positive
[[232, 353], ..]]	[[908, 359], ..]]	Positive	Positive
[[231, 352], ..]]	[[907, 359], ..]]	Positive	Positive
[[233, 352], ..]]	[[907, 359], ..]]	Positive	Positive
[[234, 352], ..]]	[[907, 356], ..]]	Positive	Positive

For further analysis, the video was tested again and ran for 10 minutes. The Facial Expressions information are used and altered as follows: Positive values are set to 5, Neutral to zero and Negative to -5. These values are then used to draw a graph of Facial Expression/ Emotion versus Frame. The Frame mentioned in both graphs refers to the number frames, as the model detects the facial expressions over each frame.

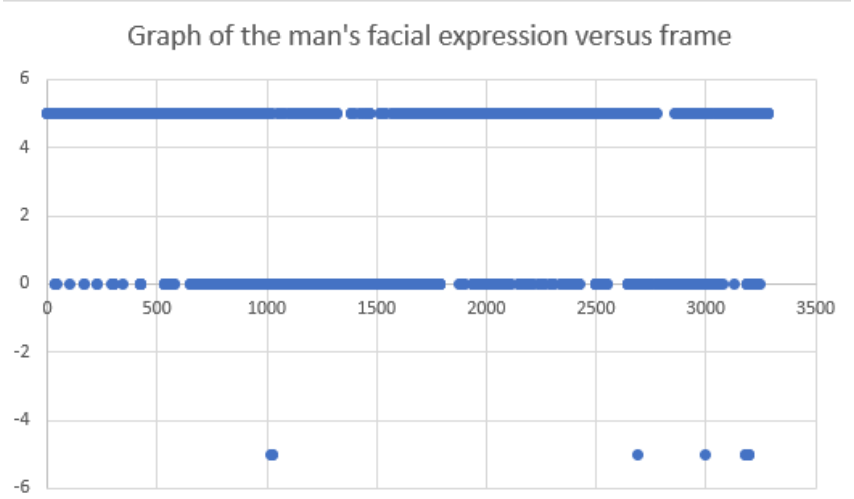


Figure 4.12: A graph of Facial Expression of person 1 (man) versus Frame.

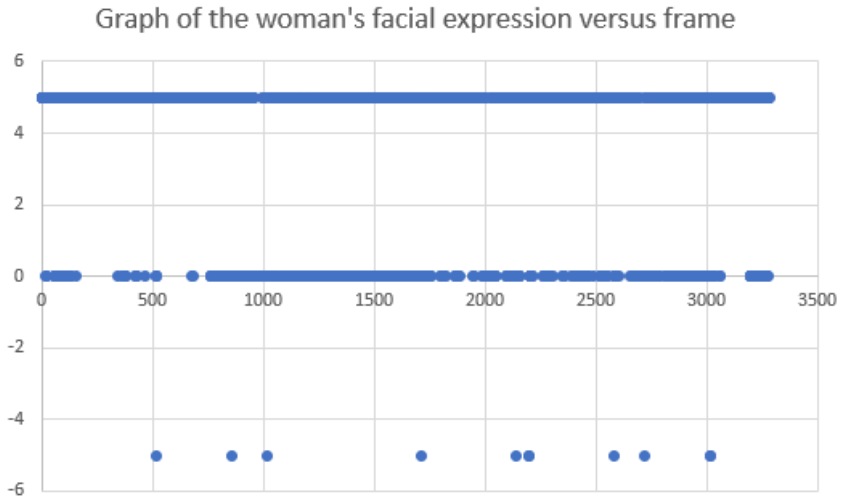
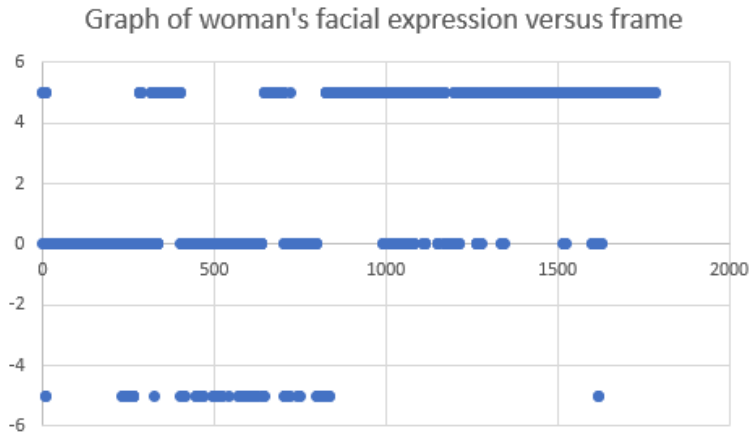


Figure 4.13: A graph of Facial Expression of person 2 (woman) versus Frame.

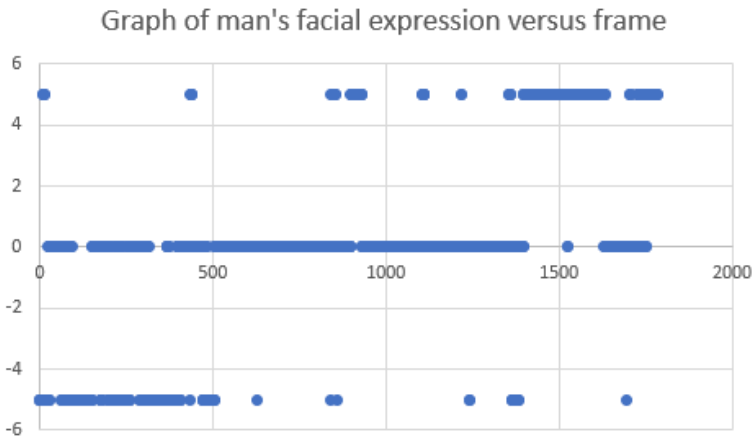
Figure 4.12 and 4.13 illustrate the facial expression of each participant over the number of frames. It can be seen that the facial expressions fluctuated between Positive and Neutral for both participants. There are also some small peaks to the Negative facial expressions. Hence, another video will be used for more understanding.

Figure 4.15 demonstrates an experiment in which a couple have challenged each other on a relationship problem [3]. Although the video includes true facial expressions of both participants as they had both an informal interaction and unhappy moments, only intense moments were used for this experiment. From part (a) to (d), the couple talked about the problem at hand, and then the situation took another turn as they started to share happy moments, which can be represented in part (e) and (f).

From the previous video, the model showed some negative peaks during the interview, and this video showed some positive peaks during the conversation when it was supposed to be negative. The csv file is obtained and evaluated after compiling the video. The evaluation of each participant was separated and a comparison is made for each sample which can be seen in Figure 4.14. Part (a) shows the result of a woman's facial expressions displaying peaks of positive expressions during a sad moment and some peaks of negative expressions during a happy moment. Although the male participant showed that he had retained neutral facial expression for a longer period of time, the overall expression was identical.



((a)) A graph of woman's facial expression plotted against frame.

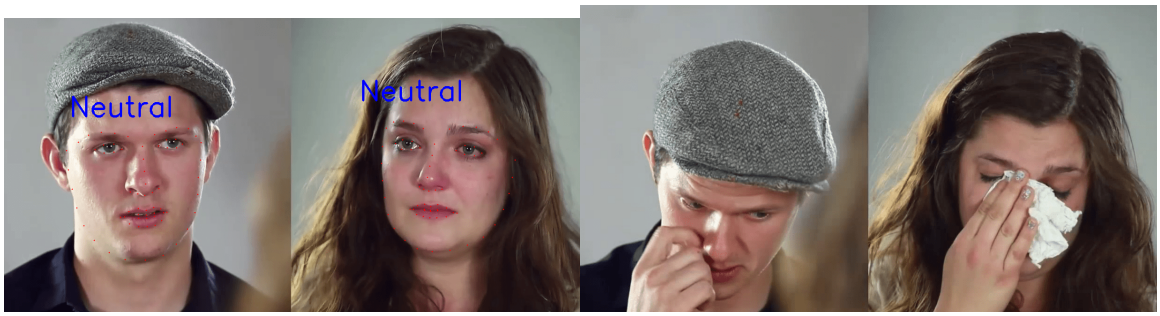


((b)) A graph of man's facial expression plotted against frame.

Figure 4.14: Comparison graph for each participants

((a)) Both participants were showing real negative facial expressions.

((b)) Different facial expressions of both participants.



((c)) Both participants were quiet and looking at each other.

((d)) The model could not detect their faces as their hands were acting as an obstacle.

((e)) Both participants started talking on a different topic.

((f)) Both participants exhibit Positive facial expressions as they were talking on some happy moments.

Figure 4.15: Some major part of the video that showed intense moments and the flow from a sad situation to a happy one.

4.2 Discussion

During the training part, CNNs and mini-Xception model were implemented and the accuracy was compared. A summary of the trained models is presented in Table 4.8 and the data show that the first two CNN models were not sufficiently effective. Although the performance accuracy of Mini-Xception model around 65% is not the perfect accuracy, the model allows improving it by fine tuning while training.

Table 4.8: Summary of the trained DNN models with their validation accuracy.

Model	Validation Accuracy (%)
CNN model using 3 Convolutional block layers and 1 fully-connected block layer	29.79
CNN model using 2 Convolutional block layers and 1 fully-connected block layer	56.25
Mini-Xception model	65.0

One of the challenges of the training was that the model had failed during training process and the entire thesis needed to be restarted. The training of the first CNN model using 3 convolutional block layers and 1 fully-connected block layer turned out to be one of the most time-consuming processes. The convolutional block layer did not take a lot of time to train, but the fully-connected block layer actually did. The larger the fully connected layer, the longer the training time would be.

From the second CNN model, the CNN model with 2 convolution block layers and 1 fully connected block layer had more parameters in the fully connected layer and the accuracy was higher than the previous trained model. This meant that this model took longer time to have better accuracy, which did not improve even after fine-tuning. The second model was tested and the facial expression always remained on 'Sad'.

In contrast, the proposed model (Mini-Xception model) improved with the same setting for fine tuning. The model has first improved to 67%. Due to further fine tuning, it then improved to

80%. After applying the above equations, the accuracy was calculated at 82%, which was due to a reduction in the number of classifications. The model has improved to the point where the output number of classification is reduced from 7 to 3. For example, a question with 7 answers is provided to someone, and a question with 3 answers to choose from is provided to another person. The person with 3 answers undoubtedly has a higher chance of getting the right answer if one has to choose randomly. This model is precise enough to detect the facial expressions that can be used in testing.

From the analysis, it can be understood that a trained Deep Neural Network (DNN) model is capable in a reasonable manner to predict human facial expressions in images. There are some problems that arise while using a video or webcam model, such as the blocking part of the face may not result in a face being detected, tilting the face may result in either no detection or incorrect prediction. In a testing settings these are due to posture and can be corrected by instructing the participants while conducting the video call. However, in natural everyday video call settings such instructing is not possible, so the facial detection should be able not to display an output when part of the facial information is missing.

The proposed model was also used to test two different videos with differently contextualized scenarios. Here, shortly remind how the participants were instructed for the dyadic setting. In the first video, the two participants had a formal conversation with less emotional expressions. The results showed that the two participants' facial expressions varied between Positive and Neutral, with some peaks of Negative expression. In the second video, the two participants had an informal conversation with a wider variety of facial expressions. According to the emotion detection the overall expressions fluctuated between Negative and Neutral in the first 700 frames, and then the expressions fluctuated between Positive and Neutral. There have been some positive peaks in the first 700 frames and some negative peaks thereafter.

At first, these peaks appeared to relate to the wrong prediction of the model, but then, after analyzing the events in the video more closely, it turned out that in these peak moments either one of the participant was talking. So it can be said, that the peaks were due to the facial articulation that occurs while the individual speaks. Using this assumption, the variations in the graphs become easier to understand as they clearly seem to show when the participants in the video recordings are actually talking. The same applies to the fluctuations Neutral & Positive and Negative & Neutral.

The model has shown that it is efficient enough to predict human facial expression in two different conditions, when one person is making facial expression on video image and when in dyadic two person face-to-face interactions. These answers the first two Research Questions of this thesis. Furthermore, the experimental results showed a positive correlation between context and face expression. This answers, therefore, the third Research Question of this thesis.

That said, further experiments should be done to gather more information in different contexts or scenarios in order to develop a broader understanding of different facial expression variations and how such understanding can be used in different HCI applications. For instance, it can help to create interaction situations with human and artificial virtual human in computer games or social robotics.

5 Conclusion and Future Work

5.1 Conclusion

Computer-vision based facial expression recognition (FER) can be used for the determination of momentary human emotional states. One of the most common methods is FACS, which has proven to be successful in a variety of fields. As the Deep Neural Network (DNN) is evolving and the results are becoming much more convincing, the method can be applied to FER. The previous FER work has mainly been based on application of DNN and FACS, while not much has been done on dyadic interactions. Also, only few works on FER seem to take into account the context or situation in which the participants are engaged in conversation during the image analysis. This thesis showed that the DNN model can be used for FER in dyadic interaction between two people within a particular context.

The literature review provided insight into the relationship between human psychology and virtual reality (VR). In fact, the use of FER can make the synthetic environment more immersive and even contribute to the improvement of the experience of interaction between human and a digital character in VR. Using implicit and explicit metrics provided an overview of how people react in a virtual environment, and this information can be used to advance the field. FER-2013 was a good choice among many existing databases compiled in this field as it is readily available, contains enough information to train a DNN model and the images are taken with low resolution in the wild. The related works gave a general overview on similar works that have been done. These works all contributed to the enhancement of Human Computer Interaction (HCI).

The method of research was explained in the section on Methodology whereby the task was divided into two parts. The first part was to train the DNN model while the second part was to test the DNN model. Subsequently, if the accuracy after evaluating the model was not satis-

factory, then the architecture would be modified and remodeled. Once the model had obtained the appropriate accuracy, the test part could be started and the results could be used for further analysis.

By assessing the results of the experiments for the two different videos, this thesis has shown how human facial expressions can be correlated with context. The use of Deep Neural Network (DNN) model was proved to be effective in Facial Expression Recognition (FER) which worked appropriately during the experiments.

5.2 Future Work

This work can be further used to study correlations between facial expressions and contexts. The thesis results also lay groundings for comparing other existing techniques. In doing so, a new model can be created using reinforcement learning as a continuity of this thesis work. On another note, image tracking can be improved. For instance, a depth camera can be used for better face detection and this would allow even three-dimensional (3D) reconstruction of an artificial face as cited in [58]. A depth camera will also allow the implementation of body posture analysis whereby a separate DNN model can be trained. This model would enable the computer to know if a person is covering his face during a dyadic interaction. Since body posture is also part of non-verbal cues, this would help the computer to understand the human expressions based on both the FER and body posture. This thesis work on facial recognition during two person settings is envisioned to have continuations in the future artificial character developments in virtual games and narratives.

6 Acknowledgement

The work has been supported by the EU Mobilitas Pluss grant (MOBTT90) of Dr. Pia Tikka, Enactive Virtuality Lab, Tallinn University (2017-2022).

The author would like to extend his deepest appreciation to the individuals who supported him make this thesis a reality with sincere gratitude. The author wishes to extend his sincere thanks to the following:

His supervisors, Dr. Cagri Ozcinar, Prof. Dr. Pia Tikka and Prof. Dr. Gholamreza Anbarjafari, whose knowledge, consistent guidance, generous time spent and meaningful advice helped him to make this dissertation a success;

His parents whose endless love had given him moral support;

His friends and classmates, Ahmed Mamdouh, Abdelrahman Hisham and Alexandr Syzonyuk, who have continuously assisted him to finish this work morally.

References

[1] *Accuracy, precision, recall f1 score: Interpretation of performance measures*, Accessed: 30-04-2020. [Online]. Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

[2] S. J. G. Ahn, J. Fox, and J. M. Hahm, "Using virtual doppelgängers to increase personal relevance of health risk communication", in *International conference on intelligent virtual agents*, Springer, 2014, pp. 1–12.

[3] *Ali and andrew part 1: Why did you cheat on me? — the and — glamour*, Accessed: 30-04-2020. [Online]. Available: https://www.youtube.com/watch?v=CfdlIMlPmuA.

[4] *Angry woman*, Accessed: 30-04-2020. [Online]. Available: https://www.google.com/imgres?imgurl=https%5C%3A%5C%2F%5C%2Fcdn.mos.cms.futurecdn.net%5C%2FDMUbjq2UjJcG3umGv3Qjjd-320-80.jpeg&imgrefurl=https%5C%3A%5C%2F%5C%2Fwww.livescience.com%5C%2F47688-universal-angry-face-explained.html&tbnid=wED5AOGKpGZ9rM&vet=12ahUKEwjvjc7f-5bpAhVOtqQKHQ2kADQQMygJegUIARDEAg..i&docid=UKfKz6JJ1aTshM&w=320&h=282&q=angry%5C%20face&client=firefox-b-d&ved=2ahUKEwjvjc7f-5bpAhVOtqQKHQ2kADQQMygJegUIARDEAg.

[5] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification", *arXiv preprint arXiv:1710.07557*, 2017.

[6] *Artificial human beings: The amazing examples of robotic humanoids and digital humans*, Accessed: 13-05-2020. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2020/02/17/artificial-human-beings-the-

amazing-examples-of-robotic-humanoids-and-digital-humans/#b21216a51654.

[7] R. Aylett and S. Louchart, "Towards a narrative theory of virtual reality", *Virtual Reality*, vol. 7, no. 1, pp. 2–9, 2003.

[8] R. Aylett, M. Luck, M. Coventry, and C. Al, "Applying artificial intelligence to virtual reality: Intelligent virtual environments", *Applied Artificial Intelligence*, vol. 14, Jan. 2001. DOI: 10.1080/088395100117142.

[9] M. S. Bartlett, P. A. Viola, T. J. Sejnowski, B. A. Golomb, J. Larsen, J. C. Hager, and P. Ekman, "Classifying facial action", in *Advances in neural information processing systems*, 1996, pp. 823–829.

[10] G. Bente, S. Rüggenberg, N. C. Krämer, and F. Eschenburg, "Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations", *Human communication research*, vol. 34, no. 2, pp. 287–318, 2008.

[11] F. Bre, J. Gimenez, and V. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks", *Energy and Buildings*, vol. 158, Nov. 2017. DOI: 10.1016/j.enbuild.2017.11.045.

[12] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social robotics", in *Springer handbook of robotics*, Springer, 2016, pp. 1935–1972.

[13] W. Bricken, "Virtual reality: Directions of growth notes from the siggraph'90 panel", *Virtual Reality: Directions of Growth*, p. 16, 1990.

[14] E. Cambria, "Affective computing and sentiment analysis", *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[15] M. Cavazza, F. Charles, and S. J. Mead, "Interacting with virtual characters in interactive storytelling", in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, 2002, pp. 318–325.

[16] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d database for facial expression analysis and biometric applications", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5117–5126.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[18] J. F. Cohn, "Foundations of human computing: Facial expression and emotion", in *Proceedings of the 8th international conference on Multimodal interfaces*, 2006, pp. 233–238.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 886–893.

[20] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients", *Pattern recognition letters*, vol. 32, no. 12, pp. 1598–1603, 2011.

[21] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0", in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 524–528.

[22] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015", in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 423–426.

[23] *Dlib toolkit*, Accessed: 26-04-2020. [Online]. Available: `http://dlib.net/`.

[24] P. Ekman, *Darwin and facial expression: A century of research in review*. Ishk, 2006.

[25] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[26] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[27] *Face classification and detection*, Accessed: 26-04-2020. [Online]. Available: `https://github.com/oarriaga/face_classification/tree/master`.

[28] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition", *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[29] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *et al.*, "Challenges in representation learning: A report on three machine learning contests", in *International Conference on Neural Information Processing*, Springer, 2013, pp. 117–124.

[30] L. N. Gray and I. Tallman, “A satisfaction balance model of decision making and choice behavior”, *Social Psychology Quarterly*, pp. 146–159, 1984.

[31] D. M. Gross, “Defending the humanities with charles darwin’s the expression of the emotions in man and animals (1872)”, *Critical Inquiry*, vol. 37, no. 1, pp. 34–59, 2010.

[32] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review”, *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[33] N. Haines, Z. Bell, S. Crowell, H. Hahn, D. Kamara, H. McDonough-Caplan, T. Shader, and T. P. Beauchaine, “Using automated computer vision and machine learning to code facial expressions of affect and arousal: Implications for emotion dysregulation research”, *Development and psychopathology*, vol. 31, no. 3, pp. 871–886, 2019.

[34] Y. Huang and S. M. Khan, “Dyadgan: Generating facial expressions in dyadic interactions”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–18.

[35] L. E. Ishii, J. C. Nellis, K. D. Boahene, P. Byrne, and M. Ishii, “The importance and psychology of facial expression”, *Otolaryngologic Clinics of North America*, vol. 51, no. 6, pp. 1011–1017, 2018.

[36] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard, “Understanding and predicting bonding in conversations using thin slices of facial expressions and body language”, in *International Conference on Intelligent Virtual Agents*, Springer, 2016, pp. 64–74.

[37] W. L. Johnson, J. W. Rickel, J. C. Lester, *et al.*, “Animated pedagogical agents: Face-to-face interaction in interactive learning environments”, *International Journal of Artificial intelligence in education*, vol. 11, no. 1, pp. 47–78, 2000.

[38] L. Kaiser, A. N. Gomez, and F. Chollet, “Depthwise separable convolutions for neural machine translation”, *arXiv preprint arXiv:1706.03059*, 2017.

[39] P. Kopp, D. Bradley, T. Beeler, and M. Gross, *Analysis and improvement of facial landmark detection*, Mar. 2019. DOI: 10.13140/RG.2.2.10980.42886.

[40] B. A. Kumar, “The parenting experience.: A soul machines like demonstration (™)”,

[41] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, “Presentation and validation of the radboud faces database”, *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[42] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch, "The effect of avatar realism in immersive social virtual realities", in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, 2017, pp. 1–10.

[43] B. Lee, G. D. Hope, and N. J. Witts, "Could next generation androids get emotionally close?relational closeness' from human dyadic interactions", in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, 2006, pp. 475–479.

[44] S. Li and W. Deng, "Deep facial expression recognition: A survey", *IEEE Transactions on Affective Computing*, 2020.

[45] ——, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition", *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[46] P. N. Lopes, M. A. Brackett, J. B. Nezlek, A. Schütz, I. Sellin, and P. Salovey, "Emotional intelligence and social interaction", *Personality and social psychology bulletin*, vol. 30, no. 8, pp. 1018–1034, 2004.

[47] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, IEEE, 2010, pp. 94–101.

[48] D. Lundqvist, A. Flykt, and A. Ohman, "Karolinska directed emotional faces [database of standardized facial images]", *Psychology Section, Department of Clinical Neuroscience, Karolinska Hospital, S-171*, vol. 76, 1998.

[49] E. Lurie-Luke, "Product and technology innovation: What can biomimicry inspire?", *Biotechnology advances*, vol. 32, no. 8, pp. 1494–1505, 2014.

[50] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets", in *Proceedings Third IEEE international conference on automatic face and gesture recognition*, IEEE, 1998, pp. 200–205.

[51] R. Maurya, *Mauryaritesh/facial-expression-detection*, 2018. [Online]. Available: `https://github.com/MauryaRitesh/Facial-Expression-Detection`.

[52] E. Mendez and E. Marcum, "Deep learning with biomimicry", *Journal of Design and Science*, 2019.

[53] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[54] G. A. Moore, A. L. Hill-Soderlund, C. B. Propper, S. D. Calkins, W. R. Mills-Koonce, and M. J. Cox, "Mother–infant vagal regulation in the face-to-face still-face paradigm is moderated by maternal sensitivity", *Child Development*, vol. 80, no. 1, pp. 209–223, 2009.

[55] G. A. Moore, C. J. Powers, A. J. Bass, J. F. Cohn, C. B. Propper, N. B. Allen, and P. M. Lewinsohn, "Dyadic interaction: Greater than the sum of its parts?", *Infancy*, vol. 18, no. 4, pp. 490–515, 2013.

[56] *Multi-class metrics*, Accessed: 30-04-2020. [Online]. Available: `https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2`.

[57] S. Y. Oh, J. Bailenson, N. Krämer, and B. Li, "Let the avatar brighten your smile: Effects of enhancing facial expressions in virtual environments", *PloS one*, vol. 11, no. 9, 2016.

[58] V. Pereira Pardinho *et al.*, "Volumetric intelligence: A framework for the creation of interactive volumetric captured characters", 2019.

[59] J. S. Pillai and M. Verma, "Grammar of vr storytelling: Analysis of perceptual cues in vr cinema", in *European Conference on Visual Media Production*, 2019, pp. 1–10.

[60] A. du Plessis and C. Broeckhoven, "Looking deep into nature: A review of micro-computed tomography in biomimicry", *Acta biomaterialia*, vol. 85, pp. 27–40, 2019.

[61] T. Raksarikorn and T. Kangkachit, "Facial expression classification using deep extreme inception networks", in *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, 2018, pp. 1–5.

[62] D. P. Salgado, F. R. Martins, T. B. Rodrigues, C. Keighrey, R. Flynn, E. L. M. Naves, and N. Murray, "A qoe assessment method based on eda, heart rate and eeg of a virtual reality assistive technology system", in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 517–520.

[63] S. I. Serengil, *Facial expression recognition with keras - sefik ilkin serengil*, 2020. [Online]. Available: `http://sefiks.com/2018/01/01/facial-expression-recognition-with-keras/`.

[64] W. R. Sherman and A. B. Craig, *Understanding virtual reality: Interface, application, and design*. Elsevier, 2002.

[65] F. Silva, M. Sanz, J. Seixas, E. Solano, and Y. Omar, "Perceptrons from memristors", *Neural Networks*, vol. 122, p. 273, Feb. 2020. DOI: 10.1016/j.neunet.2019.10.013.

[66] J. M. Susskind, A. K. Anderson, and G. E. Hinton, "The toronto face database", *Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep*, vol. 3, 2010.

[67] K. Tanaka, S. Onoue, H. Nakanishi, and H. Ishiguro, "Motion is enough: How real-time avatars improve distant communication", in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, 2013, pp. 465–472.

[68] J. Tao and T. Tan, "Affective computing: A review", in *International Conference on Affective computing and intelligent interaction*, Springer, 2005, pp. 981–995.

[69] *The top 12 social companion robots*, Accessed: 13-05-2020. [Online]. Available: https://medicalfuturist.com/the-top-12-social-companion-robots/.

[70] *Therapist guide to coronavirus: Teletherapy and covid 19*, Accessed: 30-04-2020. [Online]. Available: https://www.youtube.com/watch?v=vUi1_hNP_E0.

[71] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the mmi facial expression database", in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, Paris, France, 2010, p. 65.

[72] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater, "Building expression into virtual characters", 2006.

[73] J.-N. Voigt-Antons, E. Lehtonen, A. P. Palacios, D. Ali, T. Kojić, and S. Möller, "Comparing emotional states induced by 360$\Gamma\{\backslash$circ$\}videosviahead - mounteddisplayandcomputerscreen$", *arXiv preprint arXiv:2004.01532*, 2020.

[74] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review", *Computational intelligence and neuroscience*, vol. 2018, 2018.

[75] *Woman with neutral emotion*, Accessed: 30-04-2020. [Online]. Available: https://encrypted-tbn0.gstatic.com/images?q=tbn%5C%3AANd9GcSRZ8zf_IiDR-CFjdggDmMUmSgNhLeqVfObYk3apX9eElt3ex08&usqp=CAU.

[76] A. S. Won, J. N. Bailenson, and J. H. Janssen, “Automatic detection of nonverbal behavior predicts learning in dyadic interactions”, *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 112–125, 2014.

[77] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, and A. High-resolution, “3d dynamic facial expression database, ieee inter”, in *Conf. on Automatic Face and Gesture Recognition, Amsterdam, the Netherlands*, 2008.

[78] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3d facial expression database for facial behavior research”, in *7th international conference on automatic face and gesture recognition (FGR06)*, IEEE, 2006, pp. 211–216.

[79] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “From facial expression recognition to interpersonal relation prediction”, *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

II. Licence

Non-Exclusive licence to reproduce thesis and make thesis public

I, **Abdallah Hussein Sham**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until the expiry of the term of validity of the copyright.

   1.2 make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpacedigital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows,by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

   of my thesis

   "Facial Expression Recognition using Neural Network for Dyadic Interaction "

   supervised by Dr. Cagri Ozcinar
   Prof. Pia Tikka
   Prof. Dr. Gholamreza Anbarjafari

2. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 20.05.2020