ASHESI UNIVERSITY

EMOTION RECOGNITION USING IMAGE PROCESSING

UNDERGRADUATE THESIS

B.Sc. Computer Science

QUEEN FRIMPONG

2020

ASHESI UNIVERSITY

EMOTION RECOGNITION USING IMAGE PROCESSING

UNDERGRADUATE THESIS

Undergraduate Thesis submitted to the Department of Computer Science, Ashesi University in partial fulfilment of the requirements for the award of Bachelor of Science degree in Computer Science.

Queen Frimpong

2020

DECLARATION

I hereby declare that this Undergraduate Thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:

……

Candidate's Name:

……
Date:

……

I hereby declare that preparation and presentation of this Undergraduate Thesis were supervised in accordance with the guidelines on supervision of Undergraduate Thesis laid down by Ashesi University.

Supervisor's Signature:

……

Supervisor's Name:

……
Date:

……

Acknowledgements

To the Holy Ghost, who is my Ultimate Inspiration.

To my supervisor, Dr Justice Appati, whose encouragement and academic advice helped me to undertake this research.

To the CSIS Department faculty, whose advice and contribution at the beginning and throughout the final year helped me immensely.

To my parents, who supported me with their words of faith, wisdom and inspiration.

Thank you and God bless you immensely.

Abstract

Emotion recognition is an active field of research that has seen a lot of interest over the past decade. Historically, people's emotions were analysed and determined through human observation and psychological counselling and later evolved to electrophysiological and largely intrusive methods such as Electroencephalography (EEG) because of how complex of a task it is and how extensive its application could be. Currently, with the entrance of machine learning and computer vision-related technologies, computers and robots can now be trained to learn and predict the emotions of human beings either in real-time or with their static facial images. In this research, emotion recognition is explored with respect to its three widely recognised stages: face detection, feature extraction and emotion recognition. At each of these stages, different image processing methods and learning techniques are explored and tested. A Convolutional Neural Network was trained and tested and recorded an accuracy of 50.7%. A Support Vector Machine was also trained and tested and recorded an accuracy of 81.5%. Both classifiers were trained on 7 emotion categories. The results show that it is possible for computers to predict the emotions of humans using image processing techniques and deep learning models.

Keywords:

Convolutional neural network; deep learning; emotion recognition; image processing; machine learning

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Lee and Zheng [20] argued that humans experience rich emotions on the inside most of the time; however, their expressions do not portray these emotions. Also, emotions, according to Cohn [7], are not directly observable but are inferred from expressive behaviour, self-report, physiological indicators, and context. This statement by Cohn implies that human beings possess emotions which are visible and invisible, that facilitate interactions and communication among other human beings.

Azcarate et al. [2] argued that the interactions between humans are majorly through speech, body gestures and or a display of emotions, and the manifestations of these emotions are by visual, vocal and other physiological means. Therefore, a person's emotional state can be expressed, and one of the primary and useful indicators of emotion is the human face [11, 17]. When a person is in pain, or is happy, sad, angry, disgusted (termed as basic emotions) or feeling any emotion whether basic or combined[1], the face is a medium that reflects the innermost feelings of people powerfully [10, 17]. Accordingly, there exists a relationship between facial expressions and a person's real-time emotional state and this emotional state can be captured and detected by artificial and computing systems [10]. Over the years, emotion detection has become much more fascinating [10, 11, 15, 28, 33] because it has progressed from the domain of psychological diagnosis and invasive technologies [6] to affective computing and technologies. Some of the technologies include computer vision, deep learning, artificial intelligence and neural networks [10, 11, 16]. With interest and research in the emotion recognition field steadily rising with functions or applications in healthcare [27, 32], the car industry and humanoid robotics [28], human-to-human interactions [33] and human-to-computer interactions [11] can be improved significantly.

[1] Combined emotions are combinations of basic emotions

As human-to-human interactions and human-to-computer interactions evolve, other application areas also begin to see improvement. These application areas such as driver stress testing, teaching and learning systems, neuro-biofeedback, user profiling and video conferencing also start to see growth [11, 16, 33].

Lee and Zheng [20] attempted to detect invisible human emotion by observing and capturing a subject's haemoglobin concentration changes. The authors were successful in the implementation of this method; however, this remains one of the few works that look at invisible emotion detection. Barros et al. [3] also in a bid to improve the social interaction skills of social robots, proposed a deep neural model for internal and external emotion appraisal. This deep neural model gets as close to the detection of inner emotion in that it creates an emotion perception unit trained to learn and recognise human emotions and behaviour that are beyond mere expressions. Other methods for emotion recognition such as galvanic skin response (GSR) and electromyography (EMG) require substantial funding because of the cost and complexity involved in executing them in real life and how invasive these methods can be. Therefore, this thesis will explore the novel methods used and tested by researchers for emotion recognition and propose a more accurate emotion recognition model which predicts what people are feeling based on their facial emotions.

1.2 Problem Statement

People experience and express different emotions that are conveyed either by the face or by body gestures, speech or some other way. However, sometimes the emotions a person expresses on their face is not immediately captured or appreciated by other people and this may lead to poor communication and interaction. Dagar et al. report that facial expressions contribute 55% to the effect of a spoken message [9]. Thus, it is essential to identify how people are feeling or the emotions that are being expressed at a time. Since emotions play such a vital role in communication and decision-making, humans naturally

will want computers, machines and other emerging technologies to be able to understand the emotional needs of humans to foster functional human-to-computer interactions [30, 32].

1.3 Objective

Literature [18] has shown that it is possible to use an image processing methodology that balances cost and complexity to detect emotions in humans. Thus, the goal of this research is to use image processing and learning techniques to recognise emotions in humans that people cannot immediately tell. If emotions motivate most of our behavioural decisions [30], then the accurate recognition of these emotions that influence people's behaviours can contribute a lot to research. Research seems to suggest that facial expressions provide a more precise prediction of emotions than physiological factors and indicators [39]. In this research, a more effective emotion recognition and classification method that can accurately detect emotions in people without relying on physiological factors will be explored.

1.4 Outline of Methodology

Image processing, according to Schowengerdt and Wang [36], is the "pixel-level processing to calibrate, rectify, and enhance images for interpretation." Thus, to retrieve useful information from an image, image has to be performed. Emotion detection using image processing, computer vision [17] and deep learning [6] has been studied extensively over the years in the computer science field [11, 25]. Image processing for emotion detection can be performed in three stages: face detection, feature extraction and emotion classification [1, 6, 16, 26]. These three steps can be applied on an image or video sequence and some useful information can be obtained after the process is done. The input images needed for image processing can be obtained from open-source databases and datasets that contain 2D and 3D images and video sequences [6]. Some of these databases include the extended Cohn-Kanade Dataset (CK+) [6, 15], the Compound Emotion (CE) database [6],

Binghamton University 3D Facial Expression (BU-3DFE) database [6], Japanese Female Facial Expressions (JAFFE) [4, 5, 11], MMI [6] and the Karolinska Directed Emotional Face (KDEF) [6, 21]. The purpose of this research is to propose and explore a more effective and accurate method for detecting emotions that cannot be immediately identified, the CE database consisting of 5060 images with a representation of African, Asian, Hispanic and Caucasian races may be used [6]. The CE database has images that have minimised facial obstruction and correspond to both basic emotions and compound emotions as well. Additionally, the BP4D-Spontaneous contains video sequences and images which will be an excellent fit for determining facial emotions that are naturally occurring [6], thus are not immediately identified by the natural eye.

In executing the three stages of image processing for emotion detection, some learning algorithms will be needed for pre-processing, which is getting the raw image data into understandable forms, for feature extraction and emotion classification. Some of these learning techniques or algorithms include the Haar Cascades method [1], linear matrix inequality (LMI) optimisation method [23], Gabor filtering [9, 19], histogram of oriented gradients method [27], Local binary pattern (LBP), Nearest Neighbour algorithm [17], the Support Vector Machine (SVM) and Neural Network [9]. For face detection, it mainly has to do with processing the output received from the pre-processing stage and detecting a face and facial regions such as the eye region, mouth region and cheek region. After the detection is completed using for instance, the Haar Cascades method (or Viola-Jones algorithm), the components or features found in these regions can be extracted as spatial and temporal elements [6] that can be analysed using LBP, or a face landmarking method. In Kim et al.'s [19] work doing emotion recognition using frontal face images, the image processing algorithm employed was in three stages namely, image processing stage, facial feature extraction stage and emotion detection stage. The image processing stage consisted of face region extraction and facial component extraction and the algorithms used at this stage were

the fuzzy colour filter and a virtual face model (VFM) based histogram analysis method. However, in this study, other methods such as the Haar classifier and HOG method may be used to attempt to arrive at a higher accuracy rate.

Finally, in the emotion detection and classification stage, the emotion or facial expression that is displayed on the face of the input image is classified according to either six [19], seven or more emotional states that the emotion classifier unit has been trained on [6]. The detection and classification in this stage can be done using the SVM, the Nearest Neighbour algorithm or even a neural network. Moon et al. employed the fuzzy classifier identification algorithm for emotion recognition in their study, and it resulted in the classification of five emotions with an accuracy of 82.7%. This study may make use of a softmax algorithm, which is mostly used in CNN-based approaches [6].

1.5 Justification

Human emotion recognition is necessary for effective communication between humans or between humans and machines [37]. According to Tarnowski et al. [39], effective emotion recognition enables people to communicate non-verbally and identify other people. Regarding communication, people want to be able to communicate, and this communication may not necessarily be in verbal forms. For instance, in intelligent education systems [1, 39], the emotional state of students or participants can be gauged automatically to adjust teaching speeds and content. Also, Moniaga et al. [27] suggest Artificial intelligence games that require persons to play based on their moods. Emotion recognition using facial emotions or expressions, therefore, cuts across many different sectors including healthcare [27], law enforcement [37], education, biological systems [11, 32], gaming [27] and psychological models.

1.6 Organisation of Study

The rest of the research is organised as follows: Chapter 2 details the related work of authors or the literature review with an emphasis on finding gaps to address, Chapter 3 provides information on the outline of the methodologies explored and used by other authors and a recommendation of possible methods to be used in this research, Chapter 4 describes the implementation and testing process of the emotion recognition and classification model. Chapter 4 also describes the experiments after implementation and results after testing while Chapter 5 concludes the research by providing the conclusion, limitations of the study and suggestions for future work.

Chapter 2: Literature Review

2.1 Introduction

Emotion recognition has been a field of great interest and intense research over the past years due to the influence of automation, computer vision, machine learning, artificial intelligence and image processing and its significance to many application areas such as security, education, and healthcare [11, 28, 30, 31, 32]. Emotion recognition can be conducted in many ways including facial expressions, tone of voice, speech, body movements, facial colour [5, 27], invasive technology such as sensors, muscle movements and most recently, involuntary micro-expressions [3, 6]. Dagar et al. [9] argue that relying on only the spoken messages of people is not enough to convey the information humans want to communicate fully; thus, facial expressions are essential. Aside from human beings needing to understand other people better, emotion recognition plays such a vital role in the human-computer interaction space, in behavioural decision-making, in entertainment and not leaving out education. This chapter will explore the methods and inventions used proposed and implemented by authors to show recent advances in the emotion recognition space.

2.2 Review of Machine learning-based Emotion Recognition Methods

Alsadoon et al. [1] proposed an emotion recognition model for distance learning environments to monitor students' real-time learning status. The authors were successful in improving the accuracy of the emotion recognition framework and reducing the average processing time of the proposed model. In 2019, Bendjillali et al. [4] proposed a method for identifying six basic facial expressions through people's emotions. The authors tested their way on two different databases (JAFFE and CK+), which yielded high degrees of accuracy in both cases. Other research works that have made some contribution to the emotion recognition area include a dynamic game balancing system by Moniaga et al. [27] and an

emotion recognition methodology for patients suffering from neurodegenerative disorders by Xefteris et al. [41].

Apart from Lee and Zheng [20] attempting to solve the problem of invisible emotion detection, Barros et al. [3] also designed a deep neural model to try to address internal and external emotion appraisal. The neural model proposed by the authors was able to recognise internal emotions such as micro and macro facial expressions and to describe emotional states based on a perception unit which aided in improving the accuracy of the entire system. This model, however, was built to be integrated with social robots which will learn human emotional behaviour over some time. Halder et al. [13] proposed a prototype system which automatically detects emotions displayed on a face in three stages: face detection, feature extraction and emotion classification. The authors, however, make use of two different methods (feature invariant approach based on skin colour and SNoW classifier using local SMQT features) for face detection and the Harris corner point detection algorithm for obtaining the features from the eye, eyebrow and mouth regions [13]. For emotion classification, the authors made use of 525 images from the Facial expressions and emotions database to train a neural network for classifying the detected emotions. A naïve bias classifier was used while the results are still being compiled for publication. Once again, Dagar et al. [9] designed an automated emotion detection model using facial expressions from live video streams. The model, the authors proposed consists of a frame extraction and face detection using the Gabor method; learning phase using Principal Component Analysis (PCA) method and the pattern classification stage where K-mean clustering was used. To test and train the proposed model, the authors make use of 20 images in the JAFFE database while using Overlapping principal components and Minimum distance as measures for testing the detected emotions on the images from the database. A limitation identified in this study is the small number of images used for training the model the authors proposed. This resulted in a low emotion recognition accuracy rate.

2.3 Review of Deep learning-based Emotion Recognition Methods

There has been an impressive amount of work done by researchers on facial expression or emotion recognition [1, 11, 28, 30, 31, 32]. Some of these feats include a machine-learning emotion classifier built by Benitez-Quiroz et al. [5] to identify visible facial colour changes corresponding with a particular emotion category. Using the linear discriminant analysis (LDA) algorithm, the authors arrived at a classification accuracy of 92.93% [5]. This study was limited in that it focused on the communication of popular emotion categories, controlled or acted and widely used by researchers in the facial expression recognition space. It does not account for unidentifiable facial expressions, emotions that cannot be described and occur naturally; in other words, involuntary emotions.

Chul [6], in his work, studied both conventional facial emotion recognition (FER) approaches and deep learning-based approaches in an attempt to compare their performance. Chul concluded that deep learning approaches, specifically the convolutional neural network outperforms the conventional approach. However, in Kanjo et al.'s [15] work, a hybrid deep learning approach for emotion classification using Convolutional Neural Network and Long Short-term Memory Recurrent Neural Network (CNN-LSTM) is proposed. The neural networks employed in Kanjo et al.'s approach are modelled on smart device sensor input showing that the adoption of deep learning approaches is effective in emotion recognition and classification. The model was compared to other traditional methods which have the characteristic of not being efficient enough to learn and classify patterns in multimodal datasets and arrived at an accuracy of 95% [15]. Vyas et al. [40] also presented a survey on detecting and classifying facial expressions using Convolutional Neural Networks. In this survey, the researchers compare various CNN-based facial emotion recognition methods while delving deep into the background and usage of CNNs.

After the research, their claims support the statement that CNN-based approaches are more suited for emotion classification.

Similar to Vyas et al. [40], emotional state classification was conducted by Tarnowski et al. [39] to identify the emotions that can be detected using a 3-NN classifier and MLP neural network based on facial expressions. The classifier and neural network were made to learn action units as input, and the outputs were the emotional states such as anger, happiness, and surprise. From the research, the researchers found out that emotion recognition for multiple users is much more useful than doing it for an individual user. The researchers had classifier accuracies of 95.5% and 75.9% for the 3-NN classifier and MLP neural network, respectively [39]. For facial emotion or expression recognition and classification, deep learning-based methods are more appropriate because of their efficiency in terms of computation and recognition capacity [6, 10]. Thus, Das and Chakrabarty [11] compared the performance of three different models of deep neural networks and reported on the model that gives the highest accuracy for recognising emotion. In the course of the study, the authors were able to detect the presence of four human emotion classes (happy, angry, neutral, sad) with significantly high accuracies. The authors were also able to compare the recognition accuracy of the Restricted Boltzmann Machine (RBM), the Deep Belief Networks (DBN) and the Stacked Autoencoder with SoftMax function (SAE+SM) to state that the SAE+SM model can learn and recognise facial emotions with an accuracy of 99.68% making it suitable for feature extraction and feature classification as well [11]. The limitation identified in this survey study is its inability to explain why the authors chose the three deep networks that were chosen and why these methods are relevant and should even be considered at all in emotion recognition.

Furthermore, Dar et al. [10] proposed a method that uses a CNN to learn features and to classify the emotions according to 22 emotion categories. The interesting aspect of this study by Dar et al. [10] is that the classification by the CNN resulted in 22 emotions: 7

basic emotions and 15 compound emotions. The authors begin with pre-processing, move on to feature extraction and normalisation and finally, expression classification using the Softmax classifier. The model was trained using images from the Dataset of Basic Emotions with accuracies of 67.62% for basic emotion detection and 33% for compound emotion.

2.4 Review of Deep learning-based methods for Empathy Prediction

Unlike other authors who experimented in emotion recognition and variations of it, Yin et al. [42] proposed a deep multimodal network for predicting empathy levels of a person engaged in a two-way conversation. Although the inputs of the deep system are facial images, audio signals and time stamps are also employed in predicting the empathy levels. Their multimodal network outperformed the reference method by 133%, meaning that the method proposed is useful in predicting empathy levels using a listener's facial image, audio signals and time stamps. One exciting thing in Yin et al.'s work is the data that was used to train the deep network. While working with a dataset that was picked initially, the authors later decided to use a subset of the data represented in the dataset [42]. This detailed data election contributed significantly to the improvement of their model.

2.5 The AFFDEX SDK

Finally, McDuff et al. [24] created the AFFDEX SDK, a real-time multi-face expression recognition system, to detect and code the emotions displayed on the faces of multiple human subjects. This system was a solution in response to the time-consuming nature of manually coded facial action units. The application succeeded in delivering real-time expression descriptions as users engaged with the toolkit. Magdin and Prikler [23] proposed a system that makes use of Affectiva's Software Development Kit (SDK) so that real-time emotion classification is possible and practical. Using a web camera for face detection, the Affectiva SDK for facial landmark detection, a Geometric feature-based approach for feature extraction and a neural network algorithm for emotion classification, the authors

trained and tested images from six different databases. Some of these databases include the CMU/VASC Image Database, Bao Face Database and Yale Face Database. To measure the performance of the proposed solution, the authors subjected the method to evaluation through all the six databases which produced individual success rates for face detection and emotion classification [23]. After the solution using the Affectiva SDK was tested with images from the six databases, the overall average accuracy rate was 84.27%. A limitation of this study involved the solution's inability to detect and classify emotions on images that had a rotation rate higher than 15%.

2.6 Conclusion of Literature Review

In conclusion, the study observed that the different authors in the emotion recognition solution domain have made great strides in exploring the different methods and approaches used in recognising human emotions. Also noted, are the methodologies used by these authors and their overlapping elements. For instance, the fundamental processes that have to be carried out for emotion detection are face detection, feature extraction and emotion classification [1, 6, 17, 26]. Also, some popularly used databases from which images are used for the recognition process include the JAFFE database, the CK+, the CE database and the CMU dataset. Therefore, the contribution of this research is to develop an effective and more accurate method for detecting emotions and classifying them using image processing and supported by learning techniques. This method may take into consideration spontaneous emotions that are felt but not captured visibly and understood on the human face. Using image processing and exploring various learning techniques will help to identify how such emotions can be detected.

Chapter 3: Methodology

3.1 Introduction

In this research, the problem of emotion recognition using image processing and some learning techniques is explored in order to suggest a more accurate and effective method for the emotion recognition process. In deciding on a more precise method for detecting emotions using facial expressions, the recent popularity of deep learning methods has made it possible and much easier [6, 17]. However, the steps that are usually performed during emotion recognition consist of pre-processing (face detection), feature extraction and emotion classification [6, 19, 20, 21]. An overview of the steps is shown in Figure 1 [21]. Some authors [6, 20, 21] argue that these three different steps or processes involved in conventional emotion recognition may make use of different methodologies and techniques such as Gabor filters for feature extraction and Support Vector Machine (SVM) for classification. Recently, methods for emotion recognition based on deep learning approaches have become quite popular and are deemed more time-efficient and accurate [6]. Concerning deep learning-based emotion recognition, Chul [6] recorded that the Convolutional Neural Network (CNN) is ranked as the most popular network model. However, in this section, the three different conventional processes for emotion recognition will be explored with regards to the different methodologies that have been used and are being used in executing them. In Figure 3.1 an overview of the three stages involved in emotion recognition is illustrated with possible examples of the algorithms employed in implementing them.

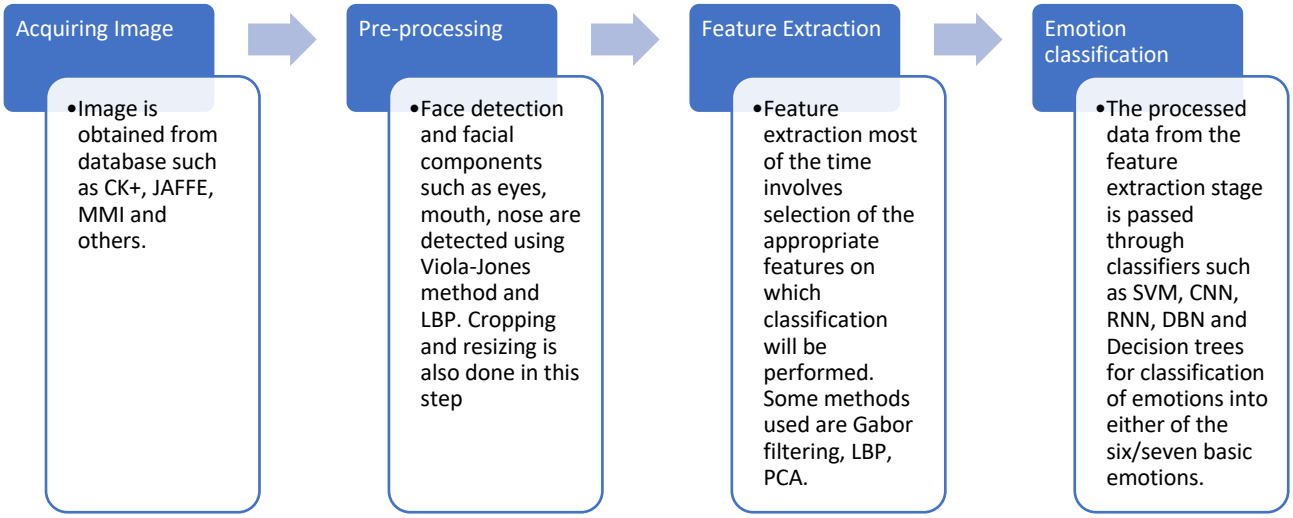

Figure 3.1: Overview of emotion recognition methodology

3.2 Image Databases

As illustrated in Figure 3.1, the emotion recognition process begins with the acquisition of an image on which emotion classification will be done. Image processing involves the processing of the image into a form from which meaningful information can be obtained. For emotion recognition and classification, there must be ample data (images) to train the model that is proposed or built [21]. Thus, these images can be obtained from accessible databases such as the Karolinska Directed Emotional Faces (KDEF), Cohn-Kanade (CK+), Japanese Female Facial Expressions (JAFFE), MMI, FER2013, Compound Emotion (CE) [6, 21, 24].

3.2.1 The Karolinska Directed Emotional Faces (KDEF)

Initially put together for medical and psychological research-related purposes, the KDEF database consists of 4900 images meant to express seven different facial expressions by 70 individuals [6, 21]. The seven expressions include anger, disgust, fear, happiness, sadness, surprise and the neutral feeling. Each image, captured from five different angles, has a size of 562 pixels x 762 pixels.

3.2.2 Extended Cohn-Kanade Dataset (CK+)

According to Li and Deng [21], the CK+ is the most extensively used database with posed images for emotion recognition systems. The CK+ dataset is a detailed dataset with 593 images from 123 subjects. The emotions represented include both posed and spontaneous emotions from a majority of female subjects and remaining male subjects [6]. The feelings include anger, contempt, disgust, fear, happiness, sadness and surprise. Also included in the CK+ dataset is the age range of the images. Each image has a pixel resolution of 640 x 480.

3.2.3 Japanese Female Facial Expressions (JAFFE)

The JAFFE database contains 213 images of size 256 pixels x 256 pixels by 10 Japanese females expressing seven facial emotions. Six of the facial emotions are basic emotions, and the seventh emotion is a neutral facial expression. The six emotions are anger, disgust, fear, happiness, sadness and surprise. Since the total number of image samples are few, it becomes challenging using the JAFFE database for training and testing purposes [21].

3.2.4 MMI

Compared to KDEF and JAFFE, MMI consists of more than 2900 video sequences and static images in total by 75 human subjects. Specifically, the video sequences are 238 in number represented by 28 subjects. The frames in the video sequences have annotations to indicate the phase and emotion being expressed by the captured action units. A challenge with using MMI is the presence of occlusion (such as the presence of glasses and scarves) in the images present [21].

3.2.5 FER2013 Database

The FER2013 database is an extensive database that collects data automatically by using the Google image search API [21]. It contains images each resized to 48 x 48 pixels

and corresponding to seven expression categories- anger, disgust, fear, happiness, sadness, surprise and neutral. The images found in the FER2013 include 28,709 training images, 3589 validation images and 3589 test images [21].

3.2.6 Compound Emotion (CE)

The CE database consists of 5060 images corresponding to 22 emotion categories from 130 female subjects and 100 male subjects. Each image in the CE database has a pixel resolution of 3000 x 4000 and represents a majority of the races around the world.

3.3 Pre-processing

In the pre-processing step, different processes need to be performed on the image in to get it in a form on which emotion classification can be done successfully. The image is obtained from image databases such as the KDEF, FER13 and then some of the following processes are performed on the image- noise removal, resizing and cropping, grayscaling, image normalisation and face alignment [6, 14, 21]. Noise removal involves the removal of unnecessary information on the picture, such as hair, glasses, scarves, which can negatively affect the success rate of emotion recognition. Noise removal can be achieved using various image processing filtering techniques such as Linear filter, Min filter, Max filter, Median filter and Gaussian filter [31]. Resizing is the adjustment of the size of an image to obtain smaller or larger pixel resolutions while cropping involves cutting the edges of the image leaving it as a “tight frame around the face” [21]. Grayscaling converts an RGB image to a grayscale representation in order to reduce the dimensions of the image. OpenCV made it possible to convert RGB images to grayscale using its *cvtColor()* function. Its implementation is shown in Eqn 1. The image annotation process is simply assigning a label or caption to data or in a study such as this, an image. Since the datasets used had already labelled data, image annotation was not necessary.

$$\text{RGB[A] to Gray: } Y \leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \quad (1)$$

3.3.1 Linear filter

The Linear filter is used for reducing random noise and can be represented in Eqn 2. The output pixel is calculated as a sum of the neighbouring input pixels.

$$r(j,k) \; = \; \sum_s^k \; = \; -k \sum_t^k \; = \; -kf(-s,-t)l\,(j+s,k+t) \quad (2)$$

3.3.2 Min filter

The Min filter is a technique that is used to find the darkest points in an image and enhances those points by using the minimum intensity level of the neighbouring input pixels. It is used for removing salt noise and is also known as the 0th percentile filter technique. It is represented in Eqn 3.

$$f^{(x,y)} \; = \min\{g(s,t)\} \quad (3)$$

where

$$(s,t)\epsilon\, Sxy$$

3.3.3 Max filter

The Max filter is the direct opposite of the Min filter. It is used in enhancing the brightest points in an image by replacing the value of the pixel using the maximum intensity level of its neighbourhood pixels. It is used for removing pepper noise and is also known as the 100th percentile filter, and it is represented in Eqn 4.

$$f^{(x,y)} \; = \max\{g(s,t)\} \quad (4)$$

where

$$(s,t)\; \epsilon\, Sxy$$

3.3.4 Median filter

The median filter is a non-linear filter which is mostly used for noise removal [31]. It is a filter which is used for smoothing the image by replacing the pixel value with the average intensity value, which is calculated using Eqn 5.

$$f^{(x,y)} = median\{g(s,t)\} \tag{5}$$

where

$$(s,t)\ \epsilon\ Sxy$$

3.3.5 Gaussian filter

The Gaussian filter is a linear filter that is used in reducing noise or blurring the image. It blurs the edges of the image explicitly and reduces image contrast as well. It is represented in Eqn 6.

$$H[u,v] = \frac{1}{2\pi\sigma^2} exp\ \frac{u^2-v^2}{(2\sigma^2)} \tag{6}$$

3.4 Face detection

Face detection or face registration as referred to by Mollahosseini et al. [26], involves finding the facial region (face), facial components such as cheeks, eyes, nose and mouth and landmark points in the image being used. Melaugh et al. [25] used the Viola-Jones method for face detection, which involves cropping the image and selecting the facial region and detecting the eyes, nose and mouth. While Mollahosseini et al. used a Supervised Descent Method (SDM) for facial landmark extraction as a facial detection step, the authors argue that doing the face detection step well can increase the overall accuracy rate for an emotion recognition model as it did in the case of the authors. To describe a simple way of detecting faces in images, He et al. [14] put together a simplified decision tree which is shown in Figure 3.2. In the decision tree, the authors explain the questions that are asked when face detection is about to occur. The image is first checked to make sure there is a face

or facial region present. If the image does not have a face, it is discarded. However, if the image does not seem to have a face because of uncertainty, the picture is taken through the next step. In the second stage, the image is queried again to check if it has a face present. If it does not have a face present, it is discarded. However, if there seems to be a face, it is accepted and taken through the subsequent stage. This process goes on until the parts of the image that do not form part of the face are removed, and the parts that are the face remain [14].

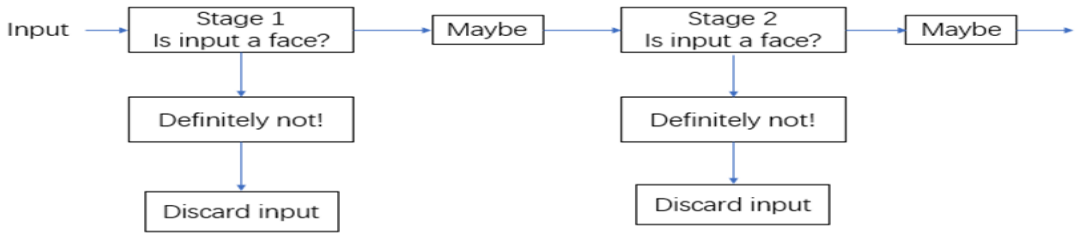

Figure 3.2: Decision tree of the face detection algorithm

3.4.1 Viola-Jones (V-J) Face Detection algorithm

The V-J method is a machine learning-based face detection algorithm also known as the Haar feature-based cascade classifier. It is used for detecting objects in positive and negative images[2]. This algorithm uses a cascade function that is trained on these positive and negative images. It works by cropping an image to a tight frame around the image and leaving the facial features and their relative muscle positions [25]. The V-J method uses the cascade function trained on images that match the goal of emotion recognition and pictures that are used as a differentiating factor for appropriate photos [25].

3.4.2 Supervised Descent Method (SDM)

[2] Positive images contain faces while negative images do not contain faces.

The SDM is a method that uses scale-invariant feature transform (SIFT) features for feature mapping and trains a descent method using linear regression on a training set in order to extract points considered as facial landmarks [26]. The facial landmarks are used to register faces to an average face in an affine transformation and finally, a solid rectangle around the average face is considered as the face region [26].

3.5 Feature Extraction

Feature extraction, according to Dagar et al. [9], defines some attributes or elements which, when associated together, represent a particular facial expression or emotion among the universal six. In determining the features that will be extracted from the detected facial region, Gabor filters, Histogram of Oriented Gradients (HOG), Local Gradient Code (LGC), Principal Component Analysis (PCA), Local Binary Pattern (LBP) and Local Directional Pattern (LDP) can be used [6, 19, 22]. In Dagar et al.'s automated framework for detecting emotions, the Gabor filter for feature extraction was used for two purposes: localization of feature points and for feature vector generation [9]. These two purposes, as claimed by Dagar et al., can be generalised to the feature extraction systems for emotion recognition models. Mollahosseini et al. [26] described the feature extraction stage as consisting of feature selection as well as feature vector generation. Thus, in Melaugh et al.'s [25] paper, the authors made use of a Gabor filter for the transformation of the facial components into shapes and features, HOG for the generation of clear features and PCA for reduction and scaling before classification.

3.5.1 Gabor filters

The Gabor filter is a technique used for analysing changes in lighting and texture in images by highlighting the prominent and desired features. The Gabor filter presents the

orientation of facial features by turning smiles into triangular shapes, as shown in Figure 3.3 [25].

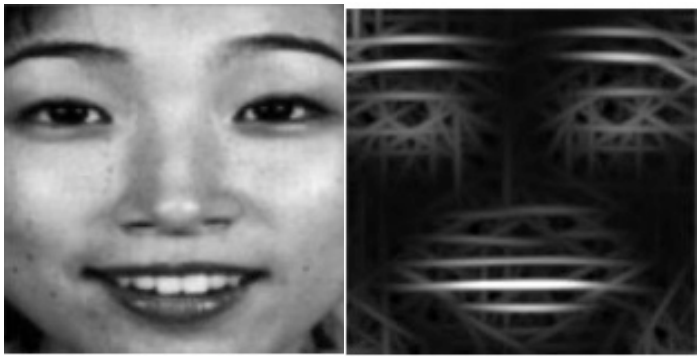

Figure 3.3: Gabor filter transforming facial features in shapes; the smile on the mouth converted into a triangular shape.

The equation of the Gabor filter is shown in Eqn 7.

$$f(x, y, \lambda, \theta, \psi, \sigma, \gamma) = exp(\frac{(xcos\theta + ysin\theta)^2 + \gamma^2 (-xsin\theta + ycos\theta)^2}{2\theta^2}) cos(2\pi \frac{xcos\theta + ysin\theta}{\lambda} + \psi) \quad (7)$$

where

$$\mathrm{x, y} = \text{The coordinates of the image pixel being analysed}$$

$$\theta = \text{The angle of range 0 to } 180°$$

$$\varphi = \text{Phase offset}$$

$$\sigma = \text{Standard deviation}$$

$$\gamma = \text{Filter nature}$$

3.5.2 Principal Component Analysis (PCA)

The Principal Component Analysis method generates the feature vector that will be fed into a classifier to determine the classified emotion. The PCA transforms scaled data to fit a new coordinate system in order to generate feature vectors that can be directly inputted into a classifier [25].

3.5.3 Local Binary Pattern (LBP)

The LBP is a feature extraction method that labels each p equally spaced pixel value within a radius (R), denoted by g_p by "thresholding its values with the central value $g_c$and converting these thresholded values into decimal numbers" as seen in Eqn 7 [33].

$$LBP_{p,R}(X_c, Y_c) = \sum_{p=0}^{p-1} s(g_p - g_c) \quad (8)$$

where

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0\ otherwise \end{cases}$$

3.5.4 Local Directional Pattern (LDP)

In order to get better performance in the presence of variation in illumination and noise, the Local Directional Pattern (LDP) was developed by using Kirsch masks of size 3x3, convolving them with image regions of size 3x3 to get a set of 8 mask values [33]. These mask values are then ranked, and the top three will be assigned with 1 in the 8-bit binary code and the others with 0. The decimal value corresponding to this binary code will be the LDP value for the centre pixel of the selected 3x3 image region [33]. This LDP generated image is divided into blocks to create the LDP feature for the image.

3.5.5 Histogram of Oriented Gradients (HOG)

The HOG method is a feature descriptor method that takes an image or its representation and converts that image into a feature vector that can be fed into a classifier for object detection and other classification tasks. The features talked about in this method are distributions of x and y derivatives (gradients) which are also histograms. Thus, features on a human face are computed by finding the local direction of the gradients of pixels in an image.

3.6 Emotion Classification

The classification step of the entire emotion recognition process seeks to categorise the detected emotion or expression from the feature vector generated using methods such as AdaBoost [6, 15], Random Forest classifier [6], Support Vector Machine (SVM), Restricted Boltzmann Machine (RBM) and CNNs [11, 19]. In Melaugh et al.'s [25] work, the authors report that the SVM works by accepting the transformed features from a PCA (or any other feature vector generator), performs some comparisons against 21 classifiers and classifies the emotion detected in the vector. In Rasamoelina et al.'s [32] research, the classifier that was used for categorising facial expressions from still pictures was a deep learning neural network with artificial neurons and three layers. Dagar et al. [9] used a modified K-mean clustering approach as a method for classifying the emotion detected in random images obtained from the chosen database while Chul [6], reported on deep learning-based CNNs which recognize emotions based on the output of the Softmax algorithm. To be able to arrive at an effective method, some of the classification methods are discussed subsequently.

3.6.1 Restricted Boltzmann Machine (RBM)

The Restricted Boltzmann Machine (RBM) is a classifier that consists of only one visible layer and one hidden layer [11]. However, the visible layer comprises of numerous visible units, and the hidden layer also includes many hidden units with each of the units associated with their unique offsets [11]. A pictorial illustration of the RBM is given in Figure 3.4. In the image in Figure 3.4, h represents the hidden layer and hidden units while x represents the visible layer with its units.

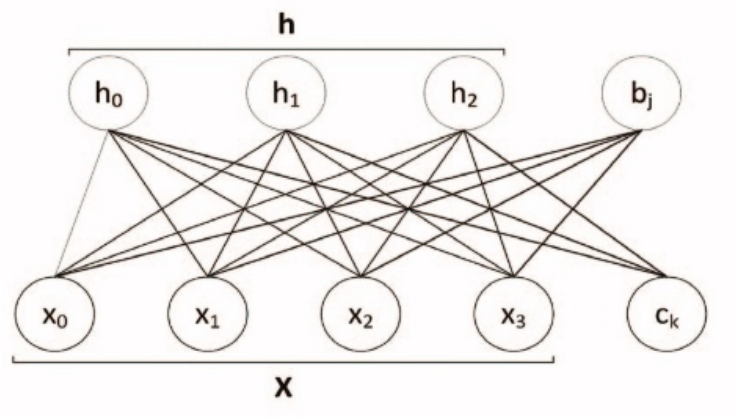

Figure 3.4: A restricted boltzmann machine

3.6.2 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are network models which fall under the deep learning category for emotion recognition. These models mostly learn features from input objects and classify them under different types. In emotion recognition, using CNN-based approaches means that the image is taken through some hidden (or convolution) layers in order to produce a feature map [6]. The feature map is then added to a fully connected neural network working behind the scenes for the detected facial expression or emotion to be classified [11]. In Figure 3.5, an image depicting the workings of the CNN is shown.

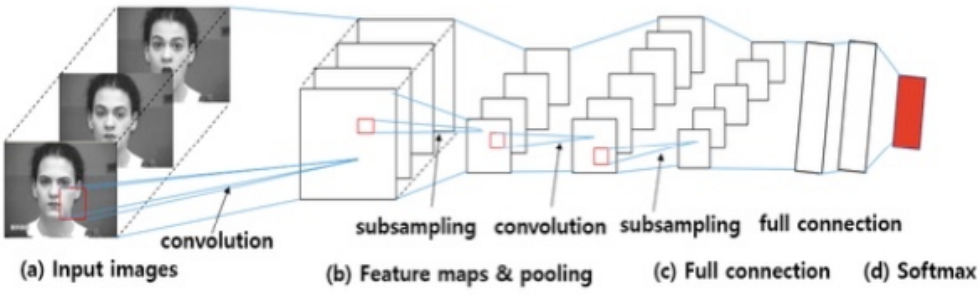

Figure 3.5: Overview of convolutional neural network process

In the image in Figure 3.5, the three images in step (a) are taken through the convolution layer of the CNN, which have embedded filters. After convolution, in step (b),

feature maps (or activation map) are created, and a process called max pooling takes place in order to reduce the spatial resolution and prevent overfitting. Step (c) describes the the fully-connected layers which are responsible for the classification. [11] The feature maps are passed through a classifier or the output layer such as a Softmax classifier so that the face expression or emotion can be recognized based on the output of the classifier [11].

An autoencoder will have to be explained to understand what the Softmax function does. An autoencoder is a non-recurrent neural network that consists of an input layer, hidden layers and an output layer, which is primarily used for encoding and decoding [11]. A Softmax function is a classifier that is used to convert likelihood score generated by an autoencoder into probability values which can be used by the function to determine the detected class (emotion) [11]. In Eqn 9, the probability values that are generated by the Softmax function are shown.

$$s(y_i) = \frac{e^{y_i}}{\sum_j e^{y_i}} \tag{9}$$

3.6.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) sometimes referred to as a Support Vector Classifier (SVC) is a machine learning classifier that can be used in both binary classification and multi-class classification problems. The main idea behind the operation of an SVM is to find a hyperplane (a dividing line) called the Maximum Marginal Hyperplane which best divides a set of spatial data points into either two or more classes. Thus, the SVM is known to be especially efficient in handling binary classification problems. The data points which are closest to the dividing line are referred to as support vectors and these support vectors in essence, determine how accurate or "correct" the classifier can be. The hyperplane is computed using the equation in Eqn 10.

$$|\beta_o + \beta^T x| = 1 \tag{10}$$

where

β = the weight vector

β_o = the bias

x = the training examples closest to the hyperplane

Chapter 4: Experiments and Results

4.1 Introduction

This chapter is dedicated to discussing the implementation of an emotion recognition model consisting of face detection, feature extraction and emotion classification modules that were described and explored in Chapter 3. In addition to details of the implementation, the experiments conducted in training and evaluating said model will also be reviewed in this chapter. Due to the nature of this study, in arriving at a final model with good enough accuracy, different methodologies for the three different stages of the emotion recognition process were studied and tested. Also, for emotion classification, two algorithms namely, a Support Vector Machine and a Convolutional Neural Network were implemented, and their accuracies recorded and compared. Thus, this chapter will also report on the results that were obtained after the training and evaluation.

4.2 Tools and Resources Used

The emotion recognition model was developed in Python and with the following packages: OpenCV, Keras, TensorFlow, Dlib and Scikit-learn. All of these packages are found in Python and freely available for use. These packages provide different machine learning, Optimization and image processing algorithms which can be easily adapted into exiting code or even be used for many classification and regression problems. Also, because of how easy and convenient they are to use and modify, many researchers make use of them in their own studies thus, the accuracies of the algorithms can also be verified.

4.2.1 OpenCV

OpenCV (Open Source Computer Source Computer Vision Library) is a set of open-source computer vision and machine learning libraries. Written natively in C++, it has Python, Java and MATLAB interfaces and has over 500 algorithms that fall in the machine

learning, computer vision and image processing domain. Specifically, it is very useful in image manipulation and processing tasks such as grayscaling of images, image transformation and resizing.

4.2.2 Keras

Keras is a neural networks API written in Python that was developed mainly for fast experimentation and ease-of-use. It can run on top of the TensorFlow, Theano and CNTK platforms and supports both convolutional and recurrent networks (CNNs and RNNs). Keras makes it possible to build, train and test convolutional neural networks which were discussed in Chapter 3. It also provides pre-trained classification models which will be discussed in this section.

4.2.3 TensorFlow

TensorFlow serves as the backend software for Keras where machine learning models can be trained and deployed faster and easier. It provides flexibility for the training of large-scale tasks by making different APIs for different needs and tasks available. It also supports usage on different devices and interfaces such as web, mobile and servers. TensorFlow is open-source and consists of machine learning and computer vision libraries for diverse ML applications and dataflow graph computations. TensorFlow, in this research, is used in training and evaluating the convolutional neural network that was used for emotion classification.

4.2.4 Dlib

Dlib is another open-source machine learning toolkit that is written in C++ and available in Python. It can be used for data analysis tasks, classification and regression problems and image processing. It has algorithms spanning from deep learning algorithms

to numerical and optimization algorithms. This library is used in face detection and feature extraction and has a face landmarking identifier that was useful in extracting the features from the detected faces in the images.

4.2.5 Sci-kit learn

Sci-kit learn is a powerful machine learning software library that is built on NumPy, SciPy and matplotlib meaning it supports classification, regression, prediction and data analysis. It is open-source and especially efficient in providing algorithms for pre-processing tasks, feature extraction and emotion classification. It provided the SVM module which contained the support vector classifier used in this study. Apart from Sci-kit learn being easy to use, it is also capable of comparing machine learning models and algorithms and providing metrics for accuracy and cross-validation.

4.3 Implementation Design

The implementation is divided into three parts; the first being face detection, the second is feature extraction and the final module is emotion classification. For the implementation of the face detection module, two different algorithms were tested out and their individual computation times were recorded in order to verify if they made any significant difference to the accuracy of the entire model. These two algorithms are the Haar feature-based Cascade method (or Viola-Jones algorithm), and the CNN method for face detection. For feature extraction, the Histogram of Oriented Gradients (HOG) feature descriptor method and Dlib's face landmarking predictor method were used. Emotion classification was performed using firstly, a Convolutional neural network adopted from the publicly available VGG-16 model and a Support vector machine from the Scikit-learn library.

The Haar Cascade algorithm is freely available in the OpenCV library and works by first converting the input image into grayscale, then the facial component in the image is detected. After the face in the image is detected, the features from the image are calculated, selected and extracted by subtracting the sum of white pixels in an image from the sum of black pixels. The feature generation and selection are aided by AdaBoost such that each feature is applied on all the images for training. From research, it is shown that when feature extraction and face detection are done well, the classifying algorithm might also perform well. When the face is detected, a rectangular box is drawn around the facial region as shown in Figure 4.1 and prediction is done on the features extracted in this region of interest. After execution, the range of times recorded are between 0.06 seconds and 0.1 seconds.

The HOG method from the Dlib software library is also quite similar to the Haar Cascade method in that it first converts the input image to grayscale format, detects the facial region in the image and overlays a bounding box around the detected face. It is a method that is more similar in definition and function to a feature descriptor. The method works by extracting useful information from an image and converting the image into a feature vector that can be fed to a classification algorithm. These features are distributions (histograms) of directions of x and y derivatives of an image. The range of computation times recorded are between 0.02 and 0.06 seconds.

The CNN face detection method is based on the Max-Margin Object detector method by King [18]. This CNN model for face detection was trained on more than 7,000 images and is also quite similar in function to the HOG method. It does not convert the image to grayscale before detection and the range of computation times recorded are between 0.4 seconds and 0.5 seconds.

The convolutional neural network that was modified and used in this thesis for the emotion classification task was adopted from the convolutional neural network VGG-16 implemented by Simonyan and Zisserman [38]. The VGG-16 is a famous 19-layer deep

CNN model that achieved above 93% accuracy when trained on the ImageNet dataset which contains more than 14 million images. In this thesis, the CNN model trained on 7 emotions is a modification of a pre-trained sequential Convolutional Neural Network (CNN) in Keras with six two-dimensional convolutional layers, three max-pooling layers and finally followed by the Softmax output layer which is flattened before the Softmax activation function is added. Table 4.1 is used to illustrate the architecture of this CNN model.

For emotion classification with the Support vector machine (SVM), first facial landmarking is performed by locating the moveable parts of the face that are responsible for the expression of any emotion. Using Dlib's 68 face landmark detector, the critical facial features pertaining to the expression of an emotion are first located. According to Khan [16], the facial regions or landmarks that Dlib's landmark detector and every other facial landmark method looks out for are the eyebrows (left and right), eyes (left and right), nose, mouth and jaws. Thus, the locations of these facial regions are located as x, y coordinates and stored in an array. After they are located, they are overlaid as blue circles on the regions of interest, and this is shown in Figure 4.2. The algorithm was trained and validated on the CK+ dataset consisting of 407 images belonging to 7 emotion classes.

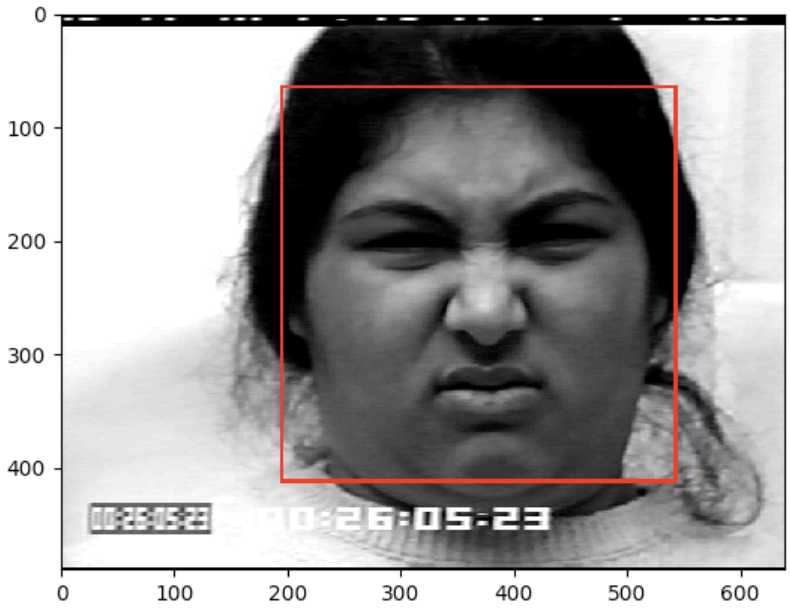

Figure 4.1: Image in face detection stage with rectangular bounding box denoting the face region

Table 4.1: Architecture of modified VGG CNN model

ARCHITECTURE
Convolutional Neural Network
CONV2D-32 RELU
CONV2D-64 RELU MAXPOOL2D
CONV2D-128 RELU MAXPOOL2D
CONV2D-128 RELU MAXPOOL2D
CONV2D-7 RELU
CONV2D-7 RELU
FLATTEN
SOFTMAX

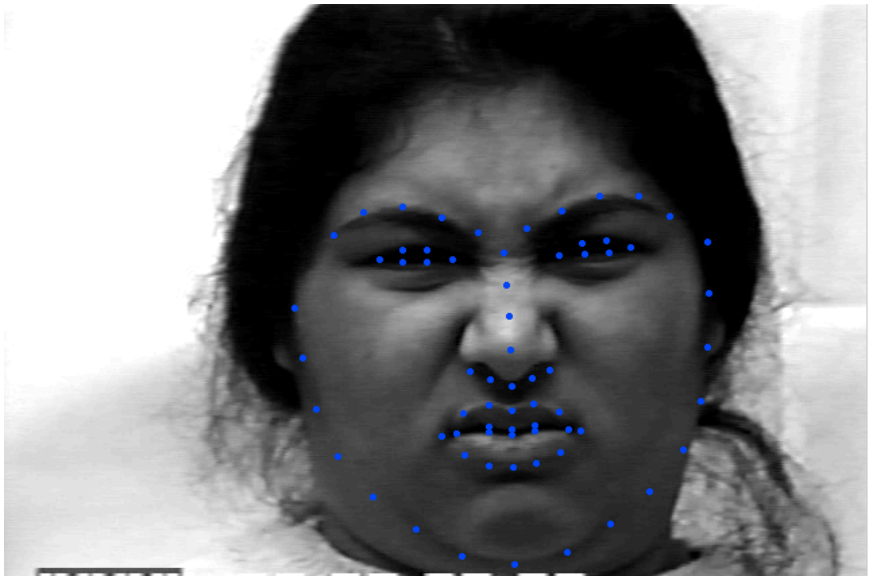

Figure 4.2: Image of blue circles overlaid on facial landmark regions (eyes, eyebrows, nose, mouth and jaw)

4.4 Results of Training

For this research, one traditional CNN model adopted from the VGG-16 was implemented, trained and validated on the FER2013 dataset. Since the training process was such that the model at each epoch could be saved, the CNN model at epoch 20, epoch 100 and epoch 150 were saved for comparison and experimentation reasons. Referring to the model at epoch 20 as model A, the model at epoch 100 as model B and the model at epoch 150 as model C, all three instances of the CNN model had 241,950 trainable parameters. They were trained on 28,709 image samples and validated on 7,178 image samples belonging to seven emotion classes[3]. The FER2013 dataset containing a total of 35,887 images and was split into *train* and *test* modules. Due to the quite small number of parameters the models had; it was trained on a 1.4 GHz Dual-Core Intel Core i5 machine with a 4GB RAM. For model A, it took the CPU approximately 500 seconds per epoch leading to about 2 hours and 7 minutes in total to train model A. Model A provided a training accuracy of 0.3056 and a validation accuracy of 0.34445 as shown in Figure 4.3. For model B, set to initially run for 150 epochs, it was stopped at the 100th epoch when the accuracy had not improved. It recorded a training accuracy of 0.4513 and a validation accuracy of 0.47977 as shown in Figure 4.4, with approximately 500 seconds per epoch making total training time about 13 hours. The final model, C, set to run for 150 epochs with approximately 500 seconds for each epoch for about 20 hours and 8 minutes, obtained a training accuracy of 0.4706 and a validation accuracy of 0.50781 at the 146th epoch illustrated in Figure 4.5. It was observed that the subsequent epochs did not result in an accuracy improvement. Therefore, model C was selected as the final model for emotion classification. However, as seen in Figure 4.5, the validation loss of 1.336 is less than the

[3] The seven classes are: angry: 0, disgusted: 1, fearful: 2, happy: 3, neutral: 4, sad: 5, surprised: 6

training loss of 1.4168 indicating that model is still underfit and may require further training.

The training results of the CNN models can be visualised in Table 4.2.

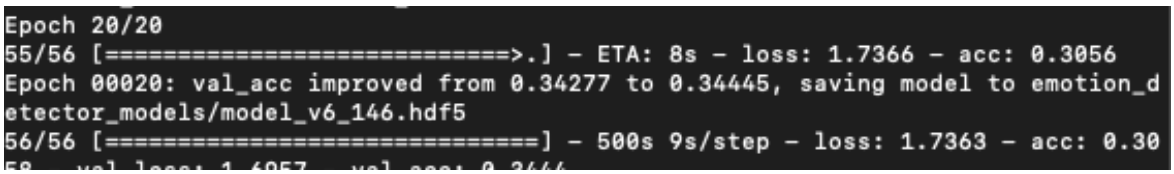

Figure 4.3: Training of CNN model A

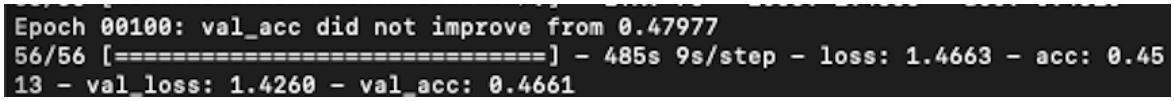

Figure 4.4: Training of CNN model B

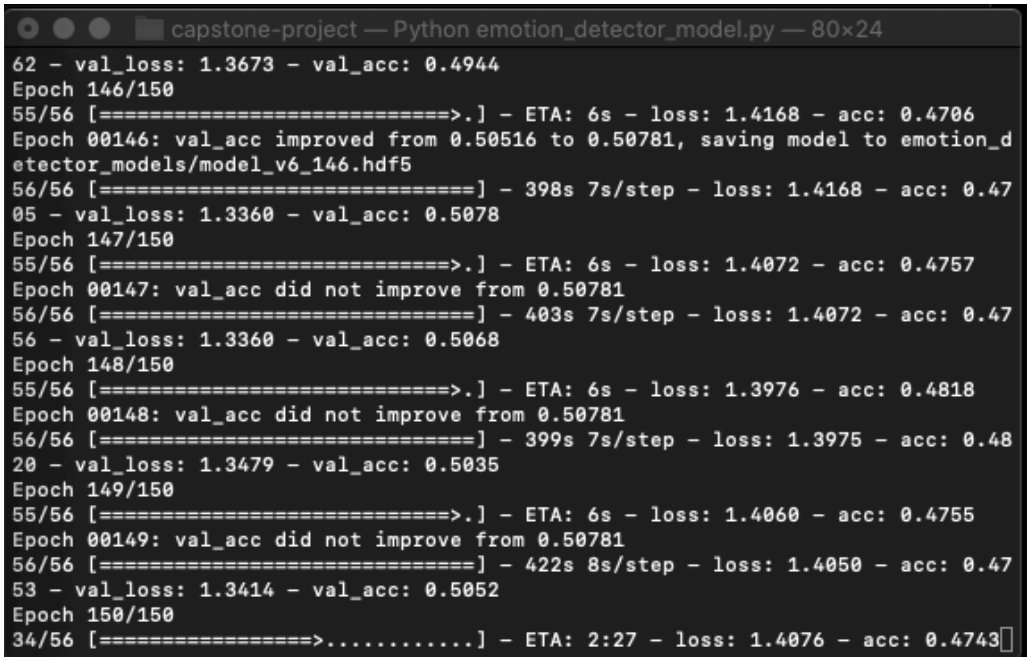

Figure 4.5: Training of CNN model C

Table 4.2: Training and validation accuracies of CNN models

CNN Model	Epochs	Training Accuracy	Validation Accuracy
A	20	0.3056	0.34445
B	100	0.4513	0.47977
C	150	0.4706	0.50781

During the training of the SVM, the publicly available CK+ dataset contained the images and their corresponding labels in separate folders, thus the images had to be sorted and put in folders according to their emotion categories. The code used in splitting the dataset into training and testing sets and for sorting the images was obtained from van Gent [12]. In 10 iterations, the SVM was trained and tested on 542 images belonging to 7 emotion classes and it achieved a classification accuracy of 81.5%. The SVM again was trained and tested for 10 iterations on 407 images for 6 (excluding the neutral emotion) emotion categories and the accuracy recorded is 83.9%. In 20 iterations, the SVM achieved an accuracy of 85.7% on 407 images belonging to 6 emotion categories. For 20 iterations again, the SVM achieved a classification accuracy of 80.3% on 542 images belonging to 7 emotion classes. The statistics reported have been displayed in Table 4.3.

Table 4.3: Classification accuracy of SVM

Number of iterations	Number of Images	Number of Emotion Classes	Classification Accuracy
10	542	7	81.5%
10	407	6	83.9%
20	542	7	80.3%
20	407	6	85.7%

4.5 Results of Testing

The testing dataset used for testing the CNN model (specifically Model C) is the CK+ Dataset. In this dataset, each of the emotion classes has images ranging from a *neutral* expression to a very extreme expression of the emotion class. As shown in Figure 4.6, the anger emotion is eventually expressed as the last image, but it begins from neutral expressions. The CK+ dataset has images according to different subjects thus, the CNN model was tested on 11 sample images from the dataset. The results of the prediction on these images are shown in Figures 4.7 to 4.17 where the predicted emotion is displayed in the console and on the image with the detected face.

Figure 4.6: Sequence of *anger* images in the CK+ Dataset

Figure 4.7: A correct prediction of the *neutral* emotion

Figure 4.8: A correct prediction of the *neutral* emotion

Figure 4.9: A correct prediction of the *neutral* emotion

Figure 4.10: A wrong prediction of the *angry* emotion

Figure 4.11: A wrong prediction of the *angry* emotion

Figure 4.12: A correct prediction of the *angry* emotion

Figure 4.13: A correct prediction of the *angry* emotion

Figure 4.14: A correct prediction of the *angry* emotion

Figure 4.15: A correct prediction of the *angry* emotion

Figure 4.16: A correct prediction of the *angry* emotion

Figure 4.17: A correct prediction of the *angry* emotion

The 11 sample images from the CK+ dataset are testing images that have not been fed to the CNN model i.e. during training and validation. Thus, the accuracy of prediction from the model will be based on how well the model was trained. It can be observed from the images in Figures 4.7 to 4.17 that, the classifier misclassified two emotions as *sad* instead of *angry*. An explanation of this observation could be the classifier's underfit nature.

Chapter 5: Conclusion and Future Work

5.1 Summary

The aim of this research was to explore the problem of identifying and recognizing emotions on human faces in static images using image processing techniques and learning techniques. This research was divided into three modules: face detection, feature extraction and emotion classification, also attempted to improve the accuracy of emotion recognition models by proposing an accurate model that explores the use of different methods in each of the aforementioned modules. In order to classify the emotion represented on the face of an image into *anger, disgust, fear, happiness, sadness, surprise* and *neutral,* a convolutional neural network model was implemented, trained and tested and achieved an accuracy of 50.7% during training on images from the FER13 dataset and testing on images from the CK+ dataset. Also, a support vector machine was trained and tested on the CK+ dataset and attained an accuracy of 81.5%.

Even though the CNN model obtained an accuracy of 50.7%, its performance during evaluation on the test images proves that the model is capable of classifying the emotions of humans. It also suggests that the increasing research and interest in emotion recognition is well founded. Regarding the SVM, even though it obtained 81% - 85% in classification accuracy, its performance on more spontaneous data such as very peak expressions of emotions, it did not perform as well, with classification accuracies ranging from 74% to 76%.

5.2 Limitations of Study

In performing this research, a restriction that may have possibly affected the accuracy or performance of the chosen model is the training time. Due to how long it took to train the CNN model (up to 20 hours), retraining the model even after it had been training for a total of 150 epochs was going to be a cumbersome process. Taking the machine (CPU-

based) it was being trained on into consideration, it explains why training was so slow. After the final model had been trained, the validation loss was still less than the training loss meaning that the model was still underfit. While the training time for the model could have been increased to fix this problem, the cost of actually training it again for 20+ hours outweighed the benefits.

Another limitation includes the datasets that were used. Some images in the FER13 and CK+ datasets were misclassified, and it is likely, this situation could have led to the classifier mistaking one emotion category for another.

5.3 Future Work

According to Pathar et al. [16, 29], CNNs can be used for feature extraction, with the convolution layers serving as feature extractors. Thus, as a suggestion for future work, a CNN can be used to extract the features from the facial images and then an emotion classification method such as Support Vector Machine or a Random Forest classifier can be experimented to reduce the computational cost and still deliver on accuracy. Also, other classification algorithms can be explored in addition to the two that were studied in this thesis.

In evaluation of the model, if a CNN is still being used for emotion classification, other more extensive databases as mentioned in Chapter 3 on Methodology such as the KDEF and JAFFE can be used. When a model is tested with images from multiple databases and the results are recorded, comparisons can be made and the model's recognition ability to generalise based on unseen images can be tested.

Finally, since very little research has been done on detecting spontaneous emotions which are essentially combinations of basic emotions and do not last for long periods on the face, research can be concentrated on this area for applications in psychological institutions and even for examination invigilation

References

[1] Abeer Alsadoon, Parikshit W.C Prasad, Ajaya K. Singh, D. Yang and Amr Elchouemi. 2018. An emotion recognition model based on facial recognition in virtual learning environment. *Procedia Computer Science*, 125, 2-10. DOI: https://doi.org/10.1016/j.procs.2017.12.003

[2] Aitor Azcarate, Felix Hageloh, Koen Sande and Roberto Valenti. 2005. Automatic facial emotion recognition. (Jan. 2005), 1-10.

[3] Pablo Barros, Emilia Barakova and Stefan Wermter. 2018. A Deep Neural Model of Emotion Appraisal. arXiv:1808.00252. Retrieved from https://arxiv.org/abs/1808.00252

[4] Ridha I. Bendjillali, Mohammed Beladgham, Khaled Merit and Abdelmalik Taleb-Ahmed. 2019. Improved Facial Expression Recognition Based on DWT Feature for Deep CNN. *Electronics*, 8, 3 (Mar. 2019), 324. DOI: 10.3390/electronics8030324

[5] Carlos Benitez-Quiroz, Ramprakash Srinivasan and Aleix Martínez. 2018. Facial colour is an efficient mechanism to visually transmit emotion. *Proc. Natl. Acad. Sci USA,* 115, 14 (April 2018), 3581-3586. DOI: 10.1073/pnas.1716084115

[6] Byoung Chul. 2018. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors,* 18, 2 (Jan. 2018), 401. DOI: 10.3390/s18020401

[7] Jeffrey F. Cohn. 2006. Foundations of human computing: facial expression and emotion. *ICMI'06 Proceedings of the 8th International Conference on Multimodal Interfaces* (Nov. 2006), 233-238. DOI: https://dx.doi.org/10.1145/1180995.1181043

[8] Enrique Correa, Arnoud Jonker, Michael Ozo and Rob Stolk. 2016. Emotion Recognition using Deep Convolutional Neural Networks. *TU Delft IN4015*, 1-12. Retrieved from https://github.com/atulapra/Emotion-detection/blob/master/ResearchPaper.pdf

[9] Debishree Dagar, Abir Hudait, Kumar Tripathy and M. N. Das. 2016. Automatic Emotion Detection Model from Facial Expression. *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).*

[10] Aymun S. Dar, Sheraz Naseer, Aihtsham Ali, Ishmal Rauf and Muhammad Ahsan. 2018. Facial emotion detection through deep covolutional neural networks. *VAWKUM Transactions on Computer Sciences,* 15, 3 (Dec. 2018), 113-120.

[11] Deepjoy Das and Alok Chakrabarty. 2016. Emotion recognition from face dataset using deep neural nets. *2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA),* 1-6. DOI: 10.1109/INISTA.2016.7571861

[12] Paul van Gent. 2016. Emotion Recognition Using Facial Landmarks, Python, DLib and OpenCV. Retrieved from: http://www.paulvangent.com/2016/08/05/emotion-recognition-using-facial-landmarks/

[13] Rituparna Halder, Sushmit Sengupta, Arnab Pal, Sudipta Ghosh and Debashish Kundu. 2016. Real Time Facial Emotion Recognition based on Image Processing and Machine Learning. *International Journal of Computer Applications,* 139, 11 (Apr. 2016), 16-19.

[14] Zhihao He, Tian Jin, Amlan Basu, John Soraghan, Gaetano Di Caterina and Lykourgos Petropoulakis. 2019. Human Emotion Recognition in Video Using Subtraction Pre-Processing. *2019 Association for Computing Machinery.* DOI: https://doi.org/10.1145/3318299.3318321

[15] Eiman Kanjo, Eman M.G. Younis and Chee S. Ang. 2019. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49, 46-56. DOI: https://doi.org/10.1016/j.inffus.2018.09.001

[16] Fuzail Khan. 2018. Facial Expression Recognition using Facial Landmark Detection and Feature Extraction via Neural Networks. arXiv:1812.04510v2. Retrieved from https://arxiv.org/abs/1812.04510

[17]Nidhi N. Khatri, Zankhana H. Shah and Samip A. Patel. 2014. Facial Expression Recognition: A Survey. *(IJCSIT) International Journal of Computer Science and Information Technologies*, 5, 1, 149-152. DOI: 10.1.1.444.2743

[18]Davis E. King. 2015. Max-Margin Object Detection. arXiv:1502.00046. Retrieved from https://arxiv.org/abs/1502.00046

[19]Moon Hwan Kim, Young Hoon Joo and Jin Bae Park. 2015. Emotion Detection Algorithm Using Frontal Face Image. *ICCAS2005.* Retrieved from https://www.researchgate.net/publication/228870902_Emotion_Detection_Algorithm_Using_Frontal_Face_Image?enrichId=rgreq-6868b037a4467b04c7e23e064d3c7475-XXX&enrichSource=Y292ZXJQYWdlOzIyODg3MDkwMjtBUzo5OTc0ODY2OTI5NjY0M0AxNDAwNzkzMzQ0OTMx&el=1_x_3&_esc=publicationCoverPdf

[20] Kang Lee and Pu Zheng. 2016. System and method for detecting invisible human emotion. (April 2016). U.S. Patent Application No. 14/868,601, Filed September 29th., 2016, Issued April 7th, 2016.

[21]Shan Li and Weihong Deng. 2018. Deep Facial Expression Recognition: A Survey. arXiv: 1804.08348v2. Retrieved from https://arxiv.org/abs/1804.08348v2

[22]Leh Luoh, Chih-Chang Huang and Hsueh-Yen Liu. 2010. Image processing based emotion recognition. *2010 International Coriference on System Science and Engineering,* 491-494.

[23]M. Magdin and F. Prikler. 2017. Real Time Facial Expression Recognition Using Webcam and SDK Affectiva. *International Journal of Interactive Multimedia and Artificial Intelligence,* 5, 1(Nov. 2017), 7-15. DOI: 10.9781/ijimai.2017.11.002

[24]Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot and Rana el Kaliouby. 2016. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. Proceedings of the 2016 CHI Conference Extended

Abstracts on Human Factors in Computing Systems, ACM (May 2016), 3723–3726. DOI: https://doi.org/10.1145/2851581.2890247

[25] Ryan Melaugh, Nazmul Siddique, Sonya Coleman, and Pratheepan Yogarajah. 2017. Gabor and HOG approach to facial emotion recognition. *Irish Machine Vision and Image Processing*, 1-8.

[26] Ali Mollahosseini, David Chan and Mohammad Mahoor. 2016. Going deeper in facial expression recognition using deep neural networks. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1-10. DOI: 10.1109/WACV.2016.7477450

[27] Jurike V. Moniaga, Andry Chowanda, Agus Prima, Oscar and Dimas Tri Rizqi. 2018. Facial Expression Recognition as Dynamic Game Balancing System. *Procedia Computer Science,* 135, 361-368. DOI: 10.1016/j.procs.2018.08.185

[28] Francesca Nonis, Nicole Dagnes, Federica Marcolin and Enrico Vezzetti. 2019. 3D Approaches and Challenges in Facial Expression Recognition Algorithms- A Literature Review. *Applied Sciences,* 9, 3904(Sept. 2019), 1-33. DOI: 10.3390/app9183904

[29] Rohit Pathar, Abhishek Adivarekar, Arti Mishra and Anushree Deshmukh. 2019. Human Emotion Recognition using Convolutional Neural Network in Real Time. *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 1-7. DOI: 10.1109/ICIICT1.2019.8741491

[30] Luis Antonio Beltrán Prieto and Zuzana Komínkova-Oplatková. 2017. A performance comparison of two emotion-recognition implementations using OpenCV and Cognitive Services API. *CSCC 2017,* 125, 02067, 1-5. DOI: 10.1051/matecconf/201712502067

[31] Sudha Rani and Nageshwar Rao. 2018. A Comparative Study of Various Noise Removal Techniques Using Filters. *Research & Reviews: Journal of Engineering and Technology, 7*, 2 (Mar. 2018), 47-52. Retrieved from https://pdfs.semanticscholar.org/710c/61387615e5bd720864e22833b71c7210af62.pdf

[32]David Rasamoelina, Fouzia Adjailia and Peter Sincak. 2019. Deep Convolutional Neural Network For Robust Facial Emotion Recognition. *IEEE,* 1-6.

[33]Rajesh Reghunadhan, K.M Pooja and Jyoti Kumari. 2015. Facial Expression Recognition: A Survey. *Procedia Computer Science*, 58, 486-491. DOI: https://doi.org/10.1016/j.procs.2015.08.011

[34]Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang and Jing He. 2019. A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. *Sensors,* 19, 1863 (Apr. 2019), 1-27. DOI: 10.3390/s19081863

[35]Minakshee Sarma and Kaustubh Bhattacharyya. 2016. Facial expression based emotion detection, A Review. *ABDU-Journal of Engineering Technology, 4*, 1, 201-205.

[36]Robert A. Schowengerdt and Han-lung Wang. 1989. A General-Purpose Expert System for Image Processing. *Photogrammetric Engineering & Remote Sensing*, 55, 9 (Sept. 1989), 1277-1284.

[37]Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu and Xinyi Yang. 2018. A Review of Emotion Recognition using Physiological Signals. *Sensors*, 18*,* 2074 (June 2018), 1-41. DOI: 10.3390/s18072074

[38]Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from https://arxiv.org/abs/1409.1556

[39]Pawel Tarnowski, Marcin Kolodziej, Andrzej Majkowski and Remigiusz J. Rak. 2017. Emotion recognition using facial expressions. *Procedia Computer Science*, 108 (June 2017), 1175-1184. DOI: https://doi.org/10.1016/j.procs.2017.05.025

[40]Ankit S. Vyas, Harshadkumar B. Prajapati and Vipul K. Dabhi. 2019. Survey on Face Expression Recognition using CNN. *2019 5th International Conference on Advanced*

Computing & Communication Systems (ICACCS), 102-106. DOI: 10.1109/ICACCS.2019.8728330

[41] Stefanos Xefteris, Nikos Doulamis, Vassiliki Andronikou, Theodora Varvarigou and George Cambourakis. 2016. Behavioural Biometrics in Assisted Living: A methodology for emotion recognition. *Engineering, Technology & Applied Science Research*, 6, 4, 1035-1044. DOI: 10.5281/zenodo.60976

[42] Shi Yin, Yonggan Fu, Can Wang, Runlong Wu, Heyan Ding and Shangfei Wang. 2019. Integrating Facial Images, Speeches and Time for Empathy Prediction. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1-3. DOI: http://dx.doi.org/10.1109/FG.2019.8756621

[43] Zhan Zhang, Liqing Cui, Xiaoqian Liu and Tingshao Zhu. 2016. Emotion detection using Kinect 3D Facial points. IEEE/WIC/ACM International Conference on Web Intelligence, 407-410. DOI: 10.1109/WI.2016.62