

AUTOMATIC FACIAL EXPRESSION
RECOGNITION USING WEIGHTED AVERAGE
ENSEMBLE DEEP LEARNING

WALA EDDINE BOURAHLA

UNIVERSITI KEBANGSAAN MALAYSIA

AUTOMATIC FACIAL EXPRESSION RECOGNITION USING WEIGHTED
AVERAGE ENSEMBLE DEEP LEARNING

WALAEDDINE BOURAHLA

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF
MASTER OF DATA SCIENCE

FACULTY SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2021

PENGESANAN EKSPRESI WAJAH SECARA AUTOMATIK MENGGUNAKAN
PURATA WAJARAN PEMBELAJARAN MENDALAM PURATA ENSEMBEL
BERPEMBERAT

WALAEDDINE BOURAHLA

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH
IJAZAH SARJANA DATA SAINS

FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI
2021

**PERAKUAN TESIS SARJANA / DOKTOR FALSAFAH
(CERTIFICATION OF MASTERS / DOCTORAL THESIS)**

Nama Penuh Pengarang
(Author's Full Name) : Wala Eddine Bourahla

No. Pendaftaran Pelajar
(Student's Registration No.) : P101459 Sesi Akademik
(Academic Session) : September 2019

Tajuk Tesis
(Thesis Title) : Automatic Facial Expression Recognition Using
Weighted Average Ensemble Deep Learning

Merujuk kepada Klausa 4.2 Dasar Harta Intelek Pelajar UKM (Tambahan), tesis adalah hak milik pelajar. Saya mengaku tesis ini sebagai:
(With regard to Clause 4.2 of the UKM Student Intellectual Property Policy (Supplementary), the thesis is the student's property. I hereby declare this thesis as:)

- RAHSIA
(CONFIDENTIAL)** Mengandungi maklumat rahsia di bawah AKTA RAHSIA RASMI 1972
(Consisting of classified information under the OFFICIAL SECRETS ACT 1972)
- TERHAD
(RESTRICTED)** Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan
(Consisting of RESTRICTED information which has been determined by the organisation/body where the research was conducted)
- AKSES
TERBUKA
/TIDAK TERHAD
(OPEN ACCESS/
NON-
RESTRICTED)** Saya membenarkan tesis ini diterbitkan secara akses terbuka, teks penuh atau dibuat salinan untuk tujuan pengajian, pembelajaran, penyelidikan sahaja.
(I allow this thesis to be published through open access, full text or copied for study, learning and research purposes only.)

Bagi kategori Akses Terbuka/Tidak Terhad, saya membenarkan tesis (Sarjana/Doktor Falsafah) ini di simpan di Perpustakaan Universiti Kebangsaan Malaysia (UKM)* dengan syarat-syarat kegunaan seperti berikut:

(For the Open Access/Non-Restricted category, I allow this (Master's/Doctoral) Thesis to be kept in the Universiti Kebangsaan Malaysia (UKM) Library with the following usage conditions:)

1. Perpustakaan UKM mempunyai hak untuk membuat salinan untuk tujuan pengajian, pembelajaran, penyelidikan sahaja.
(UKM Library has the right to reproduce the thesis for study, learning and research purposes only.)
2. Perpustakaan Universiti Kebangsaan Malaysia dibenarkan membuat satu (1) salinan tesis ini untuk tujuan pertukaran antara institusi pengajian tinggi dan mana-mana badan/ agensi kerajaan, tertakluk kepada terma dan syarat.
(UKM Library is allowed to make one (1) copy of this thesis for exchange purpose among higher education institutions and any government body/agency, subject to terms and conditions.)

DISAHKAN OLEH:
(VERIFIED BY:)

ولاء الدين بورعالم



TANDATANGAN PELAJAR
(STUDENT'S SIGNATURE)

TANDATANGAN PENYELIA /
PENERUSI JK SISWAZAH
(SUPERVISOR'S / CHAIRPERSON
SUPERVISION COMMITTEE
SIGNATURE)

197587675

DR. AZIZI ABDULLAH

KAD PENGENALAN /
NO. PASPORT
(IDENTITY CARD/PASSPORT
NO.)

NAMA PENYELIA/
PENERUSI JK SISWAZAH
(SUPERVISOR'S /CHAIRPERSON
SUPERVISION COMMITTEE NAME)

Tarikh/
Date:21 October 2021

Tarikh/
Date:21 October 2021

Pusat Sumber
FTSM

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

26 October 2021

WALA EDDINE
BOURAHLA
P101459

ACKNOWLEDGEMENT

In the name of Allah most gracious and most merciful. Praise be to Allah (SWT) and peace and prayer to be upon the Prophet Muhammad (SAW), his family, his companions, and his followers until the day of judgement.

First of all, Alhamdulillah all praises be upon Allah (SWT) for graciously bestowing me the perseverance to undertake this research. A special thank and deepest appreciation to my supervisors, Dr Azizi Bin Abdullah for his continuous support, encouragement and leadership, and for that, I will be forever grateful. Not forgetting my gratitude goes to all my lecturers which have enabled me to acquire precious knowledge.

Finally, it is my utmost pleasure to dedicate this work to my dear mother, my beloved wife, my family, and my friends who granted me the gift of their unwavering belief in my ability to accomplish this goal. The miracles of their do'a and prayers are the catalyses of my strength to complete this study. Thank you for your support and patience.

Pusat Sumber
FTSM

ABSTRAK

Penyelidikan ini adalah mengenai pendekatan automatik untuk mengesan emosi yang diekspresikan pada wajah. Emosi dan perasaan dinyatakan melalui tingkah laku, tindakan, postur, ekspresi wajah, dan suara. Pelbagai kajian telah dilakukan untuk menentukan hubungan medium dan emosi ini. Walaupun mengenali emosi dari gambar atau video adalah mudah pada mata manusia, namun ianya adalah sangat sukar bagi mesin dan memerlukan pelbagai algoritma pemprosesan gambar untuk pengekstrakan ciri ciri. Oleh itu, kajian yang dikemukakan dalam penyelidikan ini memfokuskan pada peningkatan ketepatan pengenalan ekspresi wajah dengan menerapkan ensemble tiga teknik pembelajaran mendalam untuk mengenali ekspresi wajah dari gambar. Kajian ini mengkaji kaedah untuk menangani masalah ketepatan pengesanan imej berkualiti rendah. Matlamat utama kajian ini adalah untuk meningkatkan ketepatan pengesanan set data ekspresi wajah masa nyata, yang mengandungi gambar dunia nyata yang mencabar. Pengelasan kemudian diuji menggunakan set ujian untuk menilai prestasi pengesanan. Set data Fer2013 digunakan dalam menguji kaedah yang dicadangkan dan kadar pengesanan setinggi 70% dicapai untuk pengesanan pengesanan ekspresi wajah. Ini dapat membantu dalam membuat penilaian yang bermaklumat tentang pengesanan niat, pemasaran penawaran, atau masalah yang berkaitan dengan keselamatan. Hasil eksperimen menunjukkan bahawa semua model yang ditawarkan memberikan hasil yang sama jika dibandingkan dengan pendekatan terkini yang ada untuk ketepatan pengesanan.

ABSTRACT

This research presents an automated approach for recognising the emotion expressed on a face. Emotions and feelings are expressed through behaviours, actions, postures, facial expressions, and voice. Numerous studies have been conducted to determine the link between these channels and emotions. While recognising emotions from pictures or video is easy for the human eye, it is extremely difficult for machines and needs a variety of image processing algorithms for feature extraction. Thus, the work presented in this research focuses on improving facial expression identification accuracy by applying weighted ensemble of three deep learning techniques to recognise facial expressions from images. This study looks into methods for dealing with the issue of recognition accuracy of lower quality images. The main goal of this work is to increase the recognition accuracy of the real-time facial expression dataset, which contains challenging real-world images. The classifiers are then tested using a testing set to assess the recognition performance. Fer2013 dataset is used in testing the proposed method and recognition rates as high as 70% are achieved for the recognition of facial expression recognition. This can assist in making educated judgments about intent detection, marketing of offerings, or security-related issues. The experimental results show that all of the offered models provide the same results when compared to existing state-of-the-art approaches for recognition accuracy

TABLE OF CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRAK		v
ABSTRACT		vi
TABLE OF CONTENTS		vii
LIST OF TABLES		x
LIST OF FIGURES		x
LIST OF ABBREVIATIONS		xi
CHAPTER I	INTRODUCTION	
1.1	Introduction	1
1.2	Problem Statement	2
1.3	Objectives of The Study	3
1.4	Motivation For The Study	3
1.5	Scope of The Study	5
1.6	Significance of The Study	5
1.7	Organization of The Study	6
1.8	Chapter Summary	6
CHAPTER II	LITERATURE REVIEW	
2.1	Introduction	7
2.2	Facial Expression Analysis (FEA) Terminologies	7
	2.2.1 Facial Landmarks	7
	2.2.2 Facial Action Units	8
	2.2.3 Facial Action Coding System	9
	2.2.4 Basic Emotions	10
	2.2.5 Compound Emotions	10
	2.2.6 Micro Expressions	10
2.3	Conventional Facial Expressions Recognition Approaches	10
	2.3.1 Image Preprocessing	11
	2.3.2 Feature Extraction	12
	2.3.3 Expression Classification	18

2.4	Deep Learning-Based FER Approaches	19
2.4.1	Artificial Neural Networks	19
2.4.2	Deep Learning	21
2.4.3	Convolutional Neural Network	22
2.5	Deep Learning based Work of Facial Expression	27
2.6	Datasets	32
2.6.1	FER2013 Dataset	33
2.7	Chapter Summary	34
CHAPTER III METHODOLOGY		
3.1	Introduction	35
3.2	Software Tools	35
3.3.1	Tensorflow	36
3.3.1	Keras	36
3.3	Splitting the Data	37
3.4	Data Augmentation	37
3.5	Models Building Blocks	38
3.5.1	VGG16	38
3.5.2	EmotionVGGNet	41
3.5.3	MiniGoogleNet	43
CHAPTER IV WEIGHTED AVERAGE ENSEMBLE FOR FACIAL EXPRESSION RECOGNITION		
4.1	Introduction	49
4.2	Weighted Average Ensemble	50
4.3	Chapter Summary	51
CHAPTER V RESULTS AND DISCUSSION		
5.1	Introduction	52
5.2	Experimental Setup and Implementation Details	52
5.3	VGG16 Results	52
5.4	EmotionVGGNet Results	54
5.5	MiniGoogleNet Results	56
5.6	Weighted Average Ensemble Results	58
5.6	Significance of Weighted Average Ensemble Model	59
5.6	Automatic Facial Expression Recognition system	60
5.7	Discussion and Summary	62

CHAPTER VII CONCLUSION AND FUTURE WORKS

6.1	Conclusion	64
6.2	Future Works	56
REFERENCES		66

Pusat Sumber
FTSM

LIST OF TABLES

Table No.		Page
Table 2.1	Prototypical AUs Observed in the Category of Basic and Complex Emotions	9
Table 5.1	VGG16 Model Results	53
Table 5.2	Learning Rate Values Used Over Training per Epochs for EmotionVGGNet Architecture	55
Table 5.3	EmotionVGGNet Results	55
Table 5.4	MiniGoogleNet Results	57
Table 5.5	Weighted Average Ensemble Results	58
Table 5.6	Accuracies of Used Models	59
Table 5.6	T-Values / P-Values of Used Models	59

LIST OF FIGURES

Figure No.		Page
Figure 1.1	The Six Universal Expressions of Emotion	1
Figure 1.2	Mona Lisa	4
Figure 2.1	An Illustration of Facial Landmarks (Face Images from the JAFFE Dataset)	8
Figure 2.2	Several AU Instances (images from the CK+ dataset)	8
Figure 2.3	Procedure in Conventional FER Approaches	11
Figure 2.4	Extraction of LBP histogram from a Facial Image	13
Figure 2.5	Applications of Optical Flow-Based Methods on Facial Images	15
Figure 2.6	Feature Points Displacement	17
Figure 2.7	Inner Workings of the Human Brain	20
Figure 2.8	Deep Neural Network	21
Figure 2.9	Architecture of a CNN	22
Figure 2.10	Convolutional Neural Network	24
Figure 2.11	ReLU Activation Function	25
Figure 2.12	Application of ReLU and Max Pooling	25
Figure 2.13	Training With and Without Dropout	26
Figure 2.14	FER2013 Sample Images	33
Figure 2.15	FER2013 Class Distribution	33
Figure 3.1	Framework of Facial Expressions	35
Figure 3.2	Data Augmentation Jitter Distribution	36
Figure 3.3	A Replication of Table 1 from Simonyan and Zisserman	39
Figure 3.4	VGG16 Model Algorithm	40
Figure 3.5	VGG16 Model Layers	41
Figure 3.6	EmotionVGGNet Algorithm	42

Figure 3.7	EmotionVGGNet Layers	43
Figure 3.8	The Inception Module that was First Used	45
Figure 3.9	Convolution Module of Miniception	46
Figure 3.10	Down Sample Module of Miniception	47
Figure 3.11	MiniGoogleNet Model Algorithm	47
Figure 3.12	MiniGoogleNet Model Layers	48
Figure 4.1	An Ensemble of Neural Networks Consists of Multiple Networks	49
Figure 4.2	Weighted Average Ensemble Function	50
Figure 4.3	Weighted Average Ensemble Algorithm	51
Figure 5.1	Training Accuracy / Loss of VGG16	54
Figure 5.3	Training Accuracy / Loss of EmotionVGGNet	56
Figure 5.6	Training Accuracy / Loss of MiniGoogleNet	58
Figure 5.8	Real Time FER Framework	61
Figure 6.9	Application Demo	62

LIST OF ABBREVIATIONS

AFEA	Automatic Facial Expression Analysis
FER	Facial Expression recognition
AU	Action Unit
FACS	Facial action coding system
DCT	Discrete Cosine Transform
FFT	Fast Fourier transform
HCI	Human-Computer Interaction
HCII	Human-Computer Intelligent Interaction
HMM	Hidden Markov Models
kNN	k-Nearest Neighbour
SU	Shape Unit
SVM	Support Vector Machine
HOG	Histogram Oriented Gradients
AAM	Active Appearance Model
LBF	Local Binary Feature
MLP	Multilayer Perceptron
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
CK+	Extended Cohn-Kanade
FL	Facial Landmarks
BE	Basic Emotions
CE	Compound Emotions
ME	Micro Expressions
WAE	Weighted Average Ensemble

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

An emotion is a subjective and private mental and physiological state that includes a variety of behaviours, activities, thoughts, and sensations. The book *The Expression of the Emotions in Man and Animals* by Charles Darwin is where the first research on emotions began. He believed emotions were species-specific rather than culture-specific (Grimm et al. 2005), but Ekman and Friesen defined six emotional expressions as universal in 1969 after detecting a universality among emotions despite cultural differences: happiness, sadness, anger, disgust, surprise, and fear. Figure 1.1 shows the six universal expressions of emotion.



Figure 1.1 The Six Universal Expressions of Emotion
Source: Wallho et al. (2009)

1.2 PROBLEM STATEMENT

Many factors have a role in expressing an individual's feelings. Some of them include posture, voice, facial expressions, conduct, and actions. Face expressions are more important than the other elements since they are more immediately discernible. Humans can recognise the emotions of other individuals with a high degree of accuracy while communicating with them. If it can effectively and successfully integrate previously obtained information in computer science to construct solutions for automated facial expression classification, it can achieve accuracy that is virtually identical to human perceptions.

The primary objective of this work is not only to implement an Automatic Facial Expression Recognition system but to increase the accuracy with which facial expressions can be recognised from low-quality images using weighted average ensemble of three deep learning models. Real world datasets are being gathered and analysed in which images are taken in an uncontrolled open space. Real world datasets provide a variety of challenges, including shifting illumination, occlusion, variations in head posture, and lower-quality images, all of which complicate the recognition process. Several academics are aiming to improve the recognition accuracy of their systems by tackling these challenges with facial expression datasets using Conventional and deep learning methods. An ensemble can get better predictions and achieve better performance than any single contributing model and ensemble reduces the spread or dispersion of the predictions and model performance. "To develop a weighted average ensemble of deep learning models for facial expression recognition to extract unique and distinct features from low level grey scale images in order to achieve higher recognition accuracy when compared to existing state-of-the-art research work." Is the answer of the problem statement of this thesis.

1.3 OBJECTIVES OF THE STUDY

The objectives of the study are:

- a) Study and investigate three implementation of deep learning architectures (the famous VGG16, mini VGG like architecture (in this work its named EmotionVGGNet) and a Mini GoogleNet inspired by ResNet architecture.
- b) Propose a Weighted Average Ensemble of the Three architectures.
- c) Evaluate and compare performance of the Weighted Average Ensemble.

1.4 MOTIVATION FOR THE STUDY

Facial expressions, speech, gestures, physical movement and posture, are all ways in which humans communicate their emotions. System which is capable of automatically recognising human emotions via one or more of these modalities might be useful in a number of applications, including human-computer interaction, video games, robotics, instructional software, animations, automotive safety, as well as affective computing. As a result, the creation of a reliable real-time emotion detection sector is vital, as are the applications. For instance, such a system could enable the creation of more intelligent robots capable of comprehending human emotions.

The emotions shown in the world-famous painting of Mona Lisa have sparked heated controversy in the past (Figure 1.2). According to the British weekly New Scientist (BBC. Mona lisa, 2005), she is a mix of many different emotions, with 83 percent being joyful, 9 percent disgusted, 6 percent afraid, and 2 percent angry.

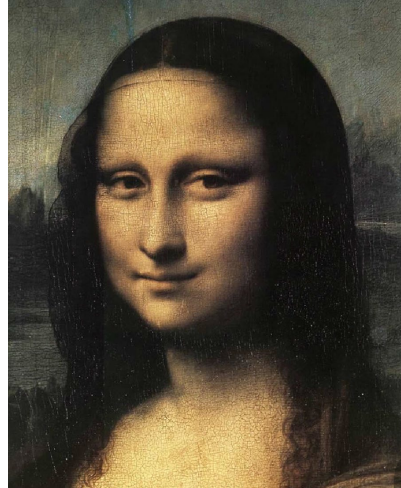


Figure 1.2 Mona Lisa
Source: BBC (2005)

Apart from the Mona Lisa, the film A.I. (Artificial Intelligence) directed by Spielberg, has makes an impressive attempt to depict the region's future prospects. The field's attractiveness has been boosted by modern trends in emotion transmission to 3D gaming avatars, accurate evaluations, as well as nonverbal interpretations.

Deep learning is a relatively new method with the potential for being extremely beneficial in a wide number of applications. They include many processing layers that allow for data representations at various levels of abstraction (LeCun et al., 2015). These layers were made up of small yet non-linear modules each one transformed a representation at a lower, more abstract level (beginning only with raw input) to a representation in a greater, much more abstract one. In terms of classification tasks, this composition may be used to learn very complex functions by amplifying features of input which are critical for discriminating and reducing unimportant variations (LeCun et al., 2015).

1.5 SCOPE OF THE STUDY

This thesis aims to develop a system that can differentiate between the seven universal emotions which are: Happy, Angry, Sad, Neutral, Surprise, Disgust, and Fear when presented with visual information. In order to avoid the difficulty of defining and recognising additional marginal emotions, the thesis has been restricted to universal emotions alone. To simplify the system's operation, it may be broken down into three stages: identifying where the face is located, extracting features, and classifying emotions. After the features of the face have been extracted, a neural network approach is utilised to identify the emotions that are encoded in the face.

1.6 SIGNIFICANCE OF STUDY

In truth, the emotional frontier is the next impediment to human study. Not only are facial expressions the most natural method for humans to convey their emotions, but they are also a crucial nonverbal communication tool. Once efficient measures for automatically recognising these facial expressions are developed, it will be able to achieve significant advancements in the area of human-computer interaction.. Face emotion recognition research is now being conducted in the goal of reaching these enhancements. Furthermore, automatic face expression recognition might be useful in a variety of applications.

Artificial intelligence has long utilised facial expression detection to acquire insight into how to correctly replicate human emotions in computers. Current revelations in this discipline have encouraged academics to redefine facial expression detection to also include avatars used in chat rooms and video conferences. Emotion recognition can enhance face recognition systems. Additionally, this work will help suspect identification systems and cognitive enhancement systems for children with brain development issues.

1.7 ORGANIZATION OF THE STUDY

The rest chapters of this thesis are as follows: Chapter 2 provides an overview of recent research in the field of facial emotion recognition. Meanwhile, chapter 3 provides the system model for the proposed solution, and also detailed explanations of every phases (face location determination, feature extraction, and emotion categorization). Chapter 4 will finish with a comprehensive evaluation of the system as a whole. Finally, Chapter 5 concludes the thesis.

1.8 CHAPTER SUMMARY

This chapter covered the study's background, problem statement, motivation, research objectives and research questions, and also the research scope and significance of study. The following chapter will explore facial expression recognition in detail, utilising both conventional and deep learning approaches.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

Face identification, facial feature extraction, and emotion classification are the three processes in the latest work pertinent to the study. In each of these categories, there has been a significant quantity of research done. These three areas address the fundamental underpinnings of the facial emotion recognition problem. Besides that, the effort being made to develop a facial database suitable for such investigations is essential.

2.2 FACIAL EXPRESSION ANALYSIS (FEA) TERMINOLOGIES

The conventional FER approach consists of three basic processes: picture preprocessing, feature extraction, and expression classification. Manual feature extraction approaches rely less on data and hardware, that is helpful for analysing limited data samples.

2.2.1 Facial Landmarks

As depicted in Figure 2.1, facial landmarks include visual highlights throughout the facial area, such as nose's alae, brow's end, as well as corner of the mouth. FLs are positioned around facial components and contours in order to capture deformations caused by head movements and facial expressions. By mapping facial markers to their point-to-point correspondences, a feature vector of a human face may be produced.

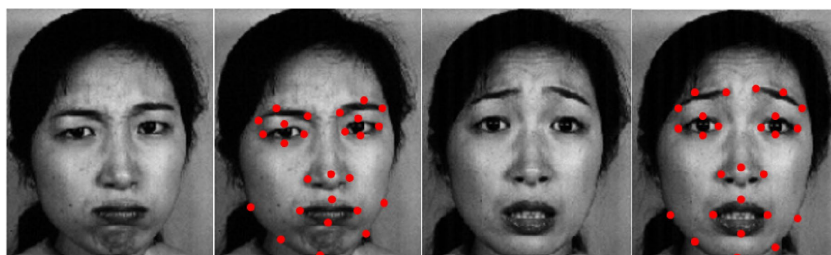


Figure 2.1 An Illustration of Facial Landmarks (Face Images from the JAFFE Dataset)

Source: Lyons et al. (1998)

2.2.2 Facial Action Units

The fundamental muscular motions that occur when a facial expression generates a certain emotion are illustrated in the 46 facial action units (Tian, et al. 2001). Some instances are shown in Figure 2.2. The FER algorithm distinguishes expression types by looking at how the detected facial AUs are combined. For instance, when an image has 1, 2, 5, and 25 AUs, this could convey the emotion "Awed."



Figure 2.2 Several AU Instances (images from the CK+ dataset)
Source: Lucey et al., 2010

2.2.3 Facial Action Coding System

Ekman and Friesen, internationally recognised psychologists, demonstrate the link between facial muscle movements and expressions through observations and biofeedback (Ekman et al., 1978). They split the entire face into several unique and

linked AUs based on anatomical characteristics and then analyse their properties. The following table summarises the regular AUs associated with the various kinds of basic and complex emotions. Today, the FACS method serves as the de facto reference standard for facial expression muscle movements, as it reliably recognises a broad variety of natural human expressions. The prototypical AUs observed as in categories of basic and complex emotions are listed in Table 2.1.

Table 2.1 Prototypical AUs Observed in the Category of Basic and Complex Emotions

Category	AUs
Happy	12,25
Sad	4,15
Fearful	1,4,20,25
Angry	4,7,24
Surprised	1,2,25,26
Disgusted	9,10,17
Happily sad	4,6,12,25
Happily surprised	1,2,12,25
Happily disgusted	10,12,25
Sadly fearful	1,4,15,25
Sadly angry	4,7,15
Sadly surprised	1,4,25,26
Sadly disgusted	4,10
Fearfully angry	4,20,25
Fearfully surprised	1,2,5,20,25
Fearfully disgusted	1,4,10,20,25
Angrily disgusted	4,25,26
Disgusted surprised	1,2,5,10
Happily fearfully	1,2,12,25,26
Angrily disgusted	4,10,17
Awed	1,2,5,25
Appalled	4,9,10
Hatred	4,7,10

Source: Benitez-Quiroz et al. (2016)

2.2.4 Basic Emotions

Ekman et al., (2003) proposes six essential human emotions: happiness, surprise, sadness, anger, contempt, and fear. These six BEs are commonly used to label FER-related datasets.

2.2.5 Compound Emotions

Compound emotions are formed by combining two fundamentals' emotions. Du et al. (2014) introduces twenty-two emotions that includes seven basic emotions, which are six basic emotions and one neutral, together with twelve compound emotions that humans frequently express, and three additional emotions which are hatred, awed, and also appealed

2.2.6 Micro Expressions

Micro expressions are involuntary facial motions that are more spontaneous and delicate (Du, S. et al., 2014). They frequently reveal a person's true and prospective expressions for a brief moment. The micro expression is incredibly brief, lasting approximately 1/25 to 1/3 of a second. Micro expression analysis is often utilised in criminal investigations and in the area of psychology.

2.3 CONVENTIONAL FACIAL EXPRESSIONS RECOGNITION APPROACHES

The traditional FER technique has a strong reliance on manual feature engineering as a distinguishing feature. The researcher must pre-process the image and choose the most appropriate technique for feature extraction and classification from the target dataset. As demonstrated in Figure 2.3, the conventional FER method is separated into three major steps: image pre-processing, feature extraction, and expression classification.

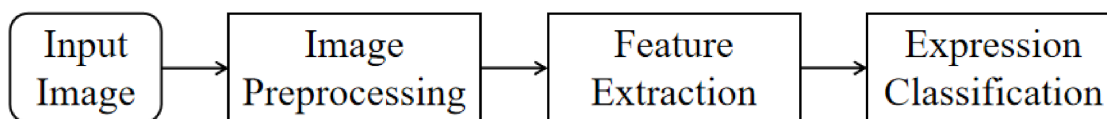


Figure 2.1 Procedure in Conventional FER Approaches
Source: Author

The traditional FER technique has a strong reliance on manual feature engineering as a distinguishing feature. The researcher must pre-process the image and choose the most appropriate technique for feature extraction and classification from the target dataset. As demonstrated in Figure 2.3, the conventional FER method is separated into three major steps: image pre-processing, feature extraction, and expression classification.

2.3.1 Image Preprocessing

This phase improves the detection ability of relevant information by removing irrelevant information from input photos. Image pre-processing has a direct impact on feature extraction and expression classification performance. Images are frequently corrupted by other signals for a variety of reasons. Even if an image is basically noise-free, it may contain complex backgrounds, such as occlusion, light intensity, and other sources of interference. Additionally, many files differ in size, with those containing color images and others containing grayscale images. Furthermore, data variability might be caused by a variety of shooting equipment. Moreover, while many datasets that include a bundle of images are different in size, other types of image datasets are composed of color images, or are made-up of grayscale images.

Furthermore, because the equipment used to shoot might result in data diversity, these objective interference elements must be handled prior to recognition in order to ensure that the target is detected. In general, the image pre-processing process can be summarised as follows:

1. The first step in preprocessing is noise reduction. At this stage, the Average Filter (AF), Gaussian Filter (GF), Median Filter (MF), Adaptive Median Filter (AMF), and Bilateral Filter are frequently used image processing filters (BF).
2. Additionally, there is face detection, which has developed into a distinct field (Viola, P.; Jones et al., 2004). Face recognition is a necessary pre-processing step in FER systems, with the goal of localising and extracting the face region.
3. Normalization of the image's scale and grayscale aims to normalise the image's input's size and colour, and ultimately to reduce calculation complexity while preserving the face's key features (Shan, Zhao et al., 2003)
4. The application of hogram equalisation to enhance the effect of the image (Tan, Sim et al., 2012)

2.3.2 Feature Extraction

When working with images, feature extraction is the process of extracting usable data or information from them. This includes things like values, vectors, and symbols. The "non-image" representations or descriptions of an image are referred to as features.

Many feature extraction methods are utilised in FER systems, including the following: Haar-like feature extraction, optical flow method, Gabor feature extraction, Local Binary Pattern (LBP) feature extraction, as well as feature point tracking and other variations on the theme. A direct impact of feature extraction on algorithm performance is possible, and this typically results in a bottleneck in the FER system's overall performance. Notably, while manually picking the most relevant feature extraction technique in conventional FER approaches, one should take into account both the applicability and the practicality of the methods under consideration.

a. Gabor Feature Extraction

The Gabor wavelet kernel function, which is based on the Fourier Transform, was devised in order to merge wavelet theory and the Gabor feature. Gabor wavelet-based FER, which is commonly employed in conjunction with other classification methods has been shown to give substantial improvements in several cases (Zhang et al., 2012). Russell et al. (1980) classified facial expression images using a series of Gabor filters

that were thought to be multi-orientation and multi-resolution in nature. To develop novel algorithms based on the Gabor feature, authors such as Yu et al. (2006) use both linear and nonlinear synthesis, as demonstrated in their paper. The Gabor-mean DWT (Discrete Wavelet Transform) provides a smaller feature vector than the frequently used Gabor-based expression classification and, as a result, reduces the dimensionality of the feature vector (Mattela et al, 2018). The Gabor wavelet, in addition, displays outstanding resistance to the transformation of multi-scale and multi-directional texture characteristics while staying insensitive to light intensity. However, because it is often used to work on global features, its major drawback is that it consumes a lot of memory.

b. Local Binary Pattern (LBP)

According to Ahonen et al. (2004), the Nearby Binary Pattern (LBP) method calculates the brightness connection between each pixel in a picture and its local vicinity. In the following step, the binary sequence is encoded in order to generate a local binary pattern. Figure 2.4 illustrates how LBP describes picture characteristics using a multi-region histogram, which is another technique used by LBP. In 2007, Feng et al. (2007) developed a method for creating local LBP histograms and associating them with FER.

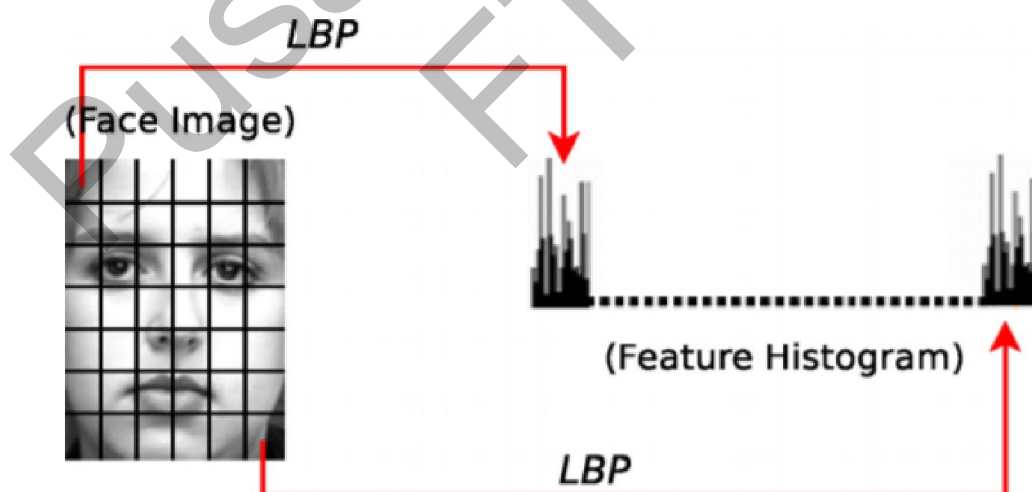


Figure 2.4 Extraction of LBP histogram from a Facial Image
Source: Ahonen et al. (2004)

It is proposed in this paper that an enhanced LBP-based method, called Complete Local Binary Pattern (CLBP), be used instead of LBP (Guo et al., 2010). However, while this improvised method beats the original LBP algorithm in terms of overall performance, it causes distortions in the image's three-dimensionality. On the other hand, Jabid and Chae (2012) proposed the LBP-based Local Directional Pattern (LDP) technique in order to improve the image's resilience while keeping the algorithm's computing cost as low as possible. Another technique is found in the Quantification of Local Phases (LPQ) by Wang et al. (2012), which is based largely on the short-time Fourier Transform with stable feature extraction and is based on the quantification of local phases (LPQ). To extract spatial information, the researchers developed an improved es-LBP (expression-specific LBP) feature and implemented the cr-LPP (class-regularized Locality Preserving Projection) method, which simultaneously reduces class independence while maintaining local feature similarity in the same region of the image (Chao et al., 2015).

Comparing the LBP operator to the Gabor wavelet, the LBP operator consumes less storage space and operates more efficiently on the same data. However, when dealing with noisy images, the findings are useless. Because it depends only on the pixel features of the image centre and its vicinity, and ignores any possible amplitude difference, the LBP operator may lose out on some valuable feature information.

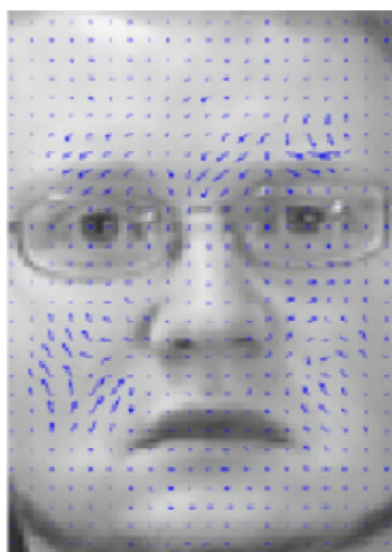
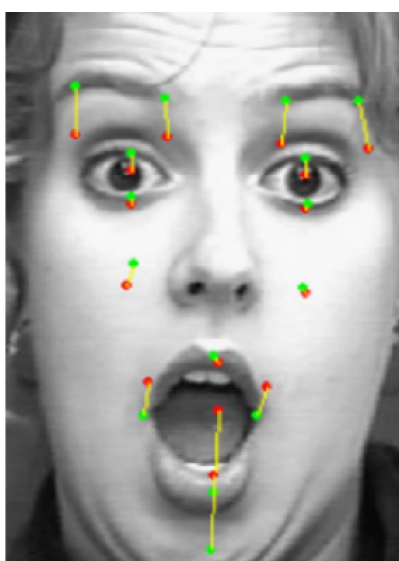
c. Active Shape Model (Asm) And Active Appearance Model (Aam)

In his 1995 work, Cootes et al. presented the Active Shape Model (ASM), which stands for Active Shape Model. Statistical models are employed to create it, and it is commonly utilised in circumstances wherein feature points of expression contours have to be retrieved from expression contours. When a global shape model is used, this model goes one step further by matching the original form of the human face and afterwards producing the local texture model that more accurately approximates the target's contour characteristics. An Active Appearance Model (AAM), on the other hand, was created as an extension of the Active Appearance Model (ASM) by integrating local texture features.

In 2004, Cristinacce et al. introduced a new approach wherein he integrated PRFR (Pairwise Reinforcement of Feature Responses) as well as AAM to identify critical feature points or landmarks of local edges including facial organs, whereas Saatci et al. (2006) delicately concatenated AAM with the SVM classifier in order to increase recognition rate.

d. Flow Method (OFM)

"Optical flow" is a phrase that refers to the pattern of apparent motion that is produced by relative motion. When the two-dimensional flow velocity is integrated with the grayscale utilising Horn–Schunck (HS) optical flow, the features linked with the continuous moving face picture arrangement are recovered (Horn et al., 1981). Yacoob et al. (1996) developed a system for analysing and expressing facial dynamics in their study. The programme creates optical flow that is produced by facial emotions in order to detect the direction of movements. By automatically detecting major changes in face expressions, Cohn et al. (1998) created an optical flow-based method for capturing emotional expression. The work of Sanchez et al. (2011), for example, may be credited with further advancements in optical flow. Their approach compares two different optical flow-based FER algorithms, which are named Feature Point Tracking as well as Dense Flow Tracking, accordingly, as seen in Figure 2.3.2.4.



Feature Point Tracking**Dense Flow Tracking**

Figure 2.5 Applications of Optical Flow-Based Methods on Facial Images
 Source: Sánchez et al. (2011)

e. Haar-like Feature Extraction

Alfred Haar created the Haar-like feature extraction, which integrates edge, centre, linear, as well as diagonal features in a single structure. Each of the rectangle regions in the feature template is separated into two equal-sized rectangle areas, white and black. These values of the feature template merely correspond to the difference in value between the pixels of the black and white rectangles. Color variation in the picture is represented by its Haar eigenvalue. In order to capture the temporal variations in the appearance of the human face, Yang et al. (2009) develop and encode dynamic Haar-like features as binary pattern features.

Because of its ability to capture the face's local grayscale variation, Haar may extract additional information regarding facial motion unit variations while the overall region illumination is stable.

f. Feature Point Tracking

As shown in Figure 2.6, the major goal of the feature point tracking technique would be to synthesise emotional expressions from the input that correlate to the displacement of the feature points in the input. The researchers at Tie et al. (2013) extracted over 20 points from the video stream as feature points for the face model, and then used these feature points in conjunction with particle filters to create a dynamic 3D expression model of recognition.

As an example, the authors of Liu (2011) provide a method for tracking feature points that is according to the Kanade-Lucas-Tomasi (KLT) transform and the Scale Invariant Feature Transform (SIFT). Using this approach, the SIFT is enhanced by spreading the feature points evenly and without any apparent aggregation of the points.

The KLT matching approach is then built in a hierarchical and iterative manner in order to allow for rapid monitoring of the match when the target displays clear attitude and size changes.



Figure 2.6 Feature Points Displacement

In FER, the phase of feature extraction is important. Consider that certain classifiers may suffer from dimensionality (i.e., the phenomenon wherein data gets sparse in a high-dimensional space) or over-fitting if the number of recovered features are excessive.

In particular, dimensionality reduction techniques are commonly employed throughout this context, since they have the potential to enhance learning performance, boost computation efficiency, and minimise memory consumption. Meanwhile, the conventional feature selection techniques, such as principal component analysis (PCA) and latent class analysis (LDA), are generally relevant to a wide range of machine learning problems. Researchers Xu et al. (2018) describe a method for choosing feature values in outlier detection and object recognition that can be used to data including binary either nominal characteristic or is described in detail in their paper.

2.3.2 Expression Classification

One critical factor determining the rate of expression recognition is the method by which the suitable classifier capable of accurately predicting facial expressions is chosen. SVM (Support Vector Machine), kNN (k-Nearest Neighbors), Bayesian, Adaboost (Adaptive Boosting), PNN and SRC (Sparse Representation-based Classifier), are among the most frequently used and widely deployed classifiers in FER systems (Probabilistic Neural Network). The next paragraphs will go over the pros, drawbacks, and comparisons in further depth.

While the kNN algorithm (Sohail et al., 2007) and (Valstar et al., 2004) are simple and straightforward to implement, the algorithm's training pace is poor. This is due to the fact that each new sample should be checked to a training set before it can be used. The next distinguishing feature of a kNN technique is also its sensitivity to the dataset's local structure. With the same weight assigned to each characteristic, classification precision is likely to be poor and unstable.

The SVM, according to Yu et al. (2006), may produce a decent compromise solution on difficult models via utilising small sampling information from the data and gain generalisation capacity. Additional to this, data that is linearly indivisible can be represented in higher dimensions by converting it to data that is linearly separable using kernel functions. Through utilising the kernel function, a system is able to efficiently process huge volumes of data while preventing dimension disruption to a certain level.

Similarly, it looks as though the AdaBoost classifier is vulnerable to noisy and anomalous data (Zhang et al., 2018). That is because it is, in certain circumstances, less likely to work better than other techniques of learning. AdaBoost is generally recognised as the best classifier available out-of-the-box (with decision trees as weak learners).

The Naive Bayes classifier, on the other hand, is very scalable. It necessitates the use of linear parameters in order to account for such high number of variables present in learning problems. The benefit would be that the classification parameters may be estimated with relatively minimal training data.

Compared to the traditional approach, SRC has a greater recognition rate, particularly when dealing with data that contain random block occlusion or random pixel distortion (Moghaddam et al., 2000). However, when dealing with data with the same vector direction distribution, SRC may be unable to categorise the data, as the sample vectors for the possible classes are spread along the same vector direction.

Traditional FER techniques tend to be less dependant on data or hardware as deep-learning-based approaches, at least in terms of overall performance. Nevertheless, since feature extraction as well as classification must have been performed manually as well as independently, both two stages are unable to be refined in tandem with one another. FER methods that are traditional in nature, on the other hand, are heavily dependent on the performance of their many constituent components.

2.4 DEEP LEARNING-BASED FER APPROACHES

Object recognition, categorization, and detection are just a few of the machine learning applications in which deep learning techniques have showed remarkable performance. Through aspects of FER, deep learning-based systems have a strong ability to reduce dependence on image preprocessing and feature extraction while also being more rigorous to settings of varying aspects, such as illumination and occlusion, allowing them to surpass conventional approaches by a significant margin. Aside from that, FER has exceptional capabilities for dealing with massive amounts of data.

2.4.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computer simulations of the brain's connected networks of neurons. Distinct areas of the human brain, organised in layers, process different pieces of information. Figure 2.4.1 shows how information enters the brain and is processed and transferred through each of these layers. As a representation of this, ANNs can have multiple layers that receive data, process it, and then send it on to the next layer.

Figure 2.7 shows a simple ANN with an input layer that receives the input data, a hidden layer that processes the data, and an output layer that makes a decision based on the collected data. Dendrites act as input terminals, neurons act as processing units and the axons act as output terminals.

Multiple nodes in ANNs replicate neurons. Nodes are connected by links that let data to flow from one node to the next, layer by layer. Weight values are assigned to these links, allowing the network to learn. Nodes receive data from other nodes, conduct simple actions on it, and then pass it on. Node value or activation refers to a node's output.

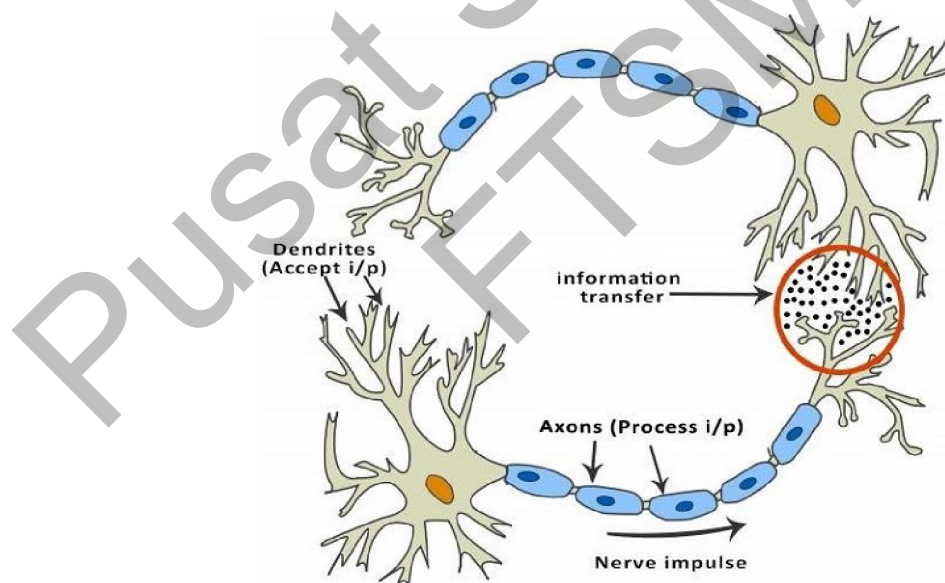


Figure 2.7 Inner Workings of the Human Brain

Source: Tutorials Point (2019)

2.4.2 Deep Learning

Neural networks are used in deep learning to perform machine learning tasks. Deep learning is a subset of machine learning, and it is becoming increasingly popular. Various deep learning designs, comprising Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), as well as the others, are depicted in Figure 2.8. These techniques are applicable to extensive areas, together with speech recognition, object detection, as well as object recognition.

Deep learning employs a cascade of nonlinear processing units, each of which takes the output of the preceding layer as input. Each layer converts the data it receives into a more abstract representation. A deep learning system learns on its own by filtering data through a number of hidden layers.

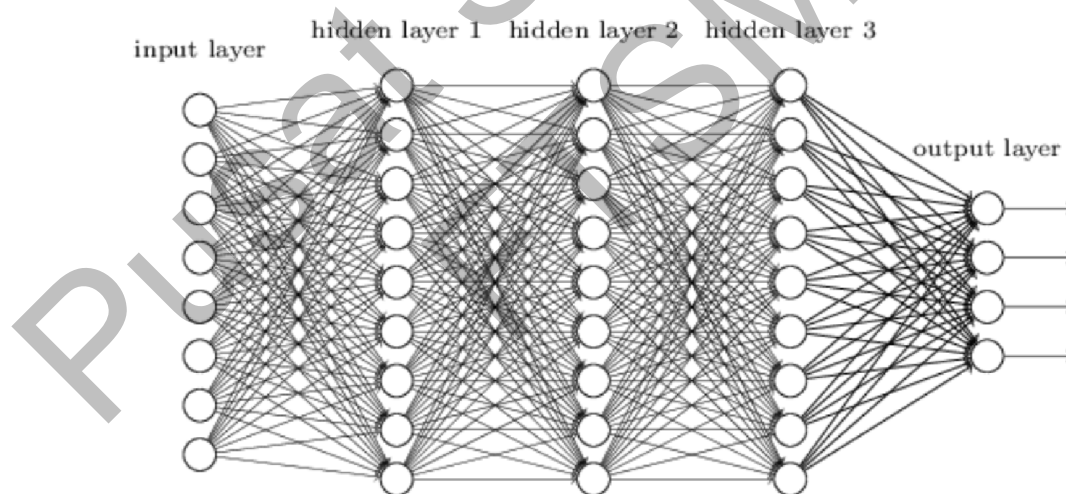


Figure 2.8 Deep Neural Network

Source: Nielsen (2015)

2.4.3 Convolutional Neural Network

LeCun et al. (1998) were the first to propose Convolutional Neural Networks (CNNs). This deep learning architecture is mostly used to categorise photos and recognise objects. They can be used to recognise faces, objects, handwritten characters, and a variety of other things.

CNNs, like neural networks, consist of a sequence of layers, the most common of which are the Pooling Layer, the Convolutional Layer, as well as the Fully Connected Layer. Figure 2.9 illustrates a CNN architecture in action.

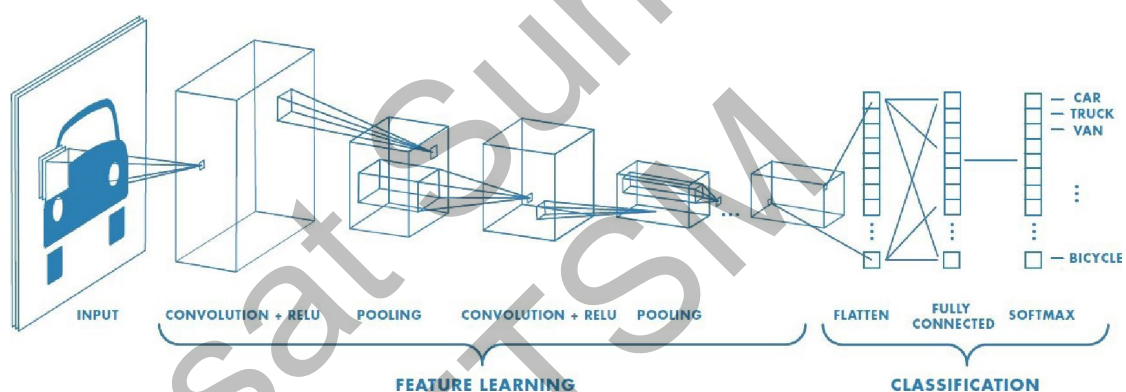


Figure 2.9 Architecture of a CNN

Source: MathWorks et al. (2019)

a. Convolutional Layers

3D arrays are used to represent RGB images. Two of the dimensions are related to the image's width and height, depicting each pixel in the image. For each pixel, the third dimension provides the three colour channels, as well as the intensity values of the red, green, and blue hues. Grayscale images have only one colour channel, which represents the amount of light in each pixel.

Multiple filters are applied to an image by CNNs, each of which selects different signals. To produce maps of an image's edges, initial layers can pass horizontal, vertical,

or diagonal line filters. A convolutional network extracts square patches of pixels from an image and filters them. A filter, also known as a kernel, is a square matrix that is smaller than the image's dimensions but equal to the size of the patches to be processed.

The image's depth (number of channels) must be the same as the filter's. The matrix values will be sent through the patch, which will aid in the discovery of patterns in the image. The filter is moved over the image from left to right and from top to bottom, starting in the top left corner and finishing in the lower right corner, starting in the top left corner and ending in the lower right corner. The amount of pixels that the filter skips while processing the next patch is referred to as the stride.

With the weights in the filter and the values in the patch of the image being processed, the filter does a dot product. The dot product result is saved in a third matrix known as an activation map. Aside from the weights, the outcome also considers one bias parameter. The stride will determine the dimensions of this matrix. A larger stride results in a smaller output matrix, which takes less time to compute.

It's possible that padding the input image with zeros around the edges will be useful. This is done to ensure that the pixels around the edges of the image are preserved and that the output volume has the same dimensions as the original image. The image can be processed with several filters, resulting in multiple activation maps that will be overlaid. The final volume will be $100 \times 100 \times 20$ if the convolution of an image resulted in an activation map of dimension 100×100 and 20 different types of patterns were processed.

Figure 2.10 shows the convolutional layer details. In blue are represented the 3 channels of an input image. The weights of two filters are shown in red. The output activation map for both filters is colored in green.

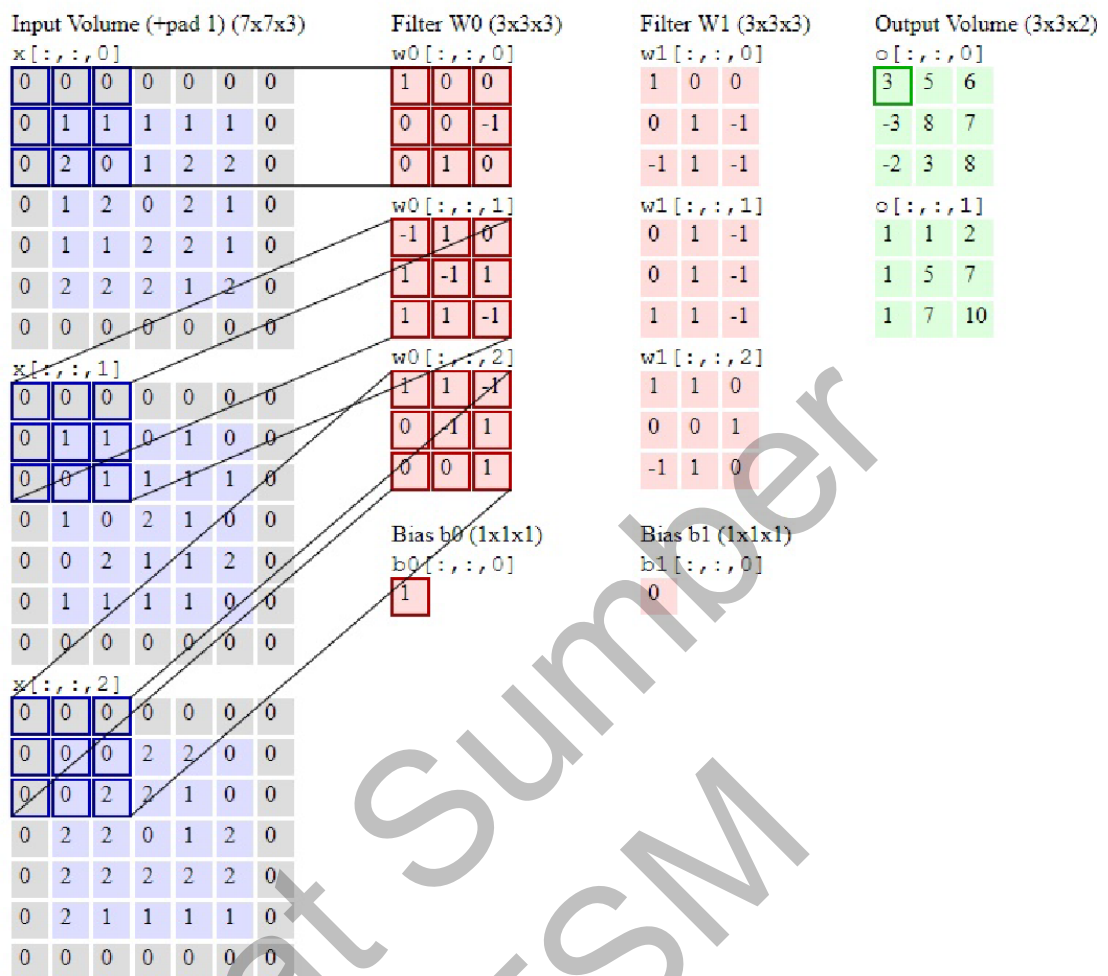


Figure 2.10 Convolutional Neural Network
Source: CS231n (2019)

g. ReLU Activation Function

Non-linearity is required in any neural network, which is obtained by passing the weighted sum of the inputs through an activation function. After each Convolution Layer, this activation function is generally employed. Figure 2.11 shows how the ReLU (rectified linear unit) removes negative values from activation maps and replaces them with zero. Because it provides a constant gradient of 1 to all input values larger than zero, it also allows for improved transmission of the error's gradient through the network.

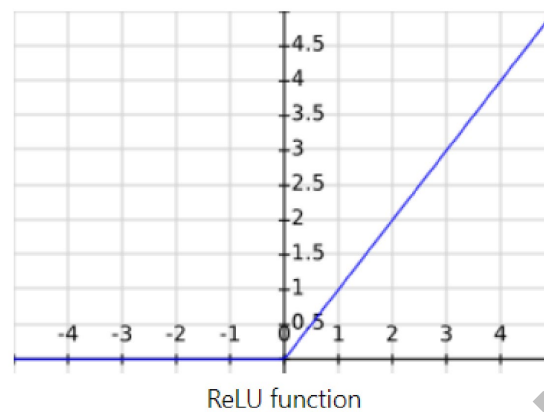


Figure 2.11 ReLU Activation Function
Source: Code Project (2019)

h. Pooling Layer

Max Pooling is the most popular form of Pooling Layer in CNNs. The activation maps generated by the Convolution Layer are fed into the Max Pooling layers as input. This approach, like convolution, is applied one patch at a time. Figure 2.12 shows how max pooling selects the highest value from a patch and converts it to another matrix containing the max pooling results of all patches.

Between successive Convolution Layers, maximum pooling is typically used. This layer minimises the representation's spatial dimension and, as a result, the amount of parameters and computations required.

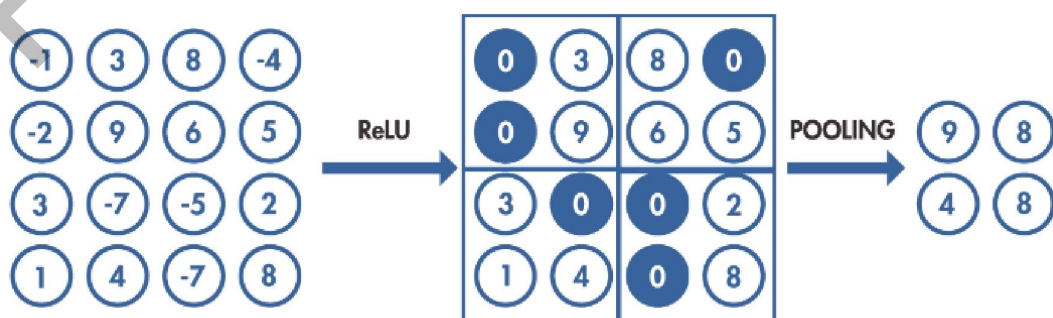


Figure 2.12 Application of ReLU and Max Pooling
Source: MathWorks et al. (2019)

i. Fully Connected Layers

By connecting each layer's neuron to every preceding layer activation, the Fully Connected Layer encourages high-level reasoning. This layer is attached to the network's end, and the input volume of any layer that comes before it is flattened to a vector. The Fully Connected Layer will then generate a 1D vector with a dimension equal to the network's number of recognised classes. The numbers in the vector represent the odds that an input image belongs to each of the classifications.

j. SoftMax Activation Function

It is used to deliver the probability for each class in the Fully Connected Layer. It produces a value in the range of zero to one, with the sum of all values equalling one.

k. Dropout

Dropout is a form of regularisation that is used to improve test accuracy while also reducing overfitting by enhancing training accuracy at the expense of accuracy during testing. For each mini-batch in the training set, dropout layers randomly disconnect inputs out from preceding layer to the next layer in the network architecture with probability p . Figure 2.13 shows the training with and without dropout. On the left is two completely linked layers of a neural network with no dropout. On the right is the same two layers with 50 percent less connections

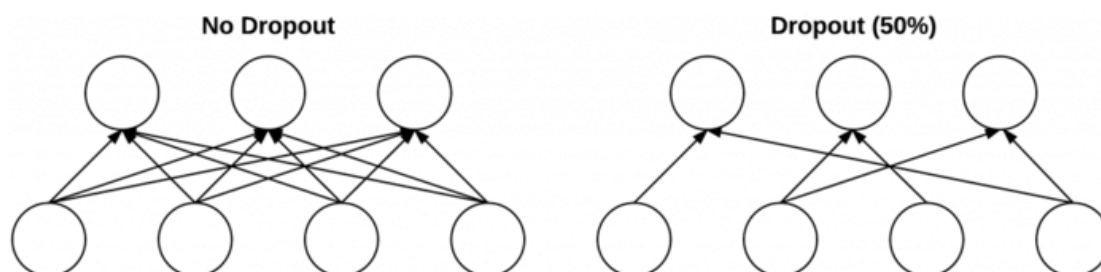


Figure 2.13 Training With and Without Dropout

I. Training

Adjustment of the filter values or weights is accomplished by backpropagation. It is divided into four stages: the forward pass, the loss function, the backward pass, and the weight update. During the forward pass, a training image travels the whole network. Weights are randomly initialised. The goal is to minimise the network's loss by performing a backward pass, identifying which weights contributed to the network's loss, and adjusting the weights accordingly.

2.5 DEEP LEARNING BASED WORK OF FACIAL EXPRESSION

Burkert et al. developed a technique based on CNNs, with a network architecture made up of four main components: its first portion does automated data preprocessing, while the succeeding sections handle feature extraction. At the network's conclusion, the fully connected layer classifies those received features according to a given phrase. The architecture is composed of fifteen layers: seven convolutions, five poolings, two concatenations, and a normalisation layer. They were unable to verify that training subjects would not be included in the testing. This is a necessary constraint for fair assessment of facial expression recognition methods (Burkert et al., 2015).

In the majority of deep learning-based FER approaches, CNN is immediately adapted for AU detection. The Facial Action System (FACS) feature representation is utilised in the CNN-based FER method described in (Breuer et al., 2017), which demonstrates networks' ability to generalise across data and tasks. When used to detect micro-expressions, the model obtains a high rate of identification.

On the other hand, for dynamic expression analysis, Liu et al. (2014) developed a deformable facial action component model. The 3D CNN contains a deformable facial parts learning module that seems to be capable of detecting and representing a specific facial action component.

In addition, Jung et al. (2015) describe a novel integration approach dubbed Deep Temporal Appearance-Geometry Network (DTAGN) for recognising facial emotion using temporal data. The Deep Temporal Appearance Network (DTAN) utilises image sequences to extract appearance features. The Deep Temporal Geometry Network (DTGN) utilises temporal FL points to extract geometry information. The combined DTAGN improves the FER's performance by maximising the usage of temporal data.

As an additional measure to boost the depth and width of the network while keeping a consistent computational budget, Mollahosseini et al. (2016) propose employing two convolutional layers as well as four inception layers in addition to the four convolutional layers. The suggested technique makes use of thinner networks than standard CNN approaches and outperforms them by a large margin in cross-database classification circumstances.

On the other hand, Li et al. (2019) investigate facial expression recognition with occlusion. Li et al. propose two ACNN (Convolutional Neural Network with Attention mechanism) variants for different face ROIs: pACNN (Patch-based ACNN) and gACNN (global-local ACNN) (Regions of Interest). Instead of using local patches in conjunction with global pictures, the pACNN utilises just local face patches in its computations. According to experimental data, ACNNs can enhance FER accuracy of occluded face images via substituting occluded patches by similar but unexcluded patches that are close in proximity to each other.

The DBN extracts unsupervised and abstract characteristics from input signals (Hinton et al., 2006). It is built on the Restricted Boltzmann Machine (RBM) (Hinton et al., 1986). The FER approach, which is based on DBN, is capable of automatically learning the abstract information contained in facial images and is activity-sensitive. DBN has been demonstrated to be a successful FER approach when coupled with several other FER components.

Furthermore, Zhao et al. (2015) provide a novel FER approach (abbreviated DBNs + MLP) that combines the DBN like an unsupervised feature learning module with MLP

as a classification module. To begin, the DBN is utilised to extract abstract features from facial images' primary pixels. Classification processing is conducted after initialising the MLP model with the DBN learning results.

Additionally, He et al. (2016) describe a FER model that incorporates LBP/VAR and DBN. The light- and rotation-resistant LBP/VAR feature is extracted first as part of the model's preprocessing. Following that, the DBN is employed for the second phase of feature extraction and classification.

The robust LDPP-PCA-GDA (Local Directional Position Pattern, Principal Component Analysis, Generalized Discriminant Analysis) features of another study on modelling and recognition are combined with DBN in another work on recognition and modelling (Uddin et al., 2017). The proposed method takes into account the direction of the highest strength and the sign of the strengths in order to extract salient features while tolerating illumination variance. Additionally, the recognition performance was superior to that of traditional methods.

An LSTM network is a type of RNN composed of LSTM units that excels at extracting temporal characteristics from successive frames. Due to the fact that previous research has demonstrated that long-range context modelling improves the accuracy of emotion analysis, certain LSTM-based FER techniques on video series have been developed (Wöllmer et al., 2013). They created an LSTM-based system to evaluate the dimensional representation of emotions in audio-visual settings. When assessing emotions, aural, verbal, and visual information are all taken into consideration in order to provide a more human-like outcome. This can be seen as a reflection of the realities of natural encounters.

On the other hand, Kim et al. (1996) use a CNN to train the spatial features of representative state frames, followed by an LSTM to understand the temporal properties of the spatial feature representation learned from the state frames. While network training was taking place, the suggested technique employed typical expression states found in facial sequences independent of the strength or length of the expression in question.

As another example, Hasani et al. (2017) used 3D Inception-ResNet (3DIR) layers in conjunction including an LSTM unit in their work to construct a 3D Convolutional Neural Network (3DCNN) architecture in their research. The Inception-ResNet module, which has been enhanced, extracts spatial relationships from expression images. These temporal connections are taken into consideration when the sequences are classified by the LSTM. It receives data from facial landmarks, which are visual signals for the presence of certain facial components.

Known as a Generative Adversarial Network (GAN), it is an unsupervised learning model composed of a generating network and a discriminative network which has been effectively employed to image synthesis to generate stunningly realistic facial images, videos, and other media (Goodfellow et al., 2014). It is not only possible to augment training data and associated recognition tasks using GAN-based models, but they are also beneficial in pose- and identity-invariant expression recognition using GAN-based models. It is described in detail in Lai et al. (2018), who describe a multitask GAN-based learning strategy for multi-view FER wherein the generator generates a frontal face image from an input non-frontal image while maintaining the identity and expression data, and the discriminator has been trained to distinguish and recognise the generated frontal face image. The viability of this facial formalisation system for FER is proven through the use of apparent head posture changes.

In addition, Zhang et al. (2018) offer a comprehensive GAN-based model (Zhang et al. 2018). It is the encoder–decoder structure of the face-picture generator that first learns to represent the identity of the face picture, which is subsequently demonstratively conveyed through the use of emotion and posture codes. The model can also create face images with any emotions and head postures on its own, which may be used in conjunction with the FER training set as a supplement.

Using GANs, Yang et al. (2018) train the generator to create six basic expressions from a face image, as well as a CNN has been fine-tuned for every identity sub-space expression classification using a CNN. A flexible method that may be utilised in conjunction with a number of CNN frameworks for FER to reduce the impact of inter-subject differences is described in detail in the next section.

Chen et al. (2018), on the other hand, propose a Representation-Learning Variational Generative Adversarial Network (PPRL-VGAN) for learning a demonstratively specified image representation from the user's identity information in place to protect the confidentiality using a demonstratively specified image representation. When it comes to extracting information from the expressive component, Yang et al. (2018) offer a technique termed DeRL (De-expression Residue Learning). A cGAN (conditional Generative Adversarial Networks) creates the suitable neutral face image for every input in order to remove expressive data.

Liu et al. (116) developed a new Boosted Deep Belief Network (BDBN) for facial expression detection that iteratively employed three training phases inside a unified loopy architecture. Their investigations chose the first frame (with neutral expression) and the final three frames from each picture sequence in order to get more CK+ samples. Extensive trials on the CK+ and JAFFE databases shown that their approach outperformed existing state-of-the-art algorithms benchmarked on these two datasets.

2.6 DATASETS

A common study design involves the use of two-dimensional static pictures in the early phases of the research process; 2D video sequences are subsequently utilised in FER studies to enhance awareness in expression throughout many dimensions. However, utilising 2D data to analyse face deep information is a difficult task. It is possible that some variables, including posture and light, will have a substantial impact on the efficacy of FER. Additionally, with the aid of some 3D-based datasets, it is possible to investigate micro-facial behaviours and differences in face structure. A number of researchers choose to use datasets from the field rather than laboratory settings in order to better fulfil the demands of practical applications.

2.6.1 FER2013 Dataset

The data collection contains 48x48 pixel grayscale images of people's faces in various poses. In order for the faces to be nearly centred in each image and cover approximately the same amount of space, they have been automatically registered. The goal is to categorise every face into each of seven categories based on the emotion exhibited in the facial expression (0 represents angry, 1 represents disgust, 2 represents fear, 3 represents happy, 4 represents sad, 5 represents surprise, and 6 represents neutral).

There are two columns in train.csv that are named "emotion" and "pixels." There are six numerical codes in the "feeling" column, each of which represents a different emotion expressed by the image and ranges from 0 to 6. Each image's "pixels" column includes a string that has been surrounded in quotation marks.

There are a total of 28,709 instances in the training set. The 3,589 cases in the public test set that was used to construct the leader board were utilised to create the leader board. The final test set, which included 3,589 samples, was utilised to decide the winner of the competition.

This dataset was developed by Pierre-Luc Carrier and Aaron Courville as part of a larger research effort that is still in progress. Because of their generosity, they have made a preliminary version for their dataset available to the workshop organisers for consideration in this competition.



Figure 2.14 FER2013 Sample Images

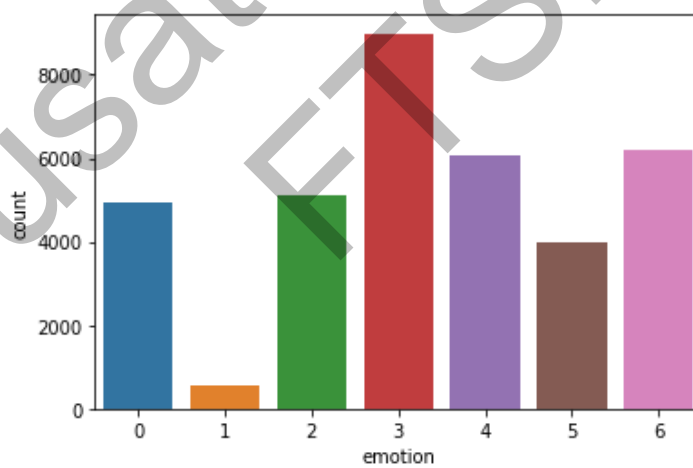


Figure 2.15 FER2013 Class Distribution

2.7 CHAPTER SUMMARY

This chapter summarised previous discussions on facial expression recognition. According to the research, CNN-based approaches enable training and feature extraction directly from raw input images, as well as combining the feature extraction and classification phases into a single step with minimal pre-processing. Simultaneously, CNN can produce promising results that rival or even surpass those obtained through more traditional methods. This is a substantial advantage of the CNN-based technique. Additionally, because they are constrained by computational time constraints, most FER techniques are impractical for use in real-time scenarios. The following chapter discusses the study's methodology.

Pusat Sumber
FTSM

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

This chapter details the research methodology used throughout this study, including the many stages of the process. A description of the components that were chosen during the creation of this system, as well as the justifications for those choices, is also provided, which includes an overview of the models that were used to recognise facial expressions.

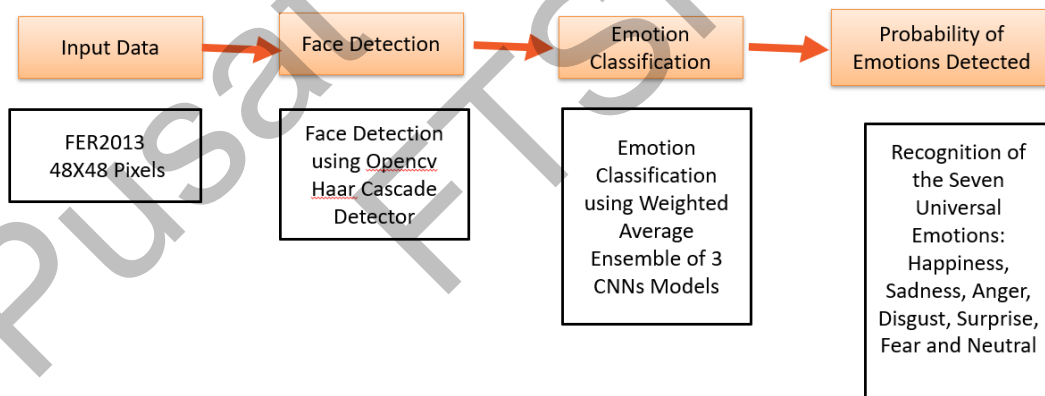


Figure 3.1 Framework of FER Classification

3.2 SOFTWARE TOOLS

An analysis of available software tools and libraries was undertaken in order to find an effective tool for implementing CNN for FER classification. Machine learning software

comes in a wide range of options. Some of them are standard machine learning tools, while others are intended exclusively for deep learning.

The software tools for machine learning have had a resurgence in the recent ten years. There is a huge range of tools accessible, and new tools are being introduced on a regular basis to keep things interesting. A software library, or at the absolute least an Application Programming Interface (API), is available for almost every widely used programming language.

A number of reasons affected the software tool's selection. To begin, the implementation language needed to be well-known and mainstream. There has to be enough learning materials available, especially in the form of tutorials. The most crucial component was having robust GPU learning support.

3.2.1 Tensorflow

As the name implies, this library focuses on effective tensor work. It was originally created for Google's internal use for machine learning tasks, but in 2015 it was made open source. Tensorflow computations are expressed as stateful dataflow graphs, allowing for efficient GPU-assisted processing. Is now marketed as one of the fastest deep learning frameworks.

3.2.2 Keras

Keras is a Python project that is both young and mature. It's a neural network API with a lot of power. It's designed to work on top of either the Theano or Tensorflow libraries. It's straightforward, with a focus on quick model development. At the same time, it is incredibly extendable thanks to the Python infrastructure.

Among similar deep learning systems, Keras undoubtedly has one of the largest communities. It features excellent documentation with numerous code examples and other tools to assist users in getting up and running quickly. Tensorflow's support for GPU-based execution models offers this capability to Keras as well.

3.3 SPLITTING THE DATA

The FER2013 dataset consists of three standard subsets: training, validation, and testing. The training set contains 28,709 images which is about 80% of the dataset, validation and testing sets contain 3589 images for each which is 10% of the total dataset. In this work, the standard splitting configuration were used to compare results obtained results of this project with the state of art works.

3.4 DATA AUGMENTATION

The term "data augmentation" refers to a group of approaches for creating new training samples from existing ones using random vibrations and perturbations while retaining the class labels. The purpose of data augmentation is to enhance the model's generalizability. Due to the continual exposure of the network to new, slightly modified versions of the input data points, the network may develop more robust features. Data augmentation is not used during testing; rather, it is used to evaluate trained networks. Generally, a gain in testing accuracy comes at the price of a modest reduction in training accuracy. Figure 3.2 shows the data augmentation jitter distribution.

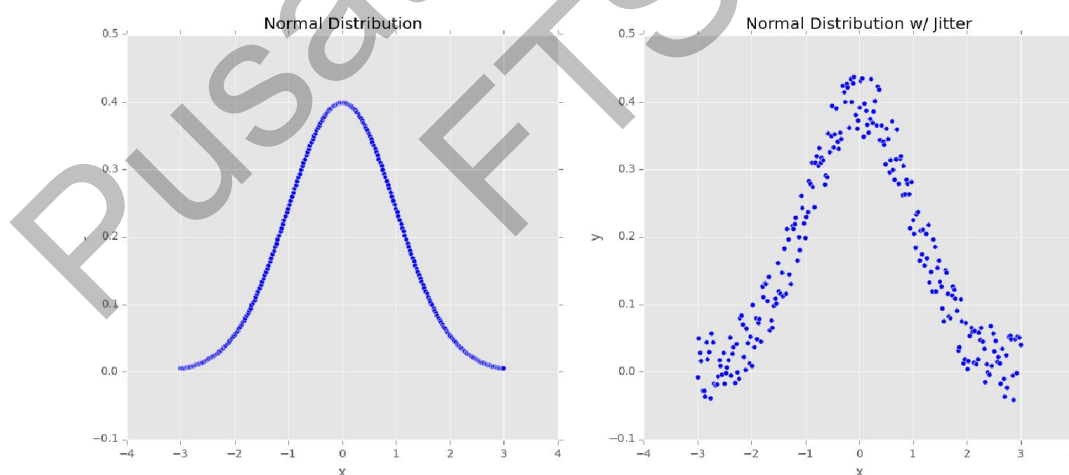


Figure 3.2 Data Augmentation Jitter Distribution

Consider a normal distribution illustrated in Figure 3.1 (left), which has a mean of 0 and a variance of 1. Because data in real-world applications seldom follows such a neat distribution, training the machine learning model on that might lead in an accurate simulation of the distribution.

To the contrary, the generalizability of the classifier can be improved by adding extra values chosen out of a random distribution onto randomly jitter points along the distribution (right). Even while the graph has a roughly normal form, it is not as perfect as the graph on the left. The model that has been trained on this data set is much more likely to generalise onto data points that did not form part of the training set.

Data augmentation makes intuitive sense in the field of computer vision. Applying simple geometric transformations to the original images, including rotations, random translations, scale changes, horizontal and shearing (in certain instances, vertical) flips, can yield additional training data.

When a small number of these modifications are applied to an input image, the appearance of the image is slightly altered, but the class label is not affected. This will make the data augmentation a natural and straightforward technique for deep learning in computer vision applications. Random colour disturbances within a given colour space (Alex Krizhevsky et al., 2012) and nonlinear geometric distortions (Lecun et al., 1998) are more advanced techniques for data augmentation in computer vision. In this thesis, data augmentation is utilised on the training set for both datasets to aid in the minimization of overfitting and the improvement of the classification accuracy of the model.

3.5 MODELS BUILDING BLOCKS

3.5.1 VGG16

When it comes to CNNs, the VGG-16 architecture is a Convolutional Neural Network (CNN) that was utilised to win the ILSVR (ImageNet) competition in 2014. Developed by Simonyan and Zisserman of the University of Oxford, this architecture came in second place in the Visual Recognition Competition (ILSVR-2014). According to most experts, it is one of the most outstanding vision model architectures ever created. Compared to other VGGs, VGG-16 is distinguished by the fact that it concentrates on 16 CNV/FC layers, which include convolutional layers of 3x3 filter with stride 1, and that it always utilises the same padding and maxpool layer of 2x2 filter with stride 2. Over the course of the design, it preserves the order of convolution and max pool layers.

Finally, it comprises two FC (Fully Connected) layers as well as a SoftMax for

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

generating output. A weighted layer is represented by the number 16 in VGG-16, which indicates that it contains sixteen layers. Figure 3.3 depicts the VGG-16 architecture, while Figure 3.3 depicts the architecture with its layers.

Figure 3.3 A Replication of Table 1 from Simonyan and Zisserman

To summarise, VGG-16 is composed of sixteen weight layers, thirteen convolutional layers with a 3×3 filter size, and three fully connected layers, with a total of sixteen weight layers. Stride as well as padding are both one pixel in all convolutional layers. Within every convolutional layer, five groups are formed, with each group being followed by a max-pooling layer. Using a 2×2 window, the stride 2 conducts max-pooling. Initially, the filter count of the convolutional layer group is 64 in the first group,

and it rises by a factor of two after each max-pooling layer until it reaches 512 in the final group. Each hidden layer contains a rectification (ReLU) nonlinearity. The SoftMax layer is the final layer in this model, and it is used for classification. It is possible to substitute another classifier for the SoftMax layer, such as a neural network, a random forest, or even a support vector machine. To prevent the network from overfitting, the dropout layer is utilised.

```
1 base_model1 = VGG16
2 Input: FER2013 dataset (35887 grey-scaled images)
3 1. Initialize parameters nb_class, x, y, epoch, lr, bs, c,
4 where nb_class = number of facial expression classes
5 x = height of the image, y = width of the image
6 epoch = number of iterations
7 lr = learning rate
8 bs = batch size
9 c = classifier
10 2. for 1: epochs
11 bs = 64 and lr = 1e-3 to 1e-4 // for FER2013 dataset
12 for 1: last_block_layer // base_model1
13 fv1 = generate feature vector for base_model1
14 acc1 = predict result of fv1 using classifier c
15 end for
```

Figure 3.4 VGG16 Model Layers

VGG16

Conv (64)
 Conv (64)
 Conv (128)
 Conv (128)
 Conv (256)
 Conv (256)
 Conv (256)
 Conv (256)
 Conv (256)
 Conv (256)
 Conv (256)
 Conv (256)
 Conv (256)
 Flatten (2304)
 Dense (1024)
 Dense (1024)
 Dense (7)

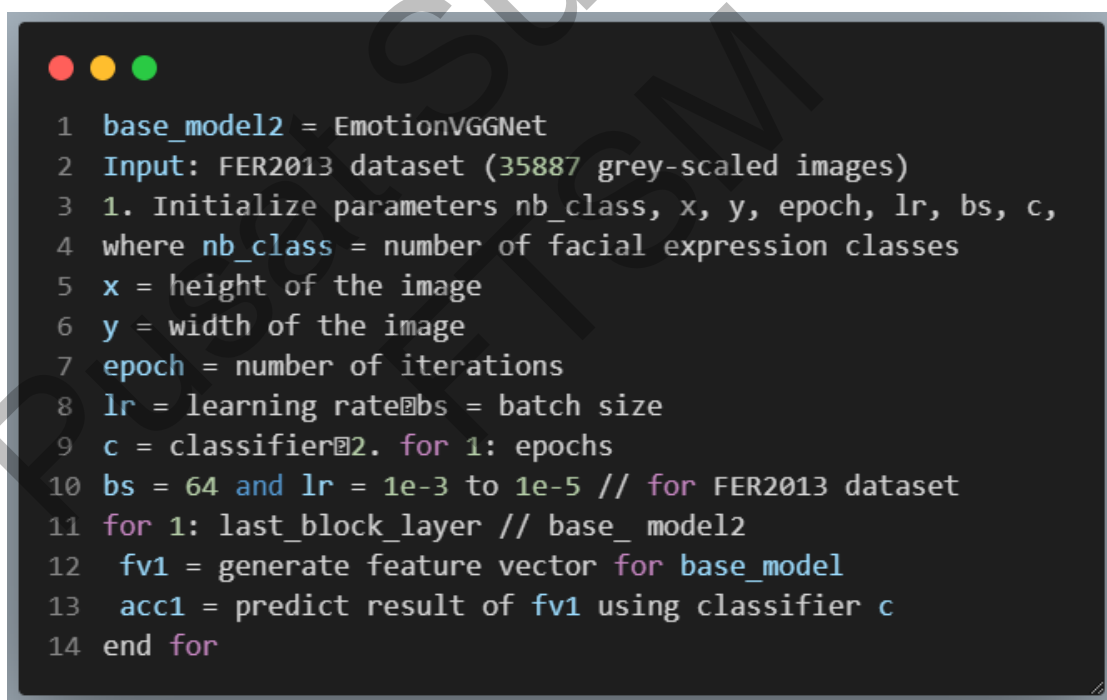
Figure 3.5 VGG16 Model Layers

3.5.2 EmotionVGGNet

The network used to classify various emotions and facial expressions is a member of the VGG family. The Convolutional layers of the network will be only 3X3. Each Convolutional layer will learn twice as many filters as we progress deeper into the network. The Sequential class was chosen because the VGG family of networks employs sequential layer application. Each convolutional block in EmotionVGGNet will be composed of (Convolutional => ReLU => Batch normalisation) * 2 => POOL layer sets. The first Convolutional layer will learn the 32 3X3 filters. Batch

normalisation was performed following RELU activation. The second Convolutional layer is identical to the first, learning 32 3X3 filters before performing an ReLU and batch normalisation. Following that, there is maximum pooling and a dropout layer with a probability of 25%. The second block of EmotionVGGNet is identical to the first, except for the increase from 32 to 64 filters in the Convolutional layers. Moving on to the next block, the same pattern was used, but this time the number of filters was increased from 64 to 128 - the more filters the model learns the deeper it gets into the Convolutional Neural Network.

Following that, the first fully-connected layer was applied. After applying 64 concealed nodes, batch normalisation and an ReLU activation function were performed. A second fully connected layer was added in the same manner. Finally, we obtained our output class label probabilities using an FC layer with the specified number of classes and a softmax classifier. Figure 3.4 shows the emotionVGGNet layers.



```

1 base_model2 = EmotionVGGNet
2 Input: FER2013 dataset (35887 grey-scaled images)
3 1. Initialize parameters nb_class, x, y, epoch, lr, bs, c,
4 where nb_class = number of facial expression classes
5 x = height of the image
6 y = width of the image
7 epoch = number of iterations
8 lr = learning rate bs = batch size
9 c = classifier
10 2. for 1: epochs
11 bs = 64 and lr = 1e-3 to 1e-5 // for FER2013 dataset
12 for 1: last_block_layer // base_model2
13 fv1 = generate feature vector for base_model
14 acc1 = predict result of fv1 using classifier c

```

Figure 3.6 EmotionVGGNet Layers

EmotionVGGNET

Conv (32)
Conv (32)
Conv (64)
Conv (64)
Conv (128)
Conv (128)
Flatten (2304)
Dense (256)
Dense (256)
Dense (7)

Figure 3.7 EmotionVGGNet Layers

3.5.3 MiniGoogLeNet

This section will introduce the MiniGoogLeNet architecture that was used in this research. The architecture was inspired by Szegedy et al 2014 's paper, Going Deeper With Convolutions.

The purpose of utilising this architecture is twofold. To begin, the model architecture is insignificant in comparison to AlexNet and VGGNet. The authors achieve this huge reduction in network architectural size (while maintaining network depth) by deleting completely connected layers and replacing them with global average pooling. The majority of weights in a CNN are contained in the dense FC layers; removing these layers results in significant memory savings. Second, Szegedy et al. create the entire macro-architecture by utilising a network in network or micro-architecture. In sequential neural networks, the output of one network feeds directly into the output of the next, but in micro-architectures, small building blocks utilised within

the larger architecture, the output of one layer might split into a number of different routes and be rejoined later.

Other significant variations, such as the Residual module in ResNet, have been influenced by micro-architectures such as Inception. The Inception module (and its variations) will be addressed next in this work. The FER2013 datasets were used to train the architecture.

b. The Inception Module (and its Variants)

State-of-the-art technology in the modern era Convolutional Neural Networks make use of micro-architectures, sometimes referred to as network-in-network modules. Micro-architectures are tiny building pieces created by practitioners of deep learning to enable networks to learn more quickly and effectively while increasing network depth. These micro-architecture building pieces, together with traditional layer types such as Convolutional, Pooling, and so on, are layered to construct the overall macro-architecture.

Szegedy et al. introduced the Inception module in 2014. The Inception module's overarching concept is twofold:

To begin, determining the size of the filter to learn at a specific Convolutional layer might be challenging. Should they be 5×5 filters? What about 3×3 filters? Should we use 1×1 filters to learn about local features? Rather than that, why not memorise all of them and let the model decide? Inside the Inception module, we learn all three 5×5 , 3×3 , and 1×1 filters (computing them in parallel) concatenating the resulting feature maps along the channel dimension. The following layer in the GoogLeNet architecture (which may or may not be another Inception module) gets these concatenated, mixed filters and repeats the process. Taken together, this method enables GoogLeNet to learn both local and abstracted features via smaller convolutions and bigger convolutions.

Second, the module may be transformed into a multi-level feature extractor by learning different filter sizes. The 5×5 filters have a wider receptive area and are

therefore more capable of learning abstract characteristics. By definition, the 1×1 filters are local. The three filters act as a counterbalance.

c. The Inception Module (and its Variants)

Take note in particular of how the Inception module divides the input layer into four different routes. The Inception module's initial branch simply learns a sequence of 1×1 local characteristics from the input. The second batch first employs 1×1 convolution as a dimensionality reduction technique, rather than as a method of learning local features. By definition, larger convolutions (i.e., 3×3 and 5×5) require more processing. As a result, by applying 1×1 convolutions to the inputs to these bigger filters, we may minimise the amount of processing required by our network. As a result, the number of filters learnt in the first branch's 1×1 Convolutional will always be less than the number of 3×3 filters learned immediately thereafter. The third branch follows the same reasoning as the second, but with the additional aim of learning 5×5 filters. We reduce dimensionality once more using 1×1 convolutions and then feed the output to the 5×5 filters. Figure 3.5 shows the inception module that was first used by Szegedy et al. (2014).

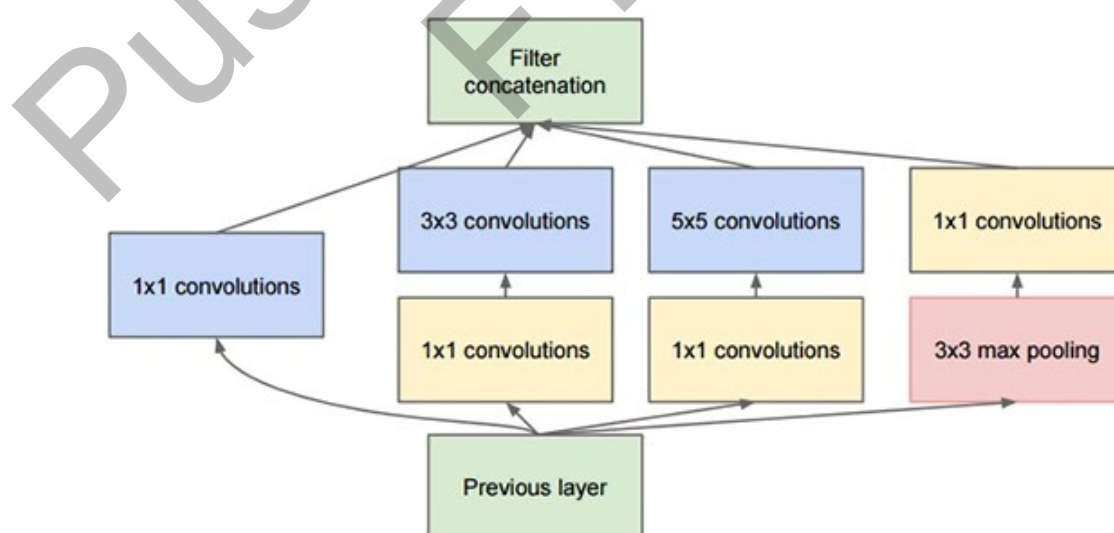


Figure 3.8 The Inception Module that was First Used

Source: Szegedy et al. (2014)

The Inception module's fourth and final branch executes 3×3 max pooling with a stride of 1×1 . Often referred to as the pool projection branch. Historically, models that utilise pooling have shown a capacity to achieve greater accuracy. In the instance of (Szegedy et al., 2014) this pooling layer was merely introduced since it was believed that it was necessary for convolutional neural networks to work relatively well. After then, the output of the pooling is fed into another set of 1×1 convolutions to learn about local characteristics.

Finally, all four branches of the Inception module converge at the channel dimension, where they are concatenated together. During implementation, special care is taken (through zero padding) to guarantee that each branch's output has the same volume size, allowing the outputs to be concatenated. The Inception module's output is subsequently sent to the network's next tier.

d. Mininception

The initial Inception module was created for GoogleLeNet with the goal of training it on the ImageNet dataset (with each input picture estimated to be $224 \times 224 \times 3$) and achieving state-of-the-art accuracy. The Inception module was streamlined for the datasets utilised in this thesis. The architecture consists of the following components:

First the convolution is applied, then a batch normalization, followed by an activation. Figure 3.6 shows the convolution module of mininception.

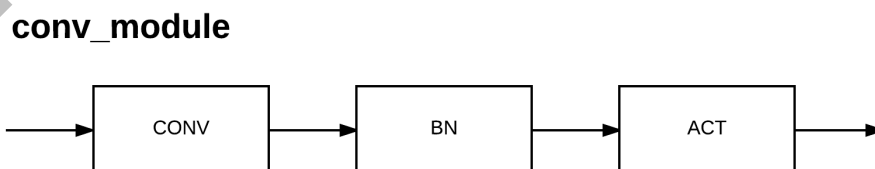


Figure 3.9 Convolution Module of Mininception

The Mininception module concatenates the outputs of two sets of convolutions, one for 1×1 filters and one for 3×3 filters (Rosebrock, 2017). No dimensionality reduction is performed prior to the 3×3 filter because (1) the input volumes will already

be reduced to minimise the number of network parameters. Figure 3.7 shows the down sample module of miniception.

downsample_module

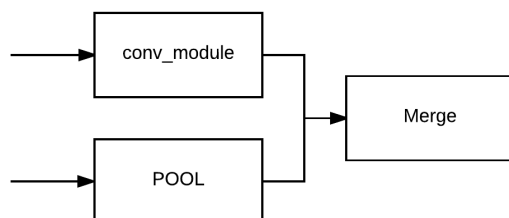


Figure 3.10 Down Sample Module of Miniception

The downsample module reduces dimensionality using convolution and max pooling, and then concatenates across the filter dimension. These construction pieces are then utilised to construct the bottom row's MiniGoogLeNet architecture. Figure 3.8 shows the MiniGoogLeNet model sample layers.

```

1 base_model3 = MiniGoogLeNet
2 Input: FER2013 dataset (35887 grey-scaled images)
3 1. Initialize parameters nb_class, x, y, epoch, lr, bs, c,
4 where nb_class = number of facial expression classes
5 x = height of the image
6 y = width of the image
7 epoch = number of iterations
8 lr = learning rate
9 bs = batch size
10 c = classifier
11 2. for 1: epochs
12   bs = 64 and lr = 1e-3 to 1e-5 // for FER2013 dataset
13   for 1: last_block_layer // base_model3
14     fv1 = generate feature vector for base_model3
15     acc1 = predict result of fv1 using classifier c
16   end for
  
```

Figure 3.11 MiniGoogLeNet Model Sample Layers

Source: Author

MiniGoogleNet

Conv (96)
Conv (32)
Conv (32)
Concatenate (64)
Conv (32)
Conv (48)
Concatenate (80)
Conv (80)
Concatenate (160)
Conv (128)
Conv (64)
Concatenate (192)
Conv (64)
Conv (32)
Concatenate (96)
AveragePooling (3)
Flatten (864)
Dense (7)

Figure 3.12 MiniGoogleNet Model Sample Layers

Source: Author