DEVELOPMENT OF AN EMOTION RECOGNITION SYSTEM BASED ON FACE IMAGES

TAN YI CHEN

UNIVERSITI SAINS MALAYSIA

2017

DEVELOPMENT OF AN EMOTION RECOGNITION SYSTEM BASED ON FACE IMAGES

by

TAN YI CHEN

Thesis submitted in partial fulfilment of the

requirements for the degree of

Bachelor of Engineering (Electronic Engineering)

JUNE 2017

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude towards my Final Year Project Supervisor, Assoc. Prof. Dr. Haidi Bin Ibrahim for his dedicated guidance and advices throughout the project. His patient guidance and useful critiques of this research work are much appreciated.

In addition, I would like to thank my research examiner Assoc. Prof. Dr. Bakhtiar Affendi Bin Rosdi. His feedback and suggestions helps me in tackling the problems better.

Lastly, I would like to show my appreciation to my family for their support and encouragement throughout my study. I would also like to take this opportunity to thank all people who have helped me in this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BU-3DFE	Binghamton University 3D Facial Expression Database
CK	Cohn-Kanade Database
CK+	Extended Cohn-Kanade Database
CNN	Convolutional Neural Network
DISFA	Denver Intensity of Spontaneous Facial Action Database
FER	Facial Expression Recognition
FER2013	The Facial Expression Recognition 2013 Database
GPU	Graphics Processing Unit
JAFFE	Japanese Female Facial Expression Database
KDEF	Karolinska Directed Emotional Faces Database
LBP	Local Binary Pattern
LBPV	LBP variance
LGBP	Local Gabor Binary Pattern
MLP	Multilayer Perceptron
PDM	Point Distribution Model
RAM	Random Access Memory
ReLU	Rectified Linear Units
ROI	Region of Interest
SFEW	Static Facial Expressions in the Wild Database
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines

PEMBANGUNAN SEBUAH SISTEM PENGECAMAN RIAK WAJAH BERDASARKAN GAMBAR WAJAH

ABSTRAK

Pengecaman riak wajah automatik mempunyai pelbagai applikasi seperti robot sosial, sistem tutor cerdas, sistem automasi rumah pintar dan applikasi lain untuk interaksi manusia-mesin. Oleh itu, penyelidikan dalam pengecaman riak wajah telah berkembang dengan pesat. Namun, kebanyakan pendekatan tidak membandingkan pengaruh cara-cara pembanyakan data dan kadar kesilapan masih lagi tinggi. Dalam projek ini, sebuah sistem pengenalan riak wajah berasaskan Rangkaian Neural Konvolusi telah dicadangkan. Sistem tersebut akan menyarikan semua maklumat berkaitan (seperti ciri-ciri) daripada gambar digital dua-dimensi dan mengelaskan gambar tersebut kepada salah satu riak semesta (seperti kegembiraan, kekejutan, kesedihan, kejijikan, ketakutan and neutral). Untuk meningkatkan ketepatan sistem yang dicadangkan, pelbagai pra-pemprosesan telah dilaksanakan. Data latihan telah dibanyakkan dengan menggunakan cara-cara seperti hingar garam-dan-ladah, hingar Gaussian, perubahan kecerahan dan pembalikan. Sistem tersebut dinilai dengan menggunakan pangkalan data yang banyak digunakan iaitu JAFFE dan CK+ dan kaedah-kaedah seperti kaedah pengesahan silang *k*-lipatan dan pengesahan silang pangkalan data. Ketepatan sistem yang dicadangkan mencapai 84.06% dengan menggunakan pangkalan data CK+ dan 77.59% dengan menggunakan kombinasi data dari pangkalan data CK+ and JAFFE. Sistem tersebut mencapai ketepatan yang tertinggi dengan menggunakan data yang ditambahkan dengan pembalikan,iaitu ditingkatkan dari 74.70% ke 77.17%.. Oleh itu, pembanyakan data telah dibuktikan bahawa dapat meningkatkan ketepatan

DEVELOPMENT OF AN EMOTION RECOGNITION SYSTEM BASED ON FACE IMAGES

ABSTRACT

Automatic facial expression recognition has vast applications such as sociable robots, intelligent tutoring system, smart home automation system and other human-machine-interaction applications. Thus, the research in facial expression recognition has been growing in interest. However, most approaches do not compare the impact of data augmentation methods and the overall error rate is still high. In this project, a facial expression recognition system based on Convolution Neural Network is proposed. The system extracts all relevant information (i.e., features) from two-dimensional digital facial image and classifies the image into one of the universal facial expression (i.e. happiness, surprise, sadness, disgust, fear, anger and neutral). To increase the accuracy of the proposed system, a number of pre-processing are carried out. The training data are augmented by using methods such as adding salt-and-pepper noises, Gaussian noise, brightness variations and flips. The system is evaluated by using widely used JAFFE and CK+ databases and methods such as *k*-fold cross-validation and cross-database validation. The proposed method achieved good result, which is 84.06% using CK+ database and 77.59% using combined data from CK+ and JAFFE databases. The system achieved the highest accuracy using data augmented with flips, which is increased from 74.40% to 77.17%. Therefore, data augmentation is proven that it can increase the accuracy.

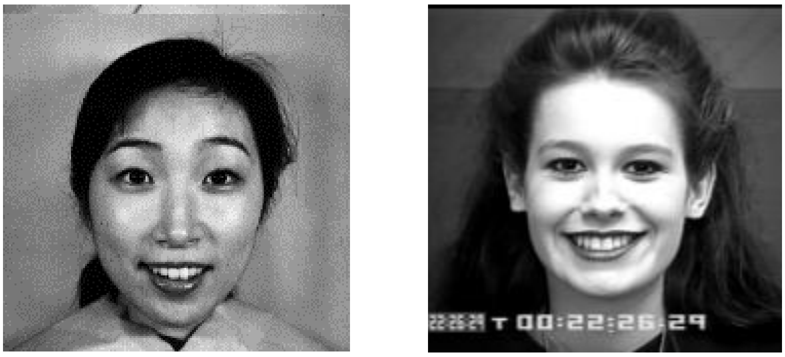
CHAPTER 1

INTRODUCTION

1.1 Background

Facial expression is one of the ways for human to express his emotion state aside from body language, action, speech, etc. Ekman and Friesen, (1971) identified and categorized six basic human emotion expressions which are happiness, sadness, anger, fear, disgust and surprise. The investigation also shows that particular facial behaviours are universally reflecting particular emotions. Besides, there are similarity traits in emotions between subjects of different age and gender.

However, the facial expression of each subject still differs from subject to subject. For example, Figure 1.1 shows subjects with happy expression from JAFFE and CK+. It can be observed that the facial expression of the images is different individually even if they are in the same emotion state. In addition, the images also vary in pose, intensity lighting, brightness, pose and background. Due to the variations, facial expression recognition cannot be performed by computers with ease yet while it still can be performed by humans effortlessly.



(a) (b)

Figure 1.1: (a) Subject with happy expression from JAFFE (b) Subject with happy expression from CK+

An automatic facial expression recognition (FER) system recognizes human's emotion or facial expression automatically. It had gained a lot of attention as it has a large variety of applications such as intelligent tutoring systems (Whitehill *et al.*, 2008), sociable robotics (Mavadati *et al.*, 2016), smart home automation system (Khowaja *et al.*, 2015), and other human-machine-interaction applications. In the intelligent tutoring system proposed by Whitehill *et al.*, (2008), the recognized student's facial expression is used as feedback signals of the learning process. It helps to detect the parts which cause students to feel confused, bored, excited. Thus, FER system is able to perform two important tasks during the lesson: to measure the difficulty as perceived by students of a delivered lesson and to detect the speed of lecture preferred by students. In smart home automation system proposed by Khowaja *et al.*, (2015), user's emotion is detected first, then home appliances are automated based on the user's need. Social robot proposed by Mavadati *et al.*, (2016) can be used for training children with Autism in recognizing facial expression using Robot-based Therapeutic Protocol.

There are two main categories for FER: a system that recognizes emotion from static images and a system that recognizes emotion from dynamic image sequences or video. Static-based methods use the feature vector consists of information about the current input image only while sequence-based methods use temporal information from the flow of images to recognize the emotion.

A lot of works has tried to develop automatic FER system with the purpose of achieving high recognition rate. One of the methods that show good result in recognizing facial expression is the neural network, a machine learning technique. In recent years, the computer can works with large databases as the computational power has drastically increased. Thus, this gives rise to the performance of the neural networks. Neural networks have the ability to learn features and classify the input data correctly after it is trained. One of the methods to train neural networks is stochastic gradient descent (Bottou, 2012). Lopes *et al.*, (2017), Liu *et al.*, (2015) and Krizhevsky *et al.*, (2012) used that method to train their neural networks model. Often, the ability of the neural networks to extract designated features depends on the amount of the training data: the accuracy of the network to extract features increases as the training data increases.

Convolutional Neural Network (CNN), which is another category of the standard neural network, is getting popular because good representations of images data can be extracted from images. The network is made up of neurons. Each neuron has learnable weights and biases, and it receives input and performs a dot product. Like other neural networks, CNNs also need a large training data for performing its designated function well. Often, CNNs detect input image features by using convolutional layer and sub-sampling layer. Convolutional layer performs convolution; Sub-sampling layer performs sub-sampling. According to Rashid, (2016), convolution is the repeated application of function across the region of the whole data set. This operation produces the map of features. On the other hand, max pooling is used in the sub-sampling operation to divide the map of feature into a group of sub-regions, which are no overlapping. In each sub-region, the maximum value is obtained. It helps reducing computation at the upper layers and responsible for translation invariance procedures.

Rectified Linear Units (ReLU) is often used as activation function of the neurons. It is a non-linear operation. It is an element-wise operation that applied per pixel. This operation replaces all negative pixel values in the feature map with zero to introduce non-linearity. The output of operation can be expressed as: Output = Max(zero, Input). Without this operation, the whole network will only act as a simple linear transformation which is unable to perform complicated tasks like digit recognition and facial expression recognition.

In a nutshell, the purpose of the project is to develop a high accuracy FER system based on CNN. Since building neural networks from scratch takes too long, the pre-trained framework such as ConvNet, Caffe, and TensorFlow are often used. The project could help in improving the well-being of the community, especially in the education field. Most of the student in Malaysia does not give feedback on the lessons even they faced difficulty in catching up a lesson. Thus, with a well-developed FER system, teachers could get real-time feedback from time to time in a lesson and adjust the speed of delivery based on the feedbacks. This could increase the quality of the delivery of lessons. On the other hand, the system could help in monitoring the emotion changes of the mentally ill patients so that doctors could plan more suitable treatments for them.

1.2 Problem Statement

In the research on automatic FER system, there are a lot of challenges and problems arise. Some of the problems had resolved. In developing and researching FER system, there is a problem to categorize human's facial expression (Rathod *et al.*, 2014). Ekman and Friesen, (1971) defined six facial expression categories as the basic emotions that are happiness, sadness, surprise, fear, anger, and disgust. Most approaches on FER are then categorized facial expression the same way as defined by Ekman and Friesen, (1971). Thus, the problem of categorizing facial expression is resolved.

However, there are challenges faced by FER system development. The recognition rate of facial expression by computer has yet achieved the level of human. It is still a challenge for computers to separate the facial expressions' feature space (Lopes *et al.*, 2017). Although facial expression by two subjects is very distinctive, the facial features can be very similar in facial spaces. Apart from that, for some subjects, expressions like "sad" and "fear" are similar. The approach by Lopes *et al.*, (2017) had shown that recognition rate of sadness and fear are lower and results in lower recognition rate.

Zhang *et al.*, (2016) had addressed high dimensionality problem in face recognition, gender and age estimation, and facial expression classification. The most significant discriminating features are still difficult to be identified even though many dimensionality reduction techniques have been proposed.

Fernandes *et al.*, (2016) had addressed the necessity of a robust method to deal with several problems in facial expression recognition such as translation, rotation, scale, orientation (pose), etc. Often, the performance of facial point extraction algorithm affected by the environmental factors such as lighting conditions heavily (Rathod *et al.*, 2014). Some methods have the ability to learn to deal with pose, environment and subject invariance but require large data set to train (e.g. CNN). Lopes *et al.*, (2017) addressed the problem of the small database (for training) could lead to lower recognition rate in CNN. Data augmentation is performed by approaches to enlarge databases but they do not compare the augmentation methods.

1.3 Objectives

The objectives of this project are:

1. To augment databases by synthesizing samples
2. To investigate the impact of data augmentation
3. To develop an accurate emotion recognition system based on face images using CNN

1.4 Scope of Research

The scope of the research is limited to review the previous successful approaches and to develop an FER system that recognizes emotion from a two-dimensional image with no rotation. The images used in the project are images from public facial expression database (i.e. JAFFE and CK+). For databases which contain sequences of images from neutral face to a specific peak expression, the first image in a sequence is used as "neutral" data while the image with peak expression is used as the data set for the emotion.

The FER system developed is based on CNN. TensorFlow, an open source software library for numerical computation is used in this project for developing the CNN model. Functions required for building CNN's layers such as the convolutional layer, pooling layer and fully connected layer are available.

The algorithm developed is to extract features from pre-processed input image and classify it into the correct facial expression. The impact of data augmentation methods on input image will be investigated. The developed algorithm will then be tested with common evaluation methods and databases for fairer comparison.

1.5 Thesis Outline

Chapter 2 presents literature review. It contains the general flow of FER system and common evaluation methods on FER system. On each stage of FER system, different methods and approaches are compared and reviewed.

Chapter 3 describes the methodology, which is focus on the development of FER algorithm. In this section, software and methods used on each stage (i.e. pre-processing, feature extraction, classification and evaluation) are discussed. The evaluation methods on the developed FER system are discussed.

Chapter 4 presents the results of this project. The accuracy of the developed FER system in different experiments is presented. Comparisons of the results with other approaches with similar evaluation methods are presented. The findings are discussed.

Chapter 5, the last chapter provides the conclusion of this project. It summarized the overall project's achievements. The future works for improvement are covered.

CHAPTER 2

LITERATURE REVIEW

2.1 General Flow of Facial Expression Recognition System

In general, the flow of FER system has four stages, which are face acquisition, pre-processing, feature extraction and representation, and classification. These stages are shown in Figure 2.1. In the first step, the FER system receives input data. The input data can be obtained from public databases which are discussed in Section 2.1.1. The subsequent step is face acquisition. It consists of face detection and head-pose estimation. The step is followed by pre-processing which involves information enhancement on the input images. Thereafter, features of facial expression presented in the images are extracted. The features can be extracted by using geometric feature-based methods (Fernandes *et al.*, 2016) or appearance-based methods (Zhang *et al.*, 2016). In geometric feature-based methods, the system works with the feature vector which represents shape and location of facial components such as eyebrows, eyes, nose and mouth; in appearance-based methods, the system works with feature vectors are extracted from the whole face. Lastly, facial expression recognition is achieved by classifying the feature obtained.

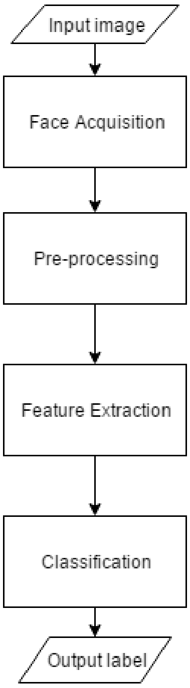


Figure 2.1: General flow of a facial expression system

2.2 Input of Facial Expression Recognition System

The input of an FER system can be an image or a sequence of images representing one of the emotions. Static-based FER system recognizes emotion from an input image and thus takes an image as the input. Sequence-based FER system takes a sequence of images as input and recognizes emotion from the change in facial expression from the flow of sequence. The input data can be obtained from the public databases. The common databases used in many approaches are discussed in the Section 2.2.1. To increase the performance, some approaches augment the data by synthesizing samples to enlarge the data set (Mollahosseini *et al.*, 2016; Lopes *et al.*, 2017) and to balance the data set (Rashid, 2016). The data augmentation methods are discussed in Section 2.2.2.

2.2.1 Public Face Databases

There are a lot of public databases available. The subjects in the databases are posing at various facial expressions. When benchmarking an algorithm, standard test databases are recommended to be used. The result can be compared directly with other approaches and thus better evaluation can be made. Databases are chosen based on the property to be tested too (e.g. BU-3DFE which contains data in different pose is used to evaluate pose invariant FER system).

The common databases used in many approaches are tabulated in Table 2.1. The table shows that most approaches use Cohn-Kanade (CK), Extended Cohn-Kanade (CK+) and Japanese Female Facial Expression (JAFFE) for training, experimenting and evaluating the models, followed by Binghamton University 3D Facial Expression (BU-3DFE) and MMI databases. The databases are briefly introduced.

Table 2.1: Databases Used in Facial Expression Recognition Approaches

Author	Database					
	JAFFE	CK+	CK	BU-3DFE	MMI	Others: *
Fasel, (2002)	✓	×	×	×	×	×
Maghami *et al.*, (2007)	×	×	✓	×	×	×
Shan *et al.*, (2009)	×	×	✓	×	×	×
Lai and Ko, (2014)	×	×	✓	×	×	×
Mayya *et al.*, (2016)	✓	✓	×	×	×	×
Mollahosseini *et al.*, (2016)	×	✓	×	×	✓	✓
Hernandez-Matamoros *et al.*, (2016)	×	×	×	×	×	✓
Zhang *et al.*, (2016)	×	✓	×	✓	✓	×
Lopes *et al.*, (2017)	✓	✓	×	✓	×	×
TOTAL	3	3	3	2	2	2

✓ = database used × = database not used

* = KDEF, CMU Multi-PIE, DISFA, GEMEPFERA, SFEW, and FER2013 databases

1. The Japanese Female Facial Expression (JAFFE)

JAFFE database (Lyons *et al.*, 1998) contains 213 grayscale images of 7 facial expressions (happy, sad, surprise, angry, disgust, fear and neutral) posed by 10 Japanese female models. The images are in TIFF (.tiff) format. The size is 256×256 pixels. The database is available at: http://www.kasrl.org/jaffe.html.

2. The Cohn-Kanade database (CK)

CK database (Kanade *et al.*, 2000) contains image sequences from 97 university students. The subject’s age are ranged from 18 to 30 years old. Among the subjects, 65 percent were female, 15 percent were African-American, and three percent were Asian or Latino. The database is available at: http://www.consortium.ri.cmu.edu/ckagree/.

3. Extended Cohn-Kanade database (CK+)

The CK+ database (Lucey *et al.*, 2010) contains 593 images sequences of 123 subjects. All sequences start from neutral face and end at a specific peak expression. The images are in PNG (.png) format. Most of the images size is about 640×490 pixels. The database is available at: http://www.consortium.ri.cmu.edu/ckagree/.

4. Binghamton University 3D Facial Expression (BU-3DFE)

BU-3DFE database (Yin *et al.*, 2006) contains 2500 facial expression 3D models and facial texture image captured at two views (about +45° and -45°) from 100 subjects. 56% of the subjects are female and 44% are male. Their age are ranging from 18 years to 70 years old with a variety of ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino. The database is available at: http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html.

5. MMI

MMI Facial Expression Database consists of more than 2900 videos and high-resolution still images of 75 subjects. The recordings contain the full temporal pattern of facial expressions from the neural expression, through a series of onset, apex, and offset phases and back to neutral face again. The database is available at: http://mmifacedb.eu/

It can be observed that the some databases are not large enough. For example, JAFFE contains only 213 static images while CK+ contains only 593 image sequences. Small training data set could affect the accuracy of an FER system (Lopes *et al.*, 2017). So, many approaches generate new images from the images from the databases by using data augmentation to enlarge the data set. The detail of data augmentation is discussed in the Section 2.2.2.

2.2.2 Data Augmentation

Data augmentation is a process to synthesize additional images indirectly from an original image. It can be done by changing its properties (i.e. rotation, brightness and etc.). It is carried out because public datasets available is not large enough or imbalance.

Mollahosseini *et al.*, (2016) and Lopes *et al.*, (2017) carried out data augmentation to increase the size of the training data. In the approach by Lopes *et al.*, (2017), the locations of the center of the eyes were changed according to the 2D Gaussian distribution with standard deviation, σ = 3 pixels and mean, μ = 0. Images with artificial rotations were generated. 70 additional images are synthesized using this method for each data. The synthetic data are used only in training phase. Figure 2.2 shows an illustration of the synthetic sample generation (Lopes *et al.*, 2017). On the other hand, in proposed method of Mollahosseini *et al.*, (2016), 10 additional images are synthesized from one image. Four regions at the corners and the center of the image are cropped to obtain the face images at different regions. Additional images are also synthesized by horizontally flip the images. The approaches augment data using one method only, and do not compare the impact of data augmentation with other methods.

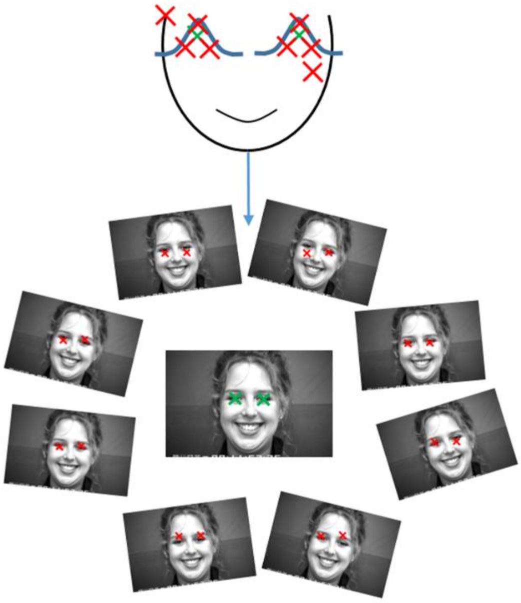


Figure 2.2: Illustration of the synthetic sample generation (Lopes *et al.*, 2017)

In the method by Rashid (2016), data augmentation is carried out for data balancing. According to the author, an imbalanced dataset's classification labels are not spread equally and this problem could act as a hinder issue for generalization. In his work, Synthetic Minority Oversampling Technique(SMOTE) (Chawla *et al.*, 2002) is used for balancing the dataset. Synthetic instances from the minority class are generated so that the number of the minority class is increased and approximately balance the majority class. This technique helps generalizes decision area for the minority class and this will improves the accuracy of the whole system.

2.3 Face Acquisition

In recognizing facial expression from image, especially image with complex scene, the information about the face’s location and head pose must be obtained before the other processes. This is achieved in face acquisition stage. In face acquisition stage, there are two steps: face detection and head pose estimation. An automatic face detector is used for face detection. As face analysis is affected by pose variations, information of the head pose face is also obtained and further processed if necessary.

2.3.1 Face Detection

In facial expression recognition system, face detection is usually the first step. The location of the face in the image is located. This step is important for the system which recognizes facial expression in the uncontrolled environment as the face location is unknown and therefore the system has to find the face region. There are approaches for facial expression recognition which performs face detection.

One of the methods to detect face is Viola-Jones algorithm. Hernandez-Matamoros *et al.,* (2016) developed an FER system which detects face image using the Viola-Jones algorithm which classifies images based on the value of simple features. Feature types used by Viola and Jones, (2001) is shown in Figure 2.3. Two-rectangle feature as shown in Figure 2.3(a) and Figure 2.3(b) computes the difference between the sums of the pixels within two rectangle regions; Three-rectangle feature as shown in Figure 2.3(c) computes the sum within two outside rectangles subtracted from the sum in a center rectangle; Four-rectangle feature as shown in Figure 2.3(d) computes the difference between diagonal pairs of rectangles. In short, the algorithm has four stages: Haar Feature Selection, creating an integral image, Adaboost Training, cascading training. The advantage of this algorithm is feature is scaled instead of the image. The disadvantage of this algorithm in face detection is it hardly copes with face image with rotation at 45 °both horizontally and vertically and thus effective for frontal face images only.

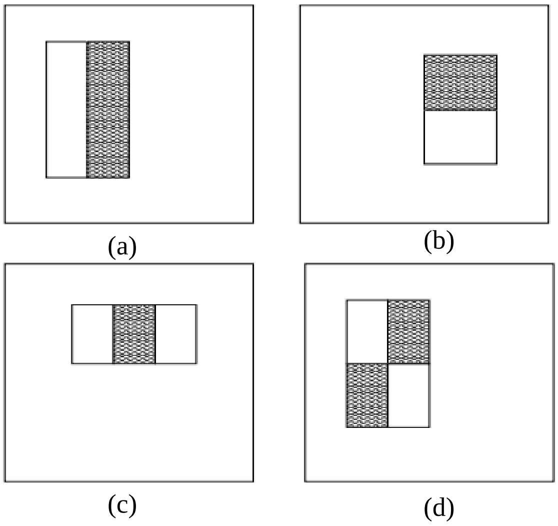


Figure 2.3: Feature types used by Viola and Jones, (2001)

Face detection can be performed by using the neural network. The method by Matsugu *et al.*, (2003) performed face detection by using CNN. Like other neural network models, the model has to be trained before it can perform its dedicated function. The CNN model is trained module by module. The sub-sampling layers are not trained, maximum value detection is performed by the layers instead. Skin area is obtained by thresholding hue data of input image in the range of [20.078, 0.255] for the full range of [20.5, 0.5]. The detection result of skin colour area is used as the input to the face detection module. The model is able to detect face present in images with complex scenes as shown in Figure 2.4 despites the difference of images in sizes, (from 30×30 to 240×240 in VGA image), pose (up to 30° in head axis rotation and in-plane rotation) and contrasts. The approaches proved that facial analysis system which utilizes CNN can deal with variability in position, size, and pose. The drawbacks of neural network are complexity in computation and training requirement.

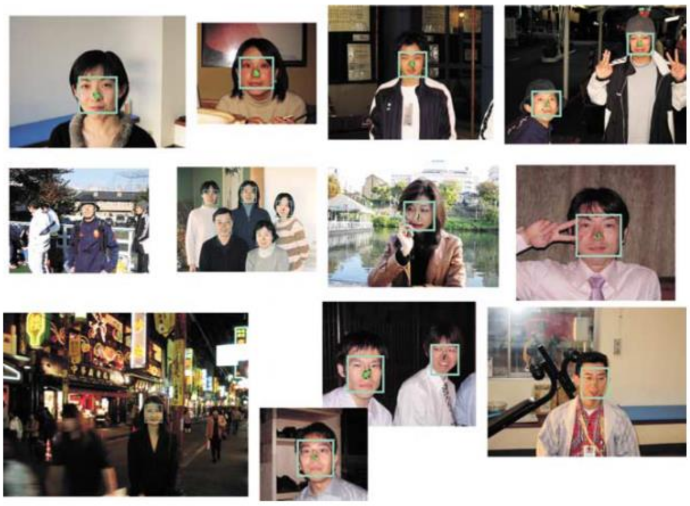


Figure 2.4: Face detection results in complex scenes (Matsugu *et al.*, 2003)

However, there are some approaches have no face detection included in the system. The approaches by Fasel, (2002), Rashid, (2016), Lopes *et al.*, (2017) and Zhang *et al.*, (2016) has no face detection. Thus, face images are given as input of the system instead. Additional information such as eye center locations may be needed. The approach by Lopes *et al.*, (2017) uses eye center locations information during preprocessing, training and testing of the system. Figure 2.5 shows the input image of the FER system with eyes center location given (Lopes *et al.*, 2017).

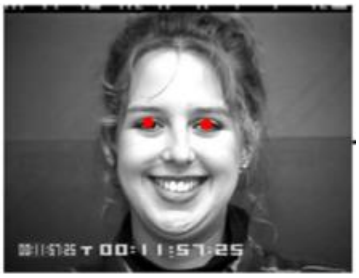
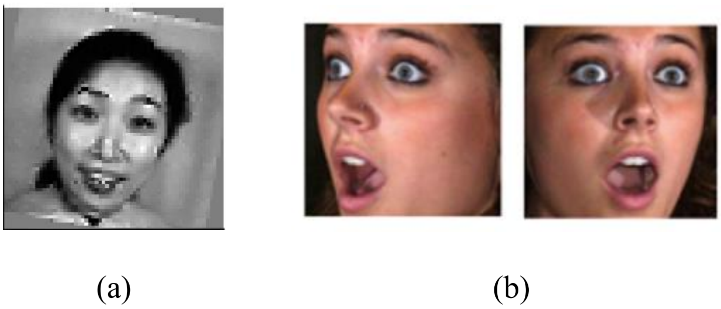


Figure 2.5: Input image of the FER system with eyes location given (Lopes *et al.*, 2017)

2.3.2 Head Pose Estimation

Head pose estimation is the next step to face detection. The head pose can be varied in term of head axis rotation and in-plane rotation in the images from databases, as well as real environments. It is important to estimation the head position before the next processes to the facial expression recognition system which extracts and classifies features with no rotation variation. In that case, the input images will be corrected to the right position for the model. Figure 2.6(a) shows face image with head axis rotation and

Figure 2.6(b) shows face image with in-plane rotation from BU-3DFE. In the method of Lopes et al. (2017), the head pose with head axis rotation is estimated by comparing the angle formed by the line segment from one eye center to another and the horizontal axis as shown in Figure 2.7.



(a) (b)

Figure 2.6: Example of head pose variation. (a) Head axis rotated face image (Fasel, 2002), (b) In-plane rotation face images from BU-3DFE

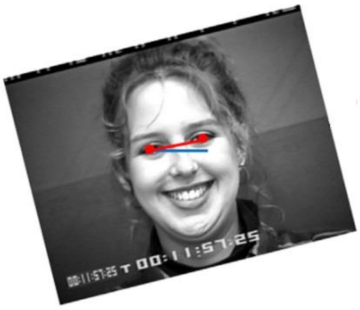


Figure 2.7: Head pose estimation (Lopes *et al.*, 2017)

2.4 Pre-Processing

Pre-processing in the FER system is a process to alter the properties of the images (in brightness, size, rotation, and etc.) before using them. Often, with suitable pre-processing involved, the accuracy of an FER can be increased drastically. Through pre-processing, noises and non-related features are removed. The work of Lopes *et al.*, (2017) proved the importance of pre-processing in CNN based FER system.

2.4.1 Rotation Correction

Rotation correction is the continued process of head pose estimation for correcting the head pose position. It is an important step for the FER system which its feature extraction method extracts features from in-plane images only. For example, method proposed by Lopes *et al.*, (2017) involved rotation correction. The information of the position of the center of both eyes is needed for this task. The angle between the line segment from one eye center to another and the reference, the horizontal axis of the image are determined. Then, rotation transformation is applied so that the angle is aligned to zero. Figure 2.8 shows an example of rotation correction.

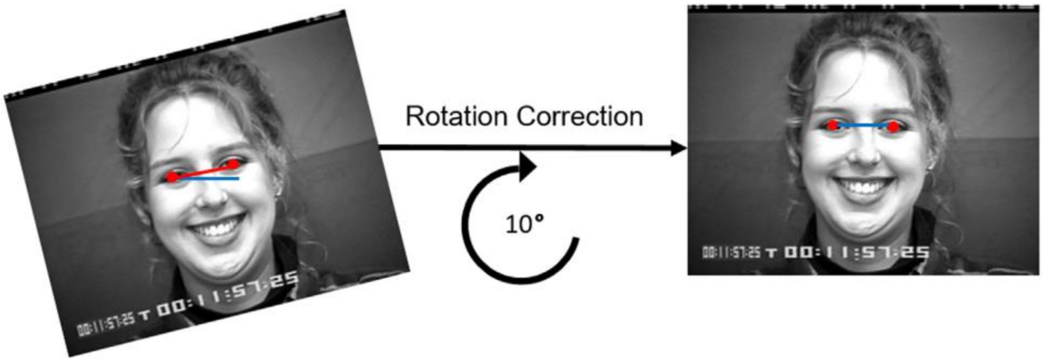


Figure 2.8: Rotation correction by Lopes *et al.*, (2017)

However, many approaches skipped this step as their FER systems are able to recognize facial expression from images which are out of plane. Head-pose invariant FER system by Fasel (2002) is one of the examples.

2.4.2 Image Cropping

Image cropping is a process to removes outer area of the selected area in an image. It is carried out to remove the unnecessary information in the background which could decrease the accuracy of the classification. The approaches by Mayya *et al.*, (2016), Lopes *et al.*, (2017) and Mayya *et al.*, (2016) performs image cropping in pre-processing.

In the approach of Lopes *et al.*, (2017), the region of the cropping is defined based on a ratio of the inter-eyes distance. The vertical cropping region is delimited by a vertical factor of 1.3 for the region above the eyes and 3.2 for the region below, a total vertical factor of 4.5. The horizontal cropping region is delimited by a factor of 2.4 (i.e. 1.2 for the left side and 1.2 for the right side). Figure 2.9 shows an example of image cropping using this method.

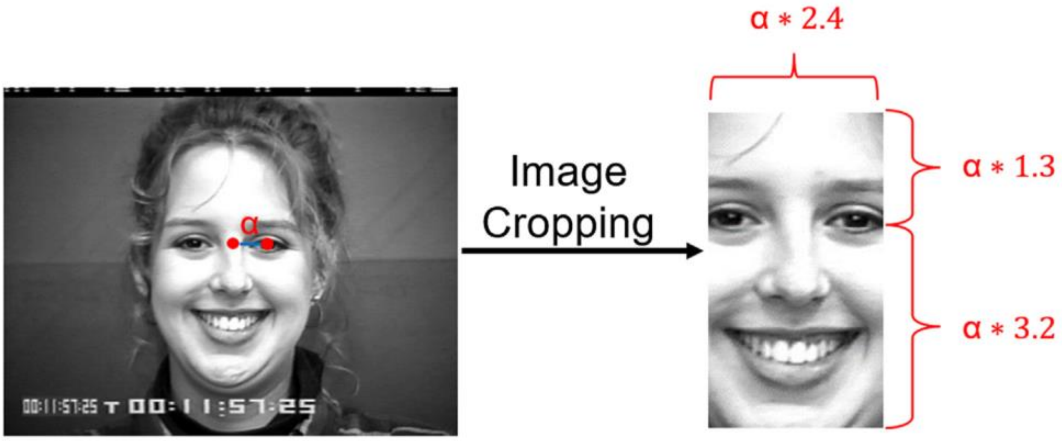


Figure 2.9: Image cropping by Lopes *et al.*, (2017)

In addition of cropping face segment found by using Viola-Jones algorithm as discussed in section 2.3.1, the method by Hernandez-Matamoros *et al.*, (2016) further crop the images with region of interest (ROI) segmentation algorithm to obtain image of the mouth segment and the forehead/eyes segment. The face region is divided into three horizontal regions (A, B and C) as shown in Figure 2.10. The forehead/eye segment is simply obtained from region A. Region C contains many other features other than the mouth. Thus, it cannot be considered as mouth segment. To obtain mouth segment, histogram equalization of the image obtained from region C is performed and the result is shown in Figure 2.10(c). The maximum value of the projective integral, D is obtained. By subtracting and adding the value D from the center point, the mouth segment is estimated and cropped as shown in Figure 2.10(b).

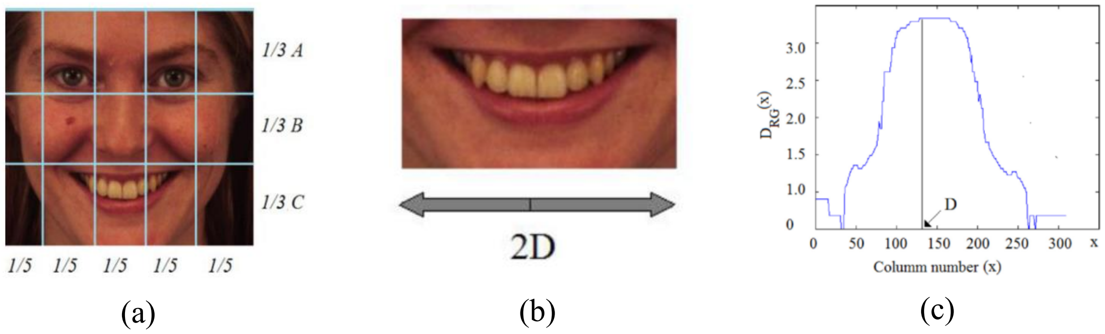


Figure 2.10: (a) Face segment division, (b) Mouth segment and (c) Horizontal projective integral of the mouth ROI (Hernandez-Matamoros *et al.*, 2016)

2.4.3 Intensity Normalization

In image processing, normalization is a process to change the range of pixel intensity values. In many FER approaches, intensity normalization is applied because the variation of the brightness and contrast increases the complexity of the problem to be solved by classifier (Lopes *et al.*, 2017).

In the method by Mayya *et al.*, (2016), the images were changed to grayscale. The procedures of the intensity normalization of methods by Lopes *et al.*, (2017) and Fasel (2002) are similar. In the method of Lopes *et al.*, (2017), subtractive local contrast normalization is performed. The process is followed by divisive local contrast normalization. A Gaussian-weighted average of every pixel's neighbour is subtracted from the value of it and the result is subsequently divided by the standard deviation of its neighbour. The neighbourhood is defined as the surrounding 7×7 pixels from the pixel. Figure 2.11 shows an example of intensity normalization by the author. The procedures described are expressed as Equation 2.1.

$$I_{normalized} = \frac{I_{in} - \mu_n}{\sigma_n} \tag{2.1}$$

where $I_{normalized}$ is the normalized pixel value, I_{in} is the original pixel value, μ_n is the Gaussian-weighted average of the neighbours of I_{in}, and σ_n is the standard deviation of the neighbours of I_{in}.

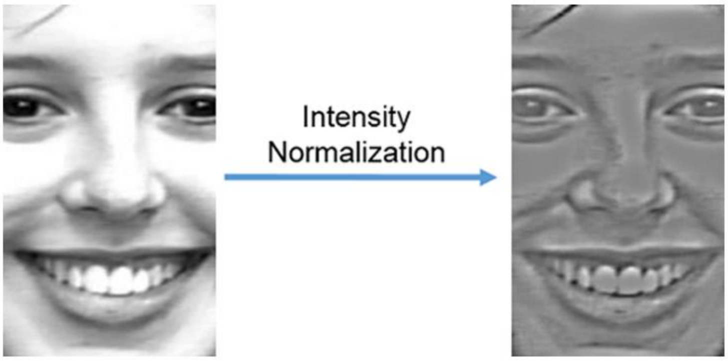


Figure 2.11: Intensity normalization by Lopes *et al.*, (2017)

On in other hand, in the proposed method of Fasel, (2002), mean value of the images, $\overline{I_{in}}$ is subtracted from every pixel of input face images, I_{in} in order to normalize the global lighting. Subsequently, the result is divided by the images' standard deviation,

σ_{in} to normalize the variances of the input variables so that the learning speed is increased. The procedures described are expressed in Equation 2.2.

$$I_{normalized} = \frac{I_{in} - \overline{I_{in}}}{\sigma_{in}} \tag{2.2}$$

2.5 Feature Extraction

After pre-processing, the processed image is fed into feature extractor. In this step, the features of facial expression are extracted from the image and represented in a numerical feature vector. Features are extracted by using the geometric feature-based and appearance-based methods as discussed in Section 2.1. An example of the geometric feature is facial landmarks (Fernandes *et al.*, 2016). Examples of appearance feature are Local Binary Pattern (LBP) (Zhang *et al.*, 2016) and Gabor Function (Hernandez-Matamoros *et al.*, 2016).

2.5.1 Point Distribution Model (PDM)

Fernandes *et al.*, (2016) used a geometrical approach for extracting features in facial expressions recognition. PDM is used to track the defined landmarks, which are the points on human face. Sets of points are representing shapes such as the contour of the face, eye, mouth, nose, lips and eyebrows. 66 landmarks as shown in Figure 2.12 are used by the author as these points are enough to describe every facial feature. After locating all the points required on the face, all Euclidean Distances between landmarks are calculated. As there are 66 landmarks, a total of 2145 distances are calculated for a face image.

There are different types of PDM which are Active Shape Models (ASM), Active Appearance Models (AAM) and Constrained Local Model (CLM). ASM applies Procrustes alignment algorithm for aligning the shapes described landmarks on databases and then applies a Principal Component Analysis (PCA) for building a linear shape model. AAM considers all texture over the whole face and builds a 2D triangulated mesh appearance model within the base mesh in addition of linear shape model. CLM is similar to AAM but only consider texture information around each landmark. PDM is able to overcome the problem of occlusion but it may require longer computation time.

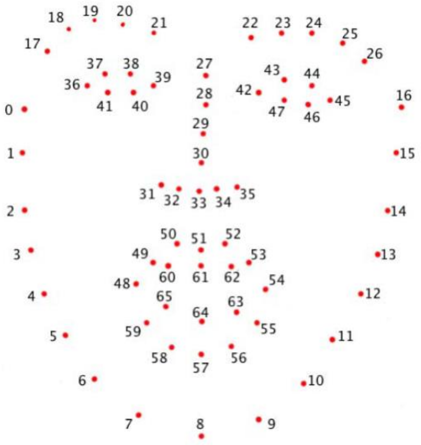


Figure 2.12: Landmarks or points used to represent facial features (Fernandes *et al.*, 2016)

2.5.2 Image Descriptor

One of the successful local image descriptor is LBP. The original LBP descriptor is introduced by Ojala *et al.*, (2002). It is used to label the pixels of an image block. The neighbourhood’s values are compared with the center value (in pixels) and result is produced in form of binary number.

Zhang *et al.*, (2016) proposed a new texture descriptor with the combination of LBP, Local Gabor Binary Pattern (LGBP) and LBP variance (LBPV) descriptors which handles illumination changes, rotations and scaling differences better. Each descriptor generates a sequence of binary representation from the 75 ×75 pixels grayscale test input image and the outputs are combined with three-parent crossover scheme to produce an off-spring.

LBP texture descriptor generates a sequence of binary output from thresholding each group of 3×3 neighbouring pixels against the center pixels. It is invariant to monotonic grayscale changes and extracts rotation invariant texture features from a local region efficiently. However, its neighbourhood contrast and global information for the texture description is lost. LGBP is the combination of Gabor filter with LBP which make representation and discriminates spatial facial information excellently. LBPV is invariant to rotation. It generates simplified joint representation by combining local spatial structure extracted by LBP and the contrast to further improve the discriminative capability of LBP.

LBPV and LGBP are chosen as the dominant parents because they have greater discriminating capability. When LBPV and LGBP disagree, LBP is used as a reference to generates offspring which has minor differences with the dominant parents but has greater discriminative capabilities. Figure 2.13 shows the outputs of all LBP operators using images from CK+ (first to the fifth column), BU-3DFE (sixth and seventh column) and MMI (eighth column). Images of the second and third column are derived from the original image in the first column with illumination changes while images of the fourth and fifth column are derived with scaling difference. It is shown that the proposed descriptor is greater in preserving the distinctiveness and differentiating local structures in the neighbouring pixels of the input image. The drawback of the proposed method is it involves 3 different visual descriptors and thus increased its complexity. Apart from that LBP is not robust on images with nearly uniform intensity as it is based on intensity differences.

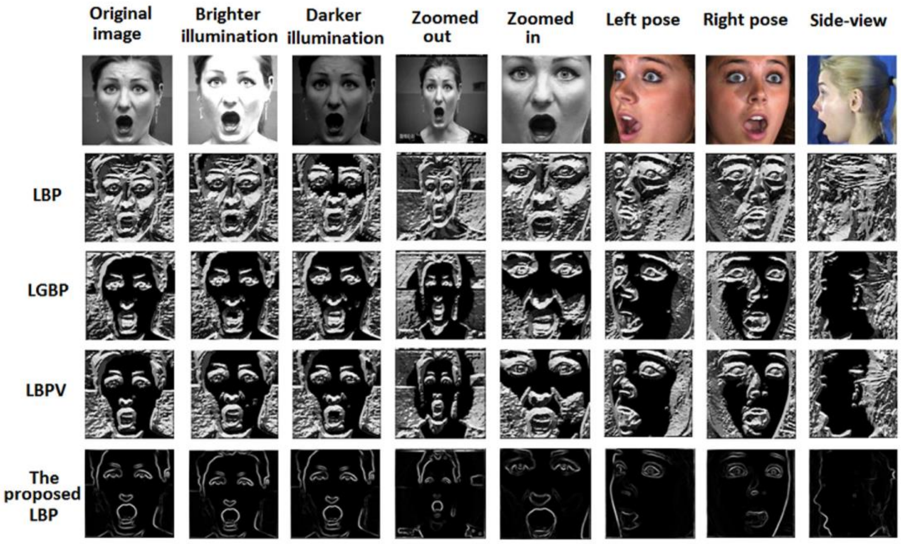


Figure 2.13: The outputs of all the LBP operators (Zhang *et al.*, 2016)

2.5.3 Convolutional Neural Networks

In CNN, feature extraction is carried out by convolutional layer to extract feature through convolution to obtain convolved feature. Kernel slides over the image and the dot product (convolved feature) is computed. Convolutional layers are characterized by the kernel's size and the number of generated maps (Lopes *et al.*, 2017). Often, pooling layer is used in between successive convolutional layers to reduce the spatial size of the representation. This reduces the amount of parameters and computation in the network.

There are a lot of approaches which perform feature extraction using CNNs. The CNN models utilize at least one convolutional layer and one sub-sampling layer. Fasel, (2002) and Lopes *et al.*, (2017) used CNN models which are simple but efficient in extracting features. Feature extraction is carried out by the first four layers in their models. There are some complex CNN based approaches. Matsugu *et al.*, (2003) utilizes modular CNNs for the task while Mollahosseini *et al.*, (2016) used a model with network-in-network CNNs. Mayya *et al.*, (2016) used a CNN model with ImageNet architecture (Krizhevsky *et al.*, 2012) for the task. The architecture of CNNs between models is distinctive from each other. The details of the models are described next.

Fasel, (2002) proposed two models of FER which extract features using CNNs. The first model is shown in Figure 2.14. The model has six layers: 2 feature extraction layers, 2 sub-sampling and two fully connected Multilayer Perceptron (MLP) layers. The kernel sizes are 5×5, 2×2, 11×11, 4×4 pixels for layer 1, 2, 3 and 4 respectively. The complex convolutional layer 3 is larger and not fully connected. The connection matrix is connected in the manner as shown in the top right corner of Figure 2.14. Larger kernel in the complex layer allows integration of features found in the preceding convolutional layer. The second proposed model is similar to the first model but it operates at three different resolutions: 5×5, 7×7 and 9×9 pixels to extract features at different sizes within a given object of interest. Figure 2.15 shows its architecture. For both models, the convolutional layers learn weight and allow dependent feature extraction while the sub-sampling layers increase the invariance of the object of interest's location dependence.

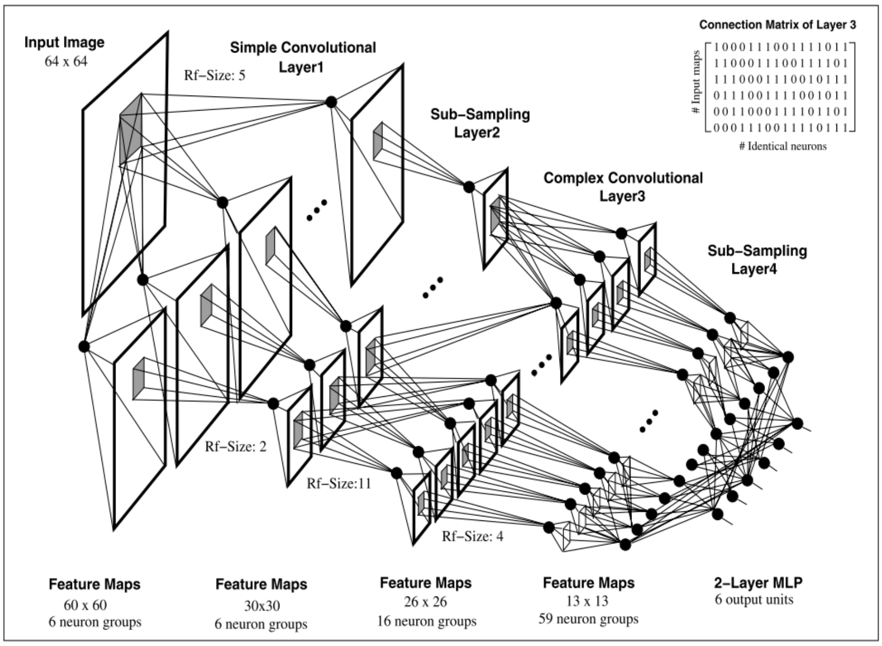


Figure 2.14: Architecture of a 5-layer convolutional neural network (Fasel, 2002)

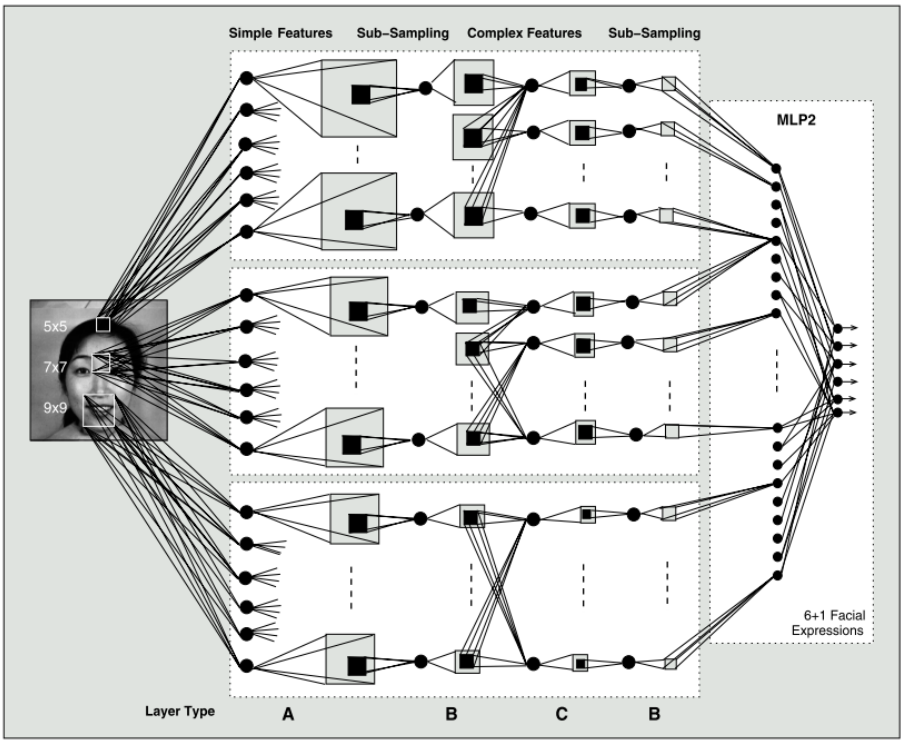


Figure 2.15: Multi-scale feature extraction convolutional neural network (Fasel, 2002)

The model used by Lopes *et al.*, (2017) has five layers as shown in Figure 2.16. The model extracts feature from input image with size of 32×32 pixels. The kernel size is 5×5, 2×2, 7×7 and 2×2 pixels for the first, second, third and fourth layer respectively. The last layer is a fully connected layer with 256 neurons. The model extracts elementary visual features like oriented edges, end-point, corners and shapes in general in the first layer (convolutional layer) and recognizing contextual elements (face elements) at third and fourth layer. Compared to the first layer, third and fourth layers handle features in a lower level. The sub-sampling layer, at layer two reduces the spatial resolution of the feature map. The layers mentioned are concatenated and as result, a high degree of invariance to geometric transformation of the input is achieved.

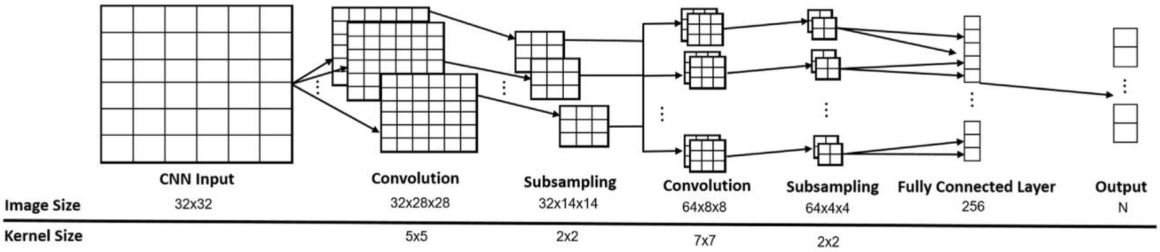


Figure 2.16: Architecture of the CNN model proposed by Lopes *et al.*, (2017)

The model used by Matsugu *et al.*, (2003) utilizes modular CNNs for extracting features. As shown in Figure 2.17, every module has its object of interest: local primitives, end stops and blobs, eyes and mouth and face for layer one, two, three and four respectively. The first layer extracts lower level features which are useful for facial expression recognition. Features like horizontal line segments and edge-like structures similar to step and roof edges representing parts of eyes, mouth, and eyebrows are extracted. This approach is complex than other CNN approaches as it consists of multiple modules but it is able to detect facial expression in complex scene.

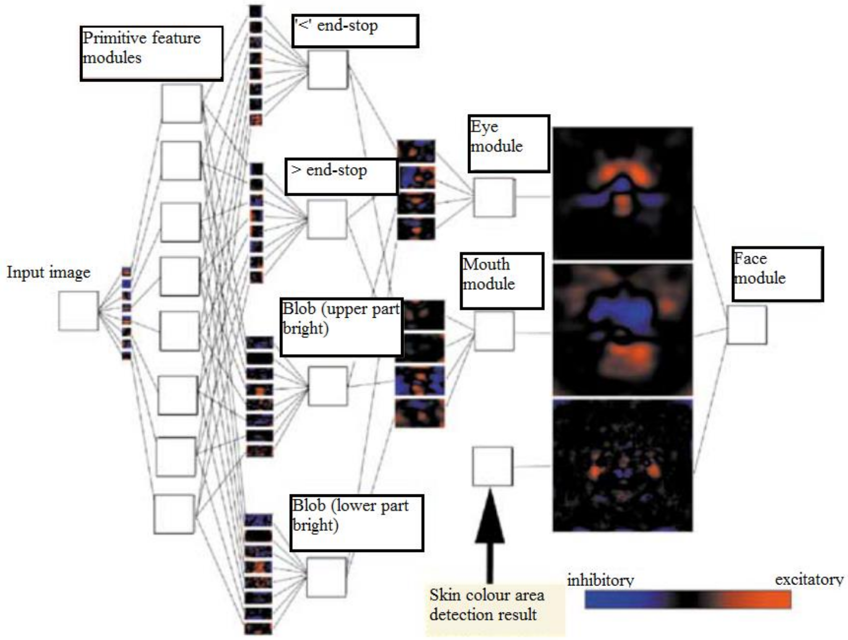


Figure 2.17: Convolution architecture of CNNs by Matsugu *et al.*, (2003)

Mayya et al. (2016) performed feature extraction using a 5-layers-CNN in their approach. The model consists of convolution, Rectified Linear Units (ReLU), Local Response Normalization (LRN) and pooling operations. 96 filters with size 11 × 11 × 3 pixels are used in this model. Feature extraction processes are described as followed: the first layer (convolution layer) extracts the low-level edge features. The following layer is ReLU. It increases the nonlinear properties of network. The third layer (pooling layer) subsampling the small rectangular blocks from previous layer and produces a single output. Max pooling process in this layer uses 3 × 3 pixels rectangular blocks with an interval of 2 pixels. LRN, the fourth layer normalizes the brightness to reduce irrelevant features and make relevant features more visible.

The advantage of CNN over other methods is information in the images is not destroyed by concentrating on limited regions or points of interest. The shared weight properties also allow CNN to be translation invariant but large training data set is needed. The drawback of CNN is large training data set is needed to be accurate.

2.6 Classification

Feature classification is the last stage of an FER system. Features which have been extracted are fed into the classifier. The classifier estimates facial expression presents in the input image based these features.

Facial expression classification can be done by using rule-based analysis and Support Vector Machine (SVM). Recently, many approaches classify facial expression by using neural networks such as Multilayer Perceptron (MLP) and CNN. The details of each method will be discussed in the following subtopics.

2.6.1 Rule-based Analysis

In rule-based analysis, classification is done based on the predefined conditions. Matsugu *et al.,* (2003) used a rule-based processing scheme for classifying facial expression. The cues (characteristic of features detected) are analyzed and saliency score of a specific facial expression is obtained. Examples of features used for classifying facial expression are the change in distance between end-stops such as the left corner of left eye and the left side end-stop of mouth) within facial components and changes in the width of line segments in lower part of eyes or cheeks. Each facial expression has a set of cues for recognition. For example, cues of happiness expression