Ain Shams University
Faculty of Computer & Information Sciences
Computer Science Department

Emotion Recognition

July 2021

Ain Shams University
Faculty of Computer & Information Sciences
Computer Science Department

Emotion Recognition

By:

Khaled Reda Ali [CS]
Khaled Gamal Mahmoud [CS]
Ahmed Mohamed Henawy [CS]
Al Amir Mahmoud [CS]
Islam Sayed Ezaat [CS]

Under Supervision of:

Prof ,Dr. Mostafa Aref
CS Department,
Faculty of computer and information science**,**
Ain Shams University.

Acknowledgement

All praise and thanks to ALLAH, who provided us the ability to complete this work. we hope to accept this work from us. We are grateful for our parents and our family who are always providing help and support throughout the whole years of study. We hope we can give that back to them. We also offer our sincerest gratitude to our supervisors, Prof. Dr. Mostafa Aref who have supported us throughout our thesis, with their patience, knowledge, and experience. Finally, we would thank our friends and all people who gave us support and encouragement.

Abstract

In real life scenario, facial expressions and emotions are nothing but responses to the external and internal events of human being. In human computer interaction, recognition of end user's expressions and emotions from the video streaming plays very important role. In such systems it is required to track the dynamic changes in human face movements quickly in order to deliver the required response system. The one real time application is physical fatigue detection based on facial detection and expressions such as driver fatigue detection in order to prevent the accidents on road. Face expression based physical fatigue analysis or detection is out of scope of this paper, but this paper reveal study on different methods those are presented recently for facial expression and/or emotions recognition using video. This paper presenting the methodologies in terms of feature extraction and classification used in facial expression and/or emotion recognition methods with their comparative study. The comparative study is done based on accuracy, implementation tool, advantages and disadvantages. The outcome of this paper is the current research gap and research challenges those are still open to solve for video based facial detection and recognition systems. The survey on recent methods is appropriately presented throughout this paper by considering future research works.

Keywords— Facial expressions, Frames, Emotions, Expressions, Fatigue, Feature Extraction, Classification

Table of Contents

List of Figures

List of Abbreviations

CNN/ ConvNet	Convolutional Neural Networks
DLA	Deep Learning Algorithm
HCI	Human Computer Interaction
HRI	Human Robot Interaction
FACS	Facial Action Coding System
FER	Facial Expression Recognition
GPU	Graphics Processing Unit
ML	Machine Learning
ReLU	Rectified Linear Unit
SIANN	Space Invariant Artificial Neural Network

1- Introduction

Emotion recognition has been researched for many years, dating back to the late 1800s. One of the first theorists to conduct research about the expression of emotion was Darwin. Darwin wrote a book describing how emotions can be involuntary in all animals and humans. He focused more on the biological reasons for displays of emotion, relating it to both animal and human species. This was the first real topic for literature critiques to look at within the field of emotion recognition. Our Human face is having a mixed emotion so we are to demonstrate the probabilities of these emotions that we have. We know that emotions play a major role in human life. At different kinds of moments or times, the Human face reflects that how he/she feels or in which mood he/she is. Humans are capable of producing thousands of facial actions during communication that varies in complexity, intensity, and meaning. Emotion or intention is often communicated by subtle changes in one or several discrete features. Nowadays the whole world seeks to improve user experience, despite rapid advances in Human-computer Interaction experience with computer systems, the need for agents to recognize and adapt to the affective state of users has been widely acknowledged. While being a critical component of human behavior, affect is nevertheless a highly subjective phenomenon influenced by a number of contextual and psychological factors including personality. Many effective studies have attempted to make personality recognition in several ways, some have employed explicit user feedback in the form of affective self-ratings while others have measured implicit user responses such as Electroencephalogram activity and heart rate for their analyses. This research attempts to extend research on facial expression recognition and psychopathic traits, specifically callousness, to an undergraduate sample. Emotion recognition is a technique used in software that allows a program to "read" the emotions on a human face using advanced image processing. Companies have been experimenting with combining sophisticated algorithms with image processing techniques that have emerged in the past ten years to understand more about what an image or a video of a person's face tells us about how he/she is feeling and not just that but also showing the probabilities of mixed emotions a face could have

Now days in many real time applications human computer interaction based systems are used to immediately and accurately track the human activities from the videos. One such area is realizing and tracking the human face expression and emotions

recognitions from the video streaming with an objective of different purpose such as physical fatigue detection. Before going to discuss about it more first we introduce about needs of facial emotion and expression recognition in upcoming sentences. In human-to-human conversation, the sound of mental, emotional, and even physical state is used in conversations about important information in addition to pronounce a communication channel and facial expressions is the notion of a person's facial expressions in its simplest form is a more subtle happy or angry thoughts , feelings or absorption of all speaker expectation from listeners, sympathy, or even what the speaker is saying no signal can provide to computing background, bring our everyday human user to remain at the forefront in the fabric will move to absorb that predict a generally establishment . It pervasive computing and ambient intelligence such as want to achieves for future computing. it's easy to naturally occurring multimodal human-human communication-focused result to the end user will need to developed to identify such interfaces and intentions and. as expressed by feelings of social and emotional signals will need to have the ability to sense future nonverbal actions and expressions. The automatic recognition Research inspired. Facial expression recognition, computer vision, pattern identification and human and computer interaction research has attracted towards notices in communities. automatic recognition of facial expressions so affective computing technologies, including intelligent tutoring systems supply the essence of the next generation computing device a form, patient monitoring systems, etc. personal wellness profiled. Human face, different age groups, genders and other physical characteristics of an individual differs from the cause. Emotions are basic to human beings in day-to-day interactions, and it use in everyday life. Emotion recognition has become an important and interesting field of study in Human-Computer Interaction (HCI), Human Robot Interaction (HRI), etc. The six basic emotions are, sickening, happy, fear, anger sad and surprise. Computer graphics, automatic driver fatigue detection, 3D/4D avatars animation in the entertainment industries, psychology, video & text chat and gaming applications are include in diverse applications. Recognition of emotions from facial expressions using videos consists of preprocessing, feature pulling out and division. Importance of facial expression system is widely recognized in social interaction and social intelligence system analysis is an active research topic since the 19th century. Suwa ET was introduced facial expression recognition in 1978 Al. creating a facial expression recognition system the main point of face detection and alignment Feature extraction and classification, image standardization.

The use of machines in society has increased widely in the last decades. Nowadays, machines are used in many different industries. As their exposure with human increase, the interaction also has to become smoother and more natural. In order to achieve this, machines have to be provided with a capability that let them understand the surrounding environment. Specially, the intentions of human being. When machines are referred, this term comprises to computers and robots. A distinction between both is that robots involve interaction abilities into a more advanced extent since their design involves some degree of autonomy. When machines are able to appreciate their surroundings, some sort of machine perception has been developed. Humans use their senses to gain insights about their environment. Therefore, machine perception aims to mimic human senses in order to interact with their environment. Nowadays, machines have several ways to capture their environment state trough cameras and sensors. Hence, using this information with suitable algorithms allow to generate machine perception. In the last years, the use of Deep Learning algorithms has been proven to be very successful in this regard. For instance, Jeremy Howard showed on his Brussels 2014TEDx's talk how computers trained using deep learning techniques were able to achieve some amazing tasks. These tasks include the ability to learn Chinese language, to recognize objects in images and to help on medical diagnosis. Affective computing claims that emotion detection is necessary for machines to better serve their purpose. For example, the use of robots in areas such as elderly care or as porters in hospitals demand a deep understanding of the environment. Facial emotions deliver information about the subject's interstate a machine is able to obtain a sequence of facial images, then the use of deep learning techniques would help machines to be aware of their interlocutor's mood. In this context, deep learning has the potential to become a key factor to build better interaction between humans and machines, while providing machines with some kind of self-awareness about its human peers, and how to improve its communication with natural intelligence

1.1 Motivation

Humans are capable of producing thousands of facial actions during communication that varies in complexity, intensity, and meaning. Many researches have suggested that the images used in the methods of assessment are becoming too familiar with the psychological research domain, so we suggest a new way of assessing the recognition of emotion in the human face. Many ideas grow better when transplanted into another mind than the one where they sprang up.

Interpersonal interaction is oftentimes intricate and nuanced, and its success is often predicated upon a variety of factors. These factors range widely and can include the context, mood, and timing of the interaction, as well as the expectations of the participants. Forgone to be a successful participant, one must perceive a counter part's disposition as the interaction progresses and adjust accordingly. Fortunately for humans this ability is largely innate, with varying levels of proficiency. Humans can quickly and even sub consciously assess a multitude of indicators such as word choices, voice inflections, and body language to discern the sentiments of others. This analytical ability likely stems from the fact that humans share a universal set of fundamental emotions. Significantly, these emotions are exhibited through facial expressions that are consistently correspondent. This means that regardless of language and cultural barriers, there will always be a set of fundamental facial expressions that people assess and communicate with. After extensive research, it is now generally agreed that humans share seven facial expressions that reflect the experiencing of fundamental emotions. These fundamental emotions are anger, contempt, disgust, fear, happiness, sadness, and surprise. Unless a person actively suppresses their expressions, examining a person's face can be one method

of effectively discerning their genuine mood and reactions. The universality of these expressions means that facial emotion recognition is a task that canal so be accomplished by computers. Furthermore, like many other important tasks, computers can provide advantages over humans in analysis and problem-solving. Computers that can recognize facial expressions can find application where efficiency and automation can be useful, including in entertainment, social media, content analysis, criminal justice, and healthcare. for example, content providers can determine the reactions of a consumer and adjust their future offerings accordingly. It is important for a detection approach, whether performed by a human or a computer, to have a taxonomic reference for identifying the seven target emotions. A popular facial coding system, used both by noteworthy psychologists and computer scientists such as Ekman and the Cohn-Kanade group, respectively, is the Facial Action Coding System (FACS). The system uses Action Units that describe movements of certain facial muscles and muscle groups to classify emotions. Action Units detail facial movement specifics such as the inner or the outer brow raising, or nostrils dilating, or the lips pulling or puckering, as well as optional intensity information for those movements. As FACS indicates discrete and discernible facial movements and manipulations in accordance to the emotions of interest, digital image processing and analysis of visual facial features can allow for successful facial expression predictors to be trained

1.2 Problem Definition

Human emotions and intentions are expressed through facial expressions and deriving an efficient and effective feature is the fundamental component of facial expression system. Facial expressions convey non-verbal cues, which play an important role in interpersonal relations. Automatic recognition of facial expressions can be an important component of natural human-machine interfaces; it may also be used in behavioral science and in clinical practice. An automatic Facial Expression Recognition system needs to solve the following problems: detection and location of faces in a cluttered scene, facial feature extraction, and facial expression classification.

1.3 Objective

The objective of the project is:

- To implement Convolutional Neural Networks for classification of facial expressions.
- Recognizing the human state (happy, sad, angry. etc.)
- Dealing with the facial expressions of human beings
- Improve the quality of human computer interaction
- Visualization of the results and performance analysis

1.4 Time Plan

1.4.1 planning phase:

in this phase we plan for project and knew problem definition and searching for dataset and write a proposal of the project

1.4.2 analysis and design phase:

in this phase read and study the related work to knew system requirements and create the design specifications

1.4.3 development phase:

in this phase we develop our module and test him to knew if it good or not then develop our model for our project.

1.4.4 release phase:

in this phase after building our model we test it and check if it have any issues and fix it till receive best model.

1.5 Document Organization

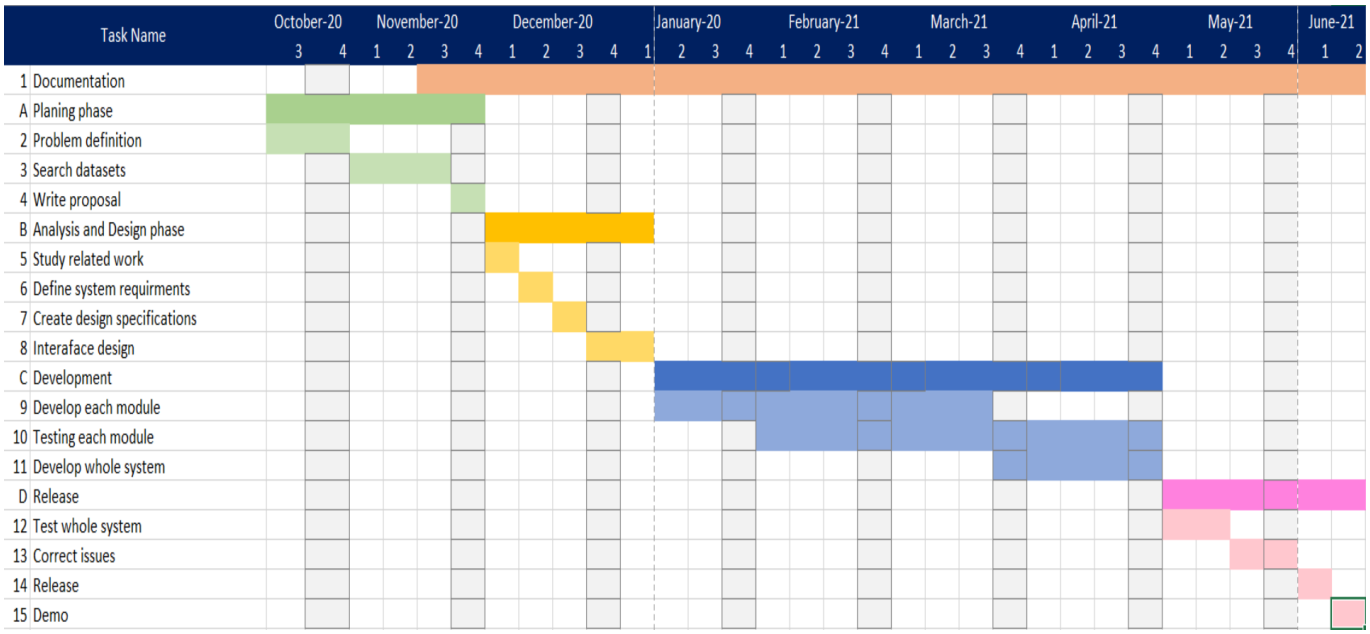


figure 1.5 Documentation organization

2- Background

A Facial expression is the visible manifestation of the affective state, cognitive activity, intention, personality and psychopathology of a person and plays a communicative role in interpersonal relations. Human facial expressions can be easily classified into 7 basic emotions: happy, sad, surprise, fear, anger, disgust, and neutral. Our facial emotions are expressed through activation of specific sets of facial muscles. These sometimes subtle, yet complex, signals in an expression often contain an abundant amount of information about our state of mind.

Automatic recognition of facial expressions can be an important component of natural human- machine interfaces; it may also be used in behavioral science and in clinical practice. It has been studied for a long period of time and obtaining the progress recent decades. Though much progress has been made, recognizing facial expression with a high accuracy remains to be difficult due to the complexity and varieties of facial expressions.

On day-to-day basics humans commonly recognize emotions by characteristic features, displayed as a part of a facial expression. For instance, happiness is undeniably associated with a smile or an upward movement of the corners of the lips. Similarly other emotions are characterized by other deformations typical to a particular expression. Research into automatic recognition of facial expressions addresses the problems surrounding the representation and categorization of static or dynamic characteristics of these deformations of face pigmentation.

In machine learning, a convolutional neural network (CNN, or ConvNet) is a type of feed- forward artificial neural network in which the connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli in a restricted region of space known as the receptive field. The receptive fields of different neurons partially

overlap such that they tile the visual field. The response of an individual neuron to stimuli within its receptive field can be approximated mathematically by a convolution operation. Convolutional networks were inspired by biological processes and are variations of multilayer perceptron designed to use minimal amounts of preprocessing.

They have wide applications in image and video recognition, recommender systems and natural language processing. The convolutional neural network is also known as shift invariant or space invariant artificial neural network (SIANN), which is named based on its shared weights architecture and translation invariance characteristics.

LeNet is one of the very first convolutional neural networks which helped propel the field of Deep Learning. This pioneering work by Yann LeCun was named LeNet5 was used mainly for character recognition tasks such as reading zip codes, digits, etc. The basic architecture of LeNet can be shown as below:

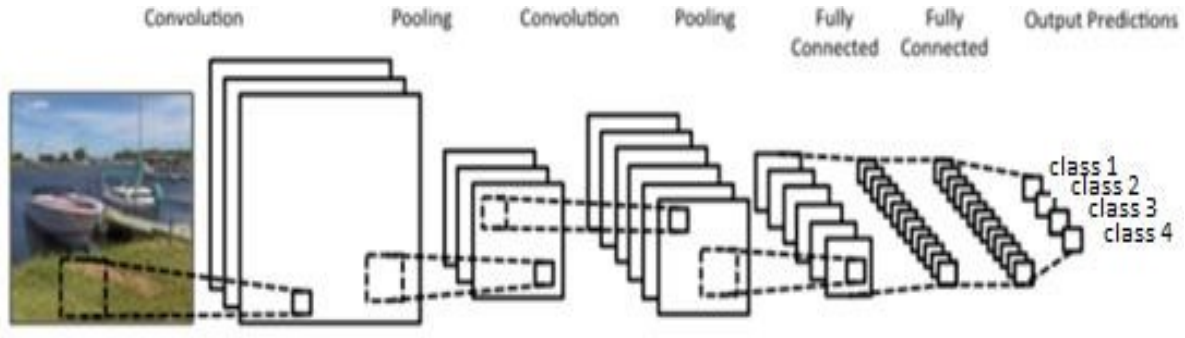


Figure 2.1: A model of CNN

There are four main operations in the Convolution Neural Network:

1- Convolution:

The primary purpose of Convolution in case of a CNN is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data. The convolution layer's parameters consist of a set of learnable filters. Every filter is small spatially (along width and height), but extends through the full depth of the input volume. For example, a typical filter on a first layer of a CNN might have size 3x5x5 (i.e. images have depth 3 i.e. the color channels, 5 pixels width and height). During the forward pass, each filter is convolved across the width and height of the input volume and compute dot products between the entries of the filter and the input at any position. As the filter convolve over the width and height of the input volume it produces a 2-dimensional activation map that gives the responses of that filter at every spatial position. Intuitively, the network will learn filters that activate when they see some type of visual feature such as an

edge of some orientation or a blotch of some color on the first layer, or eventually entire honeycomb or wheel-like patterns on higher layers of the network. Now, there will be an entire set of filters in each convolution layer (e.g. 20 filters), and each of them will produce a separate 2-dimensional activation map.

The 2-dimensional convolution between image A and Filter B can be given as:

$$C(i,j)=\sum Ma-1\sum Na-1\ A(m,n) * B(i-m,j-n) \quad (2.1)$$

where size of A is (Ma x Na), size of B is (Mb x Nb), 0 ≤ i <Ma+Mb-1^0 ≤ j < Na+Nb -1

A filter convolves with the input image to produce a feature map. The convolution of another filter over the same image gives a different feature map. Convolution operation captures the local dependencies in the original image. A CNN learns the values of these filters on its own during the training process (although parameters such as number of filters, filter size, architecture of the network etc. still needed to specify before the training process). The more number of filters, the more image features get extracted and the better network becomes at recognizing patterns in unseen images.

The size of the Feature Map (Convolved Feature) is controlled by three parameters

Depth: Depth corresponds to the number of filters we use for the convolution operation.

Stride: Stride is the size of the filter, if the size of the filter is 5x5 then stride is 5.

Zero-padding: Sometimes, it is convenient to pad the input matrix with zeros around the border, so that filter can be applied to bordering elements of input image matrix. Using zero padding size of the feature map can be controlled.

2- Rectified Linear Unit:

An additional operation called ReLU has been used after every Convolution operation. A Rectified Linear Unit (ReLU) is a cell of a neural network which uses the following activation function to calculate its output given x:

$$R(x) = Max(0,x) \tag{2.2}$$

Using these cells is more efficient than sigmoid and still forwards more information compared to binary units. When initializing the weights uniformly, half of the weights are negative. This

helps creating a sparse feature representation. Another positive aspect is the relatively cheap computation. No exponential function has to be calculated. This function also prevents the vanishing gradient error, since the gradients are linear functions or zero but in no case non-linear functions.

3- Preprocessing

- Data quality is essential to get good results
- The goal is to improve the quality of data to ensure that the measurements provided are:
    - Accurate
    - Precise
    - Complete
    - Correct
    - Consistent

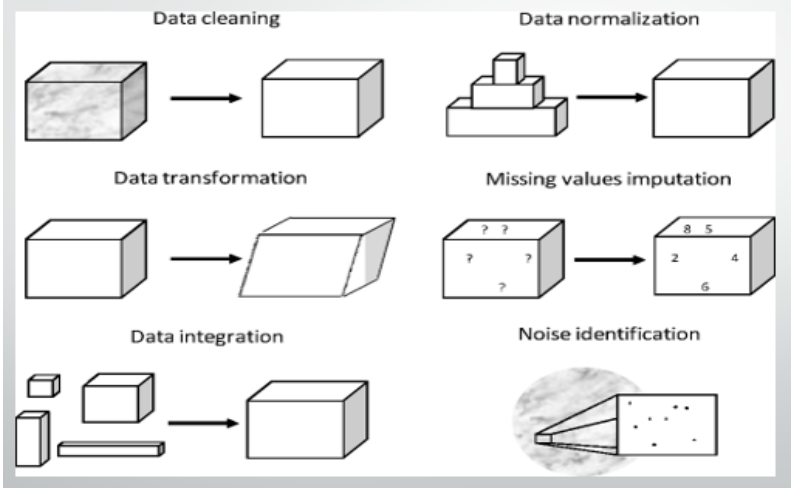


figure 2.2 preprocessing operations

- **Data Preprocessing Tasks**

  Apply sharpening filter

  - **Main objectives:**
    - Highlight transitions in intensitiesà highlight edges.
    - Remove blurring àenhance details.
    - Sharpening filters are based on spatial differentiation, which measures the rate of change of a function.
    - First and second derivatives are used for image enhancement.

Let's look at some details about First and second derivatives of sharpening:

- 1st Derivative:

The formula for the 1st derivative of a function is as follows:

$$\frac{\partial f}{\partial x} = f(x+1) - f(x)$$

figure 2.3

- 2nd Derivative:

The formula for the 2nd derivative of a function is as follows:

$$\frac{\partial^2 f}{\partial^2 x} = f(x+1) + f(x-1) - 2f(x)$$

figure 2.4

Second Derivative (Laplacian)

The Laplacian is defined as follows:

$$\nabla^2 f = \frac{\partial^2 f}{\partial^2 x} + \frac{\partial^2 f}{\partial^2 y}$$

figure 2.5

- Where in the x-direction:

$$\frac{\partial^2 f}{\partial^2 x} = f(x+1,y) + f(x-1,y) - 2f(x,y)$$

figure 2.6

- Where in the y-direction:

$$\frac{\partial^2 f}{\partial^2 y} = f(x,y+1) + f(x,y-1) - 2f(x,y)$$

figure 2.7

- Hence the Laplacian is given by:

$$\begin{aligned}\nabla^2 f = [&f(x+1,y) + f(x-1,y) \\ &+ f(x,y+1) + f(x,y-1)] \\ &- 4f(x,y)\end{aligned}$$

We can easily build a filter based on this.

Spatial Pooling (also called subsampling or down sampling) reduces the dimensionality of each feature map but retains the most important information. Spatial Pooling can be of different types: Max, Average, Sum etc. In case of Max Pooling, a spatial neighborhood (for example, a 2×2 window) is defined and the largest element is taken from the rectified feature map within that window. In case of average pooling the average or sum of all elements in that window is taken. In practice, Max Pooling has been shown to work better.

Max Pooling reduces the input by applying the maximum function over the input xi. Let m be the size of the filter, then the output calculates as follows:

$$M(x_i) = \max \{x_{i+k,+l} \; | \; |k| \leq m/2 \; , \; |l| \leq m/2 \; k, \; l \in \mathbb{N}\} \tag{2.3}$$

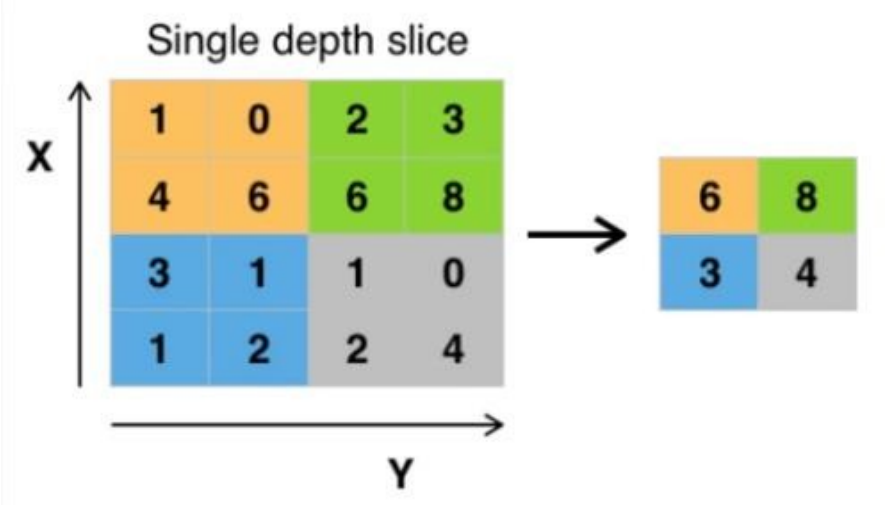


Figure 2.8: Max Pooling

The function of Pooling is to progressively reduce the spatial size of the input representation. In particular, pooling

Makes the input representations (feature dimension) smaller and more manageable

Reduces the number of parameters and computations in the network, therefore, controlling over-fitting

Makes the network invariant to small transformations, distortions and translations in the input image (a small distortion in input will not change the output of Pooling.

Helps us arrive at an almost scale invariant representation. This is very powerful since objects can be detected in an image no matter where they are located.

4- Classification (Multilayer Perceptron):

The Fully Connected layer is a traditional Multi-Layer Perceptron that uses a SoftMax activation function in the output layer. The term “Fully Connected” implies that every neuron in the previous layer is connected to every neuron on the next layer. The output from the convolutional and pooling layers represents high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset.

SoftMax is used for activation function. It treats the outputs as scores for each class. In the SoftMax, the function mapping stayed unchanged and these scores are interpreted as the un- normalized log probabilities for each class. SoftMax is calculated as:

$$f(z)j=\exp(zj) / \sum k \exp(zk)$$

where j is index for image and K is number of total facial expression class.

Apart from classification, adding a fully-connected layer is also a (usually) cheap way of learning non-linear combinations of these features. Most of the features from convolutional and pooling layers may be good for the classification task, but combinations of those features might be even better.

The sum of output probabilities from the Fully Connected Layer is 1. This is ensured by using the as the activation function in the output layer of the Fully Connected Layer. The SoftMax function takes a vector of arbitrary real-valued scores and squashes it to a vector of values between zero and one that sum to on**e**

3- Analysis and Design

3.1 System Overview

3.1.1 System Architecture

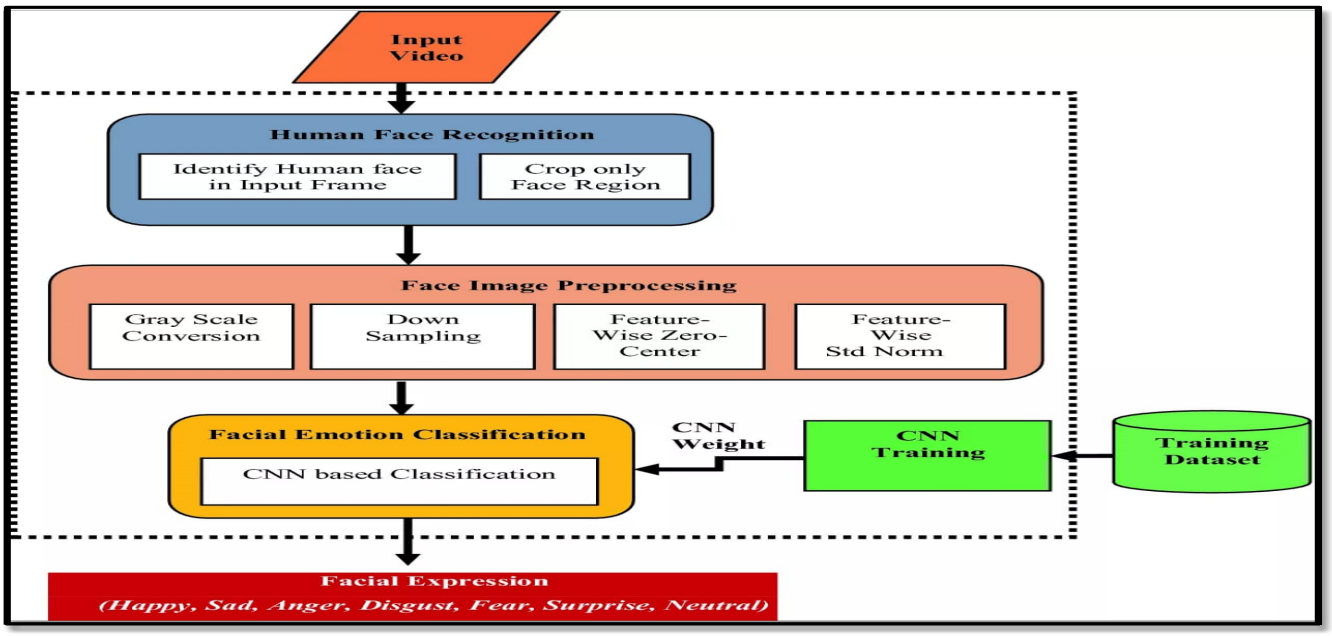


figure 3.1.1 system architecture

1- input: we start with laptop camera as input live stream

2- human face recognition: start to identify the human face from stream input and add a square in his face and don't look to any think in the screen

3- Face image preprocessing: in this phase we preprocess human face using a lot of image processing techniques like gray scale, down sampling ,feature-wise zero center...etc.

4- Facial Emotion Classification: we use our CNN classifier dependent to our trained dataset and his saved weights

5- output: recognize emotion if it (happy,sad,anger,disgust,Fear,surprise,netural)

3.1.2 System Users

A. Intended Users:

Our system is a service, not a product, we generate a system that makes many organizations benefits from this service in their work. Our aim is to improve the quality of human computer interaction in many fields.

B. User Characteristics

Developer: they need to know how to deal with the system.
Normal people: they only need to allow to open camera and do any emotion.

3.2 System Analysis & Design

3.2.1 Use Case Diagram

This is our use case as shown in the following figure:

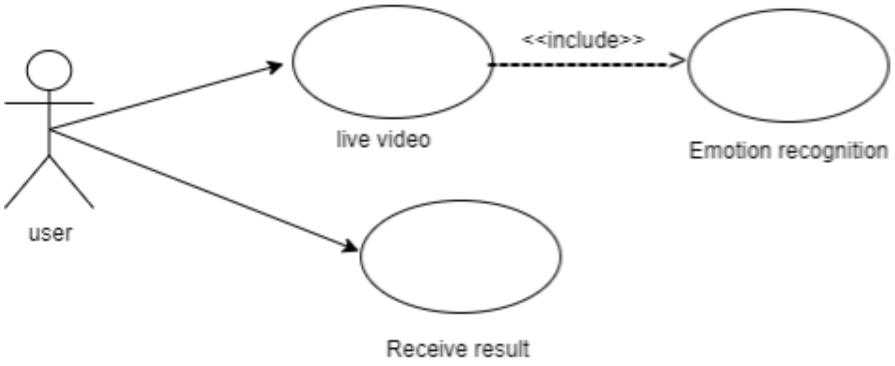


figure 3.2.1 use case diagram

functions:

1. Live video: start a live video using laptop camera and it is our input

2. Emotion recognition: using our model we classify our emotions from live video

3. Receive result: when human change his reaction from emotion to another emotion our model detects his emotion live in video streaming

3.2.2 Class Diagram

This is our class diagram as shown in the following figure:

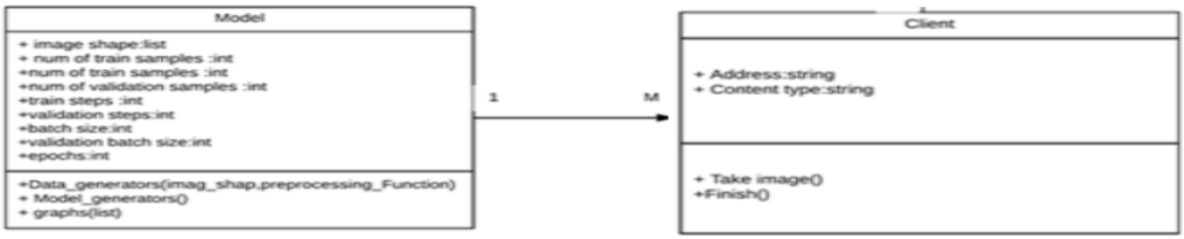


figure 3.2.2 class diagram

Description:

we have two main classes:

1- class model: it is having all work starting from preprocessing and training and testing till get the output

2- class client: in this it only how user use our project and get his output

3.2.3 Sequence Diagram

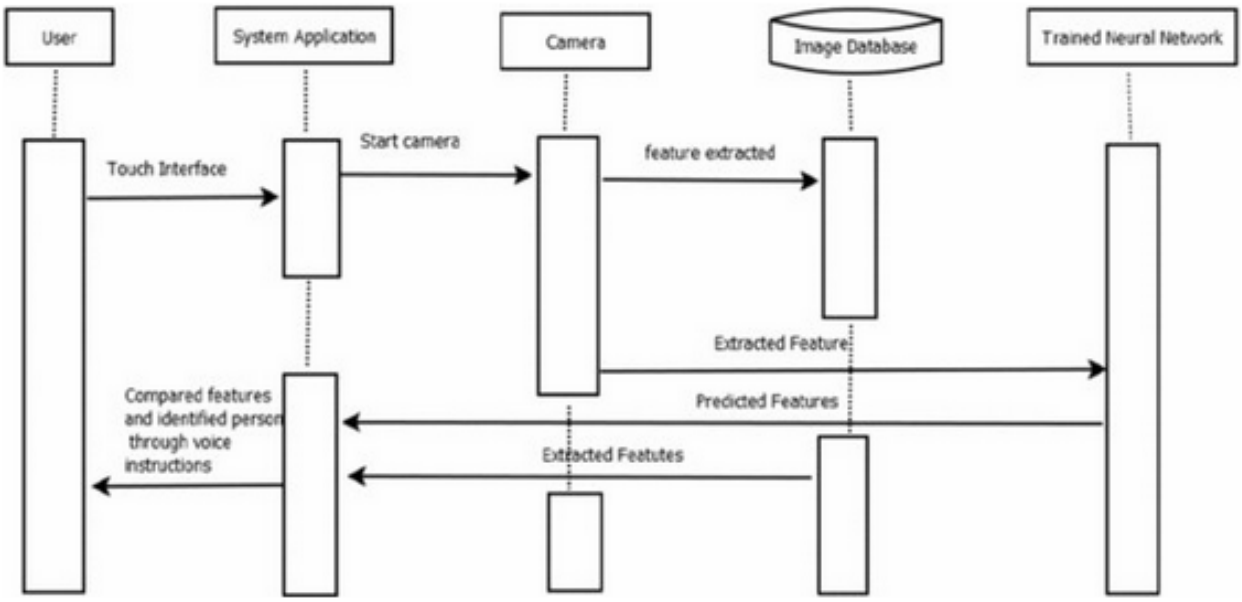


figure 3.2.3 Sequence Diagram

3.2.4 Activity Diagram

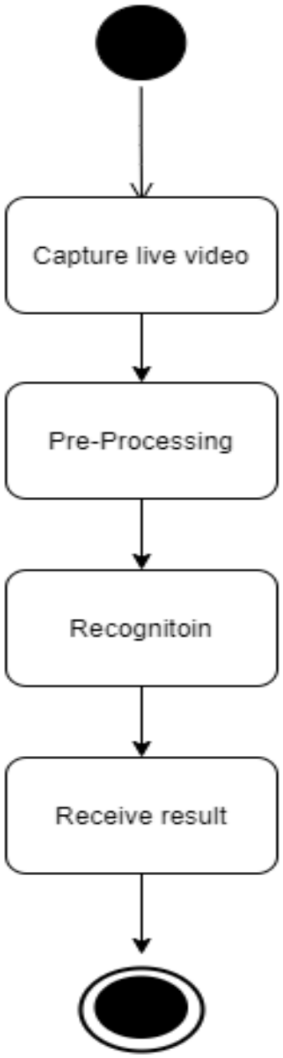


figure 3.2.4 Activity Diagram

Activities:

- start live video: receive image from user
- Preprocessing: do preprocessing functions on received image
- Recognition: recognize user emotion
- Receive result: send recognition result to user

4- Implementation and Testing

4.1 Proposed System

4.1.1 Overview of the proposed system

The overall flowchart of our proposed emotion recognition method.

- Generating emotion data set.
- In the pre-processing step, we work in sharpening images in the dataset.
- Normalization divides the dataset into an equal proportion to avoid overfitting.
- CNN developing a CNN model to classify the image.
- Testing and evaluation CNN model.
- Integrate the CNN model into API to make it easier to use.

And this figure shows our system overview:

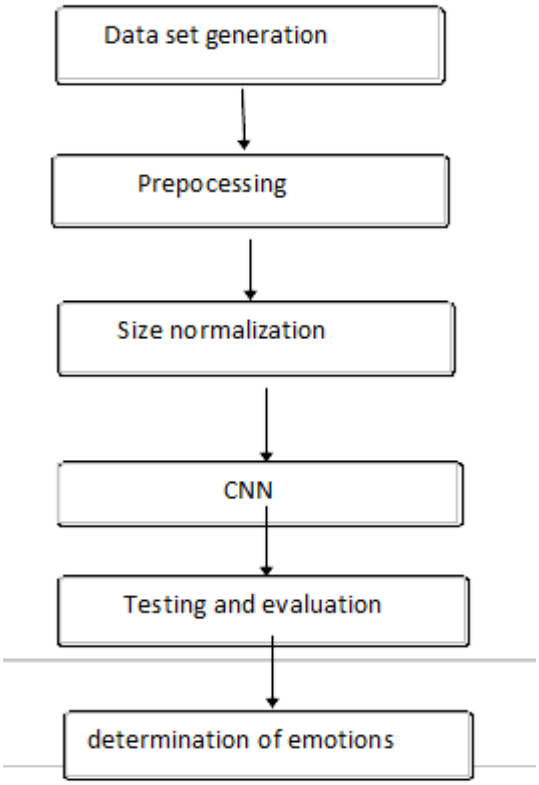


figure 4.1.1 showing system overview

4.1.2 Dataset generating

Emotion recognition dataset generation workflow

By recording videos extract frames from recorded videos by using python script and put the result in a suitable hierarchy to be suitable to CNN

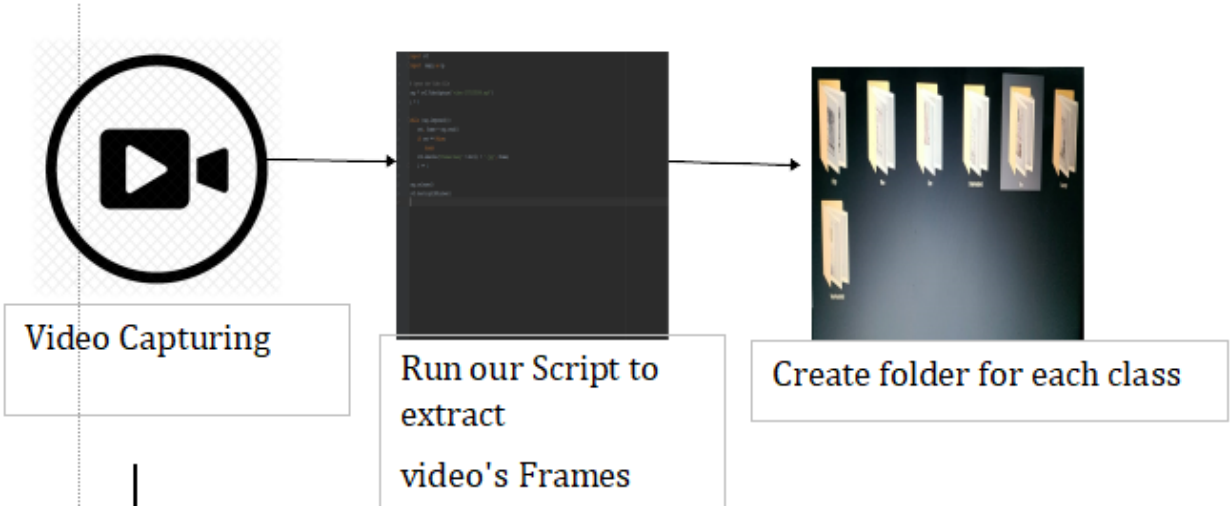


figure 4.1.2

train steps. Use image preprocessing algorithms to modify image quality. By using the Emotion recognition dataset generator, we succeed to generate and publish first Emotion recognition dataset with this size and

Normalize dataset size that we divide images into 7 classes based on emotion.

4.1.3 Convolution Neural Network

4.1.3.1 Convolutional neural network

Convolutional Neural Networks are Artificial Neural Network models that are comparable to the visual cortex, which is a fully connected, stratified network with each layer providing regions of cells that are sentient of specific fields of vision. This type of network is specifically designed for making full use of data with good spatial relations. The stratified topology of the convolutional neural network wheels out different properties of each layer, serialized as a collection of convolutional, activation, pooling, and fully connected layers, with the entire network comprising of at least one layer of convolution.

This convolutional layer takes advantage of the fact that input is made up of spatially related data, and have neurons arranged in 3 dimensions consisting of width, height, and depth of the activation volume. Mathematically, this layer computes a dot product between the weights of local receptive fields in the input and the connected region in the input volume and produces another array of numbers known as activation map or feature map. Each local receptive field converges into a hidden neuron connected to it in the succeeding layer. Since the output of the convolution produces a linear transformation of the input, it does not satisfy the universal approximation theorem which insinuates that the representational power of the network is coerced with linearity. Hence, for the network to comply with the universal approximator, the activation layer is required. The sole purpose of the activation layer is to infuse non-linearities in the network.

The output of this layer is then pooled or down sampled in the pooling layer to simplify the feature map produced by the preceding

layers. The pooling layer takes advantage of the fact that when a feature is extracted by the previous layers, the location of the feature in the feature map is not as important as its location compared to other detected features. This reduces the spatial size of the representation, consequently reducing the computational cost of the entire network and abating overfitting as well.

The entire composed stack is then consolidated into a fully connected layer, wherein each neuron is connected to every neuron in the preceding layer. The fully connected layer is usually one-dimensional and comprises all the labels that are to be classified. This layer outputs a score for each label of classification.

Since the convolutional networks are specifically designed to make use of spatial relations between the objects in the provided data, they perform better in regions of machine learning where the related spatial data needs to be manipulated, such as detection or recognition of objects. Also, as these convolutional networks explicitly assume that the inputs are images, it provides a streamlined function to encode the data and implement the network with immensely reduced parameters.

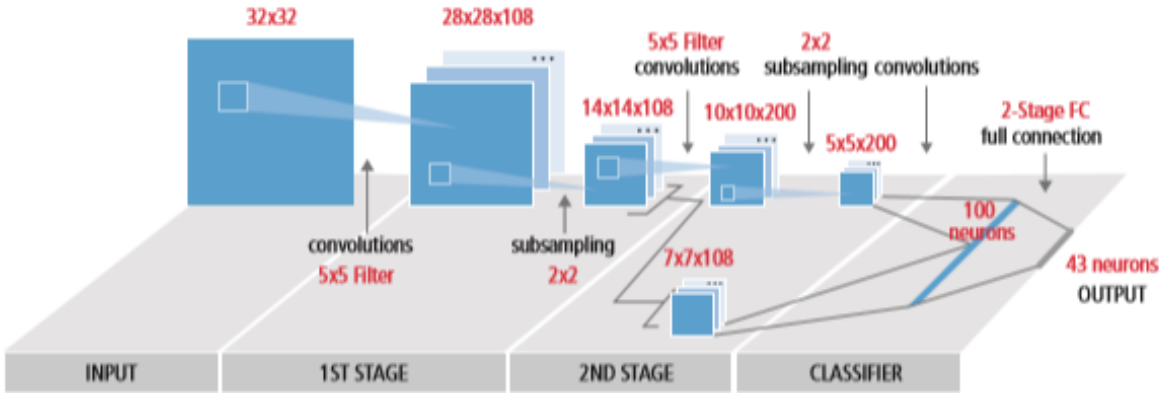


figure 4.1.3.1

4.1.3.2 Transfer learning

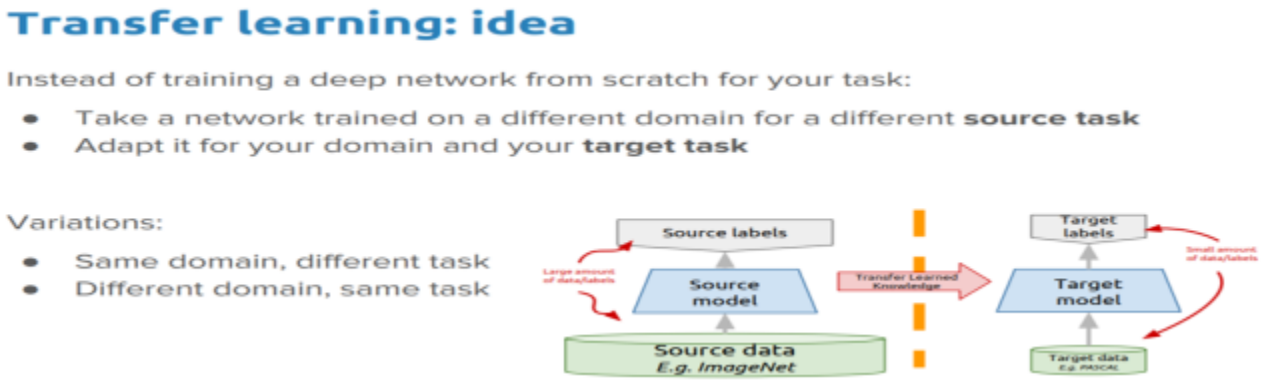


figure 4.1.3.2(1)

Fine-tuning off-the-shelf Pre-Trained Models

This is a more involved technique, where we do not just replace the final layer (for classification/regression), but we also selectively retrain some of the previous layers. Deep neural networks are highly configurable architectures with various hyperparameters. As discussed earlier, the initial layers have been seen to capture generic features, while the later ones focus more on the specific task at hand. An example is depicted in the following figure on a face-recognition problem, where initial lower layers of the network learn very generic features and the higher layers learn very task-specific features.

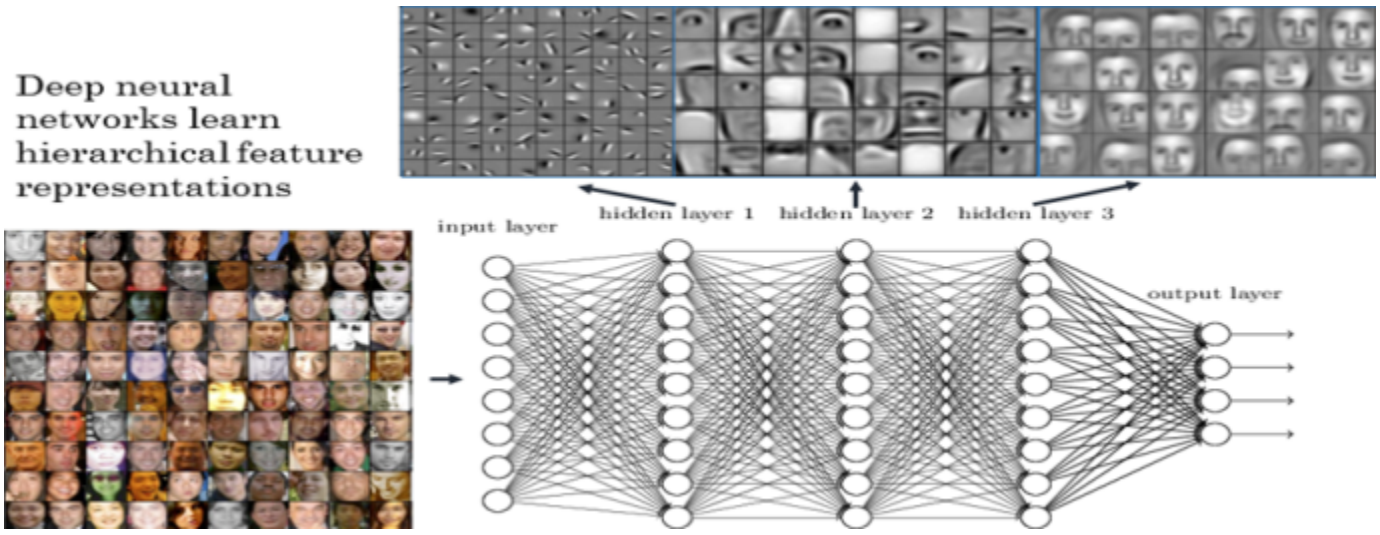


figure 4.1.3.2(2)

4.1.3.3 Select pre-train models

Selecting the most suitable Transfer Pretrained model to use, by using 3 models and train them with our data and measure its metrics. By try and test, we found that VGG16 has the best result than other models that have validation accuracy more than 92% and prediction time 0.06 minutes

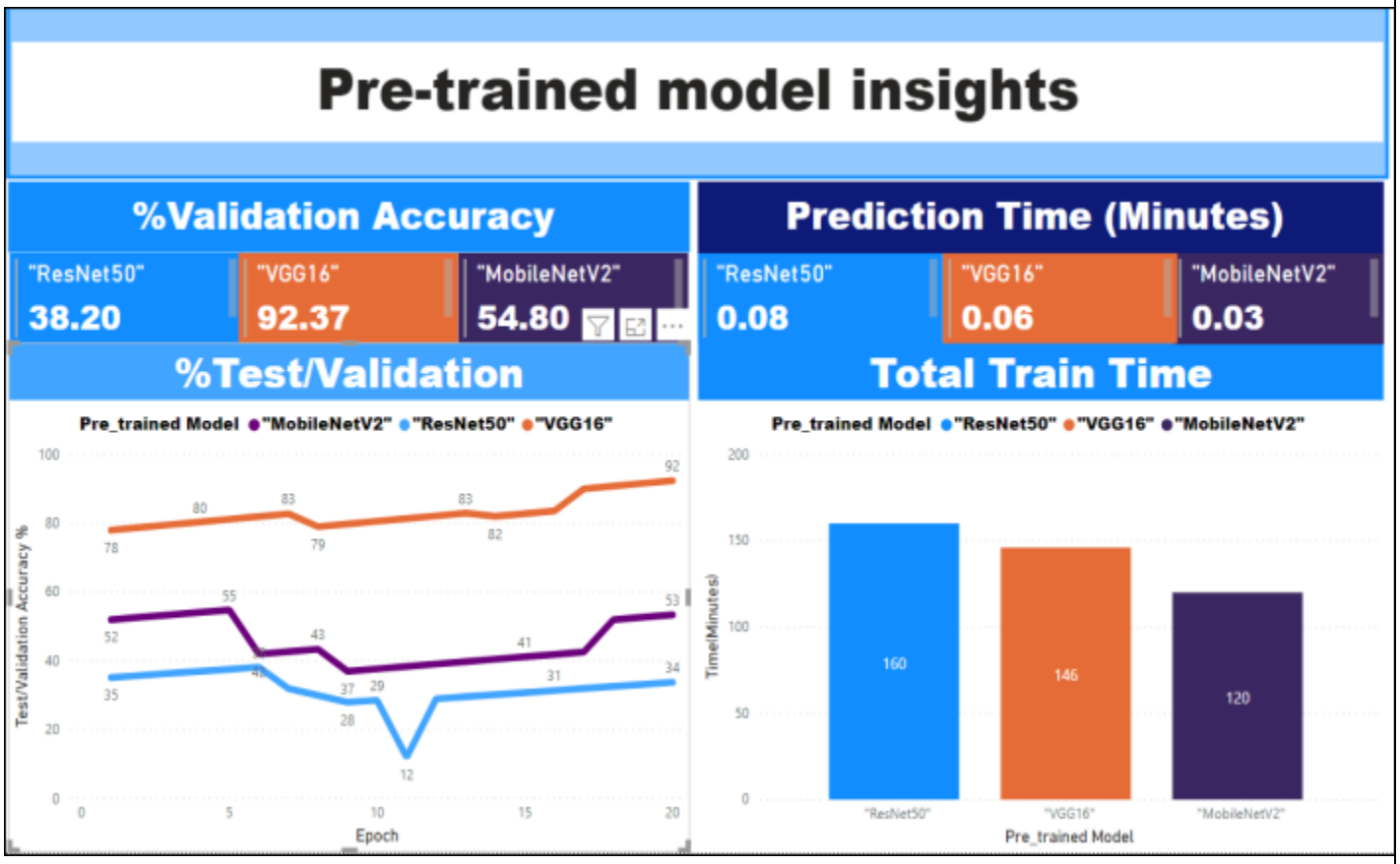


figure 4.1.3.3

4.1.3.4 VGG16

Architecture:

The input to the network is an image of dimensions *(256, 256, 3)*. The first layer has *32* channels of *3*3* filter size and the same padding. Then after apply activation function using elu then apply batch normalization then apply three step again and apply pool layer of stride *(2, 2) and* dropout, The second layer has *64* channels of *3*3* filter size and the same padding Then after apply activation function using elu then apply batch normalization then apply three step again and apply pool layer of stride *(2, 2) and* dropout, The third layer has *128* channels of *3*3* filter size and the same padding Then after apply activation function using elu then apply batch normalization then apply three step again and apply pool layer of stride *(2, 2) and* dropout , The fourth layer has *256* channels of *3*3* filter size and the same padding Then after apply activation function using elu then apply batch normalization then apply three step again and apply pool layer of stride *(2, 2) and* dropout , After that there are *2* sets of *3* convolution layers and a max pool layer. Each has *512* filters of *(3, 3)* size with the same padding. This image is then passed to the stack of two convolution layers. In these convolution and max-pooling layers, the filters we use are of the size *3*3* instead of *11*11* in Alex Net and *7*7* in ZF-Net. In some of the layers, it also uses *1*1* pixel which is used to manipulate the number of input channels. There is a padding of *1-pixel* (same padding) done after each convolution layer to prevent the spatial feature of the image, After the stack of convolution and max-pooling layer, we got a *(7, 7, 512)* feature map. We flatten this output to make it a *(1, 25088)* feature vector. After this there is *3 fully* connected layer, the first layer takes input from the last feature vector and outputs a *(1, 4096)* vector, the second layer also outputs a vector of size *(1, 4096)* but the third layer output a *1000* channels then after the output

of 3rd fully connected layer is passed to SoftMax layer to normalize the classification vector. After the output of classification vector top-5 categories for evaluation. All the hidden layers use ELU as its activation function.

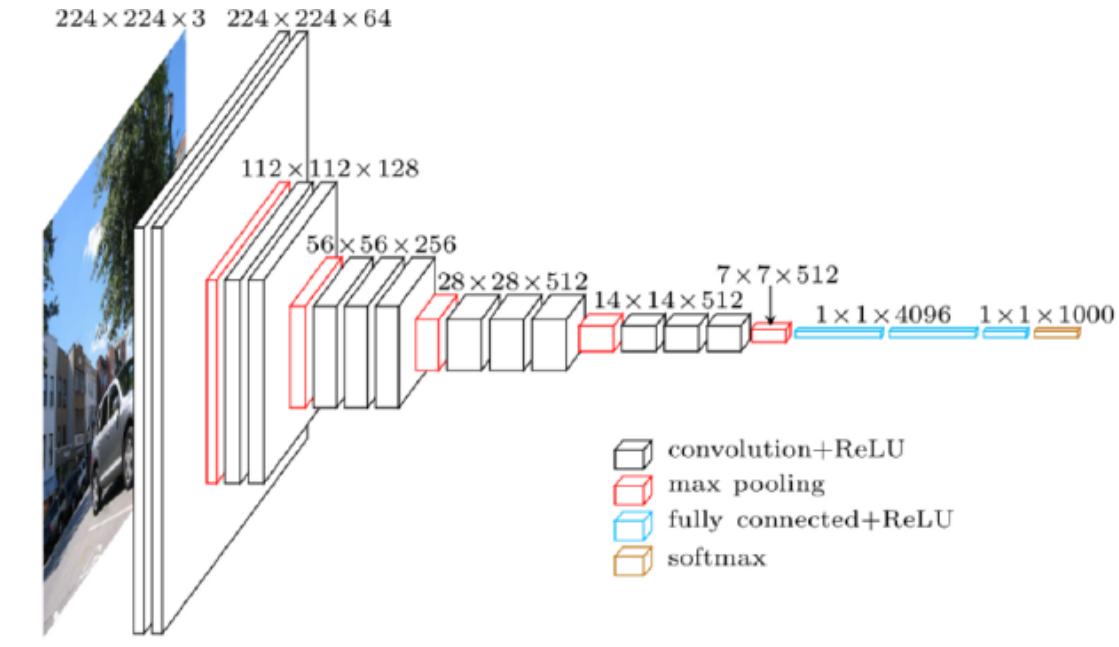


figure 4.1.3.4

4.1.3.5 VGG16 fine-tuning

By using VGG16 as our base model and did fine-tuning by removing fully connected layers and put 3 layers with 1024 neurons each and modified output layer and make it 14 neurons for each class, and retrain the model by using Emotion recognition dataset

4.1.4 Test and Evaluation

During the training phase, we divide dataset to 70% training and 30% testing and calculate validation accuracy in each epoch. The validation accuracy of our model reached 92% during the training phase.

4.1.5 Classification emotion type

After feeding an image into our model we calculate from the output layer from our model which neuron has the maximum percentage to be the right emotion type and return the class as text

figure 4.1.4

5- User Manual

- Download and install python

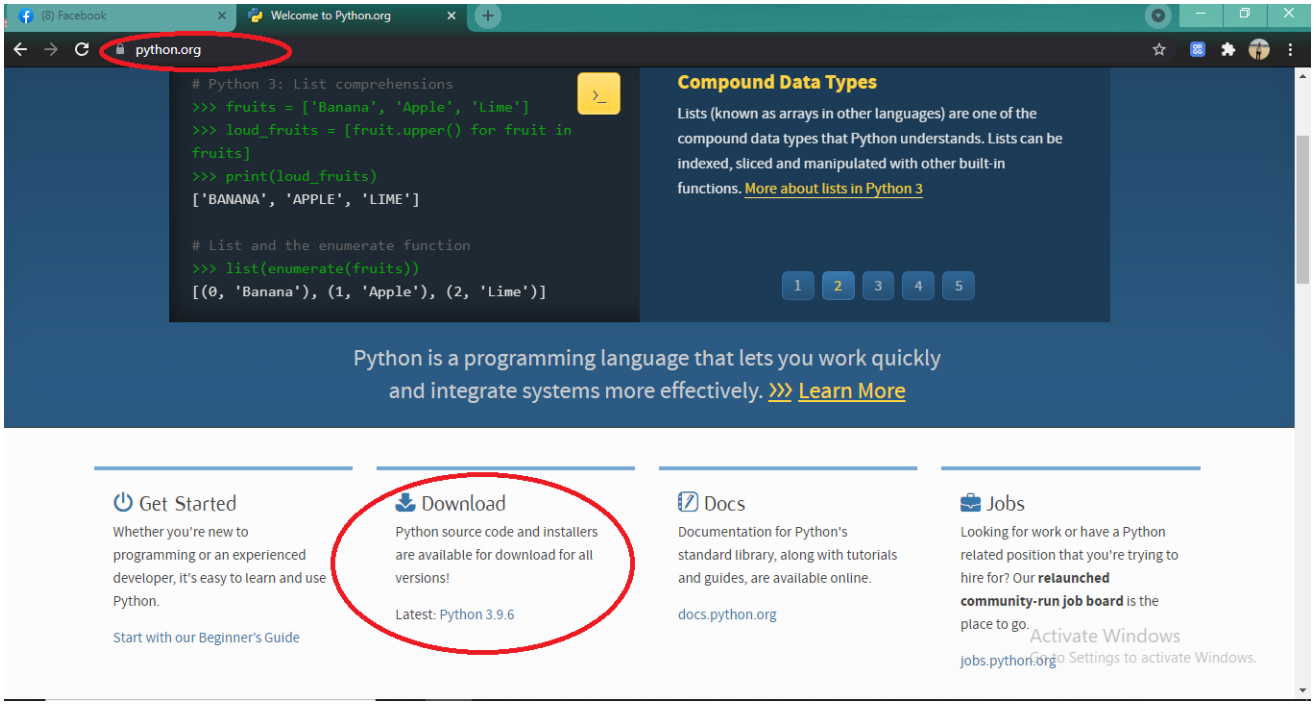


figure 5.1 - install python

- Download and install pycharm

figure 5.2 Install Pycharm

- Run the file “opencv.py”

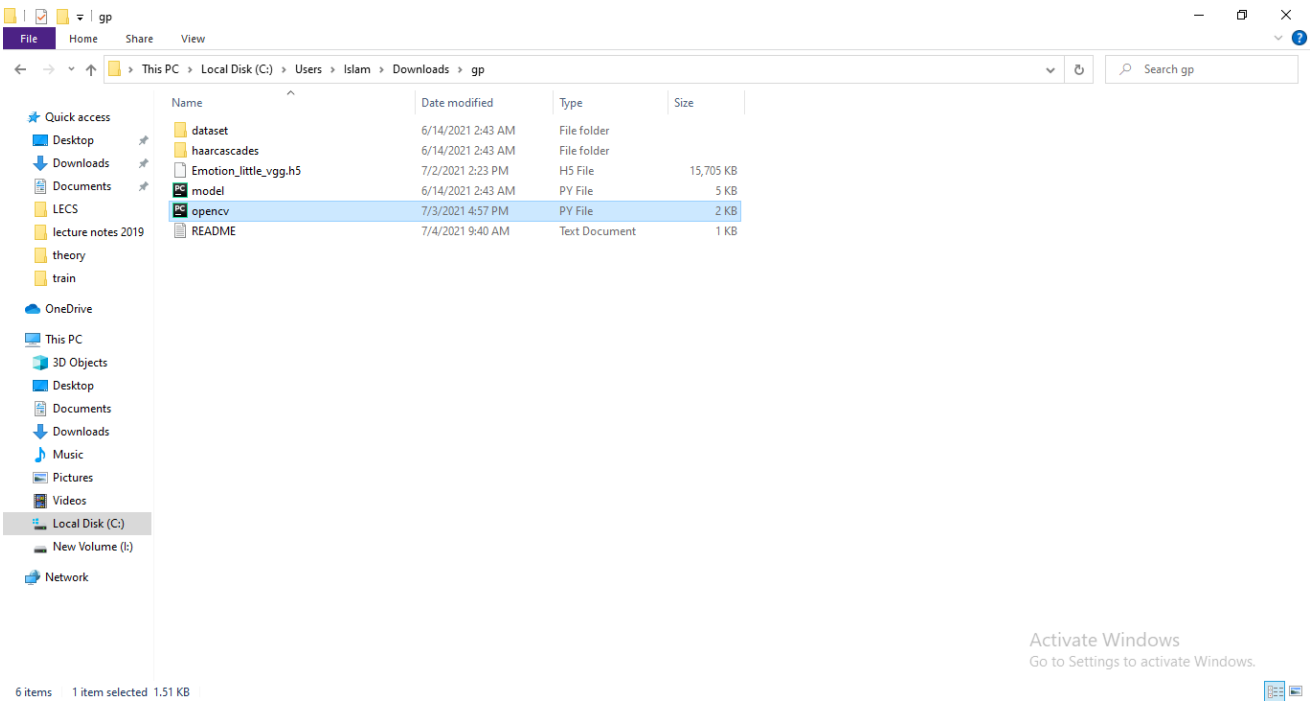
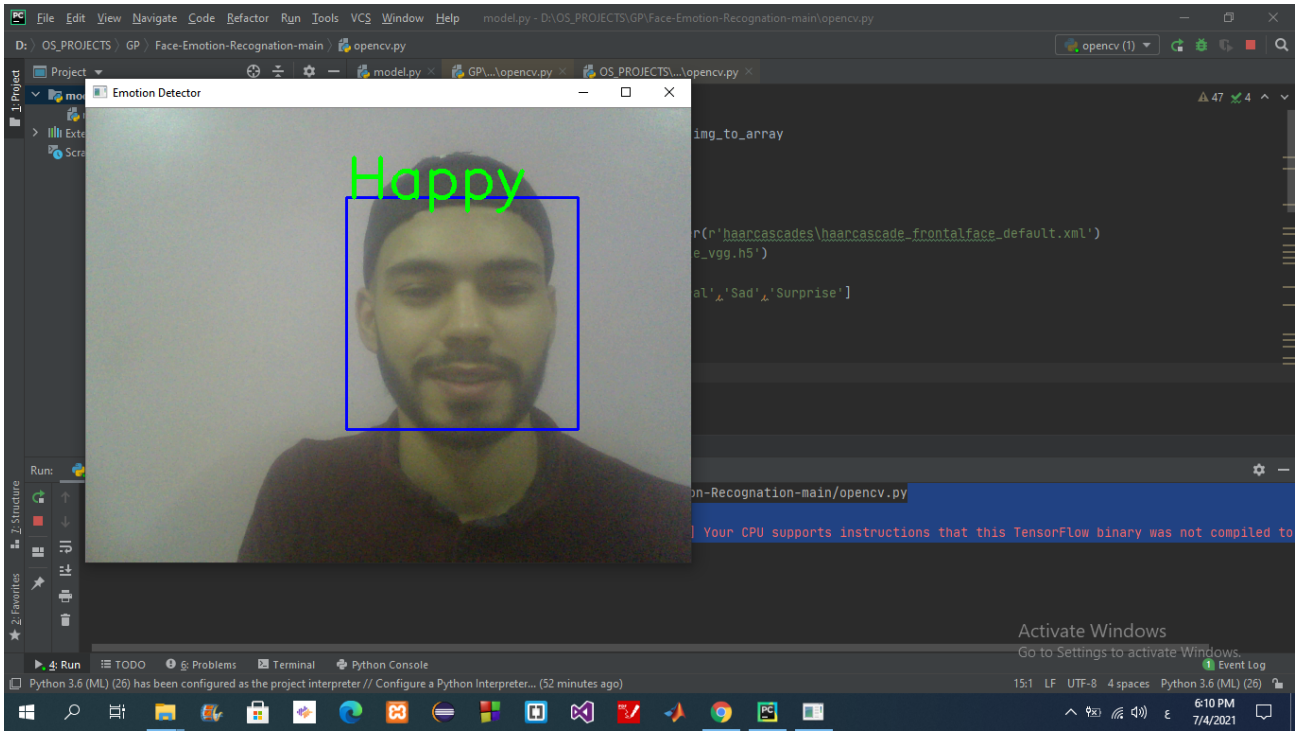
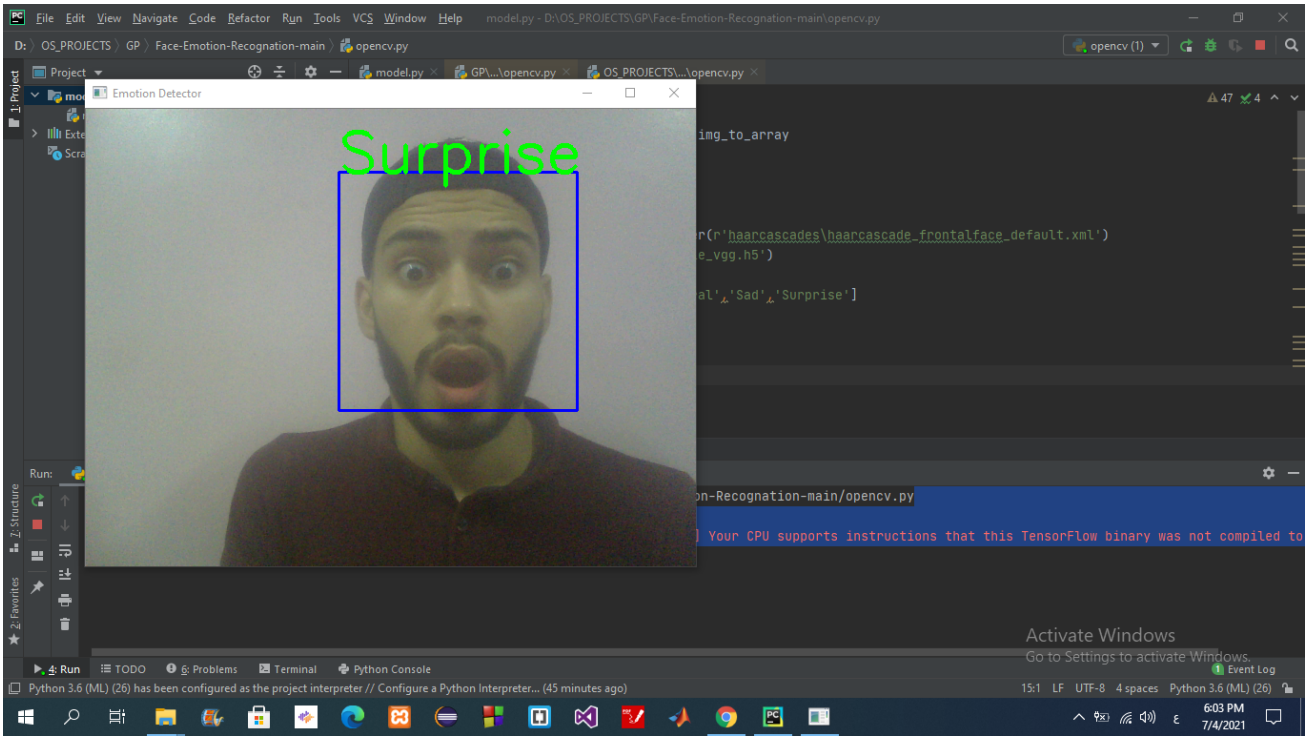
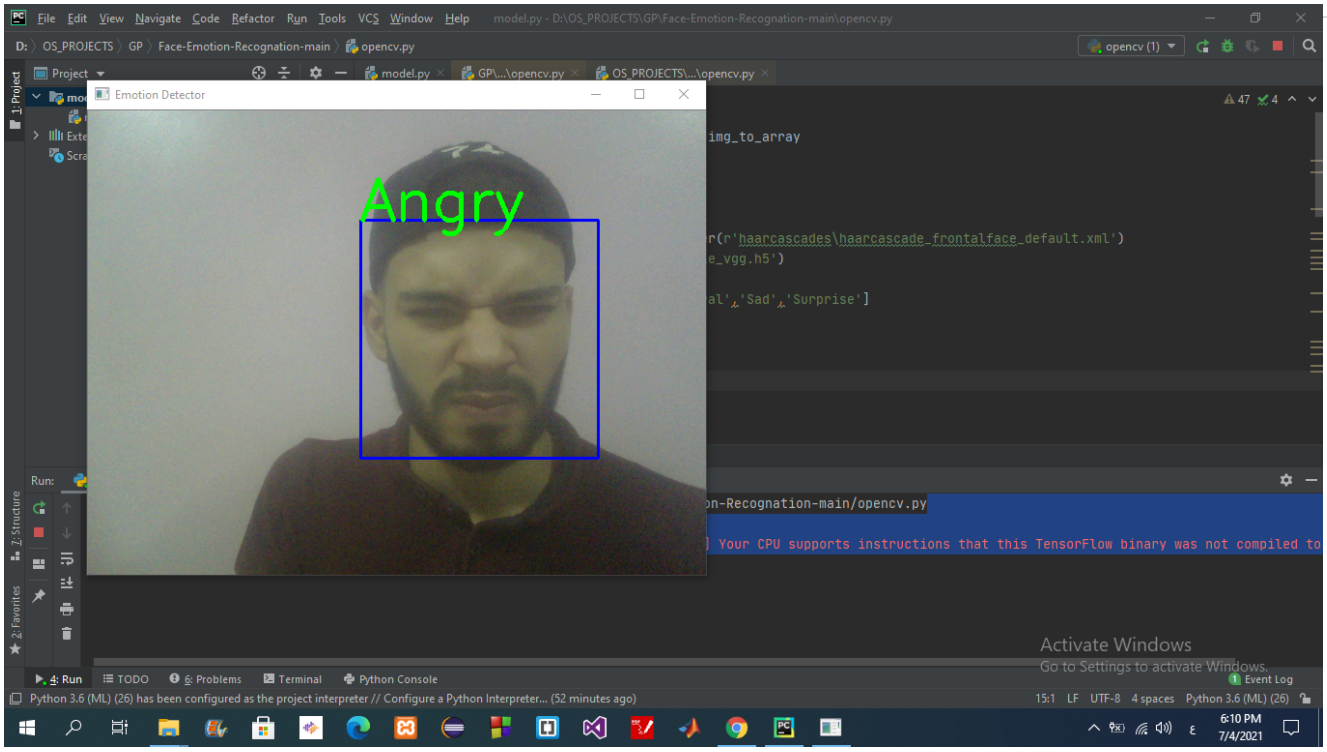


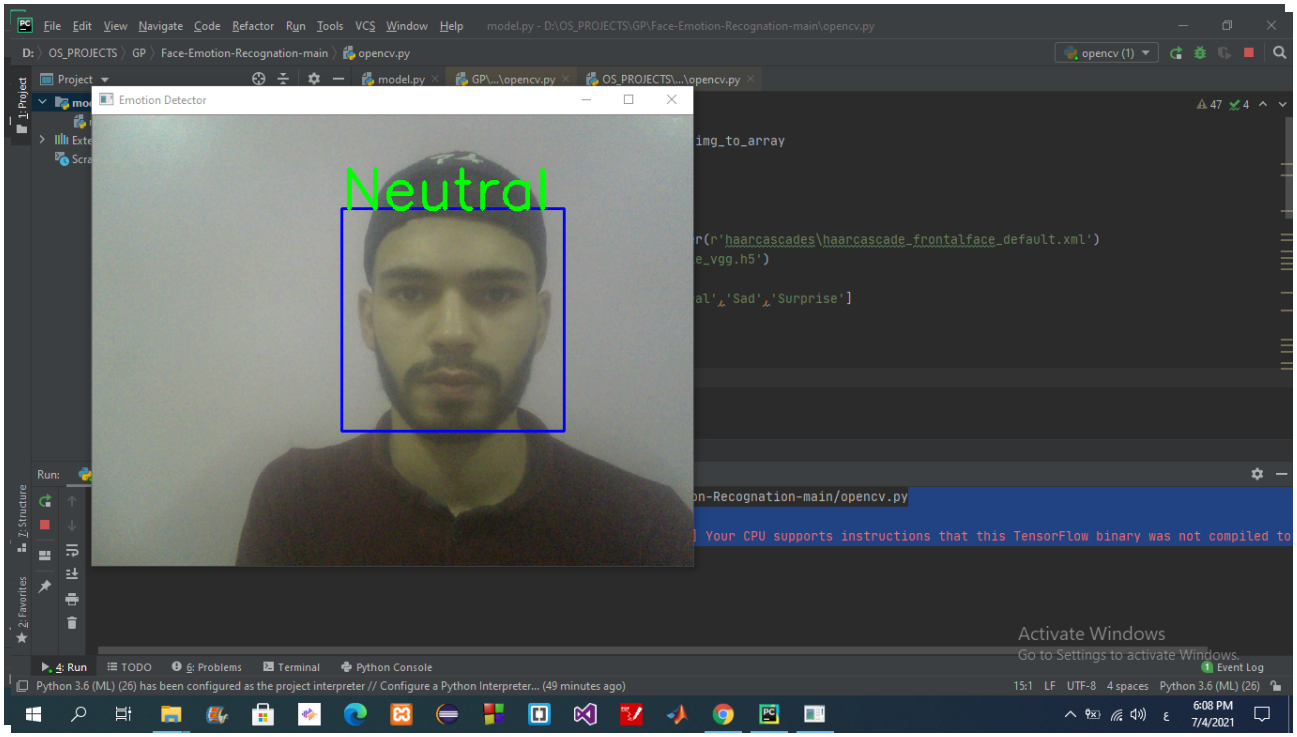
Figure 5.3

Make sure that your camera is enabled as after the running a live camera would start to detect the faces and analyze it to detect emotion






Emotion Detector
Angry
Activate Windows
Go to Settings to activate Windows.


Emotion Detector
Neutral
Activate Windows
Go to Settings to activate Windows.

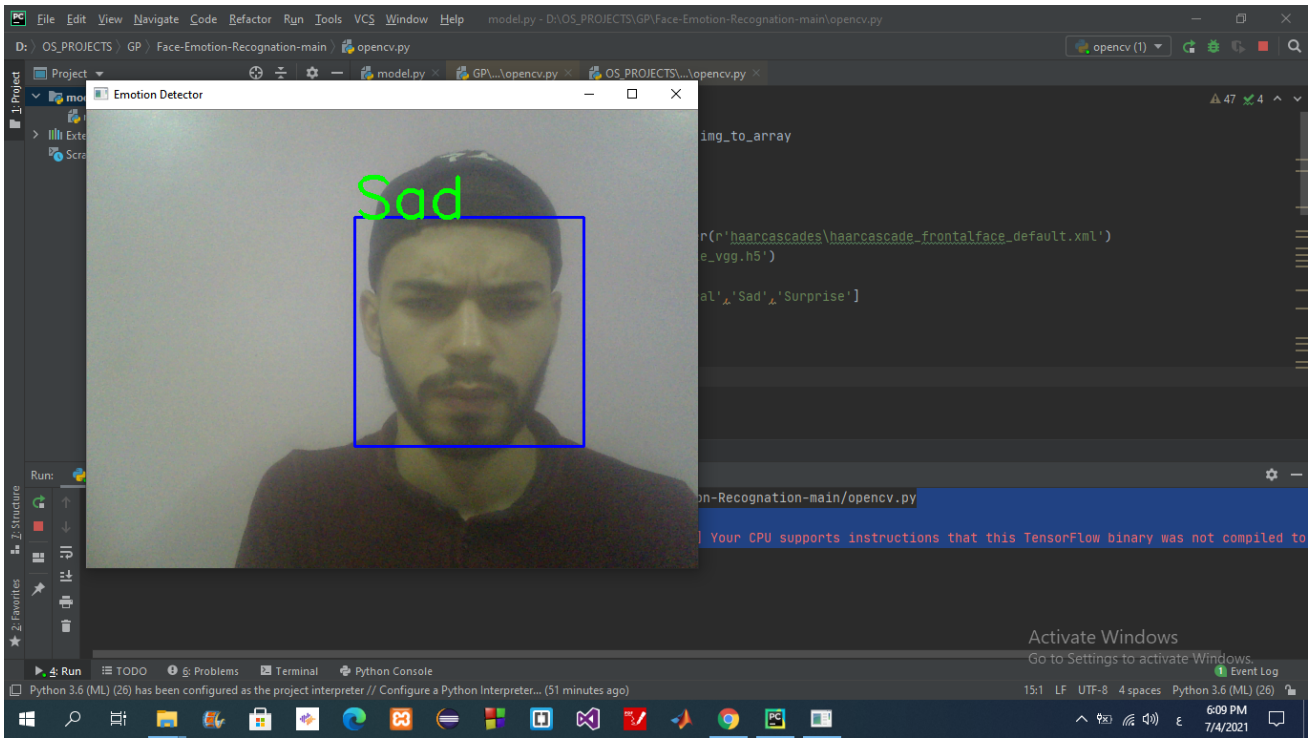

Emotion Detector
Sad
Activate Windows
Go to Settings to activate Windows.

figure 5.4

6- Conclusion and Future Work

6.1 Conclusion

Our proposed system is based on a modern solution type of deep learning CNN and that insure Performing the process as fast and robust as possible. The basic techniques utilized in our proposed system include videos dataset generator, Convolution operations, features extraction, and finally classify the image based on human face emotions. The experimental results demonstrate that the proposed method can Recognize human facial expression with high quality reaches 92% and in a short time.

6.2 Future Work

- in the future, we will work to improve accuracy and better model of our project and we will work and we will work using a dataset larger than our dataset to reduce a high performance and accuracy.
- one of our plans to use this project to build more complex model for personality detection using videos to recognize human emotion and from his speech we will detect what did the human say and from his emotions and his words he said we will detect the personality

References

1. ASCERTAIN: Emotion and Personality Recognition using Commercial Sensors Ramanathan Subramanian, Senior Member, IEEE, Julia Wache Student Member, IEEE, Mojtaba Khomami Abadi, Student Member, IEEE, Radu L. Vieriu, Member, IEEE, Stefan Winkler, Senior Member, IEEE, Nicu Sebe, Senior Member, IEEE

2. https://github.com/ohyicong/emotion-detection

3. https://ibug.doc.ic.ac.uk/resources/first-affect-wild-challenge/?fbclid=IwAR1udlobU1ohGLTdmBud36AcjZT8-g3Co9hTSi8L7NJQqLngY6-LGoCnMnw#:~:Training

4. https://github.com/DenisRang/Combined-CNN-RNN-for-emotion-recognition

5. https://github.com/omar178/Emotion-recognition

6. Multimodal Attention Network for Continuous-Time Emotion Recognition Using Video and EEG Signals,IEEE

7. https://github.com/avx99/Face-Emotion-Recognation?fbclid=IwAR0bPIYCiXKxHyJQjuA9puSnQTeXtf3Io3KZpDSgCU43dkVh0uuqPW4j6AE

8. Development of video-based emotion recognition using deep learning with Google Colab

9. https://www.researchgate.net/publication/332030751_Study_of_Video_based_Facial_Expression_and_Emotions_Recognition_Methods

10. Lee J, Kim S, Kiim S, Sohn K. Spatiotemporal Attention Based Deep Neural Networks for Emotion Recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2018, pp. 1513–7.

11. Nagarajan B, Oruganti VRM. Group Emotion recognition in adverse face detection. In: 2019 14th IEEE international conference on automatic face and gesture recognition (FG 2019). IEEE. 2019, pp. 1–5.

12. Jangid M, Paharia P, Srivastava S. Video-based facial expression recognition using a deep learning approach. In: Advances in computer communication and computational sciences. Singapore: Springer. 2019, pp. 653–60.

13. G. Mattavelli, E. Barvas, C. Longo, F. Zappini, D. Ottaviani, M. C. Malaguti, et al., "Facial expressions recognition and discrimination in Parkinson's disease", *J. Neuropsychol.*, 2020.

14. X. Su, M. Gao, J. Ren, Y. Li and M. Rätsch, "Personalized clothing recommendation based on user emotional analysis", *Discrete Dyn. Nature Soc.*, vol. 2020, pp. 1-8, Mar. 2020.

15. N. Samadiani, G. Huang, W. Luo, Y. Shu, R. Wang and T. Kocaturk, "A novel video emotion recognition system in the wild using a random forest classifier", *Proc. Int. Conf. Data Sci. (ICDS)*, pp. 275-284, 2019.

16. A. Curci, T. Lanciano, F. Battista, S. Guaragno and R. M. Ribatti, "Accuracy confidence and experiential criteria for lie detection through a videotaped interview", *Frontiers Psychiatry*, vol. 9, pp. 748, Jan. 2019.

17. J. Cai, O. Chang, X. Tang, C. Xue, et C. Wei, « Facial Expression Recognition Method Based on Sparse Batch Normalization CNN », in *2018 37th Chinese Control Conference (CCC)*, juill. 2018, p. 9608-9613, doi: 10.23919/ChiCC.2018.8483567.

18. Y. Li, J. Zeng, S. Shan, X. Chen Occlusion Aware Facial ExpressionRecognition Using CNN With Attention Mechanism IEEE Trans. Image Process., 28 (5) (2019), pp. 2439-2450mai, doi: 10.1109/TIP.2018.2886767.