MULTI-EMOTION DETECTION USING HOG DETECTOR AND DEEP LEARNING

Major Project II submitted in partial fulfilment of the requirements for the award of degree of

Master of Technology
In
Software Technology

Submitted By:

ANKITA DIXIT
(2K14/SWT/504)

Under the Esteemed Guidance of:
DR. O. P. VERMA

(Professor, Department of CSE)

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY

CERTIFICATE

This is to certify that **Ankita Dixit (2K14/SWT/504)** has carried out the Major Project II titled **"Multi Emotion Detection Using Hog Detector and Deep Learning"** in partial fulfilment of the requirements for the award of degree of **Master of Technology** in **Software Technology** by **Delhi Technological University.**

The Major Project II is a bonafide piece of work carried out and completed under my supervision and guidance. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any degree or diploma.

Dr. O. P. Verma
Professor
Department of Computer Science and Engineering
Delhi Technological University
Delhi - 110042

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude and respect to my guide **Dr. O. P. Verma**, Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi, for providing valuable guidance and constant encouragement throughout the project work. It is my pleasure to record my sincere thanks to him for his constructive criticism and insight without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members and non-teaching staff of this department for providing their valuable help and time whenever it was required.

Ankita Dixit

Roll No. 2K14/SWT/504

M.Tech (Software Technology)

ABSTRACT

Understanding the emotion of attendees in a group session is an important aspect of today's learning world. The facial expression of a human being conveys a lot of information about identity and emotional state of the person. Emotion recognition is an interesting and challenging problem which gives details about the identity and emotional state of the person.

Many experiments have been carried out in the past and different techniques are being proposed over the years for emotion detection that includes Eigen face based emotion detection and recognition, feature based emotion detection and recognition, machine learning based emotion detection and recognition and so on. However with the increasing migration of the application to be Convolutional Neural Network (CNN) most of the enterprise business operates from the CNN distance, with the advantage of enormous processing capabilities and online processing as well as device independent Application Program Interface (API).

CNN provides a platform for almost all enterprise businesses. Therefore Deep Learning Based emotion detection and recognition is extremely important in order to model any enterprise system that takes any decision based on user's emotion.

In this work we have proposed a unique Deep Learning Based emotion detection system. Multi emotions are detected using Histogram of Oriented Gradients (HOG) detector and Deep Learning. A system should be independent of the gender and age which the past systems have largely failed to provide. At the same time, our system should provide an enterprise level API's to data analytics such that recognition emotion data can be used from a large application perspective.

In this project, we intend to propose a unique emotion detection based system over a CNN to analyse the behaviour of participants during the process of the session. The system should be able to acquire multiple faces, extract their emotion and put it into a CNN-database; further a data analytic system is implemented at CNN to extract meaningful, statistical insight and aggregated behaviour of users during the session. Overall system is capable of giving an insight about whether users are happy, skeptical or unhappy during the session.

TABLE OF CONTENTS

FIGURES AND TABLES

CHAPTER 1
INTRODUCTION

1.1 PREFACE TO AUTOMATIC FACIAL EXPRESSION ANALYSIS

Group sessions/ lectures / Educational sessions are one of the leading requirements which is drawing significant attention in almost all leading industries way it be Education, marketing, Software, medical etc. This is in fact acting as an innovation in different business aspects for providing sales optimization in operational cost, better participant relationship, better services, predicting trends and so on. Traditionally, business success has been dependent on the business skills and service quality of the human resource. However, with the advancement of technology, more and more business are looking to integrate technology in various aspects of business. This will enhance the profitability as well as scalability of the business.

In the proposed work, by knowing emotional status of persons in a group, overall assessment/result of a session can be predicted/ summarized.

The different facial emotions generally observed in human are shown in fig. 1.1.

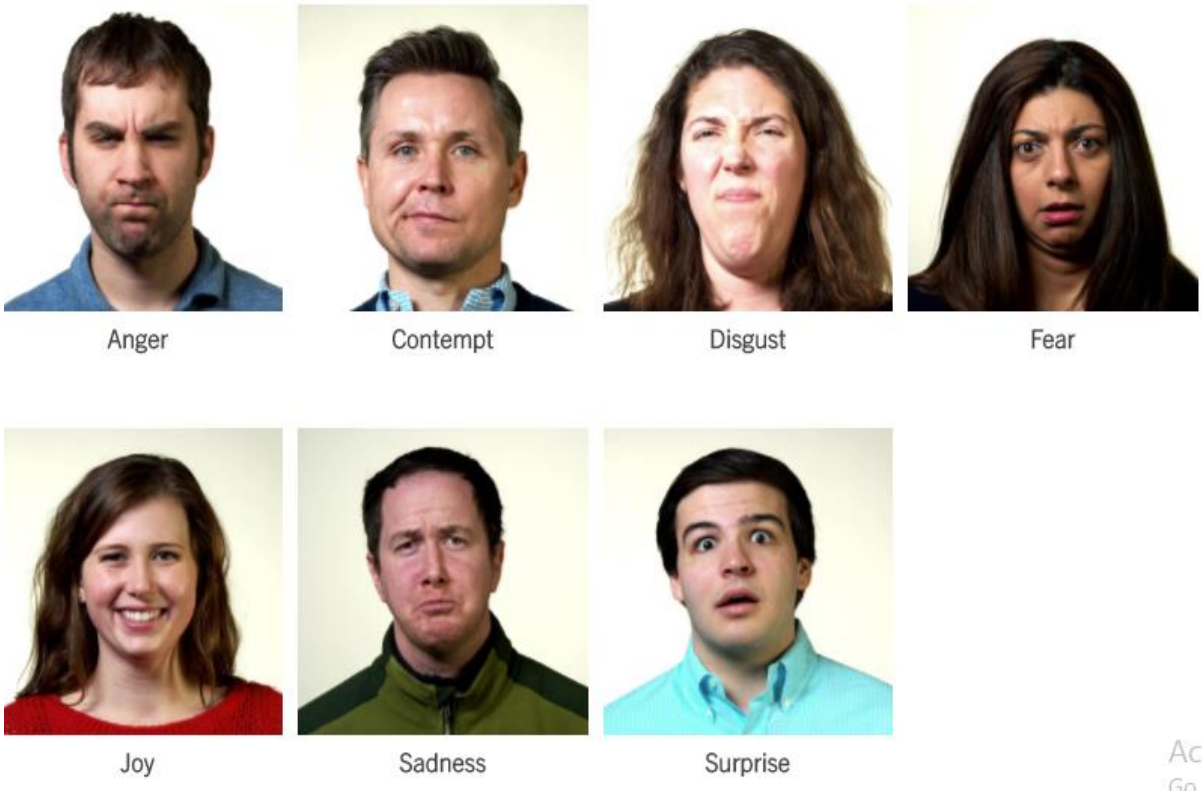

Fig. 1.1: Different facial emotions of human being

To predict the results, smile percentage is measured. To elaborate, suppose if smile percentage is 100% then the person is said to be happy. If percentage of smile is 50% then the person is considered to be in relaxed position. If the percentage of smile is low, then the person is said to be sad or depressed. If the eyes of the person is widen then the person is surprised. This way by knowing the percentage of the smile we can classify the emotions.

The emotions are classified by knowing the features of the human face. When the lip is stuck and eye brows are raised it is defined as an anger emotion. If the smile is not normal but lips position is curved then the emotion is defined as contempt. If the cheek raise and lips stuck then this emotion is defined as disgust. If the eyes are slightly widened and mouth is bit open then emotion is defined as fear.

If the smile is 100% then the emotion defined is joy. If there is chin drop, cheek drop and smile percentage is low then emotion is defined as sad. If the eyes are completely widen mouth is open then the emotion is defines as surprised.

The emotions discussed above and as shown in fig 1.1 are tabulated in Table 1.1.

Emotions	**Increase likelihood**	**Decrease likelihood**
Joy	Smile	Forehead enhance, forehead furrow
Anger	Forehead furrow, lid tighten, eye widen, chin enhance, mouth open, lip suck	Inner brow raise, forehead boost, smile.
Disgust	Nose wrinkle, upper lips enhance.	Lip suck, smile.
Surprise	Inner forehead enhances, brow increase, eye widen, jaw drop.	Brow furrow
Fear	Inner brow increase, brow furrow, eye widen, lip stretch.	Brow improves, lip corner depressor, jaw drop, smile.
Sadness	Internal forehead enhance, forehead furrow, lip corner depressor.	Brow improves, lip corner depressor, jaw drop, smile.
Contempt	Brow furrow, smirk.	Smirk

Table 1.1: Classification of emotions based on likelihood

These emotions are classified on the basis of the increase and decrease likelihood features of human face. The each emotion has different feature changes. Feature like eyes, lips, mouth, chin, cheeks are taken into consideration

These emotions are classified by knowing the features of the human face.The classified emotions are matched with the ensemble database and the final recognized emotion is taken as the result of the work.

In our study, Histogram of Oriented Gradients is used along with Support Vector Machine (HOG+SVM) approach for object detection.
This chapter involves the brief introduction of the various stages involved in Automatic detection of facial emotions / expression. The concept Histogram of Oriented Gradients (HOG) and Deep learning is also elucidated.

1.2 STAGES OF AUTOMATIC FACIAL EXPRESSION DETECTION

The Automatic Facial Expression Detection Analysis problem has been managed using the three stage approach as explained below:

1. Face Acquisition
2. Feature Extraction
3. Classification

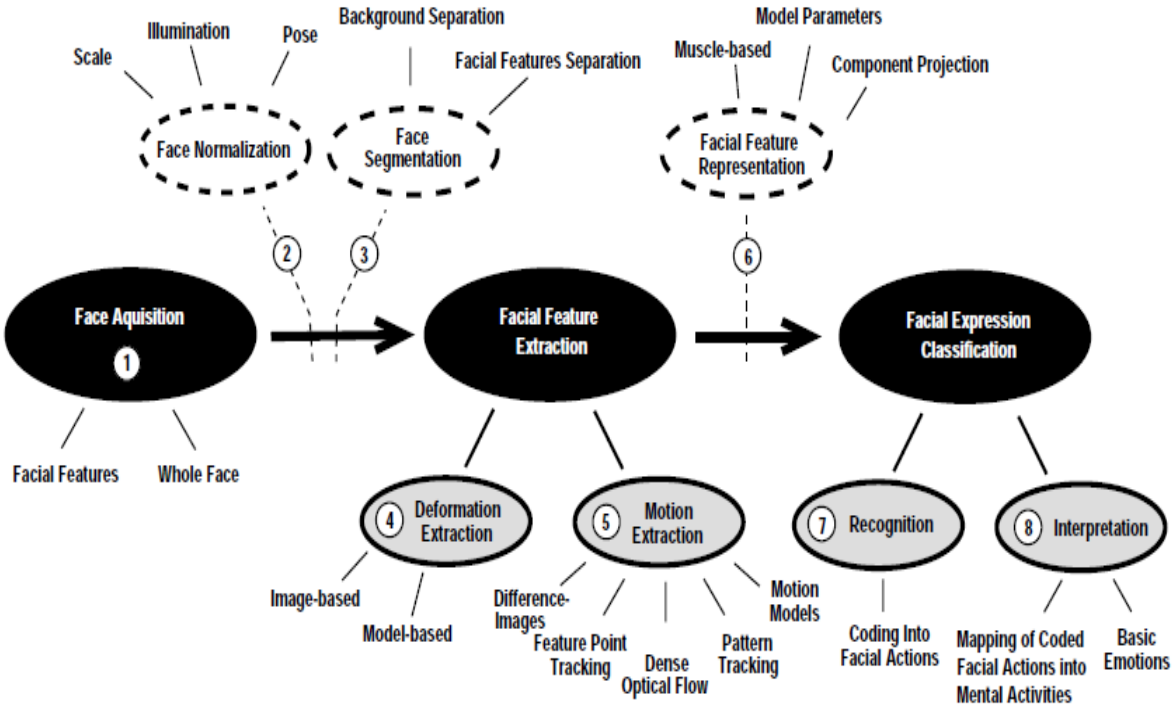

Fig. 1.2: A comprehensive Facial Expression Detection Analysis framework[8]

The first stage detects the face in the selected image. After the successful detection of face, the various expression features are extracted in feature-vector. Finally,-these feature-vectors are used in classifier for classification/recognition. The classifier can be a two class or a multi-class classifier. In the study work, we have used SVM classifier for multi-class classification in one-against-all mode.

1.2.1 Face Acquisition

For the feature extraction of facial expressions, the face in the selected images should initially be detected with all its boundaries. For human brain Face detection is a trivial task, but it has been one of the difficult problems which are handled by the machines.

1.2.2 Feature Extraction

It is well understood that it is not possible to have a universal feature extraction algorithm for all kind of recognition problem. For an effective and efficient classification by any machine learning algorithm, good discriminating features of the subject must be available. Moreover, features suitable for one problem may not work well for others. The choice of features is amongst the deciding factors of accuracy of classification/recognition.

The feature vector in case of facial expression analysis, contains either the motion information or the deformation information as displayed in the face[8]. Shape and texture changes are used to characterize the deformation based features whereas optical flow and displacement of facial points are used to define motion information.

1.2.3 Classification

Classification/recognition process is carried out to assign a particular class to the input facial expression image. It can be a two-class or a multi-class classifier. The two-class classifier outputs a yes/no, present/absent, class1/class2 etc., while a multi-class classifier is capable of discriminating between various classes. A two-class classifier can also be used for multi-class classification problem in one-against-all or one-against-one mode. In this study we have used SVM as a classifier in one-against-all mode.

1.3 HISTOGRAM OF ORIENTED GRADIENTS (HOG)

Detecting emotions of human face in an image is one of a challenging task because of their dynamic appearances and ample range of shapes/poses that they can adopt. The initial requirement is a robust feature set that allows human face to be detected clearly, even under illumination or in cluttered backgrounds. The locally normalized Histogram of Oriented Gradient (HOG) descriptors provide excellent performance as compared to the other existing feature sets.

The Histogram of Oriented Gradients method suggested by Dalal and Triggs in their seminal 2005 paper, Histogram of Oriented Gradients for Human Detection [7] demonstrated that the HOG image descriptor and a Linear SVM could be used to train highly accurate object classifiers — or in particular, human detectors.

1.4 CONCEPT OF DEEP LEARNING

Deep learning refers to the training of machines on the basis of the data available with us. Machine will automatically learn from the data. Data sets are of two types: Training Data and Testing Data. Training data are used to train machines to learn the pre requisites, like in case of emotions; machine will be trained for the type of emotion based on the available training data. Testing Data are used for determining the accuracy after comparing with the Training Data. Deep Leaning is subset of Machine Learning and used neural network to simulate human like decision making. Machine leaning has two known limitations. It does not give accurate results in Feature Extraction stage in the Automatic Facial Expression Detection Analysis stage. Further, the complex problems like object recognition or handwriting recognition, the observed results are not at par.

A *Convolutional Neural Network* (CNN), in machine learning, refers to the class of feed forward artificial neural networks that has been successfully applied in the field of visual image analysis. The feed-forward neural network was the simplest and first of the artificial neural network devised, wherein the connections between the units do not form a cycle. CNNs were inspired by biological processes and they relatively require little pre-processing if compared to other image classification algorithms.

The Network acquires information of the filters, which were hand-engineered in the traditional algorithms. This liberty from human effort and prior knowledge in feature design is indeed a major advantage. Deep Learning Based emotion detection and recognition is remarkably important in fields where decisions are taken based on human's emotion.

1.5 CHALLENGES INVOLVED IN AUTOMATIC FACIAL DETECTION ANALYSIS

There are many factors which pose a great level of challenge for automatic facial expression detection analysis using computer vision, although they are trivial for human perception. Following are some of the primarily observed challenges:

- Pose variation
- Illumination variation
- Presence of obstructing accessories like glasses, hairs etc.
- High degree of variability in the physiognomies of faces around the globe
- High degree of variability in facial expression due to cultural differences around the globe
- Scale variation
- Partial occlusion of face
- Recognizing mixture of emotions

1.6 APPLICATIONS OF AUTOMATIC FACIAL EXPRESSION DETECTION

For a partial/complete omission of human factor from the machine, the augmentation of Automatic Facial Expression Detection Analysis module in it is the most required one. Any application having this type of requirements of un-manned system becomes the potential candidate for automatic facial expression Detection analysis. Following are some of the applications:
In case of analysing the feedback of group sessions viz. Educational Lectures / Seminars/Trainings, etc. based on the facial expressions/ emotions of the people.

Video conferences and seminars and interviews can use machines capable of automatic facial expression analysis for more effective communication and inference.

In case of critical risk conditions such as flying an airplane or fighter aircraft, driving a vehicle or train etc., the detection of negative affective states such as fatigue, boredom stress etc. can help in saving of life and property.

Detection of Microexpressions can be used to detect lie or deception in case of terrorist interrogation. Microexpressions occur when a person either deliberately or unconsciously conceals a feeling. They are very brief in duration, lasting only 1/25 to 1/15 of a second.

CHAPTER 2

LITERATURE REVIEW

2.1 SOME COMMOM THREADS OF PROPOSED WORK

With the advancement of the technology more and more business are looking to integrate new research developments to their various fields to enhance the profitability as well as scalability of the business. Detection of Human expressions/ emotions plays a vital role in taking various important decisions for any firm, institute, company or say for any kind of business.
In the past also, various studies/ research has been done and various methodologies has been proposed.

Jason M. Saragih et. al[18] mention about deformable model fitting. This is widely famous in the field of computer vision. Over a decade many methods have been proposed. But the methods considered successful were the ones that have the ability to independently predict the location of model's landmark. This study proposes an optimization technique within a hierarchy of smoothed estimates; the nonparametric representations of the likelihood are maximized. The updates equation obtained from these methods are similar to that of mean shift over landmarks. But they will have a regularization, which is imposed with the help of global prior over the long motion of theirs. This work also presents extensions that are capable of handling partial occlusions and reduce computational complexity.

A.W. Senior et. al[19] discusses about the video analytics. This study work tells about a bunch of tools that help in analysis of retails business. These tools take help of video captured and transaction log. The further analysis of the collected data helps in prevention of fraud cases and increases the effectiveness of display of goods. All these features can be used with help of web browser.

Minghua Han et. al[20] discusses a customer segmentation model which is based on the analysis of behaviour of retail consumers. The segmentation is the basis for CRM (Customer Relationship Management). There are various methods based on which the segmentation can be made, viz. Purchase behaviour. This research work combines the advantage obtained from Principal Component Analysis (PCA) and Back Propagation (BP) Neural Network. The PCA is fed as an input to the feed forward neural network.

M. Venu Gopalachari et.al[25] mentions a system of filtering recommender that is personalized and collaborative with the help of domain knowledge. The current recommender systems lack the domain knowledge in predicting the products that the user might be interested in. This study work mainly focuses on incorporating domain knowledge and knowledge of usage for the recommendation of the web pages.

Anurag De et.al[28] research work has performed a comparative study on the different approaches available for identifying human emotions in real time based on the facial expression. The multiple variability of a human face such as posture, colour, texture, orientation, expression etc. has been considered by a system for identifying the facial expression. A comparative study of different approaches used for identifying real time facial expression identification is presented in this study work.

Rana Alaa El-Deen Ahmed et. al[30] mentions about a performance study of the available classification algorithm, for identifying online shopping attitudes and behaviour of the consumer using data mining. In this study, 11 different classification algorithms are considered for a performance analysis. In experimental results, decision table mentions the best accuracy and also performance when compared with the filtered classifier. The obtained result can be used by different retail industries to decide which classification algorithm suits the most.

Iftikar Ahamath Burhanuddin et. al[34] study work discuss about the model of using multi-channel data for similarity learning, so that this can be used for product recommendation and scoring. There are many ways how a customer can interact with retailer and data from all these channels can be tracked. This is achieved by considering similarities learning as a problem to learn to rank and minimizing rank loss.

Cui et. al[39] uses the distance vector and an extreme learning machine for the smile detection. Here the fact that the shape of mouth can be used to identify a smiley face is taken into consideration and a feature vector called pair wise distance vector is formed. Extreme learning machine are used to classify the smiles.

Shier et. al[40] proposes a model to recognition pain and classify intensity with the help of facial expressions. A comparative study is done on Gabor energy filter and Solomon pain intensity scale. The results presented shows that the first method gives better classification.

Shahriar Akter et. al[42] gives a systematic review on the use of big data analytics in the field of e-commerce. It proposes an interpretive framework which explores the definition aspects, distinctive characteristics, business value, type and challenges that are presented by big data analytics. It also triggers a broader discussion about the future research challenges and the available opportunities in the field.

2.2 DATABASE

Besides the study work for research problems, another task which is equally important is the database collection and customization. Some of the well-known and standardized facial expression database available: Extended Cohn-Kanade Database (CK+), JAFFE database and FEI Database. Both the CK+ and JAFFE database require customization before being used in the research work. Some scripts and functions have been developed to carry out the customization task.

Two standardized databases that have been used for the research: Cohn-Kanade Database and FEI Face Database.

2.2.1 Cohn-Kanade Database [2]

The Cohn-Kanade database is the one of the standardized and most widely used database for Automatic Analysis of Facial Expressions, both for algorithm evaluation and development. It contains 593 image sequences across 123 subjects. The subjects were 18 to 50 years old, 69% female, 81 % were Euro-American, 13% were African-American and 6% other group. The sequences comprise of images from onset, i.e. neutral, to peak expression level as the last frame. The reported emotion labels have been visually investigated and validated by emotion researches based on FACS Investigator Guide. The duration of image sequence, i.e. number of frames, also varies across subjects (10 to 60 frames). Moreover, out of 593 sequences, only 327 have been found fit for emotion labelling (non-posed emotion class). For our experiment we have cropped the images to contain the face details only. Fig 2.1 below presents a cropped image sequence from the Cohn-Kanade database.

Fig 2.1: A cropped image sequence from Cohn-Kanade database[9].

2.2.2 The FEI Face Database (http://fei.edu.br/~cet/facedatabase.html)

The FEI face database is Brazilian face database taken at Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil. The database is made up of 14 images of 200 individuals with equal number of male and female. A subset of database containing the cropped frontal images of two expressions (neutral and smiling) of each individual has been taken for Automatic Facial Expression Analysis. Thus there are 400 images of 200 subjects depicting two expressions. Fig 2.2 shows some of the expression images from FEI database.

Fig 2.2: Example images of Cropped FEI database

CHAPTER 3

AUTOMATIC FACIAL EXPRESSION DETECTION

3.1 INTRODUCTION

In this chapter we discuss some of the well understood and established methods and techniques for Automatic Facial Expression Detection analysis for human emotion recognition.

A generic process flow has been already presented in chapter 1. In our study, we have focused on the feature extraction, HOG detector, face recognition, ensemble-based classifiers and CNN used in deep learning. Hence a comprehensive literature details regarding the focussed topics will only be presented.

The first attempt on automatic analysis of facial expression was carried out by Suwa et al. [10] in 1978 using image sequences. Gradually it has been the highly researched area and a thorough survey study can be found in [8, 11].
Most of the researches in Automatic Facial Expression Analysis were performed to understand a set of prototypic emotional expressions (i.e., disgust, fear, joy, surprise, sadness, anger) proposed by Ekman [1, 3-6] as shown in fig. 3.1.

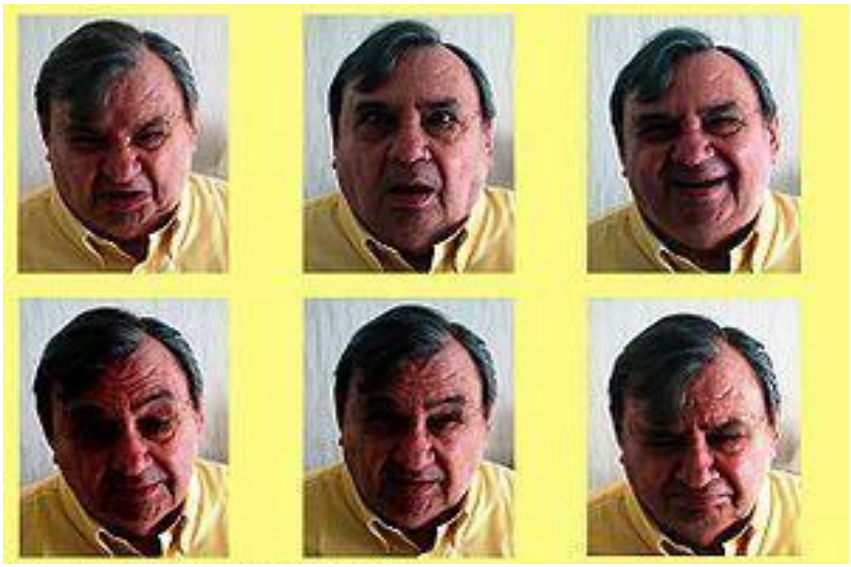

Fig 3.1 Human Emotional Expression

3.2 FEATURE EXTRACTION

Automatic Facial recognition has been a long challenging problem in analysis.
Two vital aspects of Automatic Facial Expression recognition includes facial expression representation and classifier design[5] . In case of facial expression representation, a set of features called feature vector is derived from the original

face image to effectively represent facial expressions. An optimal feature vector will maximizes the between-class variations while minimizing the within-class variations [9]. In case of sub-optimal or inadequate features in the feature vector, the best classifier may report low accuracy of recognition.

Optical Flow analysis has been used in some studies [12, 13, 44] to model muscle activities or displacement estimation of feature points. But flow estimation suffers from the disturbances caused by illumination variations and non-rigid motions of face. They are also sensitive to image registration and motion discontinuities [5]. Shapes and location of facial components are extracted using Facial Geometry analysis to represent facial expressions [3,13]. Facial movements in the image sequences (in a video) can be characterized by measuring the geometrical displacement of facial feature points between the current image frame and the initial one [9]. Valstar et al. [17] reported that the tracked facial points based feature vector is better suited for facial expression analysis. Their method is based on detection of Action Units (AU) by classifying features calculated from tracked fiducial points. An accurate and reliable facial feature detection and tracking is required in geometric feature based representations, which is not possible in many situations.

Principal Component Analysis (PCA) [45] , Linear Discriminate Analysis (LDA) [46] and Independent Component Analysis (ICA) [41] were used by Donato et al. [32] for face representation.

Ren, Nasser et. al[24] discusses various optimization techniques with aim of achieving a real time software based implementation of Voila Jones algorithms on mobile devices that have limited processing and memory capabilities. Viola et. Al[33] method is widely used for object detection in Real time. Oscar, Daniel et. al[23] discusses about the various classifiers used for the face detection and analysed their individual performance with the aim of deciding a benchmark for other approaches.

3.2.1 Viola-Jones algorithm

It possess an important property of Real-Time detection of objects by training of the images which is relatively slow, but its detection is quite fast. Viola-Jones algorithm uses principle of scanning a sub-window, which can detect presence of multiple faces in a given image. However standard image processing approach rescales the image in to different sizes and then make use of a detector of fixed size to identify multiple faces in a given image

Thus instead of rescaling or resizing input image the algorithm runs detector of varied size multiple times over the same image.
Thus the basic building block of the detector composed of the so-called integral image and rectangular frames like that of Haar Wavelet.

3.2.1.1 The scale invariant detector

In Viola-Jones face detection algorithm the first step includes conversion of an input image into an integral image. Under this to the concerned pixel's entire sum of all the pixel to the left and above it is computed. This is demonstrated in fig. 3.2.

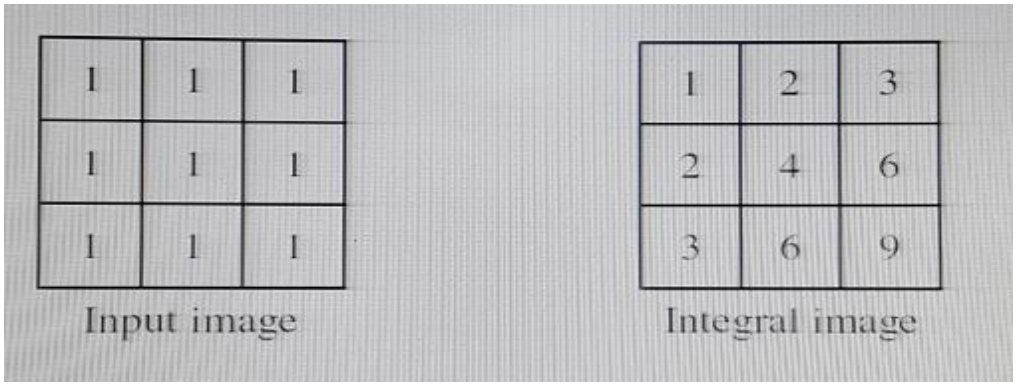

Fig. 3.2: Voila Jones - The integral image

Thus inside any given rectangle and by using any of the four values around a particular pixel, sum of all pixels can be calculated and eventually these values represents the pixels of the integral image and they coincides with the rectangular input image corners. This is demonstrated in fig. 3.3.

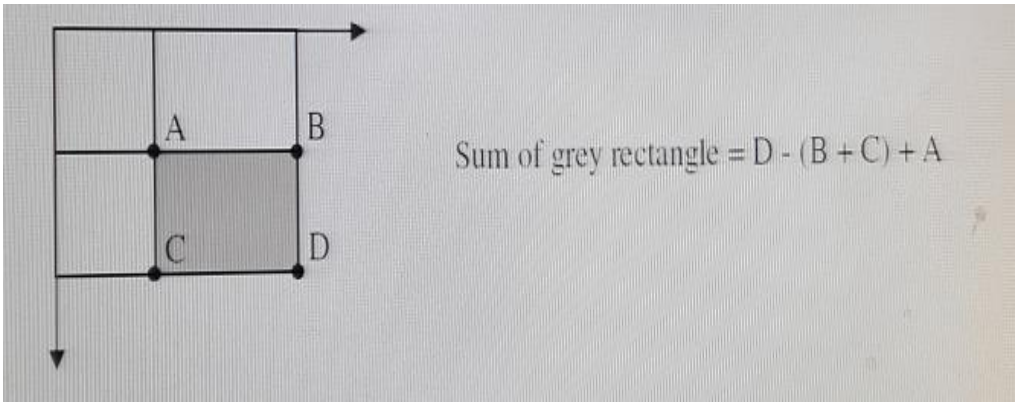

Fig. 3.3: Voila Jones - Sum calculation

As per diagram rectangle A is included in both rectangles B & C hence the sum of rectangle A will also be included in the calculation. Using features consisting of more than two rectangles the algorithm analyses the given sub-window. The different types of features are shown in fig. 3.4.

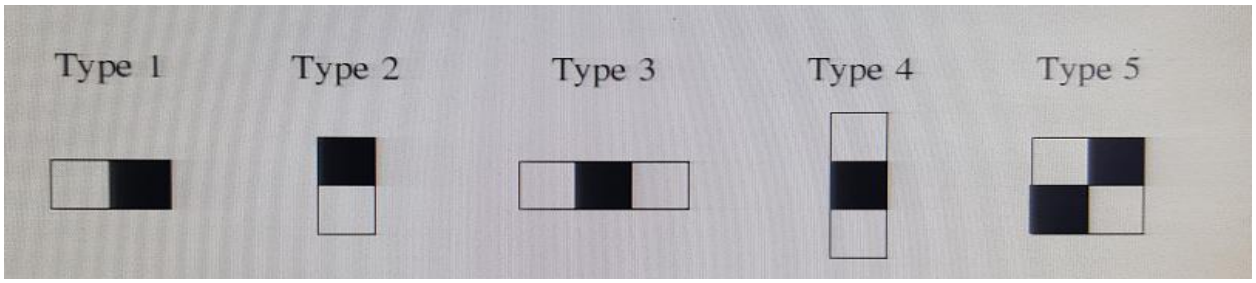

Fig. 3.4: Voila Jones: Different types of features

When sum of all the white rectangle(s) is subtracted from sum of all the black rectangle(s) we obtain a single value which represents feature of the image. Also it is empirically found that the result accuracy increased if base of resolution 24*24 pixels is used. On evaluating all the different positions and sizes of the feature as per fig. 3.4, nearly 160.000 different features can then be constructed.

3.2.1.2. AdaBoost algorithm

Out of 160.000 features few consistently gives accurate result when used on top of a face. For this to achieve these features a modified version of the AdaBoost is being used by Viola-Jones, which is machine learning boosting algorithm which from weighted combination of weak classifiers makes a strong classifier. Each feature is considered to be a potential weak classifier, as described mathematically in fig 3.5:

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{if } pf(x) > p\theta \\ 0 & \text{otherwise} \end{cases}$$

Fig. 3.5: Mathematical representation of Weak Classifier

Where x is a 24*24 pixel sub-window, f is the applied feature, p the polarity and symbol is the threshold that decides whether x should be classified as a positive (a face) or a negative (a non-face).

Out of 160.000 features, only few are potentially weak classifiers, so the AdaBoost algorithm has been tailored to select only the best feature values.

It determines the best polarity, feature and threshold. This means that the determination of each new weak classifier involves evaluating each feature on all the training examples in order to find the best performing feature. The value of the weighted error is used to choose the best feature.

Thus with the combination of the Integral Image, Modified AdaBoost algorithm and the computed efficient features, face detector can be implemented but still needs certain additional inputs.

3.*2.1.3 The cascaded classifier*

The Viola-Jones face detection algorithm uses detector of new size every time it scans the same image. Even an image contains one or more faces it is obvious that an excessive large amount of the evaluated sub-windows would still be negatives (non-faces).

This realization formulates a new problem statement i.e., It is faster to discard the non-faces then to find the faces.

This realization formulates a new problem statement i.e., It is faster to discard the non-faces then to find the faces.

Thus one single strong classifier will not be that effective since the evaluation time is independent of the input. And this raises the need of Cascaded classifier.

The cascaded classifier contains stages of strong classifier. Each stage determines whether the given sub-window is a face or not. When a non-face is identified by a sub-window of a given stage it is discarded immediately. Conversely if a face is detected it is passed to the next stage. More the number of stages an image passes through, higher is the chance face is identified by the sub-window. The concept is illustrated with two stages in fig. 3.6.

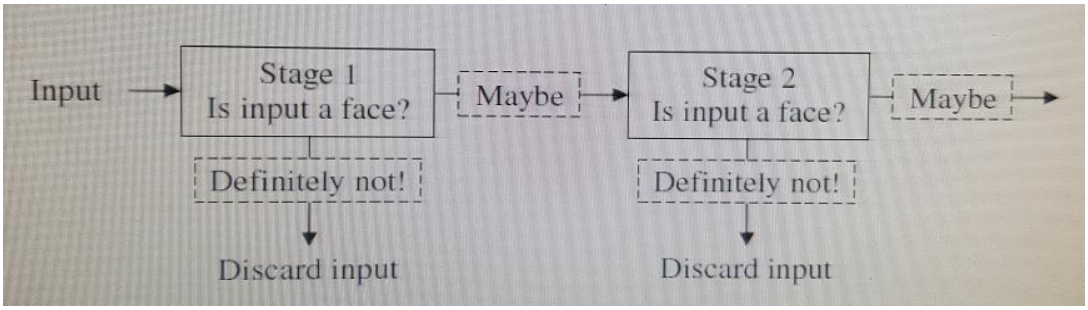

Fig. 3.6: The cascaded classifier

In a single stage classifier false negatives are accepted in order to reduce the false positive rate. However, for the first stage false positive is not a problem as succeeding stages sorts them out. Therefore Viola-Jones is prescribes to accept many false positives in the very initial stages of operation. As a result the rate of false negatives is minimum in the final stage.

Thus, when trained a given stage, say n, then the false negatives generated by the n-1 stage.

The majority of thoughts presented in the Methods' sections are taken from the original Viola-Jones paper [23].Thus using a cascade of stages Viola and Jones face detection algorithm quickly eliminates face candidates. As the stages progress the requirement at each stage becomes strict and hence more difficult for candidate to pass to the next stage. When a candidate passes all the stages successfully face is detected. This process is shown in fig. 3.7.

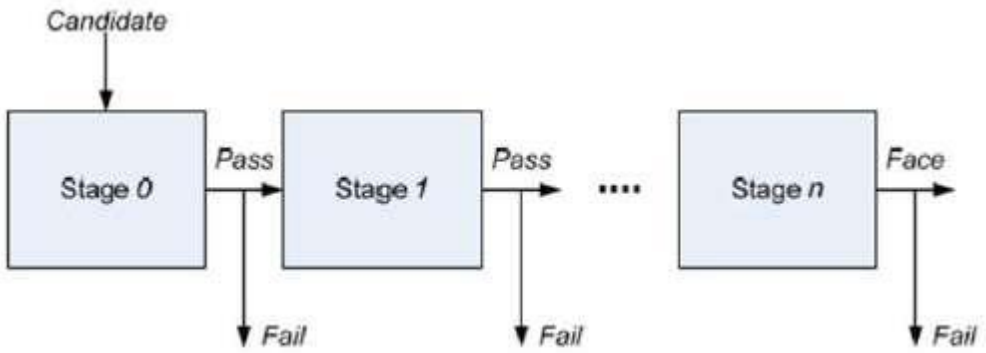

Fig. 3.7: Cascade of stages

3.3 FACIAL EXPRESSION RECOGNITION

Facial expression recognition can be considered as a machine learning problem. In one of the very simple approach, Ahonen et al. [16] and Shan et al [9] used 'template matching' for face recognition. They represented the face using LBP codes and then a template with respect to each expression was formed by averaging the LBP coded histogram of images belonging to same expression category. This was done during training. A nearest-neighbour classifier was used during classification to match the closest template. Chi square statistics (χ^2) was selected as the dissimilarity / similarity measure as follows:

$$\chi^2(S, M) = \sum_i \frac{(S_i - M_i)^2}{S_i + M_i} \quad (1)$$

where S and M are two LBP histograms. Later a weighted sub-region selection was employed in [9] with the recognition accuracy of 84.5% for 6-class expressions.
Tian[5] and Zhang et al. [5] have proposed Neural Network to classify facial expressions. Other classifiers were also used , like , Support Vector Machines (SVM) [41] Bayesian Network (BN) [5] and rule-based classifiers [11, 26].

3.4 HISTORGRAM OF ORIENTED GRADIENTS (HOG) WITH SUPPORT VECTOR MACHINE (SVM)

Recently, Histograms of Oriented Gradients (HOGs) have proven to be an effective descriptor for object recognition in general and face recognition in particular. Robust use of HOG features for face recognition plays a vital role in obtaining accurate results.

DénizaG, BuenoaJet. al[21] demonstrated the use of HOG descriptors to compensate for errors in facial feature detection due to occlusions, pose and illumination changes, In their study, they have proposed the fusion of HOG at different scales that allows to capture important structure for face recognition.

Dahmane, Meunieret. al[27] discusses about the difficulty observed with facial emotion recognition system in implementing the human expressions. The facial expression varies differently across humans. In their work, the method applied, utilizes dynamic dense appearance descriptors and statistical machine learning techniques. Histograms of oriented gradients (HOG) are used to extract the appearance features by accumulating the gradient magnitudes for a set of orientations in 1-D histograms defined over a size-adaptive dense grid, and Support Vector Machines with Radial Basis Function kernels are the base learners of emotions.

Its trivial to detect emotions out of human face that to when the human expression and nature is very dynamic. The initial basic requirement is to clearly detect human face even in presence of clutters or low or difficult illumination. HOG a type of “feature descriptor” does this very efficiently. The main aim of feature descriptor is to generate an exact clone of the object with same feature descriptor even if viewed in different conditions. This simplifies the classification task. Typically, a feature descriptor converts an image of size width x height x 3 (channels) to a feature vector / array of length n.

Within the localized region of the image (i.e., Region of interest (ROI)) the HOG descriptor technique counts occurrences of the gradient orientation.
Implementation steps of HOG descriptor algorithm:

1. Divide the image into cells (small connected regions), and compute a histogram of gradient directions (or edge orientations) for the pixels within the cell for every cell.
2. According to the gradient orientation each cell is discretized into angular bins.
3. To its corresponding angular bin, pixel of every cell's contributes to a weighted gradient.

4. A block is called Groups of adjacent cells in spatial regions. The basis for grouping and normalization of histograms is cell grouping in to Blocks.
5. A Block histogram is represented by a Normalized group of histograms. And Descriptor is the set of these block histograms.

The fig. 3.8 depicts the HOG algorithm implementation:

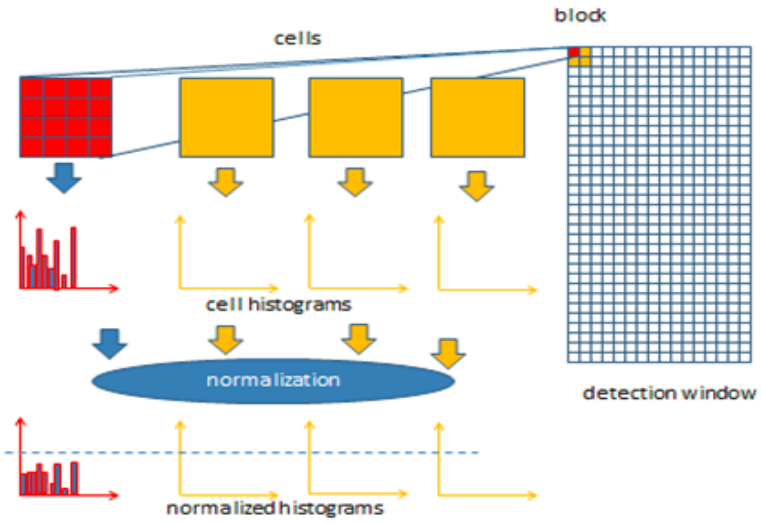

Fig. 3.8: Algorithm implementation scheme for HOG feature descriptor

Support vector machine (SVM)

SVM is basically a binary learning machine [37]. But it can also be used in case of non-linearly separable data set by mapping it to high dimensional feature space. Now once the data has been projected on the high dimensional feature space, the SVM finds a hyper plane that maximizes the margin of separation between positive and negative examples or class [9, 37] (fig 2.9) .

The equation of separating hyper plane is given as:

$$y(X) = W^T \emptyset(X) + b \tag{2}$$

where $\emptyset(.)$ is the mapping on data set X to the high dimensional feature space, W is the coefficient vector and b is the *bias* (sometimes called threshold parameter or intercept of hyperplane [9]).

Now, representing the labelled data set as $[(X_n, t_n)]_{n=1}^{N}, X_n \in R^n$ *and* $t_n \in (-1, +1)$, we have following criterion for separation

$$y(X) > 0, \forall X_n \; having \; t_n = 1 \tag{3}$$

$$y(X) < 0, \forall X_n \; having \; t_n = -1 \tag{4}$$

More formally, the hyper plane classifying the new test sample is rewritten as:

$$y(x) = sgn(\sum_{i=1}^{N} \alpha_i t_i K(x_i, x) + b) \quad (5)$$

whereα_i are the Lagrange multipliers, $K(.)$ is a kernel function and sgn is the signum function. Kernel function is the inner product of two vectors i.e .

$$K(x_i, x_j) = \emptyset(x_i).\emptyset(x_j)[9] \quad (6)$$

Domain specific kernel function can be applied in the SVM. Some of the kernel functions are as follows:

$$Polynomial\ Kernel : K(X_i, X_j) = (1 + X_i^T X_j) \quad (7)$$

$$Gaussian\ Kernel : K(X_i, X_j) = e^{\left(\frac{1}{2\sigma^2}\|X_i - X_j\|^2\right)} \quad (8)$$

$$Linear\ Kernel : K(X_i, X_j) = X_i^T X_j \quad (9)$$

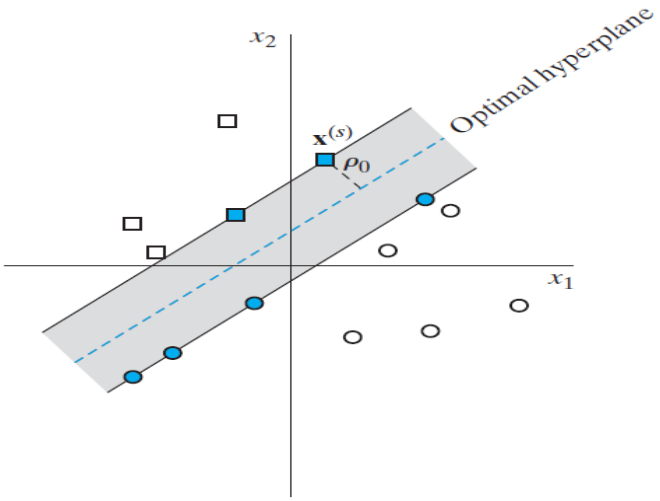

Fig 3.9: An illustration of classification by SVM. $\rho_0$is the minimal margin.
The data points in blue on both side of the optimal hyperplane are the support vectors

The SVM is a binary classifier and is used for multiclass classification. And hence Recognition or classification is done using voting scheme[31].

Methodology of SVM

In the algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features) where the value of each feature represents value of

a particular coordinate. Then, as a part of classification hyper-plane is identified that clearly differentiates the two classes fig. 3.10.

Support Vectors are simply the co-ordinates of individual observation

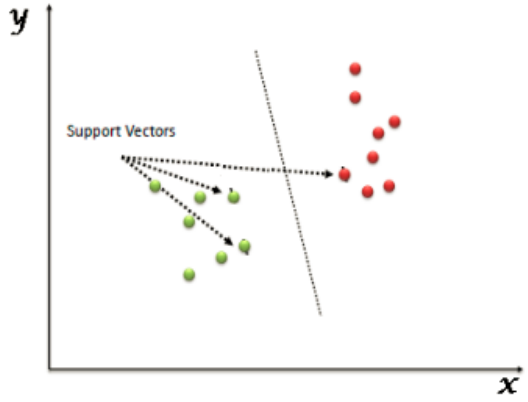

Fig. 3.10: Hyper-plane depicting the two classes

We have considered that SVM identifies the hyper-plane in 3 different scenarios, which segregates the two classes better and can better distinguishes between the classes in a neat manner.

Scenario-1: Identifying the right hyper-plane

Here, there are three hyper-planes (A, B and C). Right hyper-plane is to be identified to classify star and circle as shown in fig. 3.11.

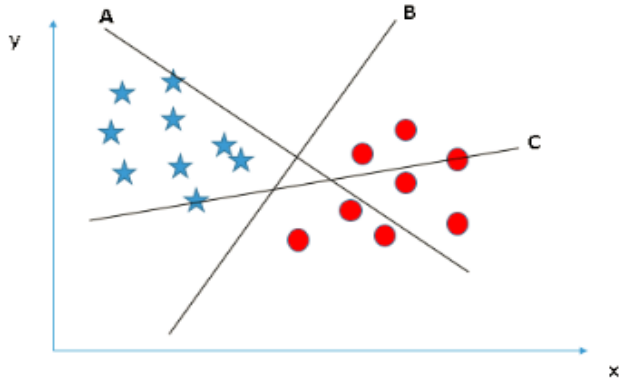

Fig. 3.11 Scenario-1: Identify the right hyper-plane

Hyper-plane "B" has excellently differentiated between the two classed i.e., Stars and Circles

Scenario-2: Identify the right hyper-plane

We have three hyper-planes (A, B and C) and all are segregating the classes well as shown in fig. 3.12.

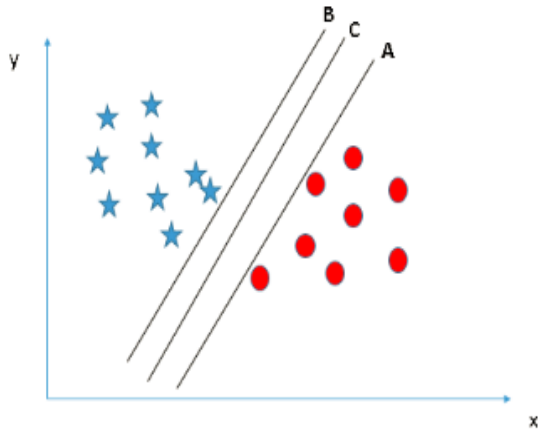

Fig. 3.12 Scenario-2: Identify the right hyper-plane

Concept of Margin: To decide the right hyper-plane we have to maximizing the distances between nearest data point (either class) and hyper-plane.

In fig. 3.13, compared to both hyperplaneA and B the hyper-plane C'smargin is high. Hence, the right hyper-plane is named as C. Another lightning reason for selecting higher margin is its robustness. If hyper-plane with low margin is selected then there might be a chance of miss-classification.

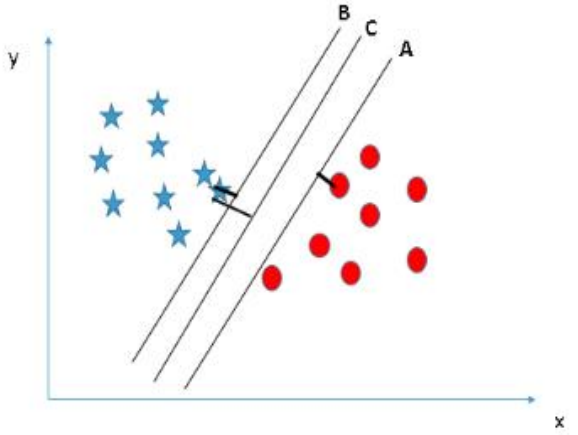

Fig. 3.13 Scenario-2: Margin distance between classes

Scenario-3: Identify the right hyper-plane

We have two hyper-planes (A and B) and all are segregating the classes well as shown in fig. 3.14.

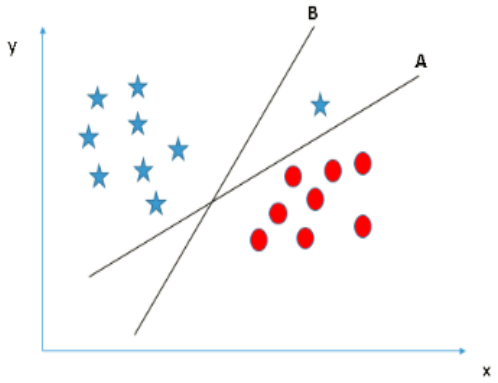

Fig. 3.14 Scenario-3: Identify the right hyper-plane

We might have selected the hyper-plane B as it has higher margin compared to A. But not, why? Because prior to maximizing margin SVM selects the hyper-plane which accurately classifies the classes. Here, hyper-plane B will result in a classification error but A has classified all correctly. Hence A is chosen as the right hyper-plane.

This approach of Histogram of Oriented Gradients (HOG) with Support Vector Machine (SVM) trains a SVM, to recognize HOG descriptors of people.

Main reasons to use HOG person detector is that instead of collecting 'local" feature it captures a "global" feature to describe a person. This means that the single feature vector is used to describe an entire person, as opposite to many feature vectors which represents smaller parts i.e., units of the person.

The concept of sliding detection window is used in HOG person detector which is moved around the image. A HOG descriptor is calculated at each position of the detector window. Later this descriptor is then displayed to the trained SVM, which then classifies it as either not a person" or a "person".

3.5 ENSEMBLE-BASED CLASSFIER

Emotions like happy, sad, surprised, joy, anger etc. are classified are being classified by Ensemble based distance classifier. The use of ensemble methodology has been considered by Researchers from various disciplines like Artificial Intelligence (AI) and statistics.

LiorRokachet. al[35] has discussed the idea of ensemble methodology, which explains the building of a predictive model by integrating multiple models as shown in fig. 3.15.

.

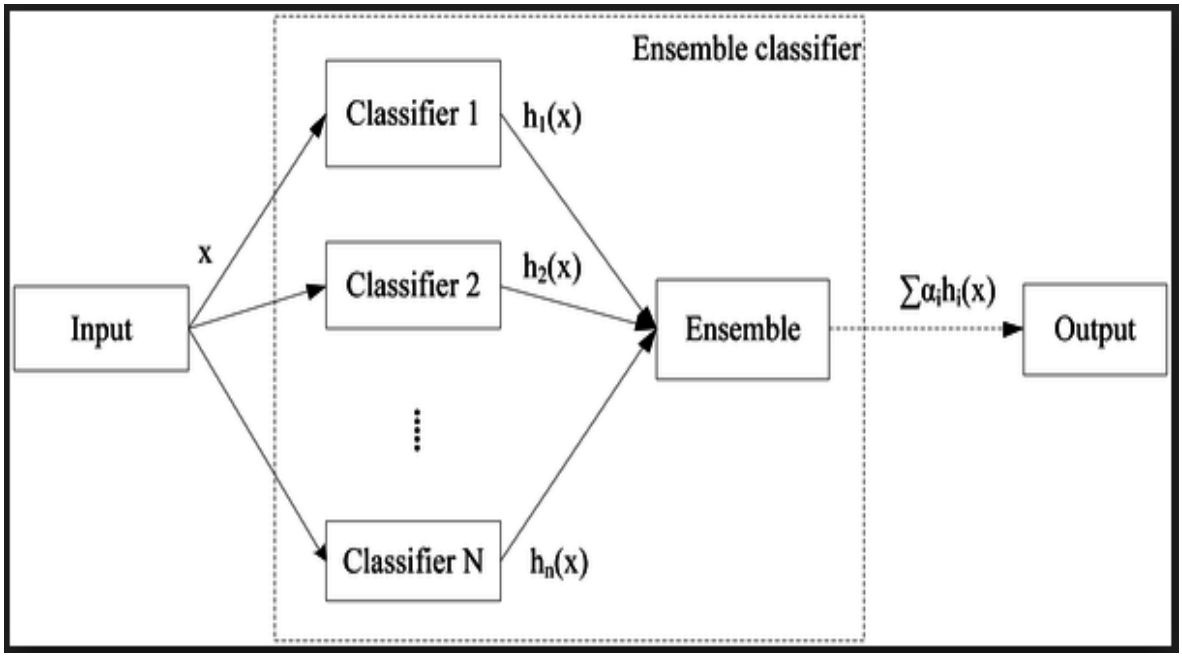

Fig. 3.15: An Ensemble-based classifiers

Supervised learning aims to classify patterns (also known as Instances) into a set of categories which are also referred to as Labels or Classes. Commonly, the classification is based on a classification models (classifiers) that are induced from an ideal set of pre-classified patterns.

The main idea behind the ensemble methodology is to weigh several individual classifiers, and combine them in order to obtain a classifier that out performs every one of them. Thus we weigh the individual opinions, and combine them to reach our final decision (Polikar 2006)[36].

Marie Jean Antoine Nicolas de Caritat, marquis de Condorcet (1743–1794) work presented the well-known Condorcet's Jury Theorem.

A strong learner is an inducer that is given a training set consisting of labelled data and produces an arbitrarily accurate classifier. A weak learner produces a classifier which is only slightly more accurate than random classification. The Theorem can also create "collection of weak classifiers create a single strong one"
To do so construct an ensemble that

(a) Consists of independent classifiers, each of which correctly classifies a pattern with a probability of $p > 0.5$; and
(b) Has a probability of $L > p$ to jointly classify a pattern to its correct class.
Sir Francis Galton (1822–911) was an English philosopher and statistician that conceived the basic concept to standard deviation and correlation. And said in Condorcet jury theorem, we can combine many simplistic predictions in order to obtain an accurate prediction.

3.6 PRINCIPAL COMPONENT ANALYSIS (PCA)

Being a dimension-reduction technique Principal Component Analysis (PCA) aims to form informative small set of linear uncorrelated variables from a pool of correlated variables

For N observations with P variables, the total number of principal components will be minimum of N-1 observations and P variables.

The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

To extract maximum variance from variables PCA seeks a linear combination of variable. It then remove the variance and perform second linear combination to explain the remaining variance and so on. This is called the principal axis method and results in orthogonal (uncorrelated) factors. PCA analyzes total (common and unique) variance.

The results of a PCA are normally discussed in terms of component scores, sometimes called factor scores and loadings.

PCA experts in providing reduced dimensional picture, without compromising with the Informative content. E.g. multivariate dataset which has high-dimensional data space (1 axis per variable).

Factor Analysis is another term of. Factor analysis uses more domain specific knowledge about the underlying data structure and provides eigenvectors of a mostly similar matrix.

3.7 CONVOLUTIONAL NEURAL NETWORK (CNN)

This section covers the popular deep learning model: Convolutional Neural Networks.

3.7.1 Introduction

It is the most popular deep learning architecture. CNN is now the go-to model on every image related problem. The main advantage of CNN is that it automatically detects the important features without any human supervision. For example, given many pictures of cats and dogs it learns distinctive features for each class by itself.
CNN is also computationally efficient. It uses special convolution and pooling operations and performs parameter sharing. This enables CNN models universally attractive and to run on any device.
It performs automatic feature extraction to achieve superhuman accuracy.

3.7.2 Architecture

All CNN models follow a similar architecture, as shown in the fig. 3.16.

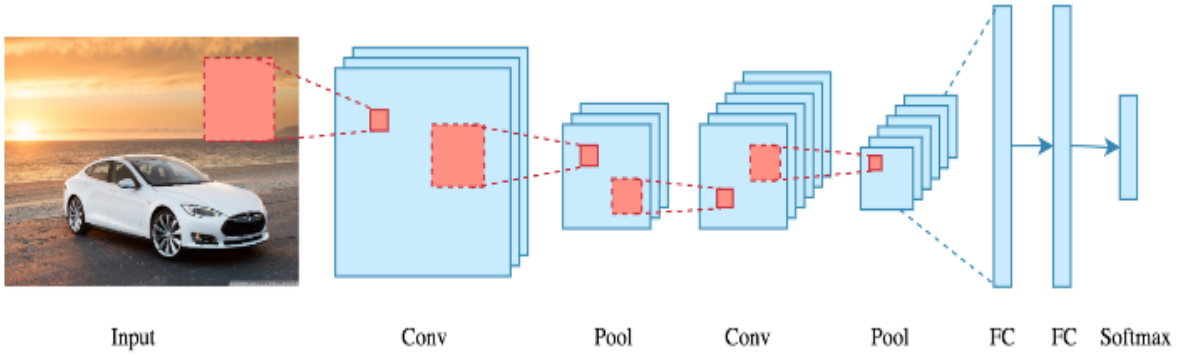

Fig. 3.16: CNN Architecture

There is an input image that we're working with. We perform a series convolution + pooling operations, followed by a number of fully connected layers. If we are performing multiclass classification the output is softmax. We will now dive into each component.

3.7.2.1 Convolution

The main building block of CNN is the convolutional layer. Convolution is a mathematical operation to merge two sets of information. In our case the convolution is applied on the input data using a *convolution filter* to produce a *feature map*. There are a lot of terms being used so let's visualize them one by one.

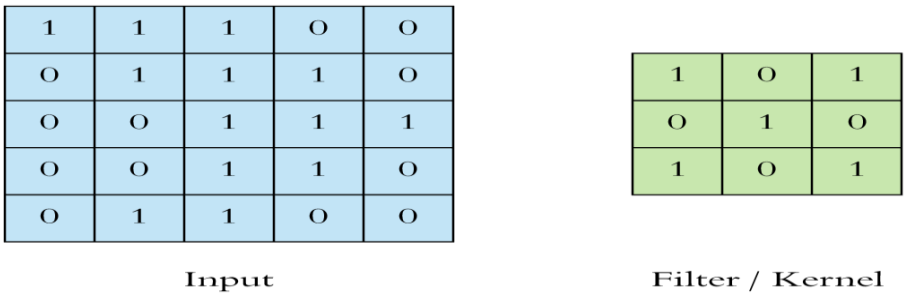

Fig. 3.17: CNN - Convolution Input and Filter

As shown in fig. 3.17, on the left side is the input to the convolution layer, for example the input image. On the right is the convolution *filter*, also called the *kernel*, we will use these terms interchangeably. This is called a *3x3 convolution* due to the shape of the filter. We perform the convolution operation by sliding this filter over the input. At every location, we do element-wise matrix multiplication and sum the result.

This sum goes into the feature map. The green area where the convolution operation takes place is called the *receptive field*. Due to the size of the filter the receptive field is also 3x3.

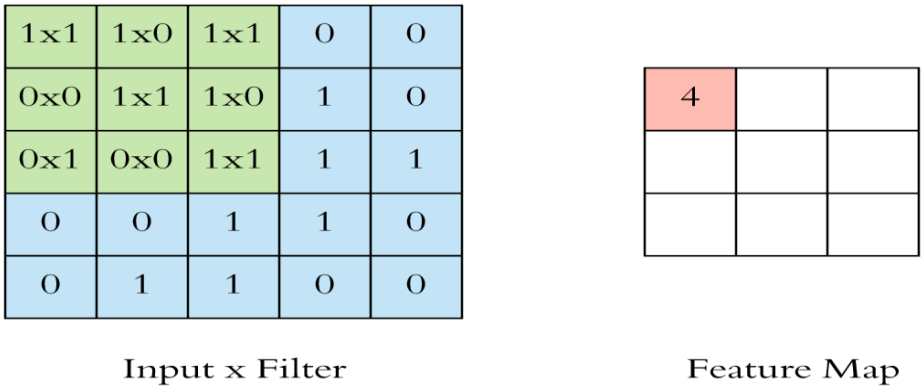

Fig. 3.18: CNN - Convolution Operation of sliding filter over input (a)

Here the filter is at the top left, the output of the convolution operation "4" is shown in the resulting feature map in fig. 3.18. We then slide the filter to the right and perform the same operation, adding that result to the feature map as well as shown in fig. 3.19.

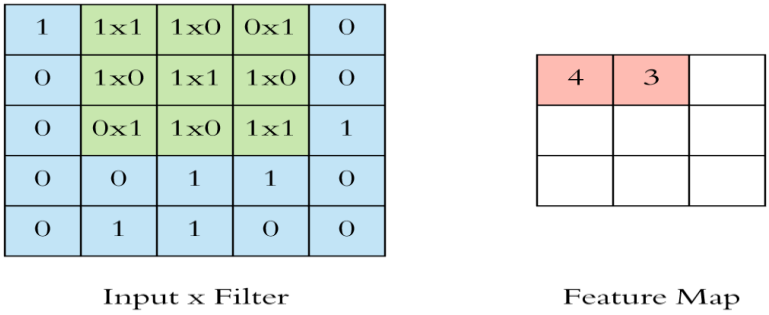

Fig. 3.19: CNN - Convolution Operation of sliding filter over input (b)

We continue like this and aggregate the convolution results in the feature map. The above stated example of convolution operation is shown in 2D using a 3x3 filter. But in reality these convolutions are performed in 3D. In reality an image is represented as a 3D matrix with dimensions of height, width and depth, where depth corresponds to color channels (RGB). A convolution filter has a specific height and width, like 3x3 or 5x5, and by design it covers the entire depth of its input so it needs to be 3D as well.

For any kind of neural network to be powerful, it needs to contain non-linearity.

Stride and Padding
Stride specifies how much we move the convolution filter at each step. By default the value is 1, as you can see in the fig. 3.20.

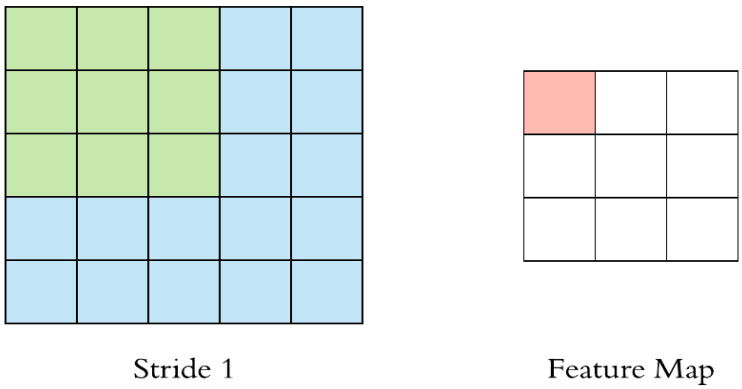

Fig. 3.20: CNN –Stride 1 and Padding

We can have bigger strides if we want less overlap between the receptive fields. This also makes the resulting feature map smaller since we are skipping over potential locations. The fig. 3.21 demonstrates a stride of 2. Note that the feature map got smaller.

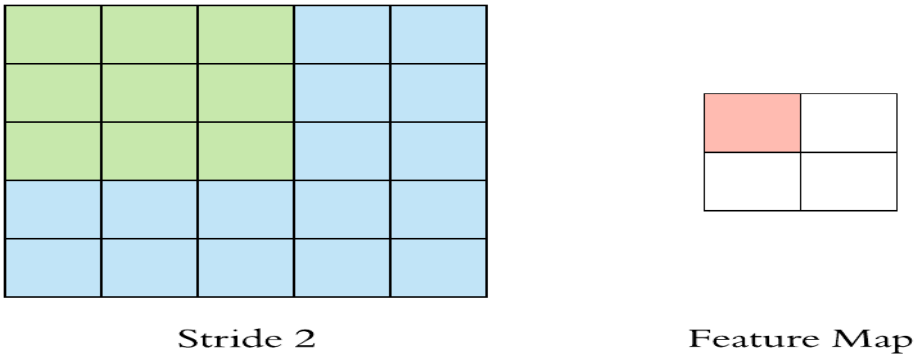

Fig. 3.21: CNN –Stride 2 and Padding

We see that the size of the feature map is smaller than the input, because the convolution filter needs to be contained in the input.

3.7.2.2 Pooling

After a convolution operation we usually perform *pooling* to reduce the dimensionality. This enables us to reduce the number of parameters, which both shortens the training time and combats over fitting. Pooling layers down sample each feature map independently, reducing the height and width, keeping the depth intact. The most common type of pooling is *max pooling* which just takes the max value in the pooling window. Contrary to the convolution operation, pooling has no

parameters. It slides a window over its input, and simply takes the max value in the window. Similar to a convolution, we specify the window size and stride.
Result of max pooling using a 2x2 window and stride 2 is shown in fig. 3.22.

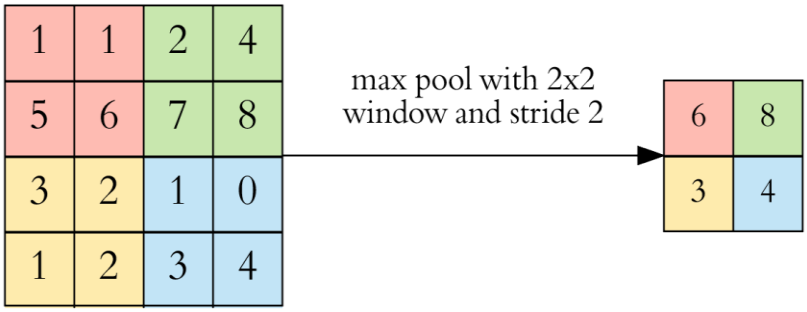

Fig. 3.22: CNN - Max pooling using a 2x2 window and stride 2

Each color denotes a different window. Since both the window size and stride are 2, the windows are not overlapping. Note that this window and stride configuration halves the size of the feature map. This is the main use case of pooling, down sampling the feature map while keeping the important information.

By halving the height and the width, we reduced the number of weights to 1/4 of the input. Considering that we typically deal with millions of weights in CNN architectures, this reduction is a pretty big deal.

In CNN architectures, pooling is typically performed with 2x2 windows, stride 2 and no padding. While convolution is done with 3x3 windows, stride 1 and with padding.

3.7.3 Intuition

A CNN model can be thought as a combination of two components: Feature Extraction part and the Classification part.

The convolution along with pooling layers performs feature extraction. For example given an image, the convolution layer detects features such as two eyes, long ears and so on. The fully connected layers then act as a classifier on top of these features, and assign a probability for the input image being a dog.

The convolution layers learn complex features by building layers on top of each other. The first layers detect edges, the next layers combine them to detect shapes, to following layers merge this information to infer that this is a nose. To be clear, the CNN doesn't know what a nose is. By seeing a lot of them in images, it learns to detect that as a feature. The fully connected layers learn how to use these features produced by convolutions in order to correctly classify the images.

3.7.4 Dropout

Dropout is by far the most popular regularization technique for deep neural networks. Dropout is used to prevent overfitting. During training time, at each iteration, a neuron is temporarily "dropped" or disabled with probability p. This means all the inputs and outputs to this neuron will be disabled at the current iteration. The dropped-out neurons are resampled with probability p at every training step, so a dropped out neuron at one step can be active at the next one.
The hyper-parameter p is called the dropout-rate and it's typically a number around 0.5, corresponding to 50% of the neurons being dropped out.
Thus we are disabling neurons on purpose and the network actually performs better. The reason is that dropout prevents the network to be too dependent on a small number of neurons, and forces every neuron to be able to operate independently.

Dropout can be applied to input or hidden layer nodes but not the output nodes. The edges in and out of the dropped out nodes are disabled as shown in fig. 3.23.
Almost all deep neural networks now incorporate dropout.

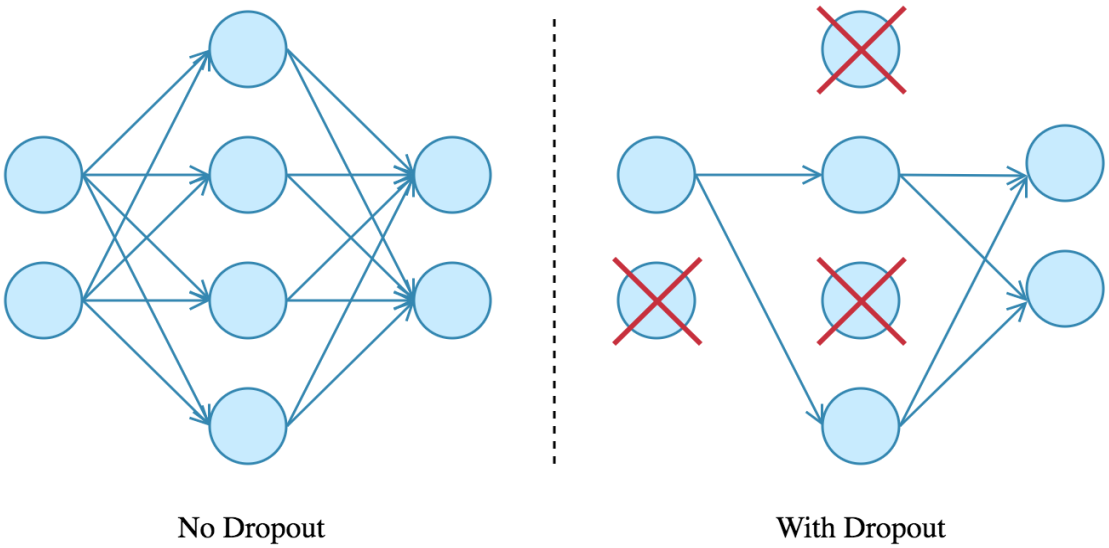

Fig. 3.23: CNN - Dropout regularization technique

3.8 METRICS FOR EVALUATING PERFORMANCE[41]

Following are some of the commonly used performance metrics of classifier:

- Accuracy (or recognition rate)
- Sensitivity (or recall)
- Specificity
- Precision

It is of always be noted that the metrics are always obtained over test data and not on training data.

3.8.1 Accuracy

Accuracy of a classifier is defined as the percentage of test set data that are correctly classified by it. Formally:

$$Accurcy = \frac{TP + TN}{P + N} \tag{10}$$

where

$$TP = number\ of\ True\ Positives$$
$$TN = number\ of\ true\ negatives$$
$$P = number\ of\ positive\ samples\ /\ data\ sets$$
$$N = number\ of\ negative\ samples\ /\ data\ sets$$

3.8.2 Error rate or misclassification rate

$$Error\ rate = 1 - Accuracy \tag{11}$$

$$Error\ rate = \frac{FP + FN}{P + N} \tag{12}$$

where

$$FP = number\ of\ false\ positives$$
$$FN = number\ of\ false\ negatives$$

3.8.3 Sensitivity

It is the true positive rate, i.e. the fraction of positive data set that are correctly classified.

$$Sensitivity = \frac{TP}{P} \tag{13}$$

3.8.4 Specificity

It is the true negative rate

$$Specificity = \frac{TN}{N} \tag{14}$$

Accuracy can be written in terms of sensitivity and specificity as follows

$$Accuracy = Sensitivity \times \frac{P}{P + N} + Specificity \times \frac{N}{P + N} \tag{15}$$

3.8.5 Class Imbalance Problem

In case of significantly less number of positive classes and large number of negative classes in the sampled data set, then the accuracy measure is not a suitable performance metric. This can be understood using an example. Following is the confusion matrix for the classes $cancer = yes$ and $cancer = no$. (here the positive class is $cancer = yes$ and negative one is $cancer = no$)

Classes	***Yes***	***No***	***Total***
Yes	90	210	300
No	140	9560	9700

Now from the above matrix, the performance metrics calculated as:

$$Accuracy = \frac{90 + 9560}{300 + 9700} = 96.5\%$$

$$Sensitivity = \frac{90}{300} = 30.0\ \%\ ,$$

$$Specificity = \frac{9560}{9700} = 96.56\ \%$$

Thus, although the classifier has high accuracy, it is mostly contributed by its ability to classify negative class and not the positive class. This is the case of class imbalance problem. In such a case, *sensitivity* and *specificity* are more significant metrics

3.8.6 Precision

It is the measure of the exactness, i.e. the percentage of data labelled positive is actually positive

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

For above confusion matrix, the value of precision is $\frac{90}{90+140} = 39.13\ \%$

Although, precision is a good metrics, the misclassified percentage cannot be obtained from it.

3.8.7 Recall

It quantifies the measure of completeness, i.e. percentage of positive data labeled as positive.

$$Recall = \frac{TP}{P} \tag{17}$$

Note that recall is same as sensitivity.

For comparison of two performance results, the measure of precision and recall are used in pairs and the comparison is carried out by fixing one measure (compare precision value at recall value of say 0.75 or vice versa).

CHAPTER 4

OVERVIEW OF WORK DONE

4.1 PROBLEM STATEMENT

Behaviour analysis on the participants in an E-education or M-education platform or any training sessions is significantly easier than doing the same for actual learning session. This is because the web and mobile based solution focus on the different matrices like search, click, time spent etc., which can be recorded for a qualitative analysis. However, a similar analysis for physical education session is extremely difficult because no data corresponding to participant's behaviour analysis is available. Therefore, for a physical education session analysis is not only a major computational challenge but also a huge analytical problem. A mathematical qualitative representation of the participant's behaviour in an educational session can significantly help in improving the participation, participant relationship, modelling the class format and overall learning session. This work focuses on addressing these challenges and offers a reliable solution for participant behaviour analysis in learning session with the help of real-time computer vision and statically aggregation.
The problem of "Human Multi-Emotion detection based on an Automatic analysis of facial Expression using Deep Learning" has been studied while focusing on the "Face Detection", "Feature Extraction" and "Deep Learning" module of the process.

4.2 OBJECTIVE OF PROPSED WORK

The main objective of this work is Facial Expressions detectionand Emotion classification of participant using Histogram of Oriented Gradients (HOG) detector and Deep Learning System.
"Face Detection" is done by using the Voila Jone'sAdaboost method. Emotion classification is done by using ensemble based model classifier which is used for improving prediction performance.The idea behind using ensemble methodology is to build a predictive model by integrating multiple models. Ensemble methods are used for improving the prediction performance. Researchers from various disciplines such as statistics and Artificial Intelligence (AI) consider the use of ensemble methodology.

The facial expressions are classified based on the features like eye brow raise, eye brow furrow, cheek raise, dimple, chin raise, eye widen, eye closure, led tighten, jaw drop, lip depressor, mouth open, nose wrinkle, smile etc.
The emotions targeted are the six universal prototypical emotional expressions (i.e., disgust, fear, joy, surprise, sadness, anger) proposed by Ekman [1,3-6] as shown in fig 4.1.

Fig. 4.1: Emotion-specified expressions of face : 1-disgust; 2- fear; 3-joy; 4- surprise; 5-sadness; 6-anger. (Cohn-Kanade database) [2]

The recognized emotions are sent to CNN by IBM Bluemix CNN platform using IBM Watson-IoT-gateway.

4.3 MOTIVATION

Often, it is difficult for a lecturer to understand the impact of a lecture in a learning session. This result in teaching session where there is a large disconnect of the lecturer from the participants. Especially in the field of education, it has been observed that distracted student results in bad learning experience. It is a proven study that if the lecturer knows the time port of the participant he can modify the teaching method to improve the participation in the class. This is the motivation of this work, where we have tried to analyse the feelings, emotions and focus of the participant in the session.

CHAPTER 5
PROPOSED WORK

This chapter explains the functional overview of study work. The proposed system is compatible for both windows and mobile. The system architecture is explained via block diagram and flow chart.

5.1 SYSTEM ARCHITECTURE

The architecture overview mainly comprises of system design of the proposed work, modules and the methodology. The block diagram as shown in fig 5.1, gives the schematic representation of the methods and components used in this work.

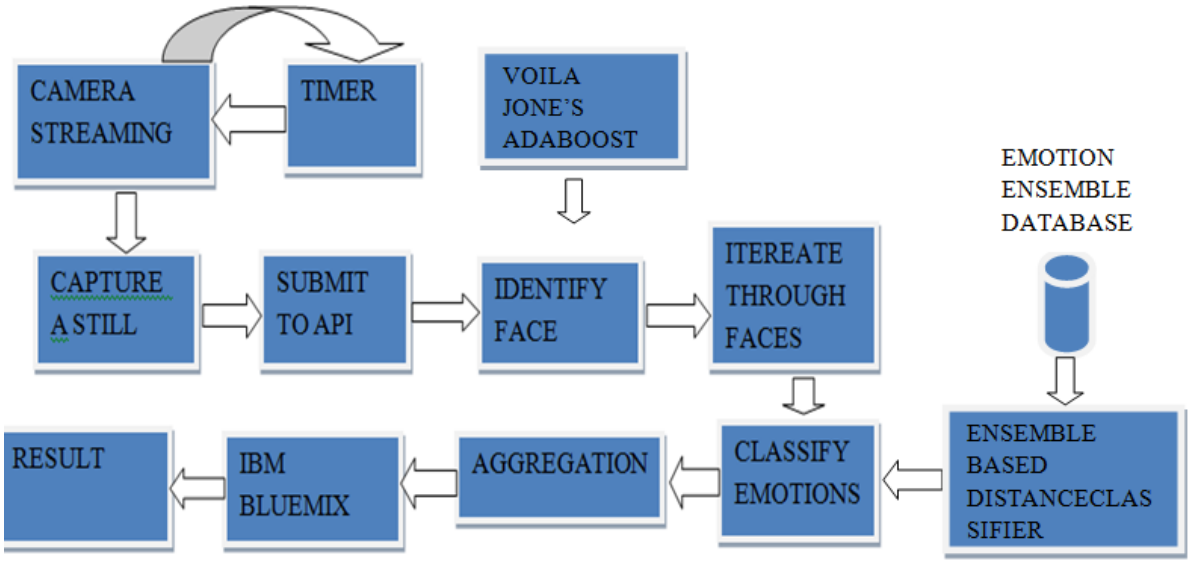

Fig. 5.1: Block diagram of proposed work

Camera streaming is done by acquiring image or video which becomes the input source of this project. Images are captured for certain time interval and frames are generated for that time. Timer manages the time for capturing the image and converting into frames. Frames are then submitted to Application Program Interface (API).
Face detection is done by analysing the geometric points on the human face. Face is identified by Voila Jones's method. This method combines many weak classifiers and make a strong classifier. The Voila Jones's method of object detection has been already explained in Chapter 2 of Literature Review.

Emotions are classified based on the feature points of the human face like chin raise, boarding of lips, widening of the mouth etc. Emotions are being classified by ensemble based distance classifier.

The emotions like happy, sad, surprised, joy, anger etc. are classified by using emotion ensemble database. The details about ensemble based distance classifies is already explained in Chapter 2.

Captured emotions are matched with the set of database and the best suited emotion is declared to be the final recognized emotion. These emotions are then aggregated and send to IBMBluemix CNN.

IBM Bluemix is a cloud Platform as a service (PaaS) developed by IBM.
It supports several programming languages and services as well as integrated DevOps (ie. software development (Dev) and software operation (Ops)) to build, run, deploy and manage applications on the cloud.

The result is stored using Message Queuing Telemetry Transport (MQTT) Data to NOSQL JavaScript Object Notation (JSON) storage Database.

MQTT (Message Queuing Telemetry Transport) is an ISO standard Publish-Subscribe-based Messaging Protocol which works on top of the TCP/IP protocol.
Under this publish-subscribe-based messaging protocol, Publish–Subscribe is a messaging pattern where senders of messages, called publishers, do not program the messages to be sent directly to specific receivers, called subscribers, but instead, it categorizes published messages into different logical classes without the knowledge/interest/demand of the subscribers.

Similarly, subscribers express interest in one or more classes and only receive messages that are of his interest, without knowledge of which publishers it belongs to and the message will be delivered. This data is interchange on the web using JSON, Where JSON is a human-readable text format that facilitates data interchange between different programming languages.

The Database Management Systems (NoSQL) is used to store data as JSON documents and are often referred to as document store databases/document-oriented database/aggregate database/simply document store/document database. Thus, NOSQL is a breed of DBMS where data is stored as JSON coming from MQTT. Display of the result can be done in tabular form of by the means of bar graph and pie chart.

5.2 WORKING PROCEDURE

The overall working procedure has been segregated into various modules:

5.2.1. Image capture

The proposed work is capable of presenting both windows and mobile based interface to capture the frames from a live classroom session. Furthermore, the system supports multiple frames through which different parts of the classroom and attendees frames can be sent to a main analysing unit through Fast Forward Moving Picture Expert Group (FFMPEG). FFMPEG has a powerful set of features relating to creating video from images or generating an image sequence from a video.

We have used an effective face Software Development Kit (SDK), which is basically a model based face and facial emotion tracking system. A presorted database of different facial emotion is used by a classifier provided by SDK to classify the emotion of a particular face. The facial index API returns the relative current location of the faces and their corresponding face ID's, this face ID's are used to initiate the independent section tracking the emotion of each of the independent face present in the frame. The emotion tracking system mainly presents the information about –

(a) Gender and age.

(b) Dominating emoji's such as smart, happiness, sadness.

(c) Different facial expressions such as broadening of the lips, shortening of the chin, closing of the eyes, boarding of the eyes and so on.

(d) It also presence an aggregated overall happiness index. These dilouserare being continuously sent to the IBM Bluemix.

5.2.2 Integration of Emotion to CNN

The captured emotions are continuously sent to the IBM bluemix CNN through MQTT protocol. The data is converted into a Json string which contains ID, emotion, emoji, facial expression. The Json array comprising of the Json string is constructed through the emotion tracking of the each of the independent faces in a frame being sent to Bluemix IOT gateway, which is a dedicated secure gateway. TheBluemix IOT gateway (Watson IOT gateway) is also linked with the object data storage. Such that the data dissipated is permanently stored in the object data storage. Bluemix Watson is a pre-built service provided by IBM to assemble application faster.

5.2.3. Analysis

Once the streaming of the data ends (the end of the class) the python script is called which reads the data stored in the object storage, and offers an aggregated result. Through the result we can come to know what is the overall emotion of the class or what is the attentiveness level of the class.

5.2.4. Facial Model

The emotion tracking is performed by fitting a facial model of the participant in a video frame as of an estimated initialization. In this case the facial model contains 70 points, which is shown in Fig. 5.2. Real Time facial model containing 70 points is shown in fig. 5.3

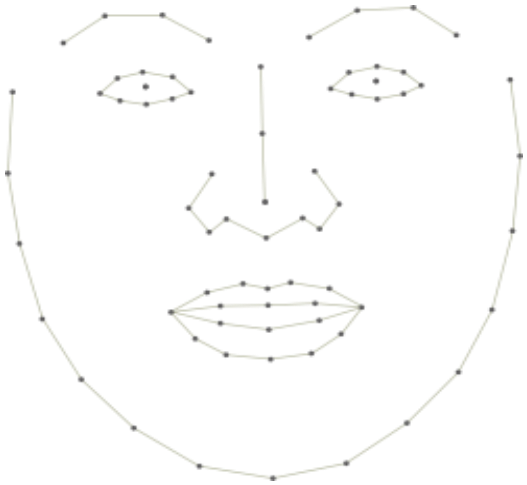

Fig. 5.2: Facial model containing 70 points

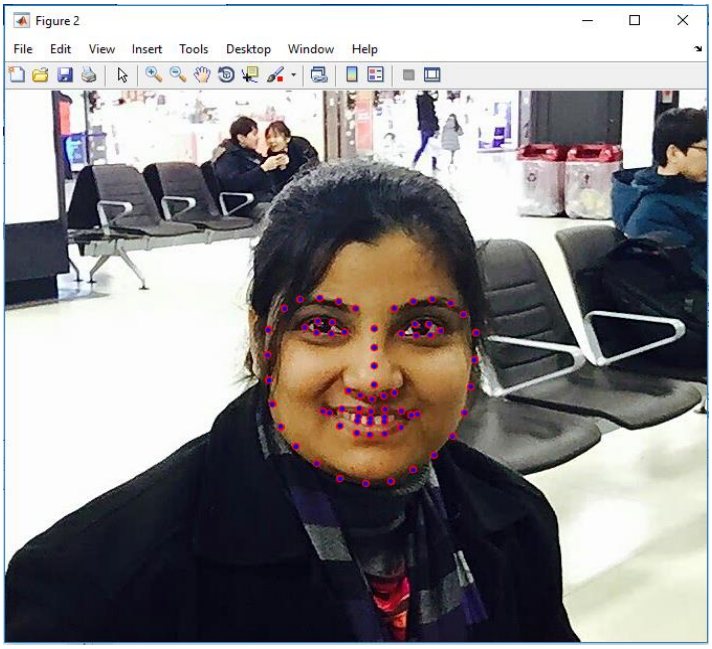

Fig. 5.3: Image showing facial land mark points

The algorithm fits the trained model by means of 70 lightweight classifiers, i.e. one classifier for each individual point in the model.

A primary approximate spot is specified, the classifiers look for a small area (thus the name 'local') around each point for an enhanced fit, and the model is then stimulated incrementally in the track giving the most excellent fit and progressively converting on the most favourable fit.

The fitting process on a gray scale face is shown in fig. 5.4.

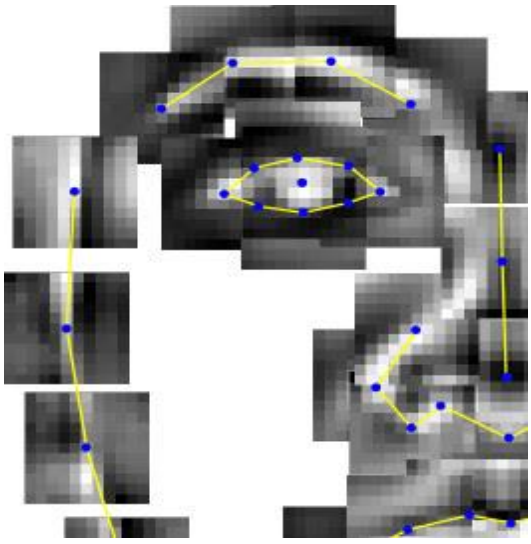

Fig. 5.4: Fitting of model data points on an image

A facial model is an annotated face data as shown in fig. 5.5 (approximated face annotation). As faces from one person to another don't vary much in terms of the geometry, a facial model is easy to construct. The model used in this work is a labeled face dataset to construct a face model.

Fig. 5.5: Approximated face annotation

To construct a model from these observations, we use Principal Component Analysis (PCA). The explanation about PCA is already explained in chapter 2.The mean points of the observations are calculated first and after that the mean points

of all the observations are found out. PCA is used to extort the variation as a linear vector set.

5.2.5. Process Flow Chart

A flow chat of a typical facial annotation extracted by model fitting process is shown in fig. 5.6.

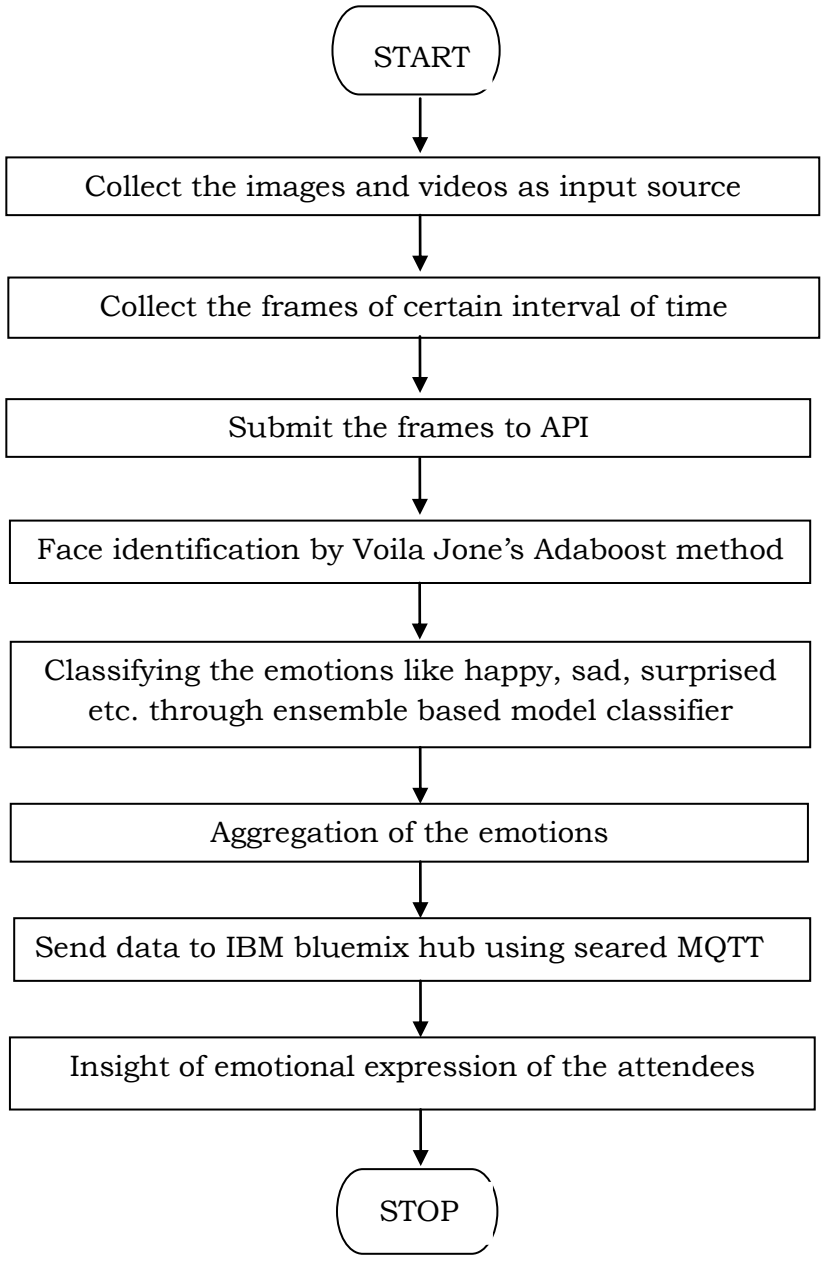

Fig 5.6: Flow chart of proposed work

Brief explanation of the process as shown in the flow chat above is explained as below:

- Image is captured by camera detector demo model and video by video detector demo model. We have an additional option to create separate application for this purpose by using android studio.
- Face is captured by using effective multi face SDK. Face is detected by tracking the geometric points of human face. Total 70 points are matched to detect the face.
- We have two algorithms i.e. appearance based algorithm and model based algorithm. We have mainly deal with model based algorithm in the study work.
- Face identification is done by using Voila Jone's method. This method has special feature of combining several weak classifiers and making strong classifier.
- Classifying the emotions based on features like eye closure, eye widen, chin raise, cheek raise, jaw drop, lip stuck, mouth open, smile, nose wrinkle, brow furrow, brow raise etc.
- Matching the emotions with ensemble based database using ensemble based distance classifier and finalizing the emotion which is best matched based on features.
- Aggregating the emotions and sending to IBM Watson Bluemix CNN.
- Displaying the results in form table or bar graph or pie chart.

CHAPTER 6
EXPERIMENTAL APPROACH

The study work has been carried out using a Laptop with Intel core i5 processor and 4 GB of AMD RAM. Simulations were developed and implemented using MATLAB version- R2015b.

6.1 WORK FLOW METHODOLOGY

This section covers the actual work flow strategy followed in the study work-

1. *face_image.m* file is executed for extracting features from facial images. It uses a feature vector called Constrained Local Model (CLM), which runs on images of faces. CLM can also run on videos of faces.

CLM features are also commonly known as "HOG" features, which are extracted linearly. This concept is already explained in Chapter 2.
The advantage of this method is that it can track even faceslooking at other directions and multiple faces in images.

2. Images are taken from 'samples' directory already created as depicted via code snippet shown in fig. 6.1. The source of the image is the free available databases as discussed in chapter 4.

We have an option also to download emotion images and save them as sample1.jpg, sample2.jpg etc. in 'sample' directory.

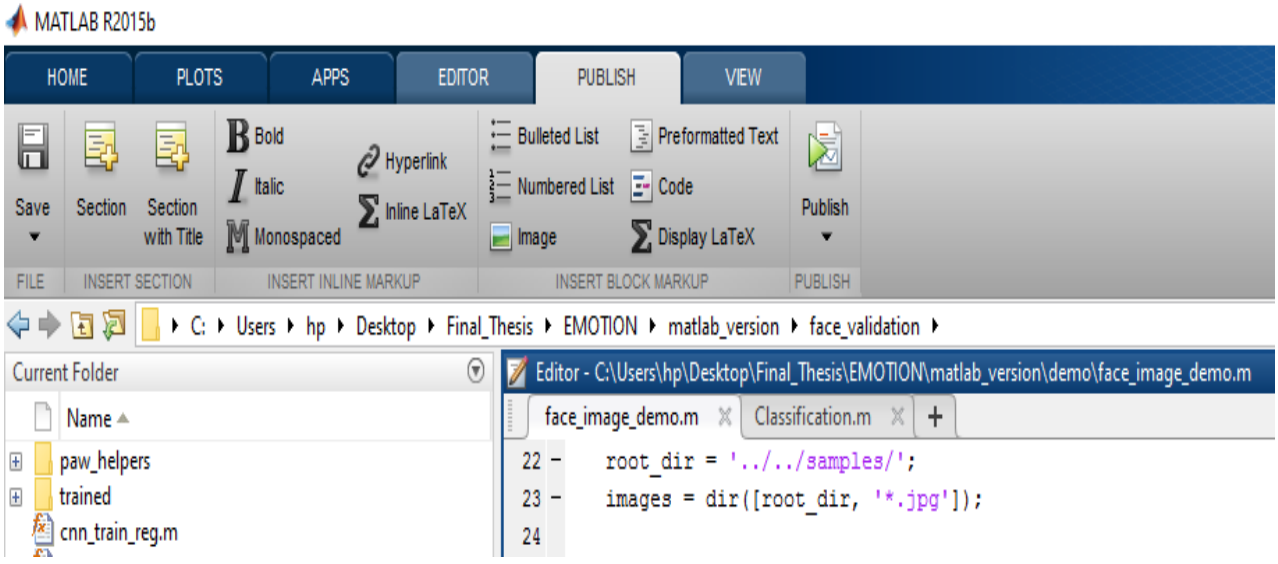

Fig 6.1: Image showing Matlab code to set 'samples' as default directly for images

3. For images, database is read. *Classification.m*file covers to code to read images from database. This covers code how classification of emotions is done.Matlab code is depicted via fig. 6.2.

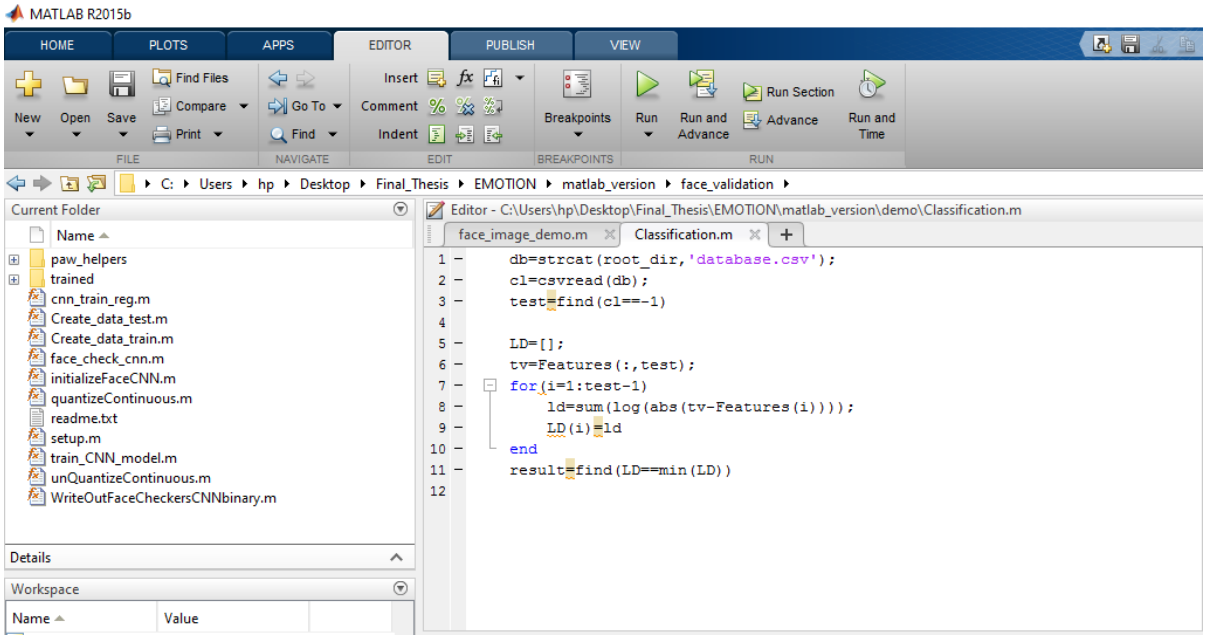

Fig 6.2: Image showing Matlab code for classification

4. After tracking a face, it extracts facial land mark points. A landmark is a distinct area of the face like the lips, eye ball, eye brow and so on as shown in fig 6.3.

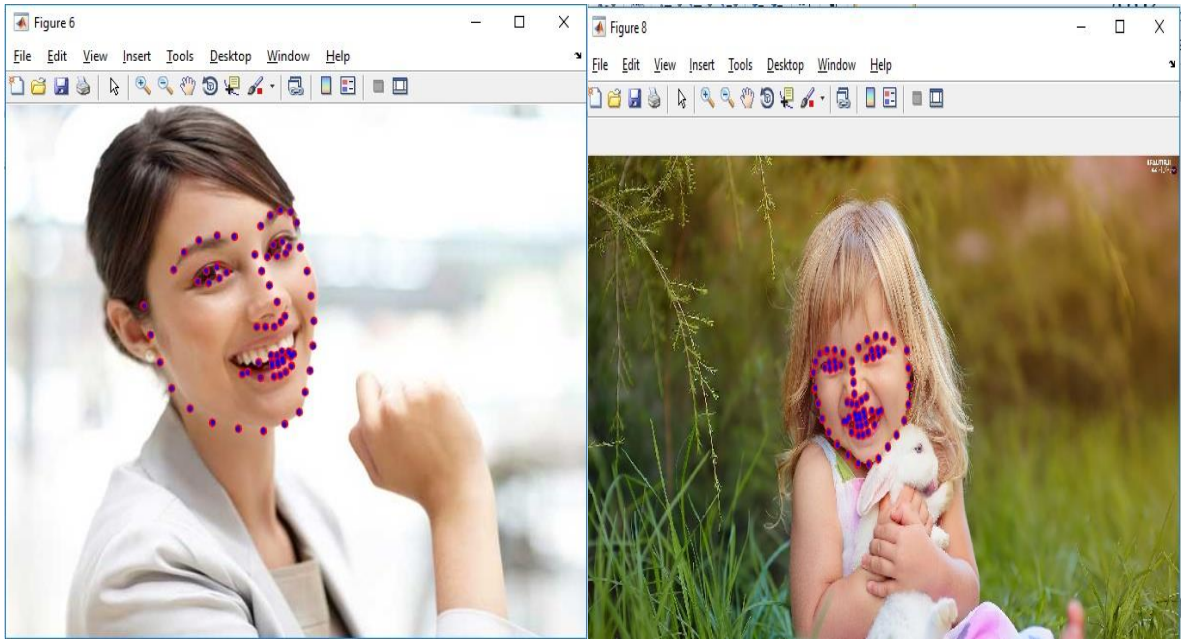

Fig 6.3: Image showing facial land mark points

5. X-Y coordinates of landmarks are contained in the file referred to as Shape.

6. Emotion can be determined, simply by training this model, and then testing against any other trained image.

7. Since the photo size, age of the person will vary; the next step is to normalize the points. This can be done by dividing x and y with their largest corresponding values. This refers to the Feature Extraction Process.
   Just like images, you can track the facial features in the Video.

8. Next is the Emotion Recognition part. The features are x, y; so cannot be used as features. The distance is taken from each point with the other point and uses it as feature vector. The objective is to detect the emotion using these features.

For eg., Let's say that sample number 1to n represents different emotion classes and n1+1 is the test sample. Database is prepared like this that last image is test image. Classes can Happy, sad, surprised, fear etc. based on emotion types.

6.2 RESULTS

Fig. 6.4 shows the proposed GUI for capturing the facial image from live video streaming.

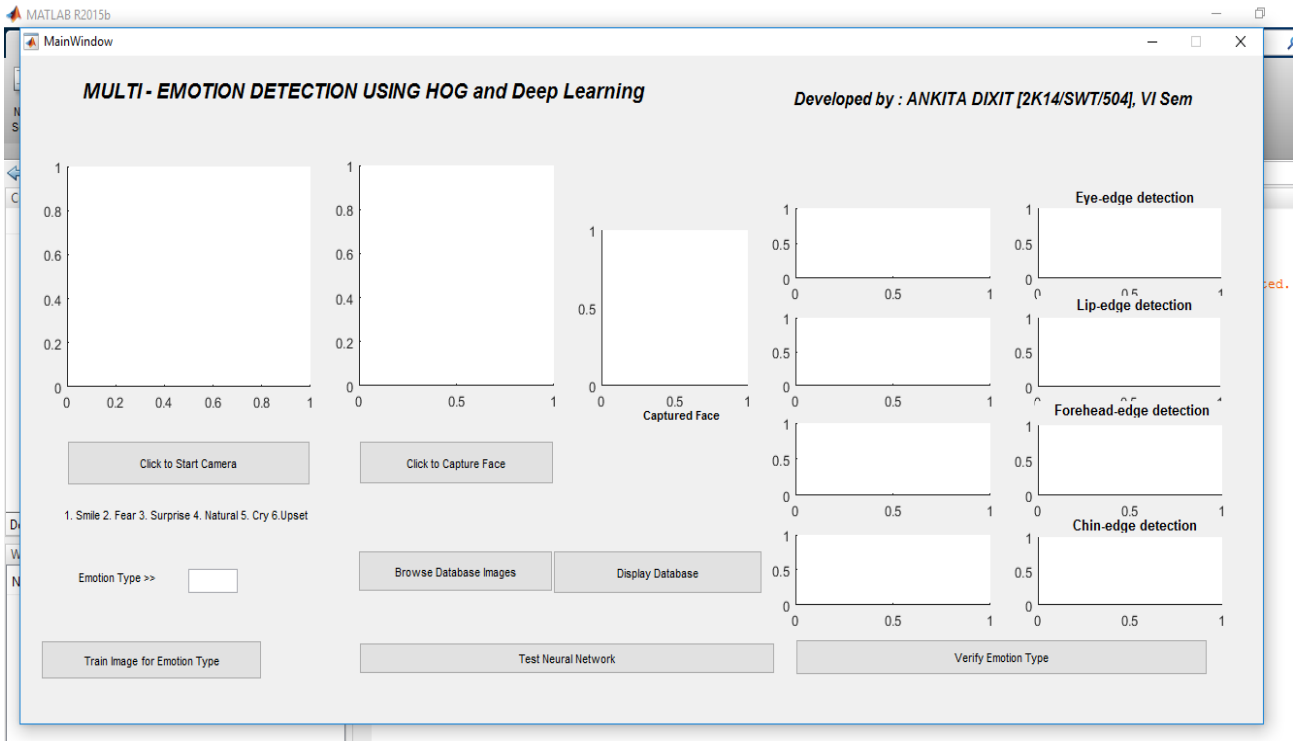

Fig. 6.4: Proposed GUI for capturing facial image

- Pressing “Click to Start Camera” the Video Camera of Laptop is activated

- Using “Click to Capture Face” : he face is extracted as shown in fig. 6.5
  - The extracted face is displayed in “Captured Face” portion.
  - The extracted facial features are displayed Viz. Forehead, eyes, nose and cheeks.
  - Voila-Jones Face Extraction Algorithm is used to extract face and facial features
  - The Sobel Edge detection is used to simply detect the edges of the extracted facial features.

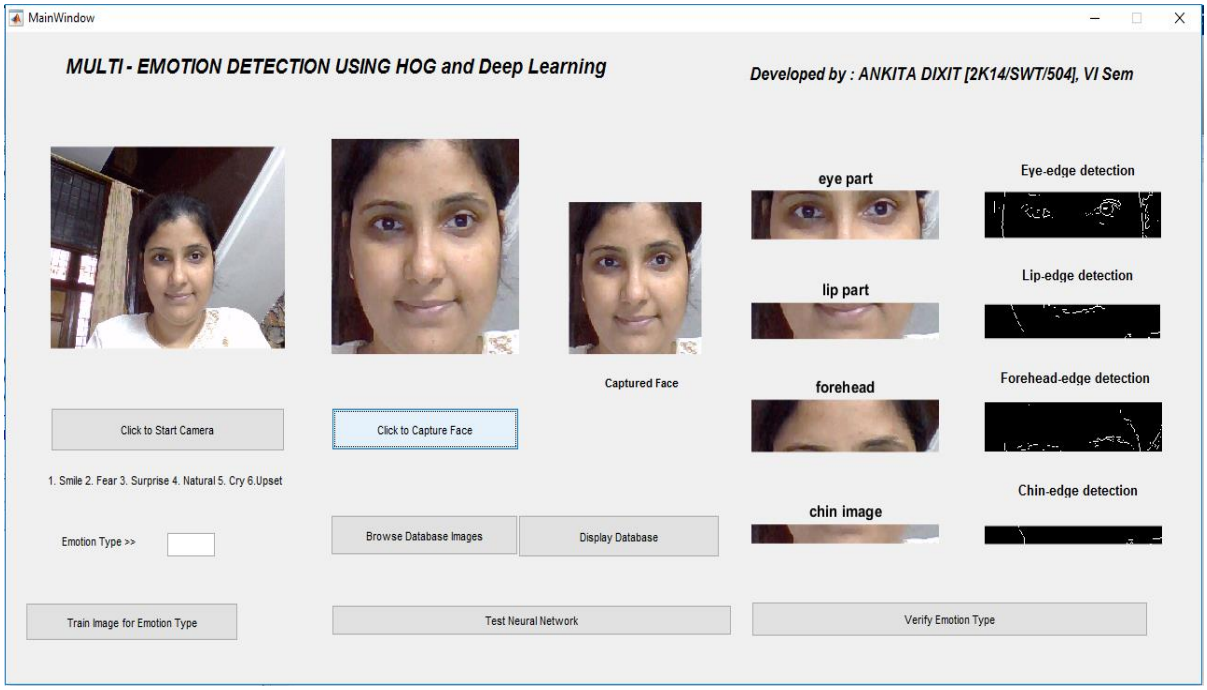

Fig. 6.5: Face extraction and edge detection of facial features

- The Captured Images are trained for various types of emotions by pressing “Train Image for Emotion Type”.

- The Pop “Training done” is displayed as shown in fig 6.6, recognizing the Emotion type initially trained ie. Selected Emotion Type is saved

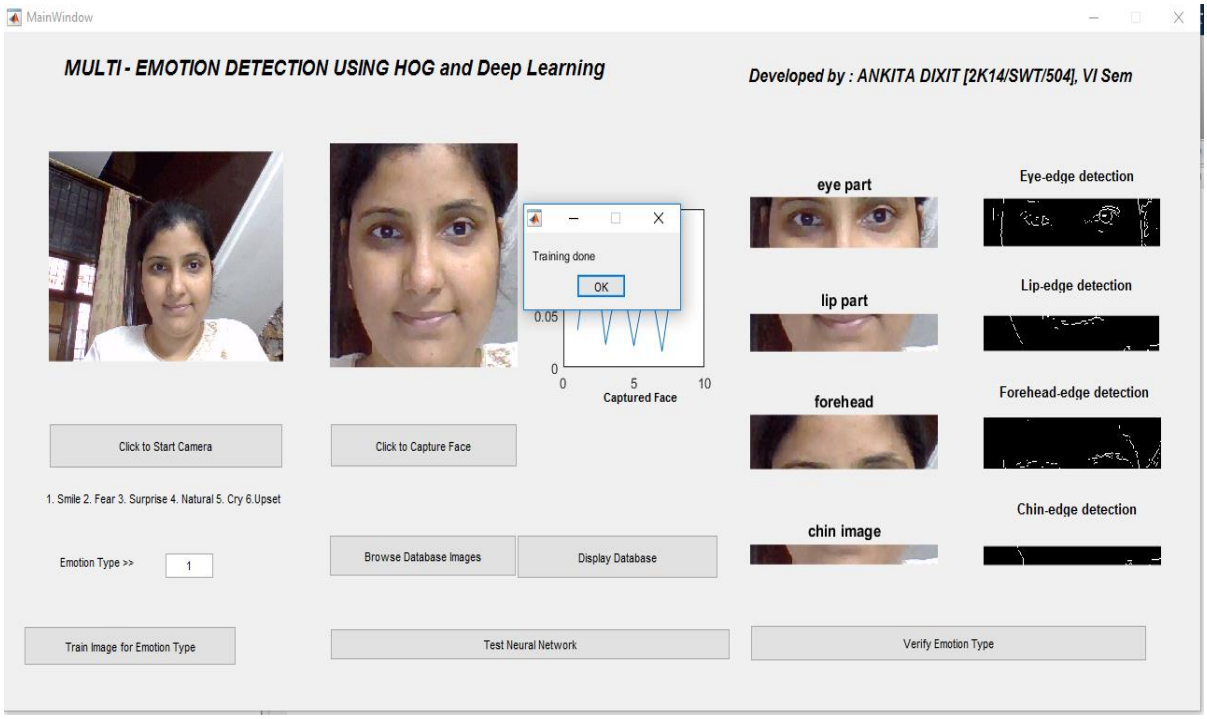

Fig 6.6: Image is trained for selected Emotion Type

- For testing the accuracy, new image is captured and the type of emotion is verified for test image by pressing "Verify Emotion Type" as shown in Fig. 6.7 HOG detection is used Emotion detection.

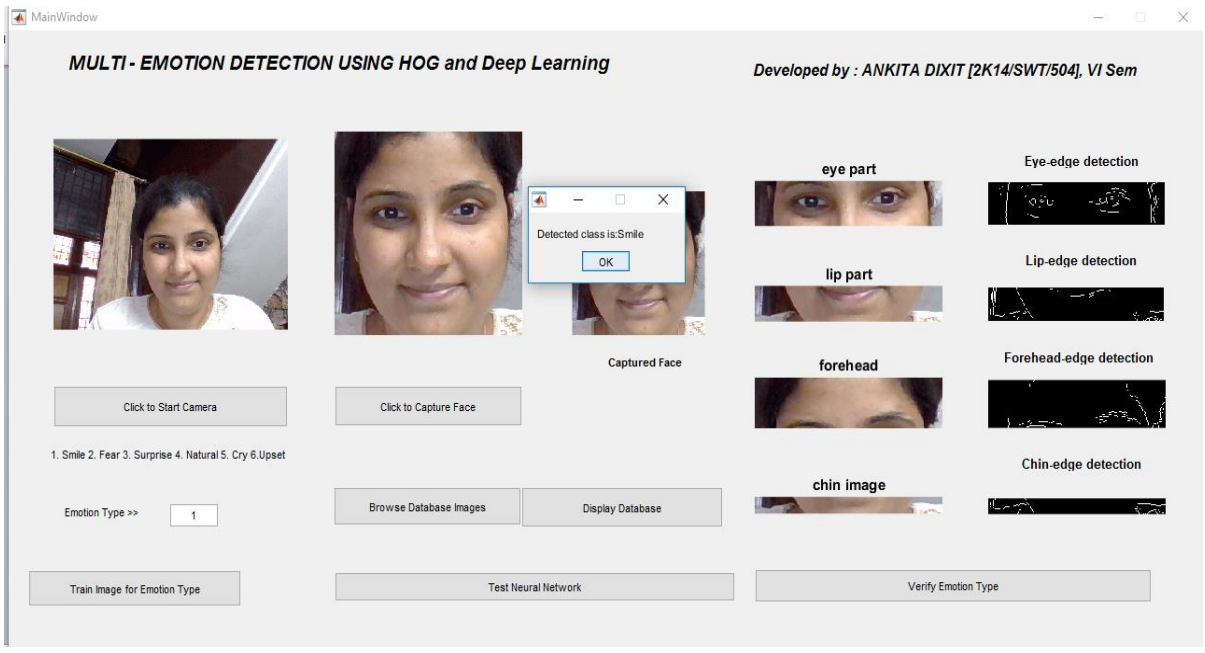

Fig 6.7: Correct Emotion is detected for test image after comparing with trained image

For experimentation, Cohn-Kanade and FEI image database are used.
We have selected 382 image sequences from the Cohn-Kanade database as these are the labeled ones among 593.

The detailed description of each database is presented in table 6.1 below.

Database	***Neutral***	***Anger***	***Contempt***	***Disgust***	***Fear***	***Happy***	***Sadness***	***Surprise***	***Total***
FEI	200	*NA*	*NA*	*NA*	*NA*	200	80	*NA*	**400**
Cohn_Kanade	382	184	96	268	112	296	136	336	**1810**

Table 6.1: Database description – number of images in each expression class

For testing, we have considered two main emotion Classes: Happy and sadness.
The experiments have been performed on Happy and sad emotion images using HOG.
Table 6.2 depict the percentage Accuracy tested on Happy and Sad emotion images used from available database as shown in Table 6.1.

Database	***Happy*** (%)	***Sadness*** (%)
FEI	86.27	89.56
Cohn – Kanade	88.03	90.66

Table 6.2 Percentage Accuracy for Happy and Sad Emotions

CHAPTER 7
DISCUSSION

7.1 CONCLUSION

It is often very difficult to know what the attentiveness of the attendees in the class is. Further it is also very difficult to analysis how well the students are focused on the class or how well the person responds to any training session. With recent advancement the facial emotion tracking technique it is not possible to track facial emotion in efficient way. However multiple emotions tracking and aggregating the result in the context of the machine learning is still an extremely challenging job. In this work, we have overcome this problem by integrating facial tracking API with emotion tracking API provided by effective SDK. The frames captured from either a mobile camera or laptop camera is given as input to the face SDKs, face API which return the number of faces. An independent session analysis of emotion of each of the faces is done and the emotion data is mitigated into the IBM Watson IOT gateway. Further at the end of the class a python scripts aggregates the data by reading them from the object storage and provides an overall insight of the class/session. Result shows that system proposed by us is extremely real time and is capable of extracting and mitigating the emotion data without any significant latency or lag. This also shows a brooded usability of the emotion tracking in the contextual learning process.

7.2 FUTURE SCOPE

The system can be further improved by incorporating live aggregation system by means of which a teaching professional or trainer or lecturer can simultaneously know overall mood of the class. This will help to take appropriate measure if the class is not attentive enough or if listeners/trainees are not happy enough with the sessions.

Emotions are a way to communicate with people. Nowadays expressions play a very important role in knowing ones feelings and mindset. Facial emotion recognition has several applications in different areas.

This proposed work can also be well implemented in mentioned area of fields:

- **Video surveillance:** Face detection and tracking is done in many areas such as malls, traffic signals, airports, railway stations etc. by fixing camera and recording videos.

- **Human computer interface:** To acquire high security the face recognition system can be used as software to unlock several devices.

- **Marketing:** It is used in marketing field to keep records of customers that have approached to them.

- **Education:** To know the expressions of the students during class and improving the teaching method so that the class becomes interesting.

- **Medical field:** To know the expression and emotions of patients who are not capable of speaking or to monitor the effect of ongoing treatments.

- **Performance Improvement:** Overall expressions of the people after some interview or some trainings, etc. can be used to predict the quality of the interview and training sessions. Based on the same, performance can be improved.

References

[1] M. F. Valstar, M. Mehu, J. Bihan, M. Pantic, and K. Scherer, "Meta-Analysisof the First Facial Expression Recognition Challenge," *Systems, Man, andCybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 42, pp. 966-979,2012.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94-101.

[3] P. Ekman, "Facial expression and emotion," *American psychologist,* vol. 48, pp. 384, 1993.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, pp. 169-200, 1992.

[5] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Handbook of face recognition," *Ch. Facial Expression Analysis, Springer, London,* pp. 487-519, 2005.

[6] Z. L. Stan and A. K. Jain, "*Handbook of face recognition*," ed: Springer, 2005.

[7] *Histograms of Oriented Gradients for Human Detection*{Navneet.Dalal,Bill.Triggs}@inrialpes.fr, Available online http://lear.inrialpes.fr

[8] B. Fasel and J. Luettin, "*Automatic facial expression analysis: a survey*" Pattern Recognition, vol. 36, pp. 259-275, 1// 2003.

[9] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: *A comprehensive study," Image and Vision Computing, vol. 27*, pp. 803-816, 2009.

[10] M. Suwa, N. Sugie, and K. Fujimora, "*A preliminary note on pattern recognition of human emotional expression*," International Joint Conference on Pattern Recognition, pp. 408–410, 1978.

[11] M. Pantic and L. J. M. Rothkrantz, "Automatic Analysis of FacialExpressions: The State of the Art," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 22, pp. 1424-1445, 2000.

[12] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from longimage sequences using optical flow," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 18, pp. 636-642, 1996.

[13] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *Pattern Analysis and Machine Intelligence,IEEE Transactions on,* vol. 19, pp. 757-763, 1997.

[14] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. PAMI,23(4):349– 361,April2001.

[15] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facialexpression understanding from image sequences," *Pattern Analysis andMachine Intelligence, IEEE Transactions on,* vol. 27, pp. 699-714, 2005.

[16] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binarypatterns," in *Computer vision-eccv 2004*, ed: Springer, 2004, pp. 469-481.

[17] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, 2005, pp. 76-76.

[18] Jason M, Jeffrey F and Cohn and Saragih Simon Lucey "*Deformable model fitting by Regularized landmark mean-shift*", International journal ofcomputer vision, vol. 91, pp. 200-215, 2011

[19] A.W. Senior, L. Brown, A. Hampapur, C.-F.Shu, Y. Zhai, R.S. Feris, Y.-L. Tian, S.Borger, C.Carlson "*Video analytics for education*", IEEE Conferenceon Advanced Video and Signal Based Surveillance, pp. 423 – 428,September 2007.

[20] Minghua Han, "*Participant Segmentation model based on educationconsumer behavior analysis*", International Symposium on intelligence information technology application workshops, pp. 914-917, Dec 2008.

[21] O.Déniza G.Buenoa J.Salidoa F.De la Torreb, "*Face recognition using Histograms of Oriented Gradients*" Available Onlinehttps://doi.org/10.1016/j.patrec.2011.01.004

[22] P. Viola, M. J. Jones, and D. Snow. Detecting pedestriansusing patterns of motion and appearance. The9thICCV ,Nice,France,volume1,pages734–741, 2003.

[23] Oscar Déniz, Daniel Hernández, Javier Lorenzo , "A comparison of face and facial feature detectors based on the Viola–Jones general object detectionframework"May 2011, Volume 22, Issue 3, pp 481–494

[24] Jianfeng Ren, Nasser Kehtarnavaz, Leonardo Estevez, "Real-time optimizationof Viola -Jones face detection for mobile platforms", Available online :https://ieeexplore.ieee.org/abstract/document/4695921/

[25] M. Venu Gopalachari and P. Sammulal "Personalized collaborative filteringrecommender system using domain knowledge", 2014 Internationalconference on Computer and Communications Technologies (ICCCT), pp. 1-6, December 2014.

[26] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *Systems, Man, and Cybernetics, Part B Cybernetics, IEEE Transactions on,* vol. 34, pp. 1449-1461, 2004.

[27] Mohamed Dahmane Jean Meunier, "*Emotion recognition using dynamicgrid-based HoG features*" Available onlinehttps://ieeexplore.ieee.org/abstract/document/5771368/

[28] Anurag De, Ashim Saha, "A Comparative Study on different approaches ofReal Time Human Emotion Recognition based on Facial ExpressionDetection", 2015 International Conference on Advances in Computer Engineering and Applications, pp. 483 – 487, March 2015.

[29] D. A. Pollen and S. F. Ronner, "Phase relationships between adjacent simple cells in the visual cortex," *Science,* vol. 212, pp. 1409-1411, 1981.

[30] Shereen Morsy, M.Elemam Shehab, Rana Alaa El-Deen Ahmed andNermeen Mekawie "Performance study of classification algorithms forconsumeronline shopping attitudes and behavior using data mining",2015Fifth International

conference on communication systems and network technologies (CSNT), pp. 1344-1349, April 2015.

[31] G. Zhao and M. Pietikäinen, "Boosted multi-resolution spatio temporal descriptors for facial expression recognition," *Pattern recognition letters,* vol. 30, pp. 1117-1127, 2009.

[32] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 21, pp. 974-989, 1999.

[33] P. Viola, M. J. Jones, "*Robust Real-Time Face Detection*", InternationalJournal of Computer Vision 57(2), 137–154, 2004

[34] Ashish, Aravind Sankar, Payal Bajaj, Sumit Shekhar, Iftikhar AhamathBurhanuddin and Dipayan Mukherjee "Similarity learning for productrecommendation and scoring using multi-channel data" 2015 IEEE 15thInternational conference on Data mining workshops, pp. 1143-1152,November 2015.

[35] LiorRokach, L. ArtifIntell Rev (2010) 33: 1, DOI https://doi.org/10.1007/s10462-009-9124-7© Springer Science+Business Media B.V. 2009

[36] Polikar R (2006) Ensemble based systems in decision making. IEEECircuits Syst Mag 6(3): 21–45

[37] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification ofsingle facial images," *IEEE Transactions on Pattern Analysis and MachineIntelligence,* vol. 21, pp. 1357-1362, 1999.

[38] Hoa Zhang, Carson K. Lengu, Adam G.M Pazdor and Fan Jiang, "A datascience model for big data analytics of frequent patterns", 016 IEEE 14thIntl conf on dependable and secure computing, pp. 866-873, October2016.

[39] Dongshun Cui, Guang-Bin Huang, Tianchi Liu, "Smile Detection UsingPair-wise Distance Vector and Extreme Learning Machine", 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2298 – 2305, July 2016.

[40] W. Shier, S. Yanushkevich, "Pain Recognition and Intensity ClassificationUsing Facial Expressions", 2016 International Joint Conference on NeuralNetworks (IJCNN), pp. 3578 – 3583, November 2016.

[41] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition byindependent component analysis," *Neural Networks, IEEE Transactions on,* vol. 13, pp. 1450-1464, 2002.

[42] Samuel Fosso Wamba and Shahriar Akter, "Big data analytics in E-education: a systematic review and agenda for future research" Electronic markets, Vol 26, no 2, pp 173-194, 2016.

[43] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in*Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, 1991, pp. 586-591.

[44] M. Yeasin, B. Bullot, and R. Sharma, "From facial expression to level interest: a spatio-temporal approach," in *Computer Vision and Pattern*

Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE ComputerSociety Conference on, 2004, pp. II-922-II-927 Vol. 2.

[45] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expressionrecognition from video sequences: temporal and staticmodeling," *Computer Vision and image understanding,* vol. 91, pp. 160-187, 2003.

[46] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs fisherfaces: Recognition using class specific linear projection," *PatternAnalysis and Machine Intelligence, IEEE Transactions on,* vol. 19, pp. 711-720, 1997.

[47] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques Morgan Kaufmann Publishers Inc., 2011.

[48] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neuralnetworks and learning machines* vol. 3: Pearson Education Upper SaddleRiver, 2009.