CHAPTER 1

AIM AND OBJECTIVE

Our primary aim is to analyze the image properties and determine the mood of the user and play the song as per the mood.

The key objectives of this project can be split into two parts, the recognition of the emotion of the user and music analysis. The project is centred and focused more towards different approaches to emotion recognition and the impact of each technique used. The emotion recognition stage is heavily based on image processing and machine learning. The music analysis is done by reading the MP3 metadata of a music file.

1. Emotion Recognition:

The key aim of this section is to implement and analyse various techniques to extract features and classify the emotion of a person. The image processing step requires turning the image to grayscale and resizing it. This is followed by extracting multiple features using different techniques and adapting different classifiers to determine the mood of the user. Using these different methods and techniques, an analysis is made to determine the best solution for the emotion recognition problem based on my project.

2. Music Extractions

As per the input, which is the mood of the user, application selects the song from the respective playlist (happy/sad)

CHAPTER 2

LITERATURE SURVEY

1. Fundamentals of Image Processing

The main focus is on the fundamental concepts of image processing. Space does not permit us to make more than a few introductory remarks about image analysis. Image understanding requires an approach that differs fundamentally from the theme of this book. Further, we will restrict ourselves to two–dimensional (2D) image processing although most of the concepts and techniques that are to be described can be extended easily to three or more dimensions. Readers interested in either greater detail than presented here or in other aspects of image processing. Modern digital technology has made it possible to manipulate multi-dimensional signals with systems that range from simple digital circuits to advanced parallel computers

2. Facial expression recognition: A survey

One of the non-verbal communication method by which one understands the mood/mental state of a person is the expression of face (for e.g. happy, sad). Automatic facial expression recognition (FER) has become an interesting and challenging area for the computer vision field and its application areas are not limited to mental state identification5, security, automatic counselling systems, face expression synthesis, lie detection, music for mood, automated tutoring systems, operator fatigue detection etc. This paper deals with various face detection and feature extraction techniques.

3. Rapid Object Detection using a Boosted Cascade of Simple Features

This paper brings together new algorithms and insights to construct a framework for robust and extremely rapid object detection. This framework is demonstrated on, and in part motivated by, the task of face detection. The system achieves high frame rates working only with the information present in a single grey scale image. These alternative sources of information can also be integrated with the system to achieve even higher frame rates.

4. Object detection using Haar-cascade Classifier

Object detection is commonly referred to as a method that is responsible for discovering and identifying the existence of objects of a certain class. An extension of this can be considered as a method of image processing to identify objects from digital images. This paper describes a machine learning approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates. This work is distinguished by three key contributions. The first is the introduction of a new image representation called the “Integral Image” which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features from a larger set and yields extremely efficient classifiers. The third contribution is a method for combining increasingly more complex classifiers in a “cascade”

which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions.

5. Real Time Facial Expression Recognition in Video using Support Vector Machines

This paper proposes a method for facial expression recognition. Facial feature vectors are generated from key point descriptors using Speeded-Up Robust Features. Each facial feature vector is then normalized and next the probability density function descriptor is generated. The distance between two probability density function descriptors is calculated using Kullback Leibler divergence. Mathematical equation is employed to select certain practicable probability density function descriptors for each grid, which are used as the initial classification. Subsequently, the corresponding weight of the class for each grid is determined using a weighted majority voting classifier. The class with the largest weight is output as the recognition result. The proposed method shows excellent performance when applied to the Japanese Female Facial Expression database.

6. Facial Expression Based Music Player

In this work, a novel facial feature extraction method is proposed for automatic facial expressions recognition, which detecting local texture information, global texture information and shape information of the face automatically to form the facial features. Extracting the required input from the human face can now be done directly using a camera. This input can then be used in many ways. One of the applications of this input can be for extracting the information to deduce the mood of an individual. This data can then be used to get a list of songs that comply with the mood‟ derived from the input provided earlier. This eliminates the time-consuming and tedious task of manually segregating or grouping songs into different lists and helps in generating an appropriate playlist based on an individual's emotional features. Various algorithms have been developed and proposed for automating the playlist generation process. Facial Expression Based Music Player aims at scanning and interpreting the data and accordingly creating a playlist based the parameters provided. The scanning and interpreting includes audio feature extraction and classification to get a list of songs belonging to a similar genre or to get a list of similar sounding songs.

CHAPTER 3

INTRODUCTION

The field of science is as big as the universe itself. Every passing day there are new developments; if not big or ground breaking, but constructive and leading towards a better tomorrow. Sound and Graphics are two vast fields of Science and Engineering that not only intrigue but also attract learners to study them in detail to explore into their depths. Since then many such inventions have propelled us to this time where thinking of various ideas which might not have been possible a few decades back and more over implementing them is now possible. Now in the present time, where clicking a photo and listening to music, on the go‟ is just a part of anyone's daily life, providing any improvements in the working of such technologies that in turn make the user experience better are always appreciated. With the improvements in technology the level of sophistication in software has also increased. Also with the idea of 'keeping it simple', developing sophisticated applications is a challenge. Listening to music has been found to affect the human brain activities. Emotion based music players with automated playlists can help users to maintain a selected emotional state. This research proposes emotion based music player that create playlists based on real time photos of the user. Two emotional statuses, happy and not-happy were considered in this study. User's images were captured in real-time camera. Grey scaled images were used to compress the image files.

To solve the problem of emotion recognition a lot of work has been done in the past. To extract and determine the emotion of a user, we need to extract features from an image and use them against a trained data set to classify the input and determine the emotion.

- **Feature Extractors:**

A feature extractor is an application which extracts important points in an image. Different works have been done in the field of Computer vision for feature extractors, the most prominent ones being Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF). Each of these has different impacts on classifying the emotion of the user. I initially work with a new technique called Binary Robust Independent Elementary Features (BRIEF) before moving onto techniques such as SURF and SIFT.

- **Classifiers and Prediction:**

After extracting features from an image set of training and testing data, a feature classifier is needed to sort out and classify the testing data with relevance to the training data. A Support Vector Machine (SVM) is the most predominantly used classifier to tackle the emotion recognition problem. For experimental purposes I use an SVM and a Naive Bayes Classifier.

- **Facial Emotion Recognition:**

Several approaches have been proposed to classify human affective states. The features used are typically based on displacements of specific points or spatial locations of particular points; this technique is known as Facial Action Coding System(FACS).

- **Music Analysis:**

Using the bit stream from mp3 files, we extract metadata to determine the required information for each particular song. Using the determined emotion, create a playlist of songs for the user.

3.1 IMAGE PROCESSING

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Nowadays, image processing is among rapidly growing technologies. It forms core research area within engineering and computer science disciplines too.

Image processing basically includes the following three steps:[1]

- Importing the image via image acquisition tools;
- Analysing and manipulating the image;
- Output in which result can be altered image or report that is based on image analysis.

There are two types of methods used for image processing namely, analogue and digital image processing. Analogue image processing can be used for the hard copies like printouts and photographs. Image analysts use various fundamentals of interpretation while using these visual techniques. Digital image processing techniques help in manipulation of the digital images by using computers. The three general phases that all types of data have to undergo while using digital technique are pre-processing, enhancement, and display, information extraction.

Automatic face detection is a complex problem in image processing. Many methods exist to solve this problem such as template matching, Fisher Linear Discriminant, Neural Networks, SVM, and MRC.Success has been achieved with each method to varying degrees and complexities.The assignment given to us was to develop an algorithm capable of locating each face in a color image of the class. We were given seven training images along with the corresponding ground truth data to develop and train our algorithms on.

3.2 ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI, also machine intelligence, MI) is intelligence demonstrated by machines, in contrast to the natural intelligence (NI) displayed by humans and other animals. In computer science AI research is defined as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. Colloquially, the term "artificial intelligence" is applied

when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

The scope of AI is disputed: as machines become increasingly capable, tasks considered as requiring "intelligence" are often removed from the definition, a phenomenon known as the AI effect [2], leading to the quip, "AI is whatever hasn't been done yet." For instance, optical character recognition is frequently excluded from "artificial intelligence", having become a routine technology. Capabilities generally classified as AI as of 2017 include successfully understanding human speech, competing at the highest level in strategic game systems (such as chess and Go), autonomous cars, intelligent routing in content delivery network and military simulations.

Machine learning, a fundamental concept of AI research since the field's inception, is the study of computer algorithms that improve automatically through experience.[4]

Unsupervised learning is the ability to find patterns in a stream of input. Supervised learning includes both classification and numerical regression. Classification is used to determine what category something belongs in, after seeing a number of examples of things from several categories. Regression is the attempt to produce a function that describes the relationship between inputs and outputs and predicts how the outputs should change as the inputs change. In reinforcement learning the agent is rewarded for good responses and punished for bad ones. The agent uses this sequence of rewards and punishments to form a strategy for operating in its problem space

3.2.1 AFFECTIVE COMPUTING

Affective computing, also known as AC or emotion AI [3], is an area of study within cognitive computing and artificial intelligence that is concerned with gathering data from faces, voices and body language to measure human emotion. An important business goal of AC is to build human-computer interfaces that can detect and appropriately respond to an end user's state of mind.

Affective computing has the potential to humanize digital interactions and offer benefits in an almost limitless range of applications. For example, in an e-learning situation, an AC program could detect when a student is frustrated and offer expanded explanations or additional information. In telemedicine, AC programming can help physicians quickly understand a remote patient's mood or look for signs of depression. Other business applications currently being explored include customer relationship management (CRM), human resource management (HRM), marketing and entertainment. A computing device with emotion AI programming gathers cues about a user's emotional state from a variety of sources, including facial expressions, muscle tension, posture, hand and shoulder gestures, speech patterns, heart rate, pupil dilation and body temperature. The technology that supports emotion measurement and analysis includes sensors, cameras, big data, deep learning analytics engines. An “affect” is the experience of feeling or emotion that is characteristic to humans. Affects can be recognized not only by your facial gestures but also the tone of your voice and your body language. In order for artificial intelligence to truly interact with humans, the ability to empathize needs to be mastered. The Gartner Hype

Cycle tells us that affective computing is an area that is experiencing high levels of innovation:

Fig 1: Gartner Hype Cycle

3.3 MACHINE LEARNING

Machine learning [5] is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. Machine learning involves two broad categories: Supervised and Unsupervised learning. Supervised learning is one where we provide a model with set of input and the output related to it (the training model) and then later the machine refers this training set to predict the value for input asked. On contrary, if we only provide a machine with set of inputs, and let machine figure out all the relations, features and behavior, falls under unsupervised learning.

CHAPTER 4

TECHNOLOGIES USED

4.1 PYTHON

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations.

4.2 OPEN CV

OpenCV (Open Source Computer Vision Library) is released under a BSD license and hence it's free for both academic and commercial use. It has C++, Python and Java interfaces and supports Windows, Linux, Mac OS, iOS and Android. OpenCV was designed for computational efficiency and with a strong focus on real-time applications. Written in optimized C/C++, the library can take advantage of multi-core processing. Enabled with OpenCL, it can take advantage of the hardware acceleration of the underlying heterogeneous compute platform.

Adopted all around the world, OpenCV has more than 47 thousand people of user community and estimated number of downloads exceeding 14 million. Usage ranges from interactive art, to mines inspection, stitching maps on the web or through advanced robotics.

CHAPTER 5

PROPOSED SYSTEM

The proposed algorithm in this involves an emotion music recommendation system that provides the generation of a customized playlist in accordance to the user‘s emotional state. .The system consists of 3 modules: Emotion extraction module (EEM), Audio Playlist Module, Emotion-Audio recognition module. EEM and APM are two separate modules and Emotion-Audio recognition module performs the mapping of modules by querying the audio meta-data file. The EEM and APM are combined in Emotion-Audio integration module.

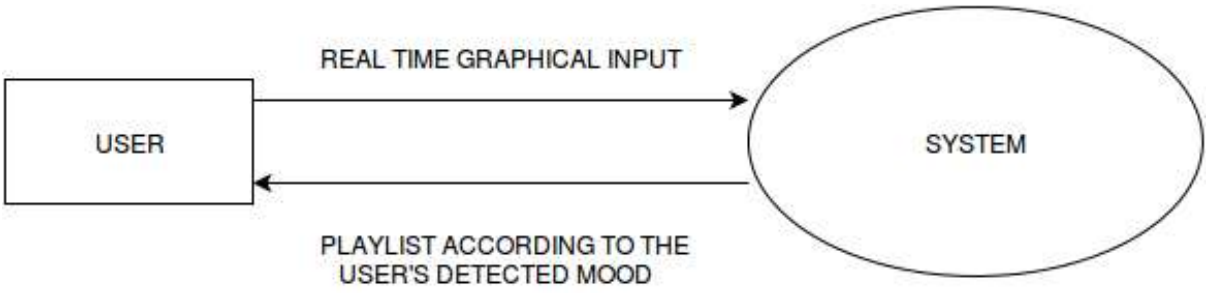


Fig 2: Outline of the model

The working can be stated as follows:

1. The user scans the memory for audio files when the application is opened.

2. After this, the songs are segregated into different playlists based on the feature extraction process.

3. The user camera is invoked with proper permissions and a real time graphical input (image) is provided to the system.

4. It then generates an output which is an emotion (mood) based on the expression extracted from the real time graphical input.

5. The emotion acts as an input and is used to select an appropriate playlist.

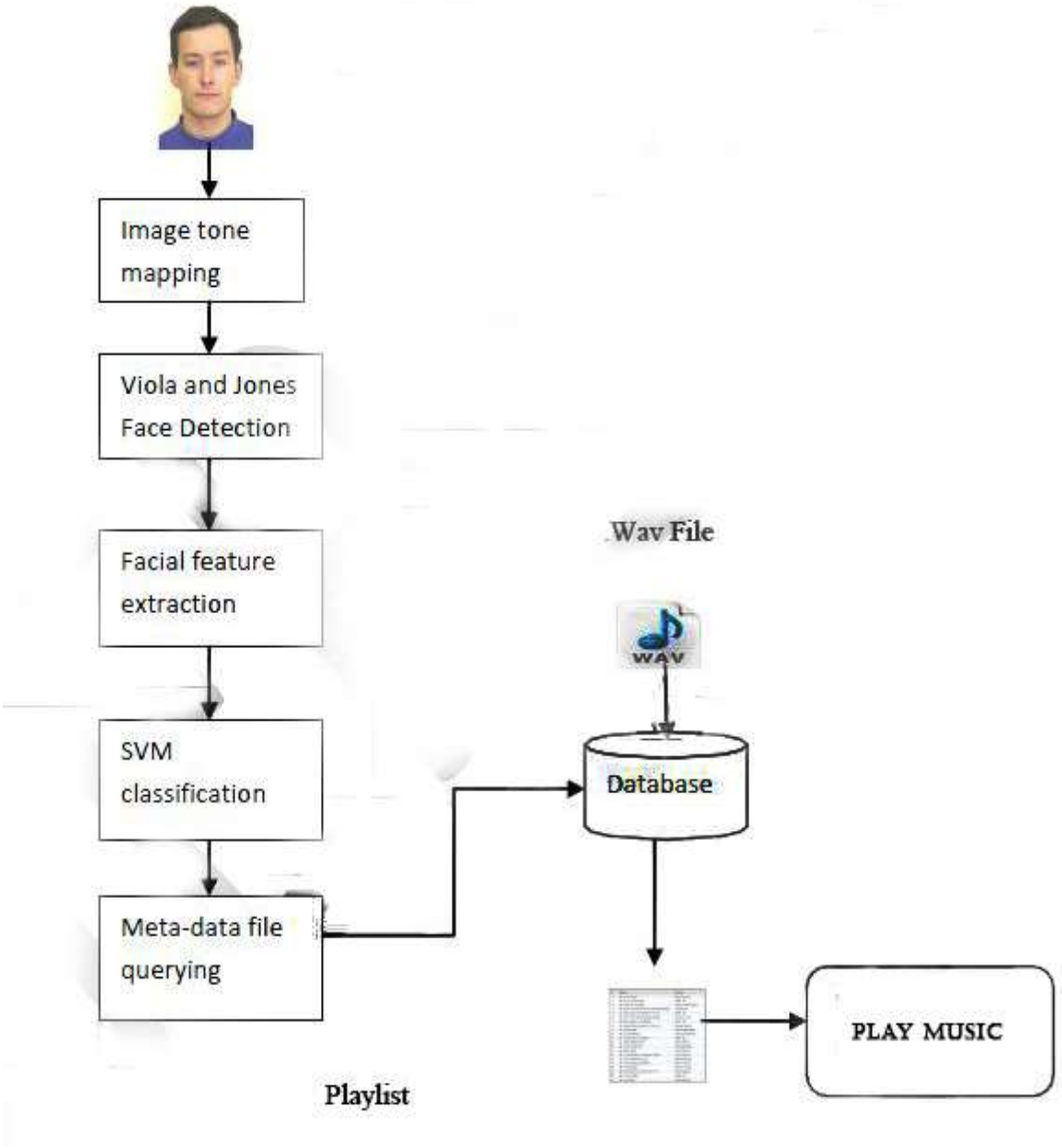


Fig 3: Block diagram of proposed system

5.1 DATA TRAINING

The training of the cascade proved to be no easy task. The first necessary bit was to gather the images, then create samples based on them and finally starting the training process. The opencv traincascade utility is an improvement over its predecessor in several aspects, one of them being that traincascade allows the training process to be multithreaded, which reduces the time it takes to finish the training of the classifier.[6]

To determine the current emotions of the user 2 things are required, an image from the user and a predetermined training dataset.

The training phase only occurs the first time in the development process and once the required data had been extracted and the classifier has been trained, the data is stored as an XML File. This file was later used to predict and classify against, rather than training the dataset on every single instance the program was run.

A dataset of 40 images per emotion are used for training purposes. The database used for training was the open source Olivetti dataset. The emotions being considered in the project are happy, sad. For each image in the dataset, the images are resized to a 260 x 360 pixels. Once the images have been resized, they are smoothed and converted to grayscale in order to reduce the "noise" and extract only important details[5].
After performing this preprocessing of images, a feature extraction algorithm is run over each of the images. Amongst the various feature extractors available, the system makes use of SIFT, SURF or BRIEF. By using these various feature extractors, different key points are located and fed to a classifier.
Several other techniques are used to improve final accuracy of the system. These are image preprocessing steps which occur before feature extraction. These techniques include only working with a few regions of interested within the face, using clustering techniques and a dimensionality reduction technique known as Principal Component Analysis (PCA).
The techniques to improve accuracy depend strongly on the choice of the feature extractor, as the results in 7 show. Clustering techniques such as the Bag of Words model, and K means clustering, proposed and used in the system do not work effectively with binary features extracted using the BRIEF feature extractor. Whereas the PCA approach has been effective with the Harris Edge Detector rather than the SURF or SIFT feature extractors. The extracted data is fed to the classifier of choice - SVM or Naive Bayes and the classifier is then trained on the data. This data is stored in an XML file to be used later.
The training can result in many types of unwanted behaviour[7]. Most common of these is either overtraining or undertraining of the classifier. An undertrained classifier will most likely output too many false positives, since the training process has not had time to properly determine which actually is positive and which is not. The opposite effect may be observed if too many stages are trained, which could mean that the classification process may determine that even the positive objects in the picture are actually negative ones, resulting in an empty result set.

5.2 LIVE LEARNING ALGORITHM

One key feature of the system is the live learning algorithm to classify music genres. After discussions with potential users, research suggested that each person associates a music genre to a different emotion based on their own preferences. The user has is to sort out which genre's he or she listens to based on the particular emotions. These preferences are stored and used later in the system.
The elementary step to emotion recognition from a live camera feed is to first accurately locate and extract the face of a person. The predominantly used technique for this task is to use a Cascade Classifier: a *cascade of boosted classifiers working with particular features*. [11]

The two Cascade classifiers the system uses are Haar Cascade Classifier and Local Binary Pattern (LBP) [7] Cascade Classifier. Both work completely differently with LBP being slightly more efficient and faster than the Haar classifier. As soon as the camera feed opens up, the classifier locates the face based on primarily difference of pixel intensities. Once the face is located, the user has to select the option to save the image once he is happy with the emotion he is portraying.
The saved image from the user goes through the exact image preprocessing steps that the training data goes through. These steps are done to maintain the consistency of the system.

Once the image has gone through these required steps, the key points and features are extracted using the chosen feature extractor. The data points about extracted as cartesian X Y coordinates and stored in a matrix. This data is fed to the classifier of choice (SVM) and it is compared against the training data. Data Classification is done and the emotion of the user is predicted.

5.3 EMOTION EXTRACTION MODULE

One of the non-verbal communication method by which one understands the mood/mental state of a person is the expression of face . Automatic facial expression recognition (FER)[6] has become an interesting and challenging area for the computer vision field. One of the non-verbal communication method by which one understands the mood/mental state of a person is the expression of face . The module can be broken down into Face detection and Feature extraction modules.

The facial component detection detects the ROI for eyes, nose, cheeks, mouth, eye brow, ear, fore-head, etc. The feature extraction step deals with the extraction of features from the ROIs. The most popular feature extraction techniques are, but not limited to, Gabor filters, Local Binary Patterns (LBP), Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), Local Gradient Code (LGC), Local Directional Pattern (LDP). The classifier step classifies the features into the respective facial expression classes based on classification methods and some of the most popular classification methods are SVM (Support Vector Machine) and NN (Nearest Neighbour). Detecting the region of interest represents an essential part of any recognition system. Ideally, this process has to be performed automatically and with a very low false positive rate. One of the most famous frameworks for object detection that is currently being
used, is called Viola-Jones.[1]

5.3.1 VIOLA-JONES ALGORITHM

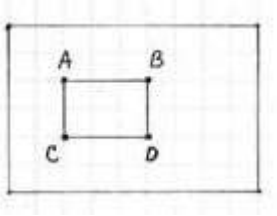
In, Viola and Jones[9] proposed a new algorithm for object detection, widely used for face detection. Their novel approach attained better results compared to previous methodologies, achieving fast detection and a low false positive rate.

- First ever real-time face detection system.
- Enable a fast and accurate detection:
- There are three ingredients
    - integral image for feature computation,
    - Adaboost for feature selection[9]
    - attention cascade for computational resource allocation
- Gives multiple detections
- Post-processing step is also proposed to reduce detection redundancy

The Viola-Jones algorithm is a widely used mechanism for object detection. The main property of this algorithm is that training is slow, but detection is fast. This algorithm uses Haar basis feature filters, so it does not use multiplications.[8]

The efficiency of the Viola-Jones algorithm can be significantly increased by first generating the integral image.

$$II(y,x) = \sum_{p=0}^{y} \square \sum_{q=0}^{x} Y(p,q)$$



The integral image allows integrals for the Haar extractors to be calculated by adding only four numbers. For example, the image integral of area ABCD is calculated as $II(y_A,x_A) - II(y_B,x_B) - II(y_C,x_C) + II(y_D,x_D)$. Detection happens inside a detection window. A minimum and maximum window size is chosen, and for each size a sliding step size is chosen. Then the detection window is moved across the image as follows:

1. Set the minimum window size, and sliding step corresponding to that size.
2. For the chosen window size, slide the window vertically and horizontally with the same step. At each step, a set of *N* face recognition filters is applied. If one filter gives a positive answer, the face is detected in the current widow.
3. If the size of the window is the maximum size stop the procedure. Otherwise increase the size of the window and corresponding sliding step to the next chosen size and go to the step 2.

Each face recognition filter (from the set of *N* filters) contains a set of cascade-connected classifiers. Each classifier looks at a rectangular subset of the detection window and determines if it looks like a face. If it does, the next classifier is applied. If all classifiers give a positive answer, the filter gives a positive answer and the face is recognized. Otherwise the next filter in the set of *N* filters is run. Each classifier is composed of Haar feature extractors (weak classifiers). Each Haar feature is the weighted sum of 2-D integrals of small rectangular areas attached to each other

The first stage of the system consists of computing and extracting the so called. Haar like features, which correspond to the rectangle patches. These templates are applied on top of a 24x24 image of a face under all scales and locations. Because computing the feature values would be an expensive operation, a new concept called 'Integral Image', was introduced, which allowed for constant time computations. This median representation enabled a fast and easy way for obtaining the feature values. However, deriving all the possible features would be very expensive. Therefore, a feature selection process was proposed, which applies a modified version of the AdaBoost [9] technique. This machine learning boosting algorithm was used to create a strong classifier, out of a series of weak classifier and a scheme of associated weights. Lastly, a cascade of classifier was used, in which the first classifier are simple and used to discard non-faces, and the stronger classifier were used for sub-windows that might be faces.

5.3.2 HAAR CASCADE

Object Detection using Haar feature-based cascade[10] classifiers is an effective object detection method proposed by Paul Viola and Michael Jones in their paper, "Rapid Object Detection using a Boosted Cascade of Simple Features" in 2001. It is a machine learning based approach where a cascade function is trained from a lot of positive and negative images. It is then used to detect objects in other images. Each feature is a single value obtained by subtracting sum of pixels under the white rectangle from sum of pixels under the black rectangle.

A Haar-like feature considers neighbouring rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. An example of this would be the detection of human faces. Commonly, the areas around the eyes are darker than the areas on the cheeks. One example of a Haar-like feature for face detection is therefore a set of two neighbouring rectangular areas above the eye and cheek regions .[4]

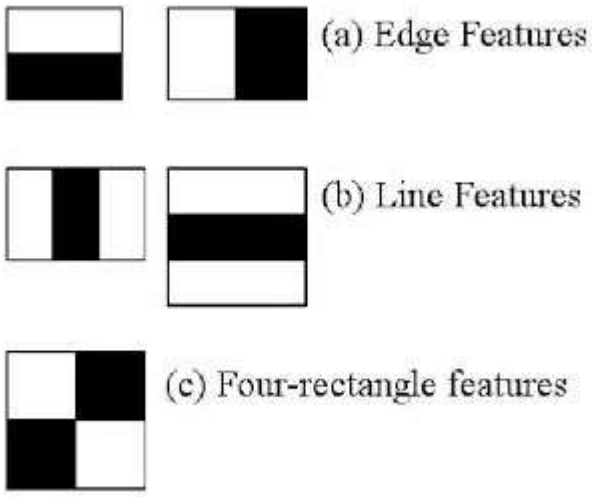


Fig 4: Haar like features

Now, all possible sizes and locations of each kernel are used to calculate lots of features. For each feature calculation, we need to find the sum of the pixels under white and black rectangles.

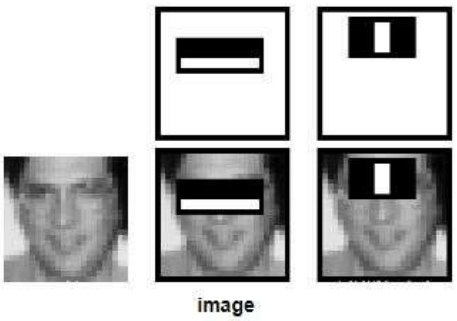


Fig 5:Implementation of Haar like features

The first feature selected seems to focus on the property that the region of the eyes is often darker than the region of the nose and cheeks. The second feature selected relies on the property that the eyes are darker than the bridge of the nose. But the same windows applied to

cheeks or any other place by **Adaboost**. For this, we apply each and every feature on all the training images. For each feature, it finds the best threshold which will classify the faces to positive and negative. We select the features with minimum error rate, which means they are the features that most accurately classify the face and non-face images.

5.3.3 CASCADE CLASSIFIER AND ADABOOST

The cascade classifier consists of a list of stages, where each stage consists of a list of weak learners[10]. The system detects objects in question by moving a window over the image. Each stage of the classifier labels the specific region defined by the current location of the window as either positive or negative – positive meaning that an object was found or negative means that the specified object was not found in the image. If the labelling yields a negative result, then the classification of this specific region is hereby complete and the location of the window is moved to the next location. If the labelling gives a positive result, then the region moves of to the next stage of classification. The classifier yields a final verdict of positive, when all the stages, including the last one, yield a result, saying that the object is found in the image. A true positive means that the object in question is indeed in the image and the classifier labels it as such – a positive result. A false positive means that the labelling process falsely determines, that the object is located in the image, although it is not. A false negative occurs when the classifier is unable to detect the actual object from the image and a true negative means that a nonobject was correctly classifier as not being the object in question. In order to work well, each stage of the cascade must have a low false negative rate, because if the actual object is classified as a non-object, then the classification of that branch stops, with no way to correct the mistake made. However, each stage can have a relatively high false positive rate, because even if the n-th stage classifies the non-object as actually being the object, then this mistake can be fixed in n+1-th and subsequent stages of the classifier

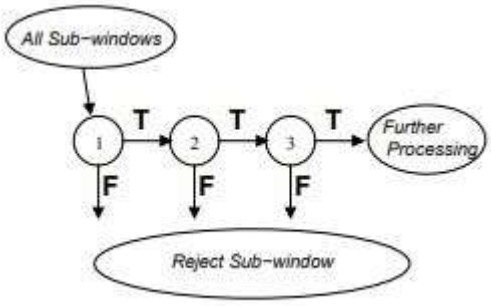


Fig 6 :Model of cascade classifier

5.3.4 SUPPORT VECTOR MACHINES

The concept of Support Vector Machines[4] was firstly introduced by Vapnik et al, and presently, they are one of the most widely used methods for pattern classi_cation. An SVM is a supervised learning model, because it uses labelled examples in its training process, examples which correspond to only two categories. This property makes the algorithm able to only tackle binary classification tasks. The model analyses the training examples and tries to derive a boundary that will linearly separate the data points into their corresponding classes. One of the most important feature of this method, is that it does not only look for a separation

boundary, but for the 'best boundary'. This is done by maximizing the margin, which is the width by which the separation boundary can be increased until it hits a data point. The difference between separating data with any border and separating it with SVM's optimal boundary is shown in Figure

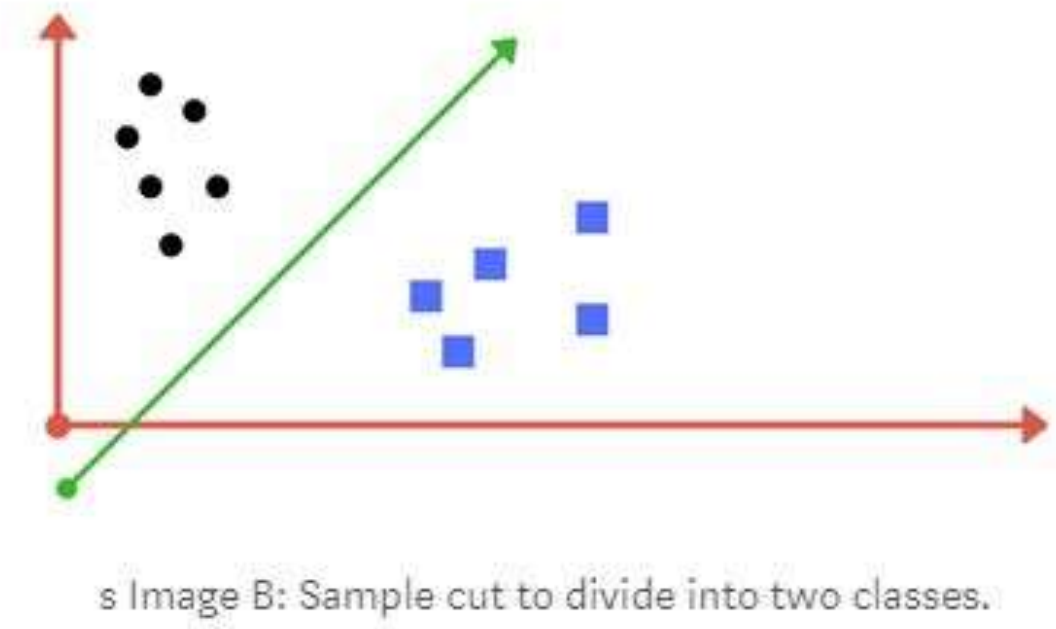
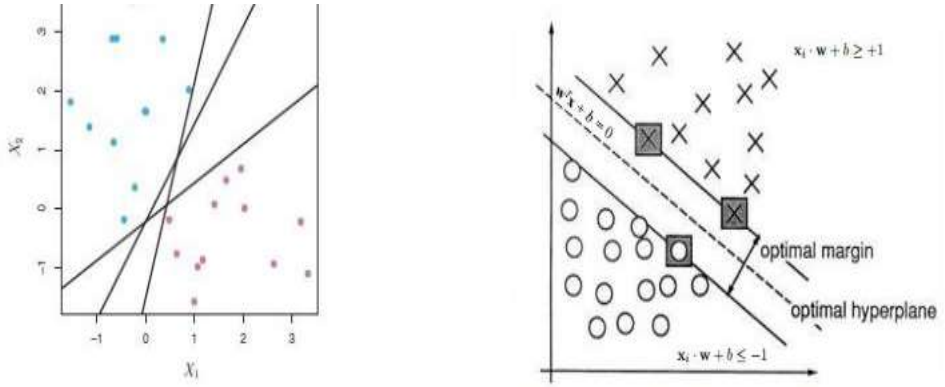


Fig 7:Hyperplane of SVM

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra, which is out of the scope of this introduction to SVM.A powerful insight is that the linear SVM can be rephrased using the inner product of any two given observations, rather than the observations themselves. The inner product between two vectors is the sum of the multiplication of each pair of input values.
For example, the inner product of the vectors [2, 3] and [5, 6] is 2*5 + 3*6 or 28.The equation for making a prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

$$f(x) = B0 + sum(ai * (x,xi))$$

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm



(A) Example of possible boundaries to separate 2D data into two classes.

(B) Separating 2D data with the maximum-margin boundary.

Fig 8 :Seperation boundaries for 2D data

As depicted in Figure, the linear classifier that separates the data has the following mathematical form:

$f(x) = w^T x + b$

where w is the normal to the separation hyper plane, known as the weigh vector, b is the bias and x is the vector of training examples, corresponding to the classes y={1, -1}.

The goal is finding the best values for w and b, that correspond to the maximum-margin boundary, such that each training example xi can be described as:

$xi \cdot w + b >= +1$ if $y_i = +1$

$xi \cdot w + b <= -1$ if $y_i = -1$

Finding such function is not trivial because most of the time, the data is simply not linearly separable. For this, SVM uses something called 'kernels', which can be regarded as complicated functions, which map the data points into a higher dimensional space, where eventually, a hyper plane would be able to separate the examples. The chosen kernel can be: a polynomial kernel or a radial basis function.

SVM algorithms separate the training data in feature space by a hyperplane defined by the type of kernel function used. They find the hyperplane of maximal margin, defined as the sum of the distances of the hyperplane from the nearest data point of each of the two classes. The size of the margin bounds the complexity of the hyperplane function and hence determines its generalization performance on unseen data. Here , the hyperplane divides the region into two and classifies the emotion into two categories as happy and sad emotions. The labels of the classifier are set as smile and nonsmile. The classification takes place by checking on the label that has a true value ,that is, if label smile is true then the emotion is recognized as happy else sad.

5.3.5 FEATURE EXTRACTION METHODS

A. Local binary pattern

Local Binary Pattern (LBP) [9] was proposed for texture analysis and later extended and applied to other applications. The LBP assigns a label to each pixel in the P-neighborhood (P equally spaced pixel value within a radius (R), denoted by gp) by thresholding its values with the central value (gc) and converting these thresholded
values into decimal number given by eqn. (1).

$$LBP_{p.R}(X_C, Y_C) = \sum_{p=0}^{p-1} s(g_p - g_c) \quad where, \; s(x) = \begin{cases} 1, & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

This algorithm has its roots in the 2D texture analysis of images. The basic idea is to summarize the local structure of an image by comparing each pixel with its neighbor. The proposed idea is to take a pixel in the center and threshold it against its neighbors. If the value is less than its neighbor, denote it with a 0 or else denote it with a 1. Hence you end with a binary value for each pixel, known as the Local Binary Patterns (LBP).[9]
A useful extension to the original operator is the so-called uniform patterns[8]. These patterns improved the efficiency of the system by reducing the length of the feature vector and hence implementing a simple rotation invariant descriptor. The system makes use of these uniform patterns to determine face locations from a camera feed.

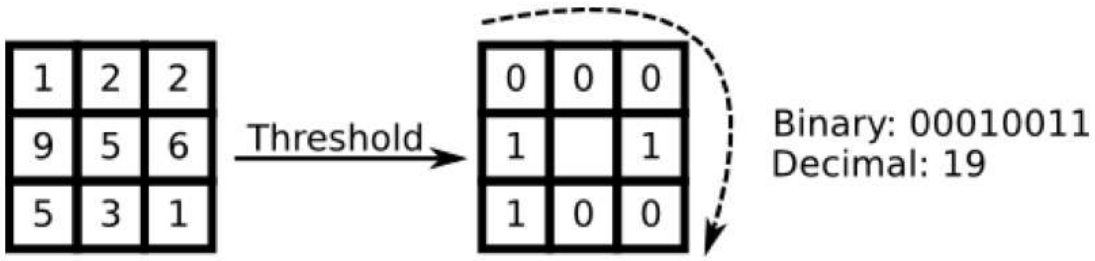


Figure 9- A typical example of a local binary pattern

Other related approaches are (i) Local Binary Pattern from Three Orthogonal Planes (LBP-TOP)22 by Yanjun Guo et al. to detect the micro-expressions from sequential images, and (ii) combination of dynamic edge and texture information named as LOE-LBP-TOP (Local Oriented Edges) by Gizatdinova et al.[28]

B.Local gradient code

LGC[9] is based on the relationship of neighbouring pixels (while LBP only compares the central pixel value with the neighbouring pixel value) . The optimized LGC-HD (based on principle of horizontal diagonal) and LGC-VD (based on the principle of vertical diagonal) .

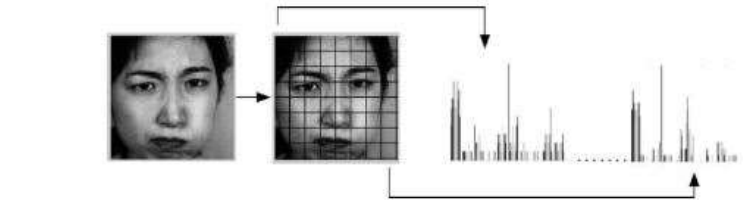
$$LGC = s(g_1 - g_3)2^7 + s(g_4 - g_5)2^6 + s(g_6 - g_8)2^5 + s(g_1 - g_6)2^4 + s(g_2 - g_7)2^3 + s(g_3 - g_8)2^2 + s(g_1 - g_8)2^1 + s(g_3 - g_6)2^0$$

$$LGC - HD_p^R = s(g_1 - g_3)2^4 + s(g_4 - g_5)2^3 + s(g_6 - g_8)2^2 + s(g_1 - g_8)2^1 + s(g_3 - g_6)2^0$$

$$LGC - VD_p^R = s(g_1 - g_6)2^4 + s(g_2 - g_7)2^3 + s(g_3 - g_8)2^2 + s(g_1 - g_8)2^1 + s(g_3 - g_6)2^0$$

C. Local directional pattern

In order to get better performance in the presence of variation in illumination and noise, Local Directional Pattern has been developed by Jabid et.al in 2010. In this method, eight Kirsch masks of size 3x3 are convolved with image regions of size 3x3 to get a set of 8 mask values. These mask values are then ranked and the top three
will be assigned with one in the 8 bit binary code and the others with zero. The decimal value corresponding to this binary code will be the LDP value for the centre pixel of the selected 3x3 image region. This LDP generated image is divided into blocks and histogram of blocks is concatenated to avail the LDP feature for the image[6]
Other variant of LDP are, (i) Local Directional Pattern Variance (LDPv)24 which combines the texture and contrast, and (ii) LDN25 which encodes the directional information along with sign.



Facial image representation using spatially combined LDP

(a)Original Image; (b)LBP Image; (c)LDP Image; (d)LGC Image

146	148	126
145	138	137
145	130	129

1	1	0
1	195	0
1	0	0

0	0	0
0	28	1
0	1	1

1	1	1
0	237	1
1	0	1

Central pixel values corresponding to original, LBP, LDP, LGC image

Fig 10. Illustration of various methods

D. SIFT

SIFT is amongst the founding feature extraction algorithms. It is highly efficient being invariant to image rotation, scaling and partially invariant to camera illumination. SIFT has its roots in three elements, the SIFT descriptor, the key point location, and the orientation. [18] The match for each key point is found by identifying its nearest neighbour, which is defined as the key point within a minimum distance threshold. To ensure a correct match Lowe suggests rejecting all matches in which the distance ratio is greater than a threshold.

Automatically locating landmark points from an arbitrary view facial image is very challenging .To overcome this problem, the system makes use of a dense SIFT feature description to describe facial images. To be precise, the system divides the entire face into patches and extracts a 128 bit SIFT feature vector from each region . These features are then used in the training and testing of images.

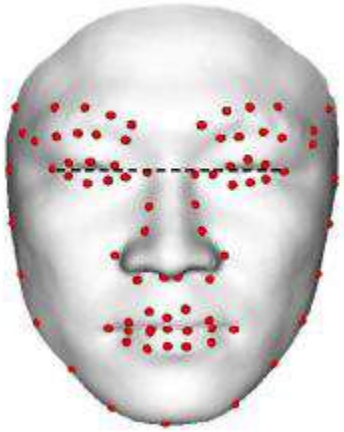


Fig 11: *Facial landmarks extracted from SIFT*

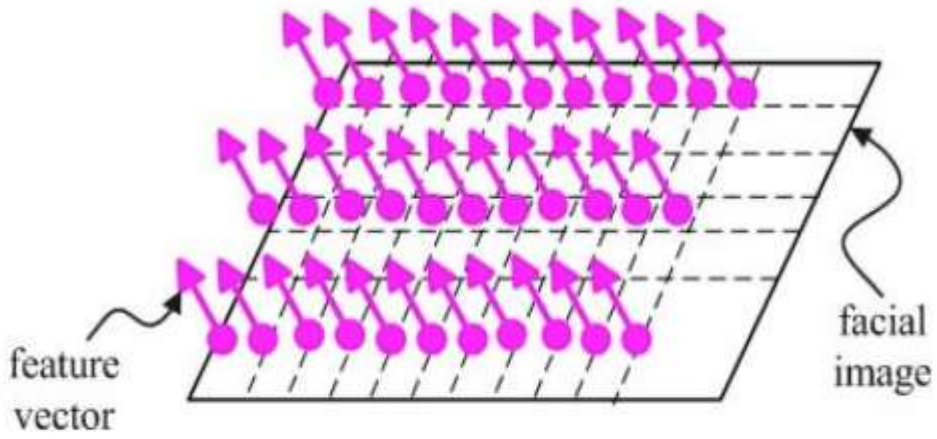


Fig 12: Facial region divided into patches and each patch produces a SIFT vector.

E. SURF

The properties of SURF lie within scale invariance, strong robustness and a strong distinction between feature points. In comparison to SIFT the SURF operator has greatly improved computer speeds[12]. The algorithm consists of 4 main parts:

1. Generating an Integral Image
2. Fast Hessian detector (interest point detection)
3. Descriptor orientation assignment (optional),
4. Descriptor generation.

The key aspect of SURF algorithm is the use of an intermediate image, the Integral image. This integral image is subsequently used by all parts of the algorithm later on.
The image is convolved with the squares, rather than the Gaussian average, as convolving of an image with the square is much faster if an integral image is used[12].
The integral image is defined as:

$$S(x,y) = \sum_{i=0}^{i<=x} \sum_{j=0}^{j<=y} I(i,j)$$

Due to its high accuracy rates, the SURF algorithm has its roots based in the determinant of the Hessian Matrix. Using the Integral Image, we calculate the Hessian matrix, as a function of both space x = (x; y) and scale σ. Given a point x = (x, y) in an image I, the Hessian matrix H (x, σ) in x at scale σ, is defined as follows [18]:

$$\mathrm{H}(\mathrm{x},\sigma) = \begin{bmatrix} L_{xx}(\mathrm{x},\sigma) & \cdots & L_{xy}(\mathrm{x},\sigma) \\ \vdots & \ddots & \vdots \\ L_{xy}(\mathrm{x},\sigma) & \cdots & L_{yy}(\mathrm{x},\sigma) \end{bmatrix}$$

F.Principal Component Analysis

PCA is a renowned technique used in pattern recognition for dimensionality reduction[15]. As these patterns contain redundant information, mapping them into feature vectors can reduce or even completely get rid of the redundancy all the while preserving most of the intrinsic information content of the pattern itself. A face image, with size N × N in 2-dimension can also be considered as one dimensional vector of dimension N2[16]. Given facial images share the same prominent characteristics, and being similar in overall configuration, they will not be distributed randomly in our image space. Therefor they can be represented in a relatively low dimensional image sub space.
PCA has its roots in identifying underlying trends in a data set and therefor the main idea is to find vectors that best represent the distribution of faces within our image space. These vectors define the subspace of face images, which we call “face space” [16]. These vectors are known

as Eigen faces because they are the eigenvectors of the covariance matrix of the original face image. The number of PCA components selected for the system was 90, as it gave a reasonably high accuracy rate without compromising the speed of the system.

5.3.6 REGION OF INTEREST

- Detection of eye and eye-brow region

According to the measured anthropometric landmark distances, the upper part of the face contains only eyebrows and the middle part contains eyes and nose. The following steps have been performed in the process of eye and eye-brow region detection[16].

After performing the morphological operations and binarization to the middle part of the face image, sequential search has been performed to find the eye corners. The Fig. 4 shows the eye candidates after performing morphological operations and binarization to the middle part of the face image. As we are only concentrating on eye region, we have discarded the nose area. After finding the eye corners and eye-centers, it is essential to find the distance between eye and eye-brow. According to the anthropometric measurement, the upper facial part contains eye-brow. Fig. 5 represents the eyebrow after binarization. Using sequential search from the lower part of the Fig. 5, we have measured the center of the eye-brow, so that the distance can be measured from the eye-center and eye-brow center. Based on the eye-corners and distance between eye and eye-brow centers, one rectangle has been plotted on the whole eye and eye-brow region

. • Detection of Mouth region The following steps have been taken to detect the mouth region. The morphological and binarization operation on the lower facial part of the face image produce the images like shown below in Fig. 6.

5.3.7 OLIVETTI DATASET

This dataset contains a set of face images taken between April 1992 and April 1994 at AT&T Laboratories Cambridge.

There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

The image is quantized to 256 grey levels and stored as unsigned 8-bit integers; the loader will convert these to floating point values on the interval [0, 1], which are easier to work with for many algorithms.

The “target” for this database is an integer from 0 to 39 indicating the identity of the person pictured; however, with only 10 examples per class, this relatively small dataset is more interesting from an unsupervised or semi-supervised perspective.

Fig 13: Sample pictures of Olivetti dataset

5.3.8 k-FOLD CROSS-VALIDATION

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

STEPS

- Split the dataset into K **equal** partitions (or "folds")
- Use fold 1 as the **testing set** and the union of the other folds as the **training set**
- Calculate **testing accuracy**
- Repeat steps 2 and 3 K times, using a **different fold** as the testing set each time
- Use the **average testing accuracy** as the estimate of out-of-sample accuracy

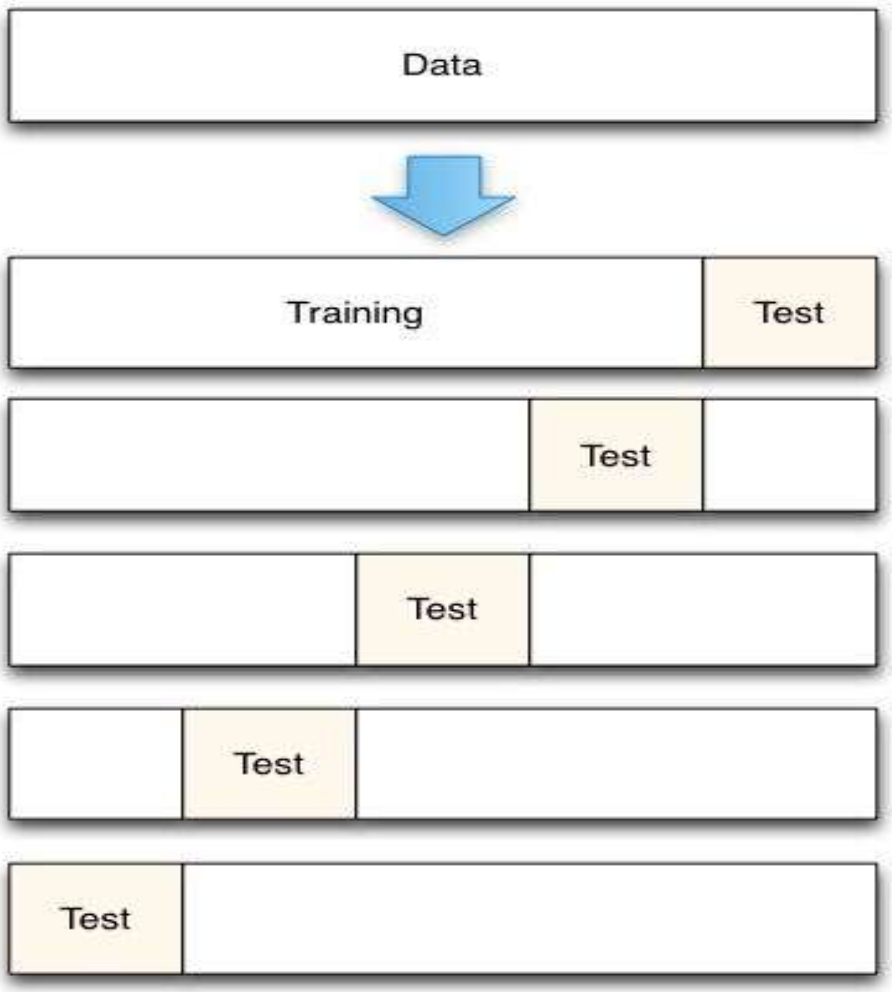


Fig:14: 5 -fold cross-validation

5.4. AUDIO PLAYLIST MODULE

The songs are classified into different emotions and stored in appropriate files under the significant names. An excel file is then defined and the folders are entered into it. The program has access to this excel file which further reads the folder path according the emotion retrieved from the emotion extraction block. Once the folder is accessed the program starts to play the songs from corresponding emotions automatically.

The algorithm first checks for the windows platform and open the file that is passed as filename

```
df = pandas.read_excel("EmotionLinks.xlsx")
```

It first reads an Excel file in which corresponding musical folders are listed and the appropriate filename is chosen.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The Emotion Music System will be of great advantage to users looking for music based on their mood (happy or sad). It will help reduce the searching time for music thereby reducing the unnecessary computational time and thereby increasing the overall accuracy and efficiency of the system. Also with its additional features mentioned above, it will be a complete system for music lovers and listeners.

A wide variety of image processing techniques was developed to meet the facial expression recognition system requirements. Proposed system will be able to process the video of facial behaviour, recognize displayed actions in terms of basic emotions and then play music based on these emotions. Major strengths of the system are full automation as well as user and environment independence. Even though the system cannot handle occlusions and significant head rotations, the head shifts are allowed. In the future work, we would like to focus on improving the recognition rate of our system. The future scope in the system would to design a mechanism that would be helpful in music therapy treatment and provide the music therapist the help needed to treat the patients suffering from disorders like mental stress, anxiety, acute depression and trauma. The proposed system also tends to avoid in future the unpredictable results produced in extreme bad light conditions and very poor camera resolution. The proposed algorithm was successful in crafting a mechanism that can find its application in music therapy systems and can help a music therapist to therapize a patient, suffering from disorders like acute depression, stress or aggression. The system is prone to give unpredictable results in difficult light conditions, hence as part of the future work, removing such a drawback from the system is intuited.

The Emotion Based Music System will be of great advantage to users looking for music based on their mood and emotional behavior. It will help reduce the searching time for music thereby reducing the unnecessary computational time and thereby increasing the overall accuracy and efficiency of the system. The system will not only reduce physical stress but will also act as a boon for the music therapy systems and may also assist the music therapist to therapize a patient. The system is highly efficient and easy to use and has an extremely fast computing methods. Manual face analysis used by psychologists was quickly replaced by suitable computer software. Proposed system will be able to process the image of facial behavior, recognize displayed actions in terms of basic emotions and then play music based on these emotions. Major strengths of the system are full automation as well as user and environment independence. Even though the system cannot handle occlusions and significant head rotations, the head shifts are allowed.

The biggest limitation of the project is in the general approach taken for emotion recognition using facial cues. The camera must take a full frontal image of the user to determine the emotion of the user accurately. Even though rotation invariance is taken into account for two of the feature extractors, accuracy rates of subjects facing the camera sideways was extremely low in comparison to frontal images. Indeed SIFT depicted an average accuracy of 63% whilst SURF fared slightly better with an average accuracy rate of 68 %.

Another massive limitation faced during the testing and development of the project was a relatively small sized training dataset. For ethical reasons express permission is required to use a database of faces depicting various emotions. Hence forth a small training data set was used which resulted in a relatively low accuracy rate for the system.

To develop a music information retrieval system, initially the library Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) was determined to be appropriate. The library turned out to be unstable and within a few weeks of development, it was unfeasible to use this with windows and OpenCV. Hence the project shifted from a machine learning approach to an ordinary data extraction for music classification. This was another hindrance and limitation faced during the project development

RESULT

Following Results have been obtained

a) When the user is not smiling, the following is obtained:

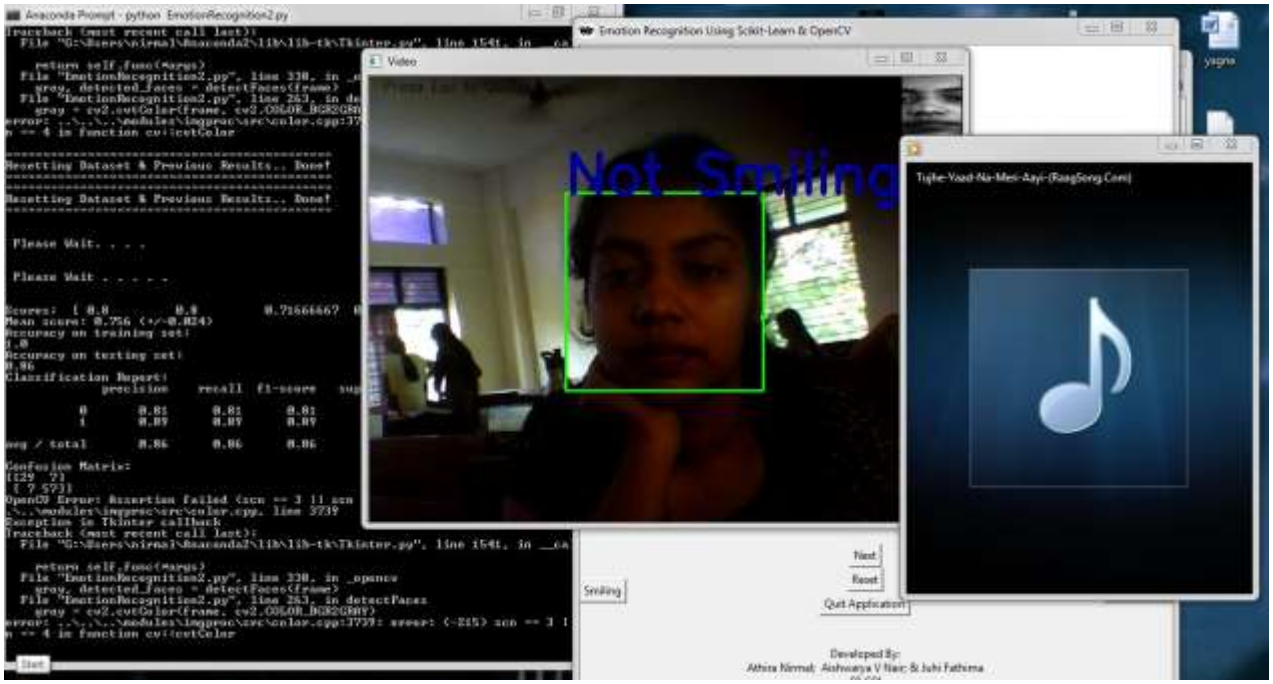
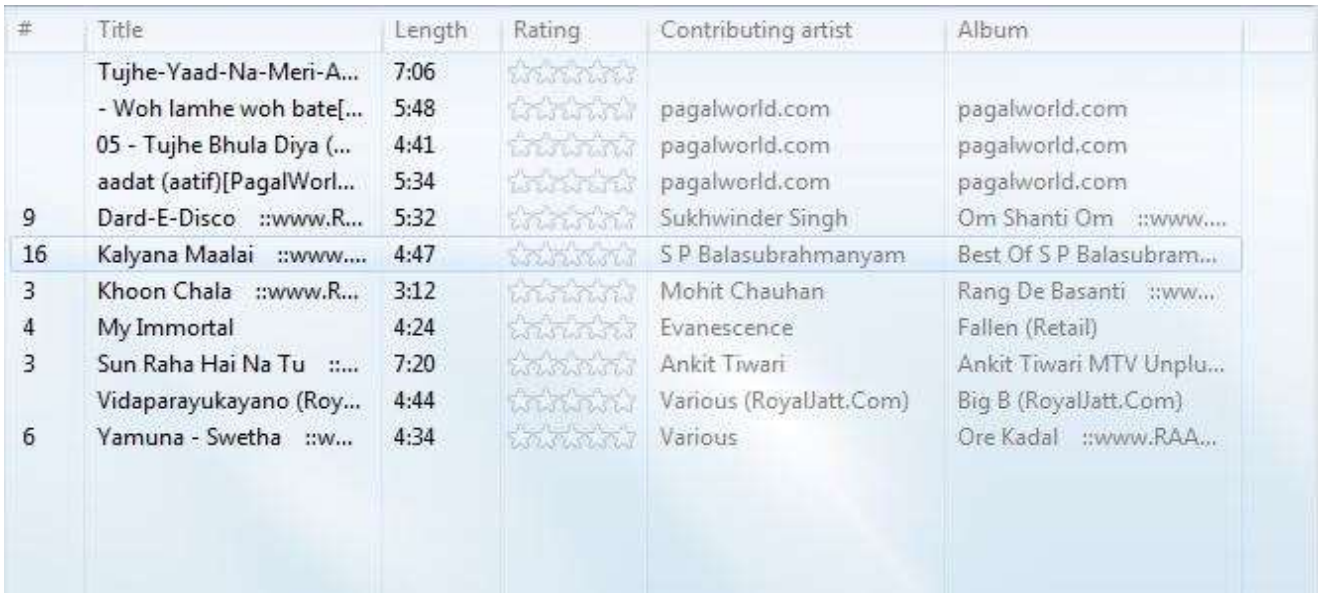


Fig:Emotion recognized as “Sad”

And instantly a media player with following songs is opened.



#	Title	Length	Rating	Contributing artist	Album
	Tujhe-Yaad-Na-Meri-A...	7:06			
	- Woh lamhe woh bate[...	5:48		pagalworld.com	pagalworld.com
	05 - Tujhe Bhula Diya (...	4:41		pagalworld.com	pagalworld.com
	aadat (aatif)[PagalWorl...	5:34		pagalworld.com	pagalworld.com
9	Dard-E-Disco ::www.R...	5:32		Sukhwinder Singh	Om Shanti Om ::www....
16	Kalyana Maalai ::www....	4:47		S P Balasubrahmanyam	Best Of S P Balasubram...
3	Khoon Chala ::www.R...	3:12		Mohit Chauhan	Rang De Basanti ::ww...
4	My Immortal	4:24		Evanescence	Fallen (Retail)
3	Sun Raha Hai Na Tu ::...	7:20		Ankit Tiwari	Ankit Tiwari MTV Unplu...
	Vidaparayukayano (Roy...	4:44		Various (RoyalJatt.Com)	Big B (RoyalJatt.Com)
6	Yamuna - Swetha ::w...	4:34		Various	Ore Kadal ::www.RAA...

Fig: The song displayed first on the list is played automatically, and the user can shuffle and change among the following category song if she wishes.

b) When the user is smiling, the following is obtained:

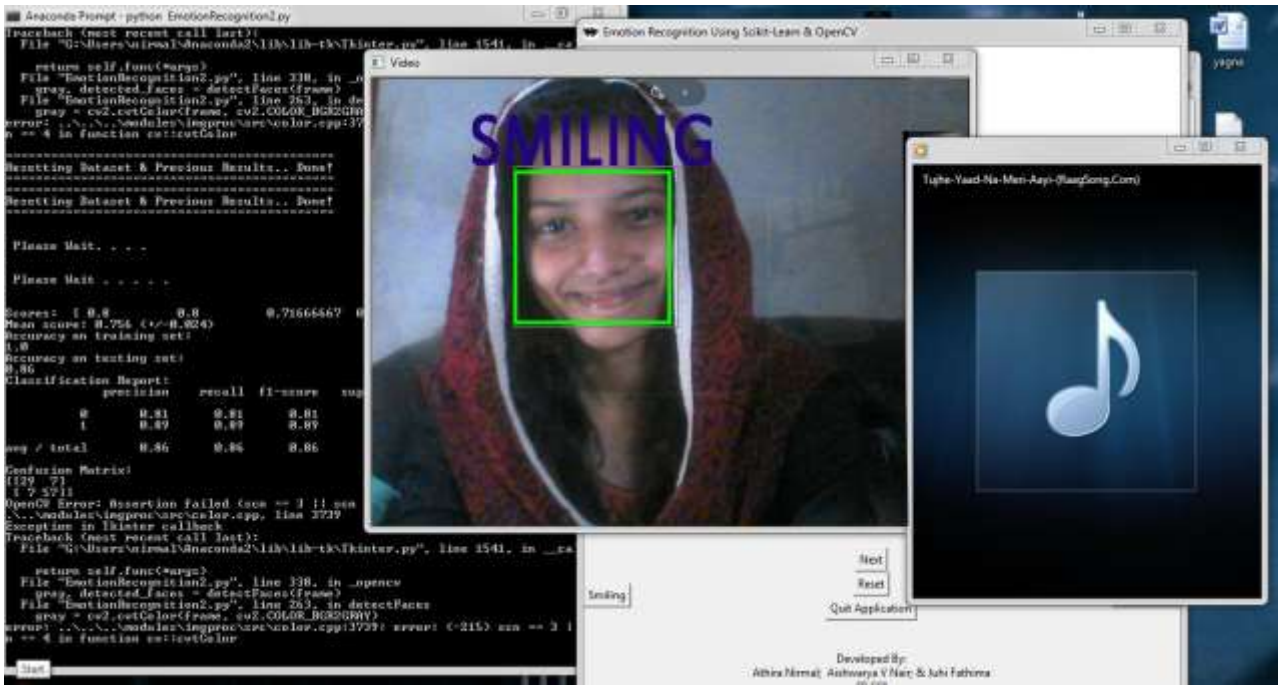


Fig:Emotion recognized as “Happy”

And instantly a media player with following songs is opened.

#	Title	Length	Rating	Contributing artist	Album
	The_Humma_Song_(Fro...	3:00			
	Love Dose - PagalWorld...	3:48		Yo Yo Honey Singh	Desi Kalakaar (2014) - P...
1	Adada_Mazhaida_NewT...	4:32		Rahul Nambiar & Saind...	Paiyaa
	BaBy - Justin BieBer	3:35			
	Befikra				
	Calvin_Harris_-_This_Is_...	3:59			
	Chundari Penne	4:00		Dulquer salman	Charlie
	Dheere_Dheere	3:32			
	Ishq_Wala_Love	4:20			
	Kar_Gayi_Chull	3:08			
	Main_Rahoon_Ya_Na_R...	5:10			
	Main_Rang_Sharbaton_...	4:23			
	Nashe_Si_Chadh_Gayi	3:58			
	Pistah_the_run_anthem...	2:30		www.KuttyWap.com	Neram
	The_Breakup_Song	4:13			
	The_Disco_Song	5:42			

Fig: The song displayed first on the list is played automatically, and the user can shuffle and change among the following category song if she wishes.

REFERENCES

[1] Fundamentals of Image Processing Ian T. Young ,J. Gerbrands, Lucas J. van Vliet ;Delft University of Technology
[2] Artificial Intelligence and Human Thinking , Robert Kowalski; Imperial College London
[3]Online : https://www.nanalyze.com/2016/04/affective-computing-and-ai-emotion-recognition/
[4]https://machinelearningmastery.com/support-vector-machines-for-machine-learning/
[5] Zhang Z, Zhang J. A new real-time eye tracking for driver fatigue detection. In: *ITS Telecommunications Proceedings, 2006 6th International Conference on*. IEEE; 2006, p. 8□11.
[6] Facial Expression Recognition System
[7] Ojala T, Pietik¨ainen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition* 1996;29:51□9.
[8] Rapid Object Detection using a Boosted Cascade of Simple Features Paul Viola, Michael Jones In: Mitsubishi Electric Research Labs 201 Broadway, 8th FL One Cambridge Center, Cambridge
[9] https://www.vocal.com/video/adaboost-training-algorithm-for-viola-jones-object-detection/
[10] Object detection using Haar-cascade Classifier Sander Soo In:Institute of Computer Science, University of Tartu
[11] Emotion Based Music Player Hafeez Kabani, Sharik Khan, Omar Khan, Shabana Tadvi
[12] Facial Expression Based Music Player
Prof. Jayshree Jha1, Akshay Mangaonkar, Deep Mistry, Nipun Jambaulikar, Prathamesh Kolhatkar
[13] Mood based Music Player Karan Mistry, Prince Pathak , Prof. Suvarna Aranjo
[14] A Novel Feature Extraction Method for Facial Expression Recognition ,Xiaoyi Feng, Baohua Lv, Zhen Li, Jiling Zhang :School of Electronic and Information, Northwestern Polytechnic University, Xi'an, China
[15] Real Time Facial Expression Recognition in Video using Support Vector Machines Philipp Michel, Rana El Kaliouby In: Computer Laboratory University of Cambridge ,Cambridge CB3 0FD, United Kingdom
[16] An Approach to Detect the Region of Interest of Expressive Face
Images Priya Sahaa,*, Debotosh Bhattacharjeeb, Barin Kumar Dec, Mita Nasipuri
[17] Facial Expression Recognition Using New Feature Extraction
Algorithm Hung-Fu Huang* and Shen-Chuan Tai :Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
[18]Emotion based musical player