**Face emotion based on music using deep learning**

**OBJECTIVE:**

The objective of this project is to develop Automatic Facial Expression Recognition System which can take human facial images containing some expression as input and recognize and classify it into seven different expression classes such as happy, sad, anger, disgust. based on emotion it will play music.

**ABSTRACT:**

The project presents speech emotion recognition from speech signal based on features analysis and NN-classifier. Automatic Face emotion recognition (SER) plays an important role in HCI systems for measuring people’s emotions and has dominated psychology by linking expressions to group of basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise). The recognition system involves Face emotion detection, features extraction and selection and finally classification. These features are useful to distinguish the maximum number of samples accurately and the NN classifier based on discriminate analysis is used to classify the six different expressions. The simulated results will be shown that the filter based feature extraction with used classifier gives much better accuracy with lesser algorithmic complexity than other Face emotion expression recognition approaches. Recognition is based on the stored image data of the different group of persons. Input images are of any type can be used for recognition,

1. Still images.

2. Video frames or video stills.

3. Video

**CHAPTER 1:**

**INTRODUCTION:**

With the advent of modern technology our desires went high and it binds no bounds. In the present era a huge research work is going on in the field of digital image and image processing. The way of progression has been exponential and it is ever increasing. Image Processing is a vast area of research in the present day world and its applications are very widespread. Image processing is the field of signal processing where both the input and output signals are images. One of the most important applications of Image processing is Facial expression recognition. Our emotion is revealed by the expressions in our face. Facial Expressions plays an important role in interpersonal communication. Facial expression is a non verbal scientific gesture which gets expressed in our face as per our emotions. Automatic recognition of facial expression plays an important role in artificial intelligence and robotics and thus it is a need of the generation. Some application related to this includes Personal identification and Access control, Videophone and Teleconferencing, Forensic application, Human-Computer Interaction, Automated Surveillance, Cosmetology and so on.

**CHAPTER 2:**

**LITERATURE SURVEY**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNO** | **TITLE** | **AUTHOR** | **YEAR** |
| **1** | A review of facial  Expression using  Artificial intelligence. | mohammed  mansoor | 2013 |
| **2** | Face recognition  Techniques for  Forsenic identify | Jinhua zeng | **2017** |
| **3** | Comparision of  Four subjective methods for image quality | R.mantiruk | **2012** |
| **4** | Objective  Assesment of  Image quality | r.f.wanger | 2020 |

**Chapter 3:**

**SYSTEM ANALYSIS**

**Existing system**

* Principal Component Analysis
* Geometric methods
* Support vector machine.

**Drawbacks**

* Low discriminatory power and high computational load
* In geometric based methods, the geometric features like distance between speech signals.

**Proposed method**

* Face Emotion recognition for transform features system through textural analysis and NN classifier. The system involves like anger, happy, sad disgust and anger.

**ADVANTAGE:**

* It will easily detect the face and recognize the human expression.
* In this image quality is essential for recognize the face based on facial expression automatically it will play music .

**CHAPTER 4:**

**BLOCK DIAGRAM**

Haar feature extraction

Input video

Haar cascade

Pre-processing

Database

CNN classifier

training

Play music

Happy,sad,disgust ,neutral,stress,angry

**Chapter 5**

**Modules description**

* Input video
* Pre-processing
* Haar cascade
* Haar feature extraction
* CNN CLASSIFIER

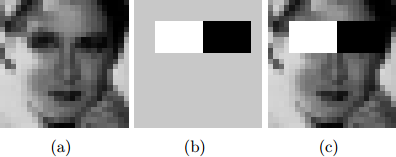
**PREPROCESSING**

* Image Pre-processing is a common name for operations with images at the lowest level of abstraction. Its input and output are intensity images. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing.
* Image restoration is the operation of taking a corrupted/noisy image and estimating the clean original image. Corruption may come in many forms such as motion blur, noise, and camera misfocus. Image restoration is different from image enhancement in that the latter is designed to emphasize features of the image that make the image more pleasing to the observer, but not necessarily to produce realistic data from a scientific point of view. Image enhancement techniques (like contrast stretching or de-blurring by a nearest neighbor procedure) provided by "Imaging packages" use no a priori model of the process that created the image. With image enhancement noise can be effectively be removed by sacrificing some resolution, but this is not acceptable in many applications. In a Fluorescence Microscope resolution in the z-direction is bad as it is. More advanced image processing techniques must be applied to recover the object. De-Convolution is an example of image restoration method. It is capable of: Increasing resolution, especially in the axial direction removing noise increasing contrast.

**Haar cascade classifier:**

**Haar – Cascades**

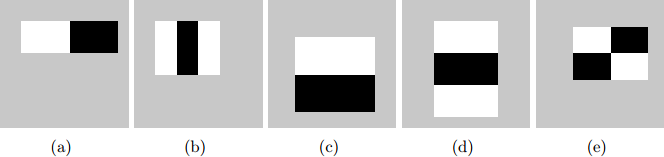
Haar- like features are  rectangular patterns in data. A cascade is a series of “Haar-like features” that are combined to form a classifier [14]. A Haar wavelet is a mathematical function that produces square wave output.



*Figure 2-2.  Haar like Features [13]*

Figure 2.2 shows Haar like features, the background of a template like (b) is painted gray to highlight the pattern’s support. Only those pixels marked in black or white are used when the corresponding feature is calculated [15].

Since no objective distribution can describe the actual prior probability for a given image to have a face, the algorithm must minimize both the false negative and false positive rates in order to achieve an acceptable performance [16]. This then requires an accurate numerical description of what sets human faces apart from other objects. Characteristics that define a face can be extracted from the images with a remarkable committee learning algorithm called Adaboost [17]. Adaboost (Adaptive boost) relies on a committee of weak classifiers that combine to form a strong one through a voting mechanism [18]. A classifier is weak if, in general, it cannot meet a predefined classification target in error terms [7]. The operational algorithm to be used must also work with a reasonable computational budget. Such techniques as the integral image and attention cascades have made the Viola-Jones algorithm [15] highly efficient: fed with a real time image sequence generated from a standard webcam or camera, it performs well on a standard PC.



*Figure 2-3. Haar-like features with different sizes and orientation [13]*

The size and position of a pattern’s support can vary provided its black and white rectangles have the same dimension, border each other and keep their relative positions. Thanks to this constraint, the number of features one can draw from an image is somewhat manageable: a 24 *×* 24 image, for instance, has 43200, 27600, 43200, 27600 and 20736 features of category (a), (b), (c), (d) and (e) respectively as shown in figure 2.3, hence 162336 features in all[13]. In practice, five patterns are considered. The derived features

are assumed to hold all the information needed to characterize a face. Since faces are large and regular by nature, the use of Haar-like patterns.

## How The HAAR – Like Features Work

A scale is chosen for the features say 24 × 24 pixels. This is then slid across the image. The average pixel values under the white area and the black area are then computed. If the difference between the areas is above some threshold then the feature matches [7].

In face detection, since the eyes are of different color tone from the nose, the Haar feature (b) from Figure 2.3 can be scaled to fit that area as shown below,



*Figure 2-4. How the Haar like feature of figure 2.3 can be used to scale the eyes*

One Haar feature is however not enough as there are several features that could match it (like the zip drive and white areas at the background of the image of figure 2.4 it is called a “weak classifier.” Haar cascades, the basis of Viola Jones detection framework

[16] therefore consist of a series of weak classifiers whose accuracy is at least 50% correct. If an area passes a single classifier, it moves to the next weak classifier and so on, otherwise, the area does not match.

**Cascaded Classifier**

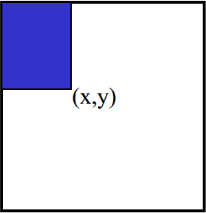
*Figure 2-5. several classifiers combined to enhance face detection*

From figure 2.5, a 1 feature classifier achieves 100% face detection rate and about 50% false positive rate. A 5 feature classifier achieves 100% detection rate and 40% false positive rate (20% cumulative). A 20 feature classifier achieves 100% d                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   etection rate with 10% false positive r                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    ate (2% cumulative)[17].Combining several weak classifiers improves the accuracy of detection.

A training algorithm called Adaboost, short for adaptive boosting [14], which had no application before Haar cascades [14], was utilized to combine a series of weak classifiers in to a strong classifier. Adaboost tries out multiple weak classifiers over several rounds, selecting the best weak classifier in each round and combining the best weak classifier to create a strong classifier [7]. Adaboost can use classifiers that are consistently wrong by reversing their decision [7]. In the design and development, it can take weeks of processing time to determine the final cascade sequence [18].

After the final cascade had been constructed, there was a need for a way to quickly compute the Haar features i.e. compute the differences in the two areas. The integral image was instrumental in this.

#### **Integral Image**



*Figure 2-6. Pixel Coordinates of an integral image*

The Integral image also known as the “summed area table” developed in 1984 came in to widespread use in 2001 with the Haar cascades [4]. A summed area table is created in a single pass. This makes the Haar cascades fast, since the sum of any region in the image can be computed using a single formula [17].

The integral image computes a value at each pixel (x, y) as is shown in figure 2.6, that is the sum of the pixel values above and to the left of (x, y), inclusive. This can quickly be computed in one pass through the image.

Let A, B, C D be the values of the integral image at the corners of a rectangle as shown in figure 2.7.

The sum of original image values within the rectangle can be computed.

= − − + - (2.1)

Only three additions are required for any size of rectangle[17]. This face detection approach minimizes computation time while achieving high detection accuracy[15]. It is now used in many areas of computer vision [4] [7].

Improving Face Detection

Face detection can be improved by tuning the detectors parameters to yield satisfactory results. The parameters to be adjusted are explained as follows.

#### **Scale Increase Rate.**

The scale increase rate specifies how quickly the face detector function should increase the scale for face detection with each pass it makes over an image. Setting the scale increase rate high makes the detector run faster by running fewer passes. If it is set too high it may jump quickly between the scales and miss the faces. The default increase rate in Open CV is 1.1. This implies that the scale increases by a factor of 10 % each pass .The parameters assume a value of 1.1, 1.2, 1.3 or 1.4.

***Minimum Neighbors Threshold***

The minimum neighbor’s threshold sets the cutoff level for discarding or keeping rectangle groups as either faces or not. This is based on the number of raw detections in the group and its values ranges from zero to four.

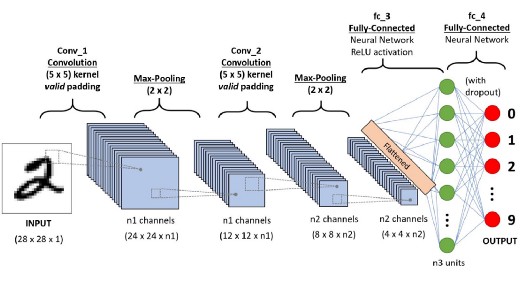
When the face detector is called behind the scenes, each positive face region generates many hits from the Haar detector as in Figure 2.8. The face region itself generates a large cluster of rectangles that to a large extend overlap. The isolated detection are usually false detections and are discarded. The multiple face region detection are then merged in to a single detection. The face detection function does all this before returning the list of the detected faces. The merge step groups rectangles that contain a large number of overlaps and then finds the average rectangle.

CNN (CONVOLUTIONAL NEURAL NETWORK)

Artificial Intelligence has been witnessing a monumental growth in bridging the gap between the capabilities of humans and machines. Researchers and enthusiasts alike, work on numerous aspects of the field to make amazing things happen. One of many such areas is the domain of Computer Vision.

The agenda for this field is to enable machines to view the world as humans do, perceive it in a similar manner and even use the knowledge for a multitude of tasks such as Image & Video recognition, Image Analysis & Classification, Media Recreation, Recommendation Systems, Natural Language Processing, etc. The advancements in Computer Vision with Deep Learning has been constructed and perfected with time, primarily over one particular algorithm — a **Convolutional Neural Network**.

## Introduction

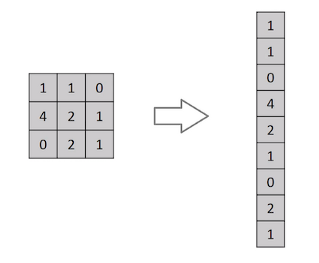


A CNN sequence to classify handwritten digits

A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area.

## Why ConvNets over Feed-Forward Neural Nets?



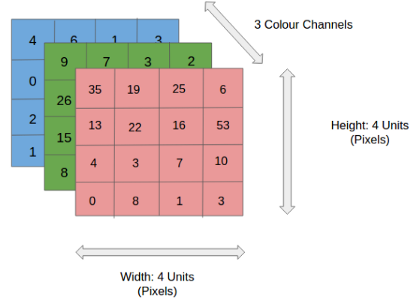
Flattening of a 3x3 image matrix into a 9x1 vector

An image is nothing but a matrix of pixel values, right? So why not just flatten the image (e.g. 3x3 image matrix into a 9x1 vector) and feed it to a Multi-Level Perceptron for classification purposes? Uh.. not really.

In cases of extremely basic binary images, the method might show an average precision score while performing prediction of classes but would have little to no accuracy when it comes to complex images having pixel dependencies throughout.

A ConvNet is able to **successfully capture the Spatial and Temporal dependencies** in an image through the application of relevant filters. The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved and reusability of weights. In other words, the network can be trained to understand the sophistication of the image better.

## Input Image

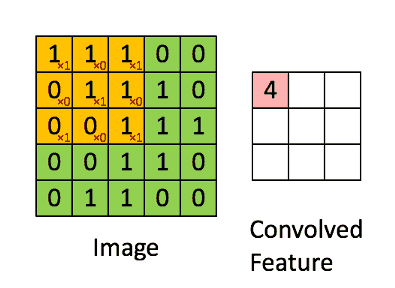


4x4x3 RGB Image

In the figure, we have an RGB image which has been separated by its three color planes — Red, Green, and Blue. There are a number of such color spaces in which images exist — Grayscale, RGB, HSV, CMYK, etc.

You can imagine how computationally intensive things would get once the images reach dimensions, say 8K (7680×4320). The role of the ConvNet is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction. This is important when we are to design an architecture which is not only good at learning features but also is scalable to massive datasets.

## Convolution Layer — The Kernel



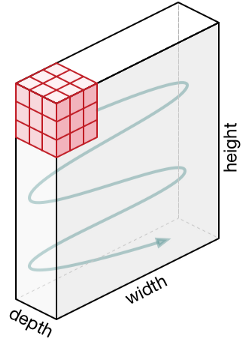
Convoluting a 5x5x1 image with a 3x3x1 kernel to get a 3x3x1 convolved feature

Image Dimensions = 5 (Height) x 5 (Breadth) x 1 (Number of channels, eg. RGB)

In the above demonstration, the green section resembles our **5x5x1 input image, I**. The element involved in carrying out the convolution operation in the first part of a Convolutional Layer is called the **Kernel/Filter, K**, represented in the color yellow. We have selected **K as a 3x3x1 matrix.**

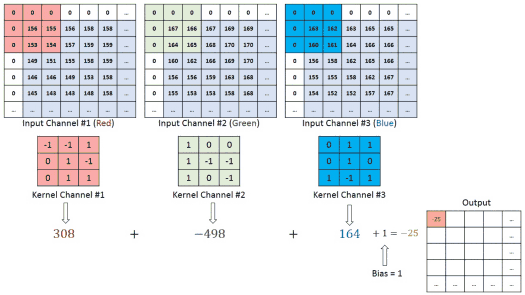
Kernel/Filter, K = 1 0 1  
0 1 0  
1 0 1

The Kernel shifts 9 times because of **Stride Length = 1 (Non-Strided)**, every time performing a **matrix multiplication operation between K and the portion P of the image** over which the kernel is hovering.



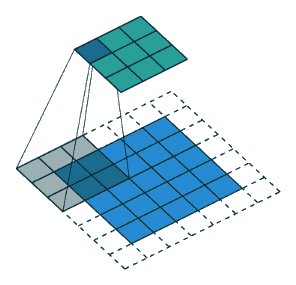
Movement of the Kernel

The filter moves to the right with a certain Stride Value till it parses the complete width. Moving on, it hops down to the beginning (left) of the image with the same Stride Value and repeats the process until the entire image is traversed.



Convolution operation on a MxNx3 image matrix with a 3x3x3 Kernel

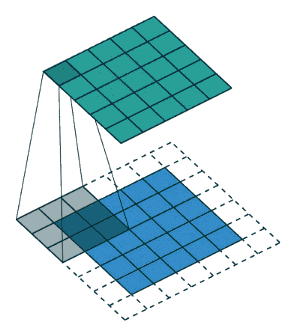
In the case of images with multiple channels (e.g. RGB), the Kernel has the same depth as that of the input image. Matrix Multiplication is performed between Kn and In stack ([K1, I1]; [K2, I2]; [K3, I3]) and all the results are summed with the bias to give us a squashed one-depth channel Convoluted Feature Output.



Convolution Operation with Stride Length = 2

The objective of the Convolution Operation is to **extract the high-level features** such as edges, from the input image. ConvNets need not be limited to only one Convolutional Layer. Conventionally, the first ConvLayer is responsible for capturing the Low-Level features such as edges, color, gradient orientation, etc. With added layers, the architecture adapts to the High-Level features as well, giving us a network which has the wholesome understanding of images in the dataset, similar to how we would.

There are two types of results to the operation — one in which the convolved feature is reduced in dimensionality as compared to the input, and the other in which the dimensionality is either increased or remains the same. This is done by applying **Valid Padding** in case of the former, or **Same Padding** in the case of the latter.

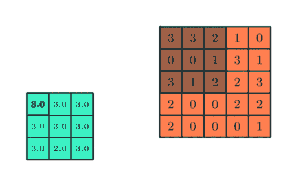


**SAME padding:** 5x5x1 image is padded with 0s to create a 6x6x1 image

When we augment the 5x5x1 image into a 6x6x1 image and then apply the 3x3x1 kernel over it, we find that the convolved matrix turns out to be of dimensions 5x5x1. Hence the name — **Same Padding**.

On the other hand, if we perform the same operation without padding, we are presented with a matrix which has dimensions of the Kernel (3x3x1) itself — **Valid Padding**.

## Pooling Layer

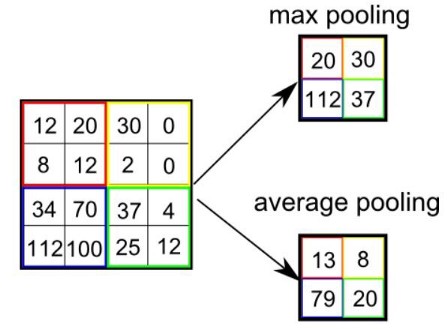


3x3 pooling over 5x5 convolved feature

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to **decrease the computational power required to process the data** through dimensionality reduction. Furthermore, it is useful for **extracting dominant features** which are rotational and positional invariant, thus maintaining the process of effectively training of the model.

There are two types of Pooling: Max Pooling and Average Pooling. **Max Pooling** returns the **maximum value** from the portion of the image covered by the Kernel. On the other hand, **Average Pooling**returns the **average of all the values**from the portion of the image covered by the Kernel.

Max Pooling also performs as a**Noise Suppressant**. It discards the noisy activations altogether and also performs de-noising along with dimensionality reduction. On the other hand, Average Pooling simply performs dimensionality reduction as a noise suppressing mechanism. Hence, we can say that **Max Pooling performs a lot better than Average Pooling**.

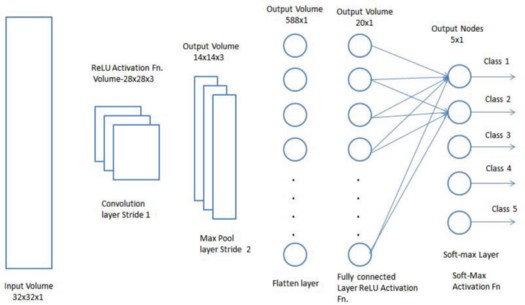


Types of Pooling

The Convolutional Layer and the Pooling Layer, together form the i-th layer of a Convolutional Neural Network. Depending on the complexities in the images, the number of such layers may be increased for capturing low-levels details even further, but at the cost of more computational power.

After going through the above process, we have successfully enabled the model to understand the features. Moving on, we are going to flatten the final output and feed it to a regular Neural Network for classification purposes.

## Classification — Fully Connected Layer (FC Layer)



Adding a Fully-Connected layer is a (usually) cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolutional layer. The Fully-Connected layer is learning a possibly non-linear function in that space.

Now that we have converted our input image into a suitable form for our Multi-Level Perceptron, we shall flatten the image into a column vector. The flattened output is fed to a feed-forward neural network and backpropagation applied to every iteration of training. Over a series of epochs, the model is able to distinguish between dominating and certain low-level features in images and classify them using the **Softmax Classification** technique.

**Neural Network:**

Neural networks are predictive models loosely based on the action of biological neurons.

The selection of the name “neural network” was one of the great PR successes of the Twentieth Century. It certainly sounds more exciting than a technical description such as “A network of weighted, additive values with nonlinear transfer functions”. However, despite the name, neural networks are far from “thinking machines” or “artificial brains”. A typical artificial neural network might have a hundred neurons. In comparison, the human nervous system is believed to have about 3x1010 neurons. We are still light years from “Data”.

The original “Perceptron” model was developed by Frank Rosenblatt in 1958. Rosenblatt’s model consisted of three layers, (1) a “retina” that distributed inputs to the second layer, (2) “association units” that combine the inputs with weights and trigger a threshold step function which feeds to the output layer, (3) the output layer which combines the values. Unfortunately, the use of a step function in the neurons made the perceptions difficult or impossible to train. A critical analysis of perceptrons published in 1969 by Marvin Minsky and Seymore Paper pointed out a number of critical weaknesses of perceptrons, and, for a period of time, interest in perceptrons waned.

Interest in neural networks was revived in 1986 when David Rumelhart, Geoffrey Hinton and Ronald Williams published “Learning Internal Representations by Error Propagation”. They proposed a multilayer neural network with nonlinear but differentiable transfer functions that avoided the pitfalls of the original perceptron’s step functions. They also provided a reasonably effective training algorithm for neural networks.

**Types of Neural Networks:**

1. Artificial Neural Network
2. Probabilistic Neural Networks
3. General Regression Neural Networks

**DTREG** implements the most widely used types of neural networks:

a) Multilayer Perceptron Networks (also known as multilayer feed-forward network),

b) Cascade Correlation Neural Networks,

c) Probabilistic Neural Networks (NN)

d) General Regression Neural Networks (GRNN).

**Radial Basis Function Networks**:

1. Functional Link Networks,

b) Kohonen networks,

c) Gram-Charlier networks,

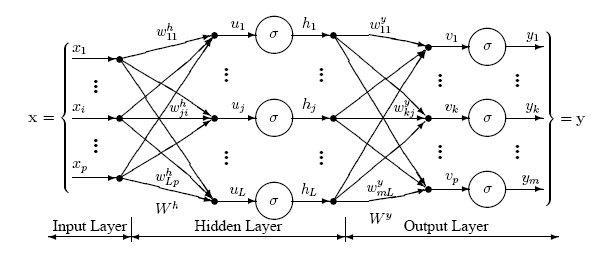
d) Hebb networks,

e) Adaline networks,

f) Hybrid Networks.

**The Multilayer Perceptron Neural Network Model**

The following diagram illustrates a perceptron network with three layers:



This network has an **input layer** (on the left) with three neurons, one **hidden layer** (in the middle) with three neurons and an **output layer** (on the right) with three neurons.

There is one neuron in the input layer for each predictor variable. In the case of categorical variables, *N*-1 neurons are used to represent the *N* categories of the variable.

**Input Layer** — A vector of predictor variable values (*x1...xp*) is presented to the input layer. The input layer (or processing before the input layer) standardizes these values so that the range of each variable is -1 to 1. The input layer distributes the values to each of the neurons in the hidden layer. In addition to the predictor variables, there is a constant input of 1.0, called the *bias* that is fed to each of the hidden layers; the bias is multiplied by a weight and added to the sum going into the neuron.

**Hidden Layer** — Arriving at a neuron in the hidden layer, the value from each input neuron is multiplied by a weight (*wji*), and the resulting weighted values are added together producing a combined value *uj*. The weighted sum (*uj*) is fed into a transfer function, σ, which outputs a value *hj*. The outputs from the hidden layer are distributed to the output layer.

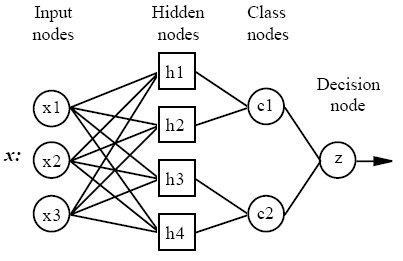
**Output Layer** — Arriving at a neuron in the output layer, the value from each hidden layer neuron is multiplied by a weight (*wkj*), and the resulting weighted values are added together producing a combined value *vj*. The weighted sum (*vj*) is fed into a transfer function, σ, which outputs a value *yk*. The *y* values are the outputs of the network.

If a regression analysis is being performed with a continuous target variable, then there is a single neuron in the output layer, and it generates a single y value. For classification problems with categorical target variables, there are *N* neurons in the output layer producing *N* values, one for each of the *N* categories of the target variable.

**Neural Networks (NN):**

Neural Network (NN) and General Regression Neural Networks (GRNN) have similar architectures, but there is a fundamental difference: networks perform classification where the target variable is categorical, whereas general regression neural networks perform regression where the target variable is continuous. If you select a NN/GRNN network, DTREG will automatically select the correct type of network based on the type of target variable.

**Architecture of a NN:**



All NN networks have four layers:

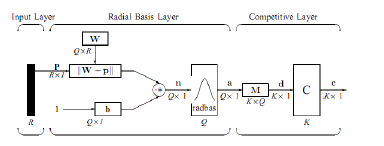
1. **Input layer** — There is one neuron in the input layer for each predictor variable. In the case of categorical variables, *N*-1 neurons are used where *N* is the number of categories. The input neurons (or processing before the input layer) standardizes the range of the values by subtracting the median and dividing by the interquartile range. The input neurons then feed the values to each of the neurons in the hidden layer.
2. **Hidden layer** — This layer has one neuron for each case in the training data set. The neuron stores the values of the predictor variables for the case along with the target value. When presented with the *x* vector of input values from the input layer, a hidden neuron computes the Euclidean distance of the test case from the neuron’s center point and then applies the RBF kernel function using the sigma value(s). The resulting value is passed to the neurons in the pattern layer.
3. **Pattern layer / Summation layer** — The next layer in the network is different for NN networks and for GRNN networks. For NN networks there is one pattern neuron for each category of the target variable. The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron’s category. The pattern neurons add the values for the class they represent (hence, it is a weighted vote for that category).

For GRNN networks, there are only two neurons in the pattern layer. One neuron is the denominator summation unit the other is the numerator summation unit. The denominator summation unit adds up the weight values coming from each of the hidden neurons. The numerator summation unit adds up the weight values multiplied by the actual target value for each hidden neuron.

1. **Decision layer** — The decision layer is different for NN and GRNN networks. For NN networks, the decision layer compares the weighted votes for each target category accumulated in the pattern layer and uses the largest vote to predict the target category.

For GRNN networks, the decision layer divides the value accumulated in the numerator summation unit by the value in the denominator summation unit and uses the result as the predicted target value.

The following diagram is actual diagram or propose network used in our project



**1) Input Layer:**

The input vector, denoted as **p**, is presented as the black vertical bar.Its dimension is *R × 1*. In this paper, *R = 3*.

**2) Radial Basis Layer:**

In Radial Basis Layer, the vector distances between input vector p and the weight vector made of each row of weight matrix **W** are calculated. Here, the vector distance is defined as the dot product between two vectors [8]. Assume the dimension of **W** is *Q×R*. The dot product between **p** and the *i*-th row of **W** produces the *i*-th element of the distance vector ||**W-p**||, whose dimension is *Q×1*. The minus symbol, “**-**”, indicates that it is the distance between vectors. Then, the bias vector **b** is combined with ||**W- p**|| by an element-by-element multiplication, .The result is denoted as **n** = ||**W- p**|| **.**.**p**. The transfer function in NN has built into a distance criterion with respect to a center. In this paper, it is defined as *radbas(n)* = 2 *n e*- (1) Each element of **n** is substituted into Eq. 1 and produces corresponding element of **a**, the output vector of Radial Basis Layer. The *i*-th element of **a** can be represented as ai = *radbas*(||**W**i **- p**|| **.**.**b**i) (2) where **W**i is the vector made of the *i*-th row of **W** and **b**i is the *i*-th element of bias vector **b**.

Some characteristics of Radial Basis Layer:

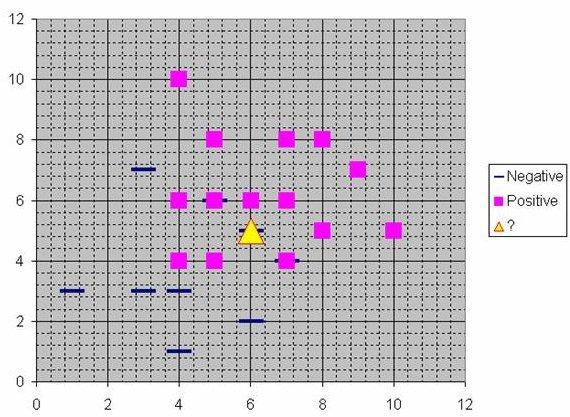
The *i*-th element of a equals to 1 if the input **p** is identical to the *i*th row of input weight matrix **W**. A radial basis neuron with a weight vector close to the input vector **p** produces a value near 1 and then its output weights in the competitive layer will pass their values to the competitive function. It is also possible that several elements of **a** are close to 1 since the input pattern is close to several training patterns. .

**3) Competitive Layer:**

There is no bias in Competitive Layer. In Competitive Layer, the vector **a** is firstly multiplied with layer weight matrix **M**, producing an output vector **d**. The competitive function, denoted as **C** in Fig. 2, produces a 1 corresponding to the largest element of **d**, and 0’s elsewhere. The output vector of competitive function is denoted as **c**. The index of 1 in **c** is the number of tumor that the system can classify. The dimension of output vector, *K*, is 5 in this paper.

**How NN network work:**

Although the implementation is very different, neural networks are conceptually similar to *K-Nearest Neighbor* (k-NN) models. The basic idea is that a predicted target value of an item is likely to be about the same as other items that have close values of the predictor variables. Consider this figure:



Assume that each case in the training set has two predictor variables, *x* and *y*. The cases are plotted using their *x,y* coordinates as shown in the figure. Also assume that the target variable has two categories, *positive* which is denoted by a square and *negative* which is denoted by a dash. Now, suppose we are trying to predict the value of a new case represented by the triangle with predictor values *x*=6, *y*=5.1. Should we predict the target as positive or negative?

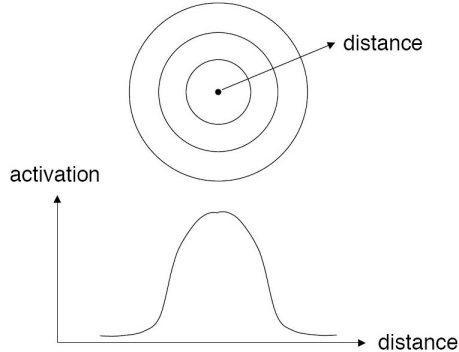
Notice that the triangle is position almost exactly on top of a dash representing a negative value. But that dash is in a fairly unusual position compared to the other dashes which are clustered below the squares and left of center. So it could be that the underlying negative value is an odd case.

The nearest neighbor classification performed for this example depends on how many neighboring points are considered. If 1-NN is used and only the closest point is considered, then clearly the new point should be classified as negative since it is on top of a known negative point. On the other hand, if 9-NN classification is used and the closest 9 points are considered, then the effect of the surrounding 8 positive points may overbalance the close negative point.

A neural network builds on this foundation and generalizes it to consider all of the other points. The distance is computed from the point being evaluated to each of the other points, and a *radial basis function* (RBF) (also called a *kernel function*) is applied to the distance to compute the weight (influence) for each point. The radial basis function is so named because the radius distance is the argument to the function.

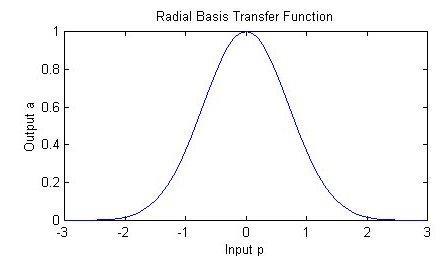
Weight = RBF (*distance*)

The further some other point is from the new point, the less influence it has.



**Radial Basis Function**

Different types of radial basis functions could be used, but the most common is the Gaussian function:



**Advantages and disadvantages of NN networks:**

1. It is usually much faster to train a NN/GRNN network than a multilayer

perceptron network.

1. NN/GRNN networks often are more accurate than multilayer perceptron

networks.

3) NN/GRNN networks are relatively insensitive to outliers (wild points).

4) NN networks generate accurate predicted target probability scores.

5)NN networks approach Bayes optimal classification.

6) NN/GRNN networks are slower than multilayer perceptron networks at

classifying new cases.

7)NN/GRNN networks require more memory space to store the model.

**Removing unnecessary neurons**

One of the disadvantages of NN models compared to multilayer perceptron networks is that NN models are large due to the fact that there is one neuron for each training row. This causes the model to run slower than multilayer perceptron networks when using scoring to predict values for new rows.

DTREG provides an option to cause it remove unnecessary neurons from the model after the model has been constructed.

Removing unnecessary neurons has three benefits:

1. The size of the stored model is reduced.
2. The time required to apply the model during scoring is reduced.
3. Removing neurons often improves the accuracy of the model.

The process of removing unnecessary neurons is an iterative process. Leave-one-out validation is used to measure the error of the model with each neuron removed. The neuron that causes the least increase in error (or possibly the largest reduction in error) is then removed from the model. The process is repeated with the remaining neurons until the stopping criterion is reached.

When unnecessary neurons are removed, the “Model Size” section of the analysis report shows how the error changes with different numbers of neurons. You can see a graphical chart of this by clicking Chart/Model size.

There are three criteria that can be selected to guide the removal of neurons:

1. Minimize error – If this option is selected, then DTREG removes neurons as long as the leave-one-out error remains constant or decreases. It stops when it finds a neuron whose removal would cause the error to increase above the minimum found.
2. Minimize neurons – If this option is selected, DTREG removes neurons until the leave-one-out error would exceed the error for the model with all neurons.
3. # of neurons – If this option is selected, DTREG reduces the least significant neurons until only the specified number of neurons remain.

**CHAPTER 5**

**UML DIAGRAM**

**Use case diagram:**

Input video

Pre-processing

User

CNN

Database hdf5

Play music

Emotions such as happy,sad,disgustst

Haar cascade feature extraction

Sequence diagram:

preprocessing

CNN

User

haar feature extraction

Input video

Emotion such as happy,sad,disgust

Play music

Activity diagram:

start

Input video

Pre-processing

Haar cascade

CNN

stop

Play music

Happy,sad,disgust,neutral etc

**Chapter 6**

**Digital image processing**

**DIGITAL IMAGE PROCESSING**

The identification of objects in an image would probably start with image processing techniques such as noise removal, followed by (low-level) feature extraction to locate lines, regions and possibly areas with certain textures.

The clever bit is to interpret collections of these shapes as single objects, e.g. cars on a road, boxes on a conveyor belt or cancerous cells on a microscope slide. One reason this is an AI problem is that an object can appear very different when viewed from different angles or under different lighting. Another problem is deciding what features belong to what object and which are background or shadows etc. The human visual system performs these tasks mostly unconsciously but a computer requires skillful programming and lots of processing power to approach human performance. Manipulating data in the form of an image through several possible techniques. An image is usually interpreted as a two-dimensional array of brightness values, and is most familiarly represented by such patterns as those of a photographic print, slide, television screen, or movie screen. An image can be processed optically or digitally with a computer.

To digitally process an image, it is first necessary to reduce the image to a series of numbers that can be manipulated by the computer. Each number representing the brightness value of the image at a particular location is called a picture element, or pixel. A typical digitized image may have 512 × 512 or roughly 250,000 pixels, although much larger images are becoming common. Once the image has been digitized, there are three basic operations that can be performed on it in the computer. For a point operation, a pixel value in the output image depends on a single pixel value in the input image. For local operations, several neighbouring pixels in the input image determine the value of an output image pixel. In a global operation, all of the input image pixels contribute to an output image pixel value.

These operations, taken [singly](http://www.answers.com/topic/singly) or in combination, are the means by which the image is enhanced, restored, or compressed. An image is enhanced when it is modified so that the information it contains is more clearly evident, but enhancement can also include making the image more visually appealing.

An example is noise smoothing. To smooth a [noisy](http://www.answers.com/topic/noisy) image, median filtering can be applied with a 3 × 3 pixel window. This means that the value of every pixel in the noisy image is recorded, along with the values of its nearest eight neighbours. These nine numbers are then ordered according to size, and the median is selected as the value for the pixel in the new image. As the 3 × 3 window is moved one pixel at a time across the noisy image, the filtered image is formed.

Another example of enhancement is contrast manipulation, where each pixel's value in the new image depends solely on that pixel's value in the old image; in other words, this is a point operation. Contrast manipulation is commonly performed by adjusting the brightness and contrast controls on a television set, or by controlling the exposure and development time in [printmaking](http://www.answers.com/topic/printmaking). Another point operation is that of [pseudo colouring](http://www.answers.com/topic/pseudocoloring) a black-and-white image, by assigning arbitrary colours to the gray levels. This technique is popular in [thermograph](http://www.answers.com/topic/thermography) (the imaging of heat), where hotter objects (with high pixel values) are assigned one color (for example, red), and cool objects (with low pixel values) are assigned another color (for example, blue), with other colours assigned to intermediate values.

Recognizing object classes in real-world images is a long standing goal in Computer vision. Conceptually, this is challenging due to large appearance variations of object instances belonging to the same class. Additionally, distortions from background clutter, scale, and viewpoint variations can render appearances of even the same object instance to be vastly different. Further challenges arise from interclass similarity in which instances from different classes can appear very similar. Consequently, models for object classes must be flexible enough to accommodate class variability, yet discriminative enough to sieve out true object instances in cluttered images. These seemingly paradoxical requirements of an object class model make recognition difficult. This paper addresses two goals of recognition are image classification and object detection. The task of image classification is to determine if an object class is present in an image, while object detection localizes all instances of that class from an image. Toward these goals, the main contribution in this paper is an approach for object class recognition that employs edge information only. The novelty of our approach is that we represent contours by very simple and generic shape primitives of line segments and ellipses, coupled with a flexible method to learn discriminative primitive combinations. These primitives are complementary in nature, where line segment models straight contour and ellipse models curved contour. We choose an ellipse as it is one of the simplest circular shapes, yet is sufficiently flexible to model curved shapes. These shape primitives possess several attractive properties. First, unlike edge-based descriptors they support abstract and perceptually meaningful reasoning like parallelism and adjacency. Also, unlike contour fragment features, storage demands by these primitives are independent of object size and are efficiently represented with four parameters for a line and five parameters for an ellipse.

Additionally, matching between primitives can be efficiently computed (e.g., with geometric properties), unlike contour fragments, which require comparisons between individual edge pixels. Finally, as geometric properties are easily scale normalized, they simplify matching across scales. In contrast, contour fragments are not scale invariant, and one is forced either to rescale fragments, which introduces aliasing effects (e.g., when edge pixels are pulled apart), or to resize an image before extracting fragments, which degrades image resolution.

In recent studies it is shown that the generic nature of line segments and ellipses affords them an innate ability to represent complex shapes and structures. While individually less distinctive, by combining a number of these primitives, we empower a combination to be sufficiently discriminative. Here, each combination is a two-layer abstraction of primitives: pairs of primitives (termed shape tokens) at the first layer, and a learned number of shape tokens at the second layer. We do not constrain a combination to have a fixed number of shape-tokens, but allow it to automatically and flexibly adapt to an object class. This number influences a combination’s ability to represent shapes, where simple shapes favor fewer shape-tokens than complex ones. Consequently, discriminative combinations of varying complexity can be exploited to represent an object class. We learn this combination by exploiting distinguishing shape, geometric, and structural constraints of an object class. Shape constraints describe the visual aspect of shape tokens, while geometric constraints describe its spatial layout (configurations). Structural constraints enforce possible poses/structures of an object by the relationships (e.g., XOR relationship) between shape-tokens.

**CLASSIFICATION OF IMAGES:**

There are 3 types of images used in Digital Image Processing. They are

1. Binary Image
2. Gray Scale Image
3. Colour Image

**BINARY IMAGE:**

A binary image is a [digital image](http://en.wikipedia.org/wiki/Digital_image) that has only two possible values for each [pixel](http://en.wikipedia.org/wiki/Pixel).  Typically the two colours used for a binary image are black and white though any two colours can be used.  The color used for the object(s) in the image is the foreground color while the rest of the image is the background color.

Binary images are also called bi-level or two-level. This means that each pixel is stored as a single bit (0 or 1).This name black and white, monochrome or monochromatic are often used for this concept, but may also designate any images that have only one sample per pixel, such as [grayscale images](http://en.wikipedia.org/wiki/Grayscale)

Binary images often arise in [digital image processing](http://en.wikipedia.org/wiki/Digital_image_processing) as [masks](http://en.wikipedia.org/w/index.php?title=Mask_(image_processing)&action=edit&redlink=1) or as the result of certain operations such as [segmentation](http://en.wikipedia.org/wiki/Segmentation_(image_processing)), [thresholding](http://en.wikipedia.org/wiki/Thresholding_(image_processing)), and [dithering](http://en.wikipedia.org/wiki/Dither). Some input/output devices, such as [laser printers](http://en.wikipedia.org/wiki/Laser_printer), [fax machines](http://en.wikipedia.org/wiki/Fax), and bi-level [computer displays](http://en.wikipedia.org/wiki/Visual_display_unit), can only handle bi-level images

**GRAY SCALE IMAGE**

A grayscale Image is [digital image](http://en.wikipedia.org/wiki/Digital_image) is an image in which the value of each [pixel](http://en.wikipedia.org/wiki/Pixel) is a single [sample](http://en.wikipedia.org/wiki/Sample_(signal)), that is, it carries only [intensity](http://en.wikipedia.org/wiki/Luminous_intensity) information. Images of this sort, also known as [black-and-white](http://en.wikipedia.org/wiki/Black-and-white), are composed exclusively of shades of [gray](http://en.wikipedia.org/wiki/Gray) (0-255), varying from black (0) at the weakest intensity to white (255) at the strongest.

Grayscale images are distinct from one-bit [black-and-white](http://en.wikipedia.org/wiki/Black-and-white) images, which in the context of computer imaging are images with only the two [colors](http://en.wikipedia.org/wiki/Color), [black](http://en.wikipedia.org/wiki/Black), and [white](http://en.wikipedia.org/wiki/White) (also called bi-level or [binary images](http://en.wikipedia.org/wiki/Binary_image)). Grayscale images have many shades of gray in between. Grayscale images are also called [monochromatic](http://en.wikipedia.org/wiki/Monochromatic), denoting the absence of any [chromatic](http://en.wikipedia.org/wiki/Chromaticity) variation.

Grayscale images are often the result of measuring the intensity of light at each pixel in a single band of the [electromagnetic spectrum](http://en.wikipedia.org/wiki/Electromagnetic_spectrum) (e.g. [infrared](http://en.wikipedia.org/wiki/Infrared), [visible light](http://en.wikipedia.org/wiki/Visible_spectrum), [ultraviolet](http://en.wikipedia.org/wiki/Ultraviolet), etc.), and in such cases they are monochromatic proper when only a given [frequency](http://en.wikipedia.org/wiki/Frequency) is captured. But also they can be synthesized from a full color image; see the section about converting to grayscale.

**COLOUR IMAGE:**

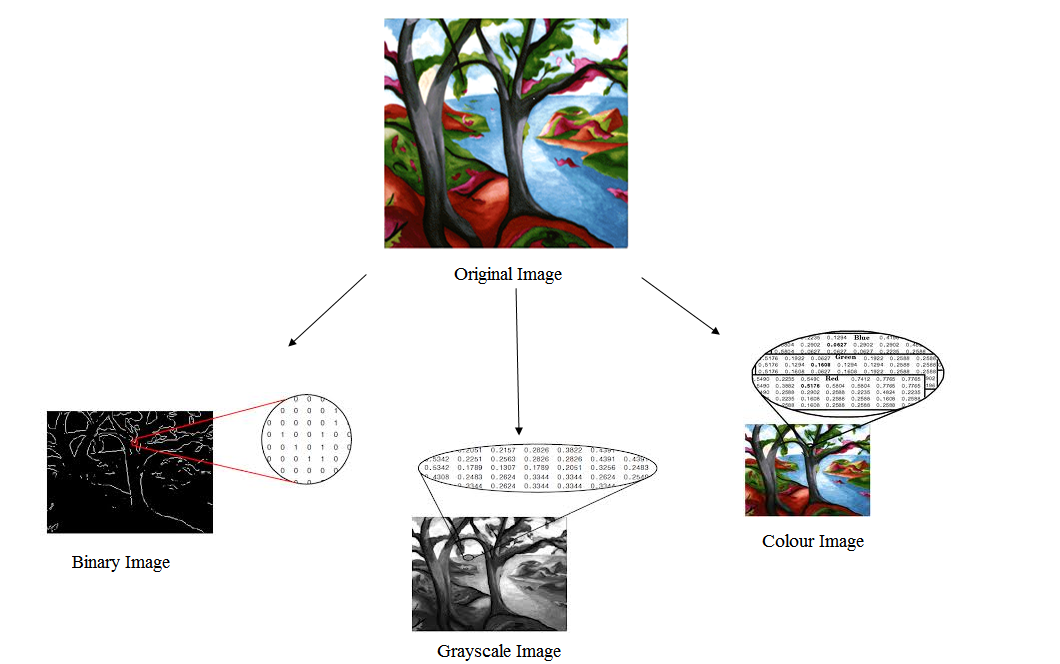
A (digital) color image is a [digital image](http://en.wikipedia.org/wiki/Digital_image) that includes [color](http://en.wikipedia.org/wiki/Color) information for each [pixel](http://en.wikipedia.org/wiki/Pixel). Each pixel has a particular value which determines it’s appearing color. This value is qualified by three numbers giving the decomposition of the color in the three primary colours Red, Green and Blue. Any color visible to human eye can be represented this way. The decomposition of a color in the three primary colours is quantified by a number between 0 and 255. For example, white will be coded as R = 255, G = 255, B = 255; black will be known as (R,G,B) = (0,0,0); and say, bright pink will be : (255,0,255).

In other words, an image is an enormous two-dimensional array of color values, pixels, each of them coded on 3 bytes, representing the three primary colours. This allows the image to contain a total of 256x256x256 = 16.8 million different colours. This technique is also known as RGB encoding, and is specifically adapted to human vision

**CLASSIFICATION OF IMAGES:**

There are 3 types of images used in Digital Image Processing. They are

1. Binary Image
2. Gray Scale Image
3. Colour Image



**Figure 1.1**Representation of a image in binary,grayscale and colour form.

**Binary image:**

A binary image is a [digital image](http://en.wikipedia.org/wiki/Digital_image) that has only two possible values for each [pixel](http://en.wikipedia.org/wiki/Pixel).  Typically the two colours used for a binary image are black and white though any two colours can be used.  The colour used for the objects inthe image is the foreground colour while the rest of the image is the background colour.

Binary images are also called bi-level or two-level. This means that each pixel is stored as a single bit (0 or 1) [Refer figure 1.1].This name black and white, monochrome or monochromatic are often used for this concept, but also designate any images that have only one sample per pixel, such as [grayscale images](http://en.wikipedia.org/wiki/Grayscale).

Binary images often arise in [digital image processing](http://en.wikipedia.org/wiki/Digital_image_processing) as [masks](http://en.wikipedia.org/w/index.php?title=Mask_(image_processing)&action=edit&redlink=1) or as the result of certain operations such as [segmentation](http://en.wikipedia.org/wiki/Segmentation_(image_processing)), [thresholding](http://en.wikipedia.org/wiki/Thresholding_(image_processing)) and [dithering](http://en.wikipedia.org/wiki/Dither). Some input/output devices, such as [laser printers](http://en.wikipedia.org/wiki/Laser_printer), [fax machines](http://en.wikipedia.org/wiki/Fax), and bi-level [computer displays](http://en.wikipedia.org/wiki/Visual_display_unit), can only handle bi-level images.

**Gray scale image:**

A grayscale Image is [digital image](http://en.wikipedia.org/wiki/Digital_image) is an image in which the value of each [pixel](http://en.wikipedia.org/wiki/Pixel) is a single [sample](http://en.wikipedia.org/wiki/Sample_(signal)), that is, it carries only [intensity](http://en.wikipedia.org/wiki/Luminous_intensity) information. Images of this sort, also known as [black-and-white](http://en.wikipedia.org/wiki/Black-and-white), are composed exclusively of shades of [gray](http://en.wikipedia.org/wiki/Gray) (0-255), varying from black (0) at the weakest intensity to white (255) at the strongest [Refer figure 1.1].

Grayscale images are distinct from one-bit [black-and-white](http://en.wikipedia.org/wiki/Black-and-white) images, which in the context of computer imaging are images with only the two [colours](http://en.wikipedia.org/wiki/Color), [black](http://en.wikipedia.org/wiki/Black), and [white](http://en.wikipedia.org/wiki/White) (also called bi-level or [binary images](http://en.wikipedia.org/wiki/Binary_image)). Grayscale images have many shades of gray in between. Grayscale images are also called[monochromatic](http://en.wikipedia.org/wiki/Monochromatic), denoting the absence of any [chromatic](http://en.wikipedia.org/wiki/Chromaticity) variation.

Grayscale images are often the result of measuring the intensity of light at each pixel in a single band of the [electromagnetic spectrum](http://en.wikipedia.org/wiki/Electromagnetic_spectrum) (e.g. [infrared](http://en.wikipedia.org/wiki/Infrared), [visible light](http://en.wikipedia.org/wiki/Visible_spectrum), [ultraviolet](http://en.wikipedia.org/wiki/Ultraviolet), etc.), and in such cases they are monochromatic in nature when only a given [frequency](http://en.wikipedia.org/wiki/Frequency) is captured.

**Colour image:**

A colour image is a [digital image](http://en.wikipedia.org/wiki/Digital_image) that includes [colour](http://en.wikipedia.org/wiki/Color) information for each [pixel](http://en.wikipedia.org/wiki/Pixel). Each pixel has a particular value which determines it’s appearing colour. This value is qualified by three numbers giving the decomposition of the colour in the three primary colours Red, Green and Blue. Any colour visible to human eye can be represented this way. The decomposition of a colour in the three primary colours is quantified by a number between 0 and 255. For example, white will be coded as R = 255, G = 255, B = 255; black will be known as (R,G,B) = (0,0,0)[Refer figure 1.1].In other words, an image is an enormous two-dimensional array of colour values, pixels, each of them coded on 3 bytes, representing the three primary colours. This allows the image to contain a total of 256x256x256 = 16.8 million different colours. This technique is also known as RGB encoding, and is specifically adapted to human vision.

**APPLICATIONS**

**Digital camera images**

Digital cameras generally include dedicated digital image processing chips to convert the raw data from the [image sensor](http://en.wikipedia.org/wiki/Image_sensor) into a [colour-corrected](http://en.wikipedia.org/wiki/Color_correction) image in a standard [image file format](http://en.wikipedia.org/wiki/Image_file_formats). Images from digital cameras often receive further processing to improve their quality, a distinct advantage that digital cameras have over [film](http://en.wikipedia.org/wiki/Photographic_film) cameras. The digital image processing typically is executed by special software programs that can manipulate the images in many ways. Many digital cameras also enable viewing of [histograms](http://en.wikipedia.org/wiki/Histogram) of images, as an aid for the photographer to understand the rendered brightness range of each shot more readily.

**Intelligent transportation systems**

Digital image processing has wide applications in intelligent transportation systems, such as [automatic number plate recognition](http://en.wikipedia.org/wiki/Automatic_number_plate_recognition) and [traffic sign recognition](http://en.wikipedia.org/wiki/Traffic_sign_recognition).

**Chapter 6**

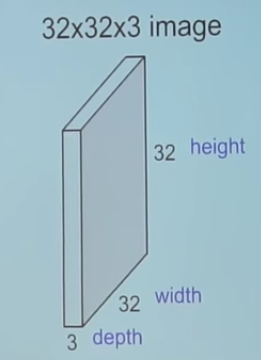
**Algorithm description**

**Convolution Neural** **Networks**

CNNs, like neural networks, are made up of neurons with learnable weights and biases. Each neuron receives several inputs, takes a weighted sum over them, pass it through an activation function and responds with an output. The whole network has a loss function and all the tips and tricks that we developed for neural networks still apply on CNNs. Pretty straightforward, right?

So, how are Convolution Neural Networks different than Neural Networks?

**CNNs operate over Volumes!**



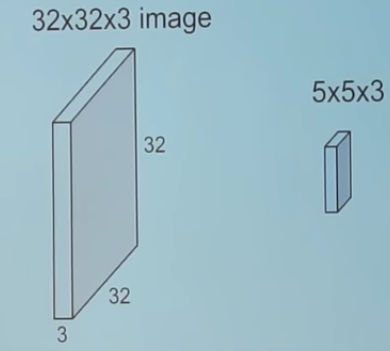
1. Example of a RGB image (let’s call it ‘input image’)

Unlike neural networks, where the input is a vector, here the input is a multi-channeled image (3 channeled in this case).

*There are other differences that we will talk about in a while.*

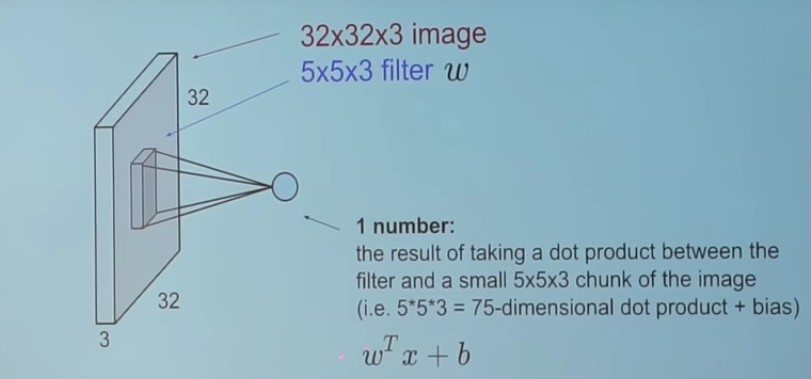
Before we go any deeper, let us first understand what convolution means.

**Convolution**



2. Convolving an image with a filter

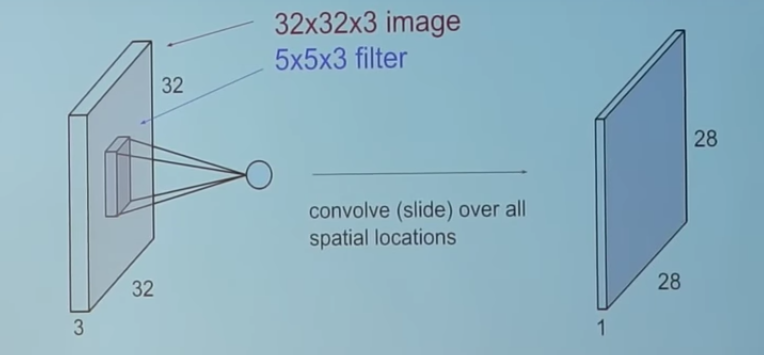
We take the 5\*5\*3 filter and slide it over the complete image and along the way take the dot product between the filter and chunks of the input image.



3. This is how it looks

For every dot product taken, the result is a scalar.

So, what happens when we convolve the complete image with the filter?

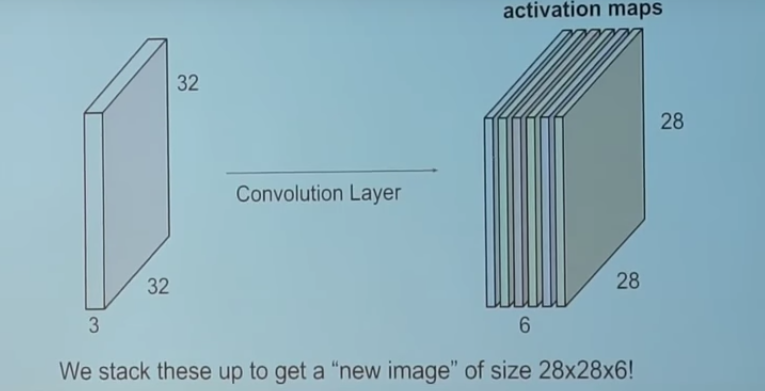


4. This!

I leave it upon you to figure out how the ‘28’ comes. (*Hint: There are 28\*28 unique positions where the filter can be put on the image*)

**Now, back to CNNs**

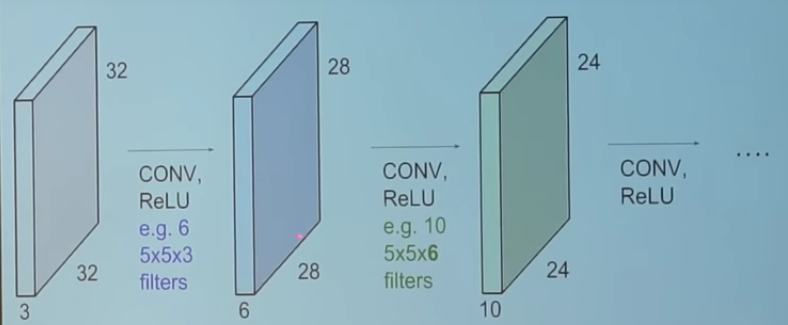
The convolution layer is the main building block of a convolutional neural network.



5. Convolution Layer

The convolution layer comprises of a set of independent filters (6 in the example shown). Each filter is independently convolved with the image and we end up with 6 feature maps of shape 28\*28\*1.

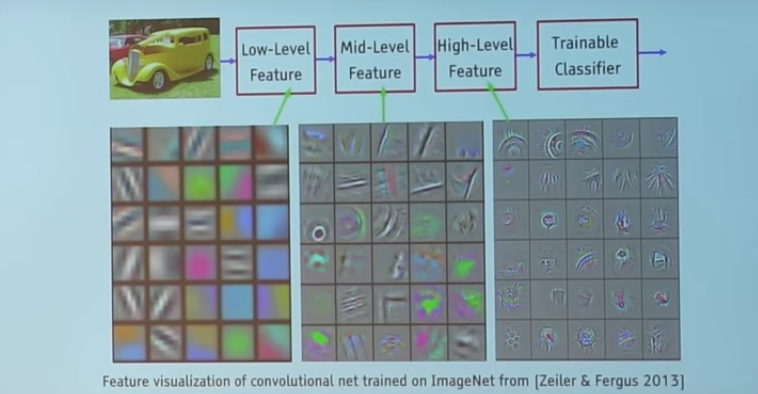
Suppose we have a number of convolution layers in sequence. What happens then?



6. Convolution Layers in sequence

All these filters are initialized randomly and become our parameters which will be learned by the network subsequently.

I will show you an example of a trained network.



7. Filters in a trained network

Take a look at the filters in the very first layer (these are our 5\*5\*3 filters). Through back propagation, they have tuned themselves to become blobs of coloured pieces and edges. As we go deeper to other convolution layers, the filters are doing dot products to the input of the previous convolution layers. So, they are taking the smaller coloured pieces or edges and making larger pieces out of them.

Take a look at image 4 and imagine the 28\*28\*1 grid as a grid of 28\*28 neurons. For a particular feature map (the output received on convolving the image with a particular filter is called a feature map*)*, each neuron is connected only to a small chunk of the input image and all the neurons have the same connection weights. So again coming back to the differences between CNN and a neural network.

CNNs have a couple of concepts called parameter sharing and local connectivity

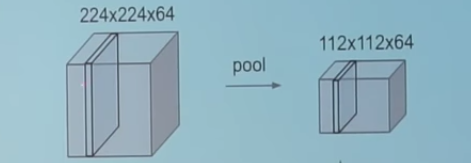
Parameter sharing is sharing of weights by all neurons in a particular feature map.

Local connectivity is the concept of each neural connected only to a subset of the input image (unlike a neural network where all the neurons are fully connected)

This helps to reduce the number of parameters in the whole system and makes the computation more efficient.

**Pooling Layers**

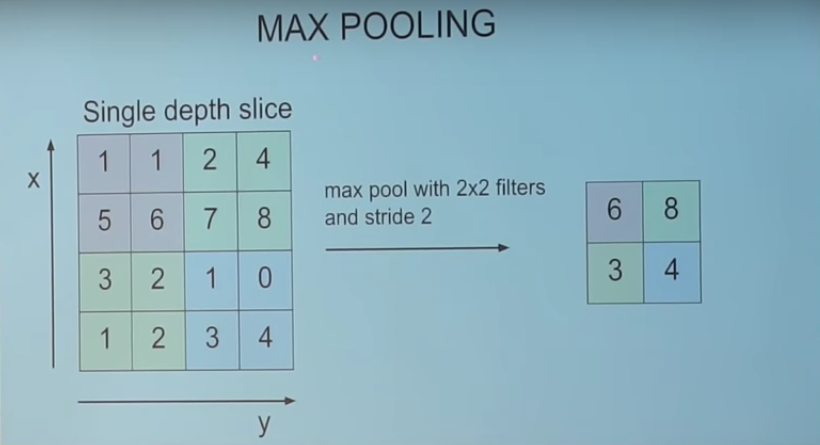
A pooling layer is another building block of a CNN.



Pooling

Its function is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network. Pooling layer operates on each feature map independently.

The most common approach used in pooling is *max pooling.*



Max Pooling

**Typical Architecture of a CNN**

Typical architecture of CNN

We have already discussed about convolution layers (denoted by **CONV**) and pooling layers (denoted by **POOL**).

**RELU** is just a non linearity which is applied similar to neural networks.

The **FC** is the fully connected layer of neurons at the end of CNN. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks and work in a similar way.

CNNs are especially tricky to train, as they add even more hyper-parameters than a standard MLP. While the usual rules of thumb for learning rates and regularization constants still apply, the following should be kept in mind when optimizing CNNs.

#### Number of filters

When choosing the number of filters per layer, keep in mind that computing the activations of a single convolutional filter is much more expensive than with traditional MLPs !

Assume layer (l-1) contains K^{l-1} feature maps and M \times N pixel positions (i.e., number of positions times number of feature maps), and there are K^l filters at layer l of shape m \times n. Then computing a feature map (applying an m \times n filter at all (M-m) \times (N-n) pixel positions where the filter can be applied) costs (M-m) \times (N-n) \times m \times n \times K^{l-1}. The total cost is K^l times that. Things may be more complicated if not all features at one level are connected to all features at the previous one.

For a standard MLP, the cost would only be K^l \times K^{l-1} where there are K^l different neurons at level l. As such, the number of filters used in CNNs is typically much smaller than the number of hidden units in MLPs and depends on the size of the feature maps (itself a function of input image size and filter shapes).

Since feature map size decreases with depth, layers near the input layer will tend to have fewer filters while layers higher up can have much more. In fact, to equalize computation at each layer, the product of the number of features and the number of pixel positions is typically picked to be roughly constant across layers. To preserve the information about the input would require keeping the total number of activations (number of feature maps times number of pixel positions) to be non-decreasing from one layer to the next (of course we could hope to get away with less when we are doing supervised learning). The number of feature maps directly controls capacity and so that depends on the number of available examples and the complexity of the task.

#### Filter Shape

Common filter shapes found in the literature vary greatly, usually based on the dataset. Best results on MNIST-sized images (28x28) are usually in the 5x5 range on the first layer, while natural image datasets (often with hundreds of pixels in each dimension) tend to use larger first-layer filters of shape 12x12 or 15x15.

The trick is thus to find the right level of “granularity” (i.e. filter shapes) in order to create abstractions at the proper scale, given a particular dataset.

#### Max Pooling Shape

Typical values are 2x2 or no max-pooling. Very large input images may warrant 4x4 pooling in the lower-layers. Keep in mind however, that this will reduce the dimension of the signal by a factor of 16, and may result in throwing away too much information.

**Chapter 7**

**Hardware and software requirement**

**Hardware requirement:**

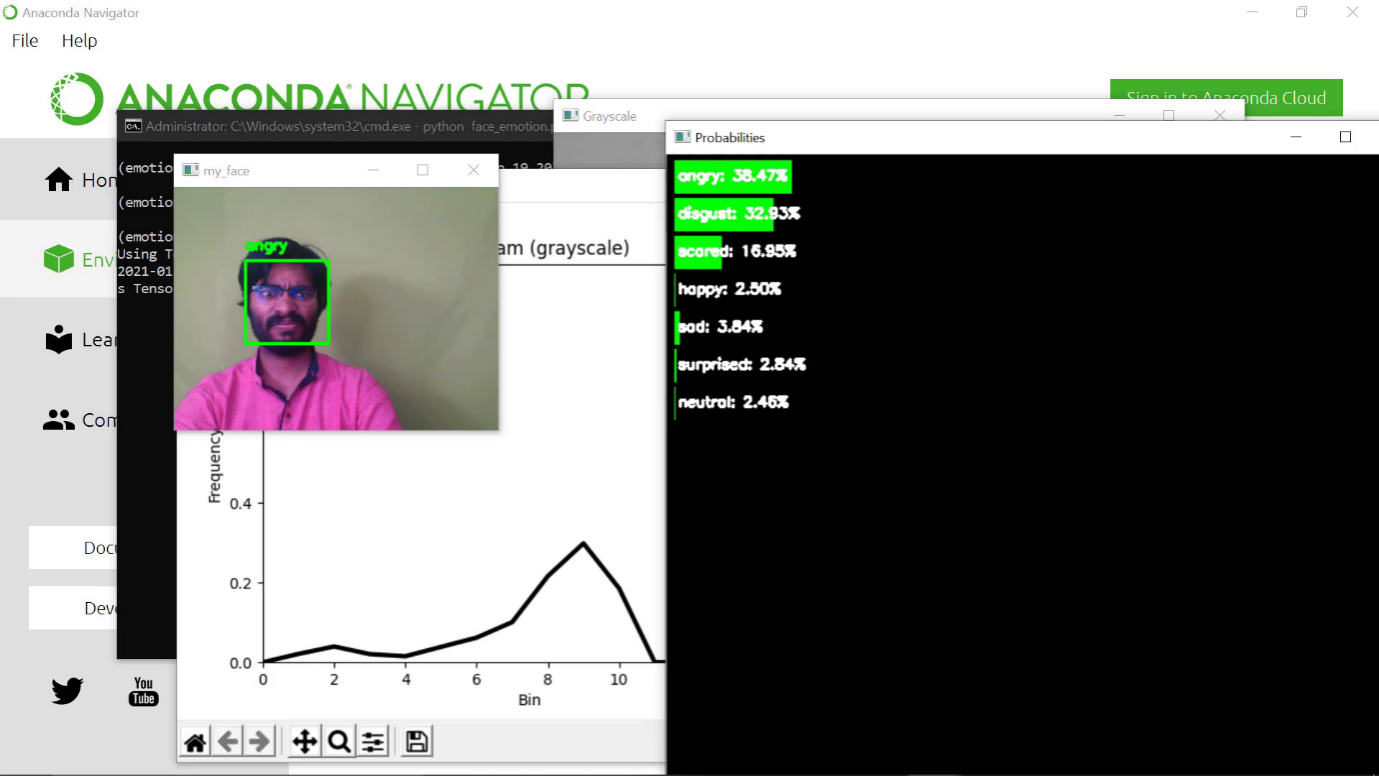
* Hdd=1tb

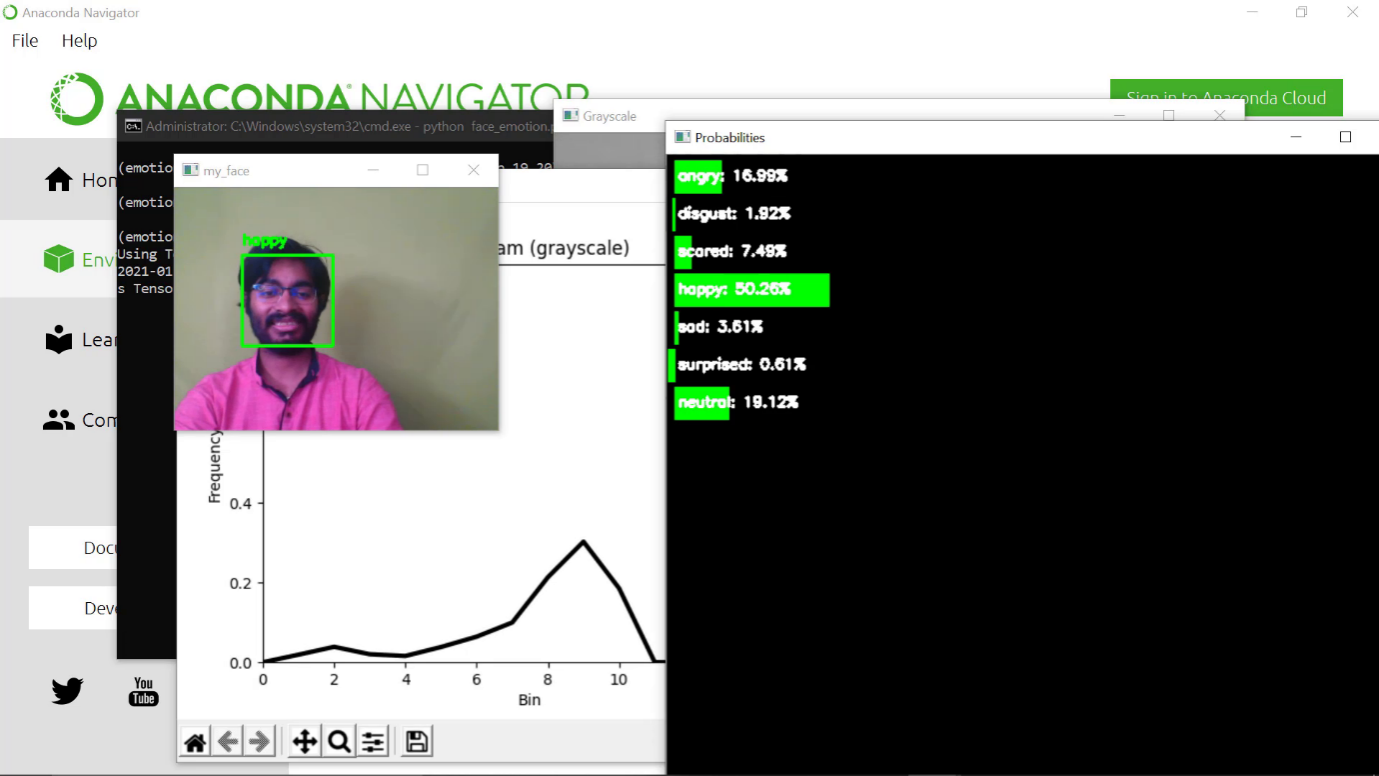
**Software requirement**

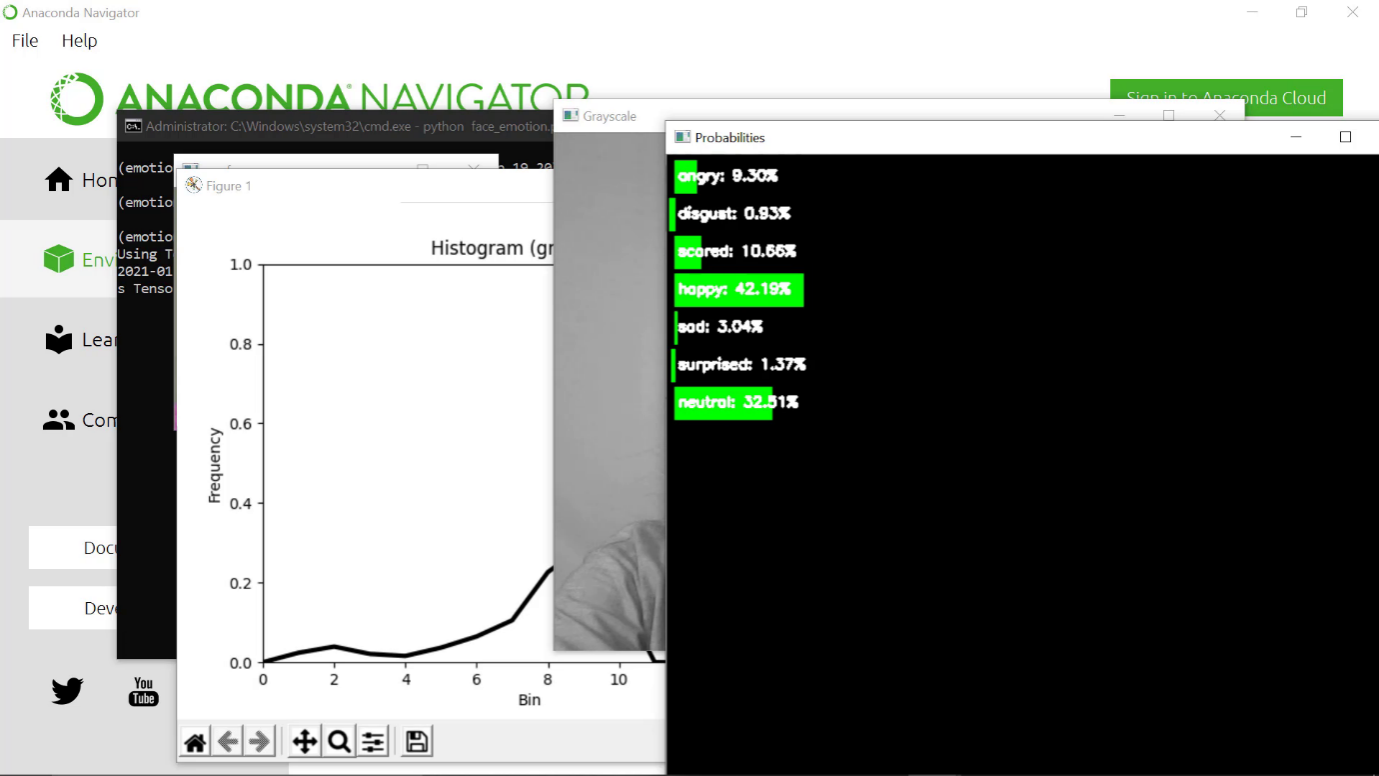
* Python ,
* anaconda navigator

**CHAPTER 8**

**RESULT ANALYSIS:**

****

****

****

**Conclusion and Future Scope**

The aim of the project is to face recognition and face emotion by using deep learning technique. In this we proposed real time video surveillance, what human face expresses it in front of camera and they were recognising the face.

**Future scope**:

In future we have increase the accuracy rate based on facial expression.

**Chapter 9:**

**Reference**

[1] P. Póo Argüelles, “Parálisis cerebral infantil,” 2008.

[2] I. C. Valdovinos Lopez, “Evaluación diagnóstica del niño con parálisis cerebral en el tercer nivel de atención,” 2009.

[3] J. M. Valdez, Medicinaௗ: organo de la Sociedad Argentina de Investigación Clínica. La Sociedad.

[4] G. Sumathy and A. Renjith, “A survey of vision and speech stimulation for cerebral palsy rehabilitation,” in 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1315–1319.

[5] M. González, D. Mulet, E. Perez, C. Soria, and V. Mut, “Vision based interface: an alternative tool for children with cerebral palsy,” in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 5895–5898.

[6] H. Rahmati, O. M. Aamo, O. Stavdahl, R. Dragon, and L. Adde, “Videobased early cerebral palsy prediction using motion segmentation,” in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 3779–3783.

[7] M. Betke, J. Gips, and P. Fleming, “The Camera Mouse: visual tracking of body features to provide computer access for people with severe disabilities,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 10, no. 1, pp. 1–10, Mar. 2002.

[8] C. Stauffer and W. E. L. Grimson, “Learning patterns of activity using real-time tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 747–757, 2000.

**Input Video**

**CNN Classifier**

**Music recommendation**

**Database**

**HDF5**

**Feature Extraction**

**NN training**

**Identification of emotion such as sad, happy, neutral, disgust, stress, scared, angry**

**Preprocessing**

**Haar cascade**