GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

FACIAL EXPRESSION RECOGNITION USING DEEP LEARNING

A project report

Submitted by

SUHAIL AHMED

(1613101757 / 16SCSE101037)

in partial fulfilment for the award of the degree

of

Bachelor of Technology

IN

COMPUTER SCIENCE

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the supervision of

Mr. Ravi Sharma,

Assistant Professor

APRIL / MAY – 2020

GALGOTIAS UNIVERSITY

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"FACIAL EXPRESSION RECOGNITION USING DEEP LEARNING"** is the bonafide work of **"SUHAIL AHMED (1613101757)"** who carried out the project work under my supervision.

SIGNATURE OF HEAD

Dr. S. Raju,

ME(CSE), PhD(CSE).

Professor & Dean,

School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Mr Ravi Sharma,

Assistant Professor

School of Computing Science & Engineering

ABSTRACT

As we move towards a digital world, experiences are real and fathomable. People need results in real-time and in this case human-computer interaction becomes useful. A lot of researches has been done in this field over the past decade. Facial expression is a key feature in expressing one's feelings, affection, and love, and they play an important role in human-computer interaction.

It is known that the average human shows seven different emotions based on the situation, namely anger, sadness, happiness, surprise, disgust, neutral, and scared. Each individual has its way of expressing its emotions can not be linked culturally.

Though there are plenty of methods and researches in the field of machine learning and artificial intelligence, this paper solely focuses on facial expression recognition using deep learning. In this experiment, various datasets have been explored including the FER 2013 and Kaggle and Karolinska directed emotional faces datasets.

This paper aims to classify a fixed number of images into the above-mentioned categories with the implementation of the FaceEX algorithm using the FER2013 database and the CNN model coupled with the Viola-Jones principle.

Keywords: Convolutional neural network, Deep belief network, Facial expression recognition, FER2013, Viola-Jones, FaceEX.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\boldsymbol{TP} = True\ positive$

$\boldsymbol{TN} = True\ negative$

$\boldsymbol{FP} = False\ positive$

$\boldsymbol{FN} = False\ negative$

$\boldsymbol{M_x}$ & $\boldsymbol{M_y} = Mean\ values\ of\ \boldsymbol{F_d}\ and\ \boldsymbol{F_{Nd}}$

$\boldsymbol{R_x}$ & $\boldsymbol{R_y} = Mean\ values\ of\ \boldsymbol{F_d}\ and\ \boldsymbol{F_{Nd}}$

$\boldsymbol{A} = Adjacent\ information$

$\boldsymbol{P_t} = Head\ pose$

$\boldsymbol{S} = State\ space$

$\boldsymbol{\pi} = Sample\ data's\ weight\ for\ state\ space$

$\boldsymbol{W} = Maximum\ weight\ for\ state\ space$

$\underset{\boldsymbol{jK}}{\rightarrow}\ and\ \underset{\boldsymbol{kl}}{\rightarrow} = Angle\ formed\ by\ two\ different\ vectors$

$\boldsymbol{\psi_\theta} = Mother\ wavelet\ of\ image\ plane$

$\boldsymbol{P_{opt}} = Optimal\ position$

$\boldsymbol{A'} = Global\ and\ local\ locations$

$\boldsymbol{IL(u,v)} = Illumination\ muscle\ details$

$\boldsymbol{M(u,v)} = Mask\ pixel\ details$

1. INTRODUCTION

1.1 Motivation

Recognizing human expressions and emotions has drawn several attention lately, as the capabilities are unknown. Many claims have been made about the positivity of facial recognition, including, targeted advertisement, crowning with an augmented and enhanced human communication by amending and learning several human emotions.

Facial image recognition is a thriving area with lots of progress being made daily, such as automatic translation systems, machine to human interaction, and controlling machine with the help of facial muscles.

With an increase in the use of technology, it is observed that human-computer interaction has become important. As a result, facial image recognition by machines has become a heavily researched field by experts over the last decade. There is a need for an application that will be able to detect and classify human expressions. This classification of emotions can then be used for understanding human nature in the field of psychology, or to assist the user as we see now.

A facial expression is a comprehensive tool that distinguishes an individual from another. Although facial expression can vary from person to person but still the underlying feelings that they showcase are the same. A significant amount of studies have been conducted on the topic of facial expression recognition considering its benefits. For example, FER can help identify if a driver is fatigued or not which could prevent a possible cause of an accident, and the same with the case in medical treatment.

Altogether a human shows seven different expressions – namely anger, sad, happy,

scared, surprise, disgust, and neutral, which varies from person to person and is not culture-specific. FER usually has three different stages – pre-processing, facial expression learning, and classification of faces based on the emotion shown.

In contrast, this paper focuses on several aspects of facial image recognition, such as existing systems, proposed methodology, challenges faced, the algorithm used, and its accuracy, datasets, and future enhancement.

1.2 Emotion recognition

Facial recognition is the art of detecting emotions from a human being and then successfully categorizing the emotions in the afore-mentioned emotion categories. Various technological advancements and developments in this area made sure that interest is not lost and continuous progress is being made. It is expected that non-verbal communication will be the next source of communication between a human being and a computer.

Emotions can be identified using various techniques, and one common way is to flatten the image and convert it into ASCII codes. Now with the help of ASCII codes, plain vector points can be mapped onto the face, and facial images can be categorized with accuracy. Having said so, it is not the only way of identifying and categorizing facial images.

As we have seen one way of identifying the emotions, it is to be noted that this process is not as easy as it sounds. Lots of factors come into play when categorizing the emotions, and readers must be aware that each human being is unique and showcases different sets of emotions and emotions are not culturally linked.

1.3 Deep learning

Deep learning is a machine learning technique that models the data that are

designed to do a particular task. Deep learning is neural networks that have wide applications in the area of image recognition, classification, decision making, pattern recognition, etc. Other deep learning techniques like multi-modal deep learning is used for feature selection, image recognition, etc.

1.4 Application of facial image recognition

Facial image recognition is used in BPO's for identifying calls based on emotions. Emotion recognition serves as the medium of communication and identifying conversational analysis, such as unsatisfied customers, customer satisfaction, etc. FER is also used in medical industries and to identify whether a driver is fatigued or not, which can help prevent possible accidents.

Before images are fed into the machine for deep learning, we must be careful about the quality of the images, for example, facial rotation, illumination, facial alignment, etc, and lots of pre-processing techniques are applied so that the images are flattened at an earlier stage to avoid problems and diminishing accuracy result at a later stage of time.

1.5 ***Pre-processing***

Before we can feed the images to the learning model, we need to make sure that all the images are identical. Identical in the sense that, the facial images are oriented properly, illumination, contrast, facial rotation, all of these play a vital role in categorizing the images and determining the accuracy. Pre-processing can help in normalizing the visual appearance and the semantics, which can help the neural networks to learn better.

1.6 ***Facial expression learning***

There can be multiple ways of training the model, but the most common way is to feed in the images multiple times so that the neural networks are familiar with the

facial images.

Suppose say for example that we have four facial images. Two images of one person and the other two of the next person showing different expressions. The first step is to feed the images of the first person only to the neural network and allow the model to categorize them of the same person. Now, this step takes multiple iterations and can be time-consuming but the results speak for themselves. After the successful classification of the images fed, the images of the next person should be allowed to train and a similar approach can be followed.

The next step is to feed two different images of both the persons. One image of person one and one image of person two. Now, the image should be instantly recognized by the neural network and then categorize them as different. If the images are not categorized differently then we need to follow every step from the very beginning.

1.7 ***Classification of facial images***

After the model is trained, the final step in FER is to categorize the facial images into one of the seven emotion categories. As discussed earlier, a normal human being shows seven different basic emotions – anger, sad, happy, scared, surprise, disgust, and neutral. In a broad spectrum, images are categorized and classified as one of the above-mentioned categories.

To classify the images correctly, the images should first be normalized so that any unwanted properties may not harm the final accuracy of categorization.

1.7.1 ***Face normalization***

There can be lots of variations to an image, especially the contrast, the facial alignment, illumination, etc, and can hence impair the accuracy of FER. Below are two methods that are used commonly in normalizing image irregularities: (A)

Contrast normalization and (B) pose normalization (facial alignment).

1.7.2 Contrast normalization

Contrast and brightness may vary in facial images due to unconstrained backgrounds. Numerous findings have shown that histogram equalization with contrast normalization results in superior facial recognition. Global Contrast Normalization (GCN), local normalization, histogram normalization has shown consistent results.

1.7.3 Pose normalization

Facial poses can significantly alter the results and neural network learning. One solution in combating this problem is to render a 3D texture model related to a class or group of faces and estimate the facial components [4][11]. Then, the front portion of the face can be normalized by projecting the frontal area of the face onto the graphical coordinate system.

2. LITERATURE SURVEY

There are a lot of studies that have been conducted on this topic, and listing each one of them is difficult. However, in Table 1, a small number of these studies have been listed whose motivation has been the same as mine.

Table 1: Summary of some of the recent papers

Dataset	Feature extraction technique	Classification technique	Recognition rates
JAFFE, and MUG	Local Fisher Discriminant Analysis (LFDA)	1-nearest-neighbour	JAFFE: 94.37% MUG: 95.24%
JAFFE	Gabor filter	Bayesian	96.23%
JAFFE, and YALE	Gabor techniques	The neural network back-propagation algorithm	JAFFE: 96.83% YALE: 92.22%

JAFFE	Gabor wavelet transform, PCA and LBP	k-NN	90%
CK+	Kernel PCA	KPCA	KPCA: 76.5% PCA: 72.3%
Private	Eigenface approach	Euclidean distance	85.38%
CK+	Active shape models	RBF kernel SVM, HMM	SVM: 70.6% HMM: 65.2%
Private	Biorthogonal wavelet entropy	Fuzzy multiclass SVM	96.77%
CK, and Berlin	Gabor filter for images, and Mel-frequency cepstral coefficients for audio signals	SVM	CK: 84.68% Berlin: 80.68% In-real-time: 81.58%
JAFFE, and CK+		CNN	JAFFE: 76.74% CK+:80.3%
CK+	Gabor	Linear, RBF	97.42%
CK+, JAFFE		CNN	CK+: 96.76% JAFFE: 82.10%
Private	DWT	Single hidden layer	89.49%

3. Existing system

There are many studies and research techniques that are being done for analyzing facial expressions. Those techniques have been implemented through geometrical features and location-based parameter distances.

Deep learning has recently become a hot research topic and lots of progress has been made in achieving state-of-the-art performance. In this section, I briefly introduce some of the existing system worth looking into.

3.1 Support Vector Machine (SVM)

A support vector machine is used to perform the classification of the given images based on its extracted features by using the two-group classification problems.

Statistical and geometrical/topological features are extracted from images to identify the components which are used to make up a particular object for analysis purpose. This extractions process involves high tolerance of distortion and variations in the styles of an object. Translations and rotation methods are involved in SVM to classify them based on segregated data from an input image. In this facial expression recognition module, Gabor and wavelet transformation are utilized from SVM. Those manipulation methods perform non-linear mapping functions and coordinate geometry to separate input data into a high dimensional feature space through a selected non-linear mapping function. SVM aims to solve the problem of inadequate images, which otherwise would affect the accuracy of the image classification in the final result severely.

3.2 Deep Belief Network (DBN)

It is used to extract a deep hierarchical representation of the training data and was developed by Hinton et al. In machine learning, DBN is a generative graphical model, composed of multiple layers of latent variables, with connections between the layers but not between units within each layer. It is especially beneficial if the images are perfectly aligned, using size and rotation. Fig. 1 shows the DBN functional structure. In the era of machine learning, DBN has many layers for image classification problems. Each layer is trained with the help of a greedy layer-wise strategy.

DBN learns probabilistically when it is trained on a set of examples without supervision. The layers then detect the features and the model can further be trained with supervision to perform classification. This deep belief network does fine-tune of the image given as input in its module. This DBN performs the operations for trimming the features through visible layers and “N” number of perceptron layers to the segregation of the image features.

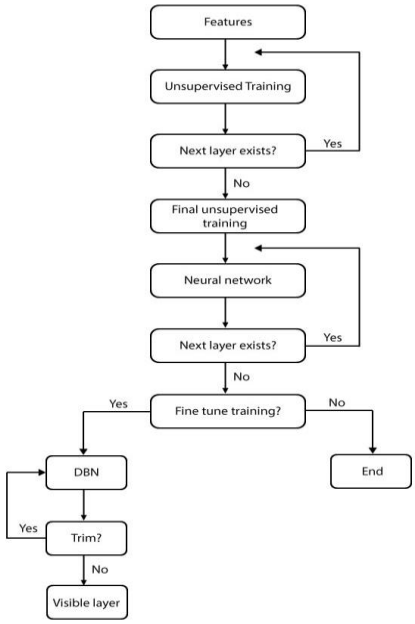

Fig 1: Layers of Deep Belief Network

3.3 GoogleNet

GoogleNet is the award-winning concept of ILSVRC 2014 (ImageNet Large Scale Visual Recognition Competition) presented by Prof. Yan LeGun's LeNet for better image classification and finally it is adopted with Google services. When compared to VGGNet, GoogleNet minimizes the error rate significantly. It started to work on a 1x1 to 5x5 layered structure for a fully connected layer. Inception module concept is used in this technique where the different single input image is passed through several convolutions and various outputs are stacked together in a single place. 22 layers are used to go deeper for image feature extractions. There are two different testings used as follows: Multi-Scale and Multi-Crop Testing.

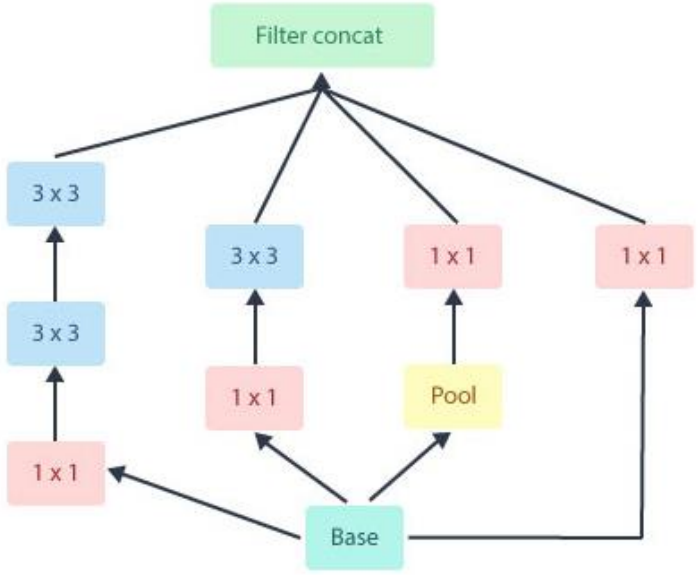

Fig 2: The inception structure of GoogleNet

As shown in figure 2, one significant characteristic of GoogleNet is that it is designed very deep. The optimization methods which GoogleNet uses are worthy of study. For example, GoogleNet has adopted a modular approach to standardizing the results, which makes it easier to modify. Besides, the average pooling was used to replace the whole connectional layer at the end, which makes the rate of success increase by 0.6.

3.4 Deep autoencoder (DAE)

DAE was first introduced to learn efficient coding for dimensionality reduction. DAE is composed of two symmetrical DBNs that have approximately 4 to 5 layers. DAE is optimized to reconstruct its inputs by minimizing the reconstruction error. The main feature of DBN is that it recovers the original undistorted input.

3.4.1 Auto-encoding – How it works

Let’s sketch out a simple encoding output.

800 (Input) → 1000 → 400 → 250 → 100 → 48

Here, the input fed is 800 pixels, and the first layer is 1000, which is larger. Now, this may seem counter-intuitive because if the original image is 800 pixels and if we stretch it to 1000 pixels, normally one's comment would be that the image will stretch and get blurry.

Contrary to this, expanding will make auto-encoding possible. This phenomenon is due to the capacity of the sigmoid-belief units. Sigmoid-belief units cannot represent as much information as real-valued data and hence in the first layer, the image needs to stretch.

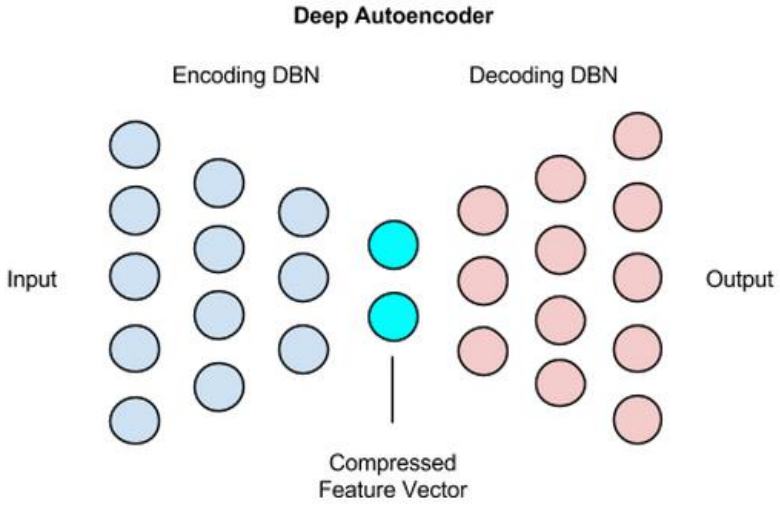

Fig 3: Deep auto-encoder process

3.5 Recurrent neural network (RNN)

A recurrent neural network is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. In addition to training the deep neural network in a single feed-forward manner, RNNs include recurrent edges that span adjacent time steps and share the same parameters across all the steps. The term recurrent justify to this model because it consists of two steps: one is a finite impulse and the other is an infinite impulse. The finite impulse can be

unrolled and replaced whereas the infinite impulse cannot be unrolled or replaced due to its nature of infinite loops.

3.6 Generative adversarial network (GAN)

GAN was first introduced by Goodfellow et al in 2014, which trains the model using vector points. Suppose that we are interested in generating a black and white image of a person of size n x n pixels. We reshape each of the data as N = n x n-dimensional vector such that the image can be understood and mapped into a vector. We can say that the N-dimensional vectors that effectively give something that looks a person's face are distributed to a specific probability.

Hence we can say that in a given training set, this technique learns to generate new data with the same statistics as the training set.

4. Proposed system

4.1 Convolutional neural network (CNN)

CNN's are deep learning algorithms and are also known as shift invariant or space invariant artificial neural networks (SIANN). They have applications in image and video recognition, image classifications, medical image analysis, and financial time series. An image is a primary input on CNN.

The main feature of CNN is to learn the filters which can help distinguish the images. The learning of filters are automatic, that means, as more number of images are fed, a CNN should be able to train itself and then categorize the images accordingly.

CNN's can be said to be analogous to the human brain. Similar to a human brain, CNN can be trained to understand the sophistication of the image, which in turn will lead to better accuracy in determining or categorizing the images.

The main goal of CNN is to resize the images in the form which is easier to process, reduces power consumption without sacrificing the quality of image prediction.

A CNN consists of three types of heterogeneous layers: Convolutional layers, pooling layers, and fully connected layers.

4.1.1 Convolutional layers

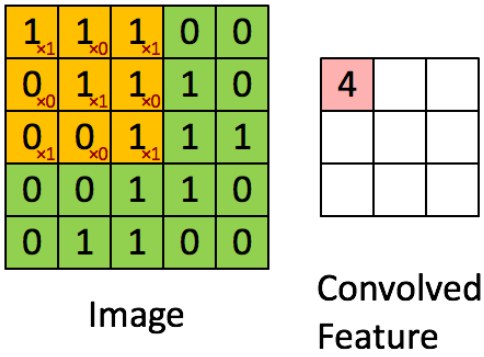

Fig 4: Selected 3x3x1 matrix

As seen in the image, the selection of a 3x3x1 matrix moves right until it traverses the complete image, and then moves down until it traverses the complete image. Please note that the movement of the matrix, that is to the left to the right and from the top to the bottom is continuous. The selected 3x3x1 matrix kernel shifts 9 times because of a stride length of 1. The convolutional layers are associated with three main benefits: local connectivity, which learns the relation between neighboring pixels and weight sharing, which reduces the number of parameters to be learned.

The main objective of the convolutional layers is to extract the high-level features of the image, such as the edges, color gradient orientation, etc.

4.1.2 Pooling layers

The pooling layers reduce the spatial size of the maps (convoluted feature) and decrease the computational power required to process the data which in turn

reduces the computational cost of the network.

There are two types of pooling layers: Max pooling and average pooling.

Max pooling returns the max value from the image whereas, the average pooling returns the average value from the image.

4.1.3 Fully connected layers

After the images are converted to a suitable format, fully connected layers flatten the images into vectors which can be understood or read by the machine. The vectors are fed multiple times and the model can distinguish or learn the ability to distinguish the images. The fully connected layer ensures that all neurons in the layers are fully connected and to enable the 2D feature maps to be converted to 1D feature maps.

4.2 Facial expression datasets

The two datasets that I have used in this experiment are Kaggle's Facial Expression Recognition Challenge and Karolinska Directed Emotional Faces (KDEF). These datasets are very popular and are open-source to which hundreds of good quality images are being uploaded everyday.

The dataset consists of 35,887 labeled images, which are divided into 3,589 test images and 28,709 train images. The dataset consists of another 3,589 private test images, on which this experiment is based on. All the images are 48x48 pixels, colored black, and white, and normalization is done automatically.

Table 2: Categorization of the images in the dataset

Label	**Number of images**	**Emotion type**
0	4593	Anger
1	547	Disgust
2	5121	Fear

3	8989	Happy
4	6077	Sad
5	4002	Surprise
6	6198	Neutral

4.3 Architecture of CNN used

4.3.1 VGG-16

VGG-16 is one of the functional and comprehensive model architecture for a convolutional neural network (CNN). This architecture uses stride1 and stride2 to for the convolution layers of 3X3 filter with padding max pool 2X2 layers instead of hyperparameters. In Fig.4, the picture is gone through a heap of convolutional layers, where the channels are utilized with an exceptionally little responsive field: 3x3. In the ensuing designs, it additionally uses 1x1 convolutional channels, which can be viewed as a straight change of the info channels. The convolutional stride is fixed to 1 pixel (px) and the cushioning is 1px for 3x3 convolutional layers. Spatial pooling is done by five max-pooling layers, which follow a portion of the convolutional layers. Max-pooling is performed over a 2x2 pixels window, with stride 2. Three Fully Connected (FC) layers follow a pile of convolutional layers (which has an alternate profundity in various models). The initial two have 4096 channels each, the third performs 1000-way, ILSVRC classification and in this way contains 1000 channels (one for each class). The last layer is the softmax layer. The setup of the Fully Connected layers is the equivalent in all networks. All concealed layers are furnished with the correction of nonlinearity [9]. It is likewise noticed that none of the systems (aside from one) contain Local Response Normalization (LRN), such standardization doesn't improve the presentation on the ILSVRC dataset, and however, it prompts for expanded memory utilization and calculation time.

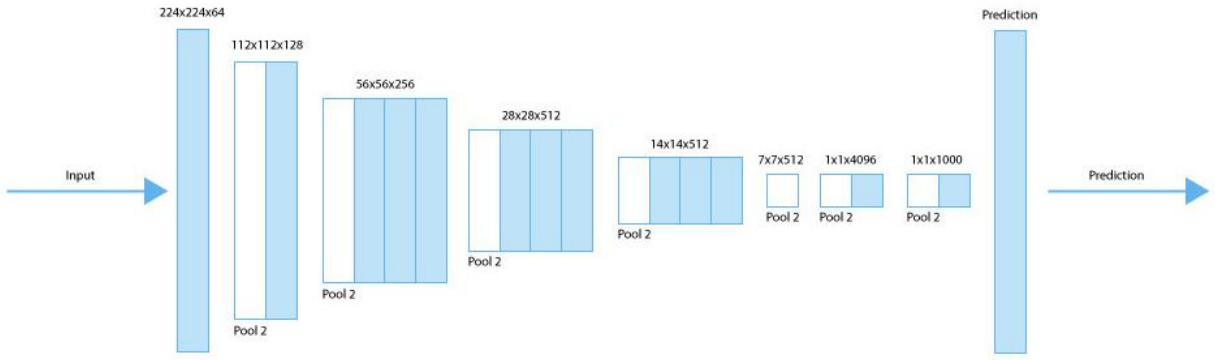

Fig. 5: VGG-16 architecture at a glance

5. The implementation procedure of facial expression

Read the input image
If (Non face image (F_{Nd}) == yes)
Read face (F_d) and nonface (F_{Nd}) data sets to identify face regions from kernel lines
Trace the face from background situations
Remove nonface regions using $f\{\theta\}$
Calculate distance between (F_d) and (F_{Nd}),
$\boldsymbol{f\{\theta\}} = \dfrac{\theta^T\{[M_x - M_y][M_x - M_y]^T + R_x + R_y\}\theta}{\theta^T R_x \theta}$
Where,
$\boldsymbol{M_x}$ & $\boldsymbol{M_y} = Mean\ values\ of\ \boldsymbol{F_d}\ and\ \boldsymbol{F_{Nd}}$
$\boldsymbol{R_x}$ & $\boldsymbol{R_y} = Mean\ values\ of\ \boldsymbol{F_d}\ and\ \boldsymbol{F_{Nd}}$
Apply coarse region detection to refine the resulting images,
$d_q(x,y)^2 = (x-y)^T A(x-y)$
Where,
$\boldsymbol{A} = Adjacent\ information$
Analysis of face and head pose estimation
$\boldsymbol{P_t} = \dfrac{\sum_{i=1}^{N} S_t{}^i \pi_t{}^i W_t{}^i}{\sum_{i=1}^{N} \pi_t{}^i W_t{}^i}$
Where,
$\boldsymbol{P_t} = Head\ pose$

$\boldsymbol{S} = State\ space$
$\boldsymbol{\pi} = Sample\ data's\ weight\ for\ state\ space$
$\boldsymbol{W} = Maximum\ weight\ for\ state\ space$
Facial changes extraction for facial expressions (Geometric facial features for shape and facial components locations)
$\boldsymbol{r^{jkl}} = \frac{d^{jk}}{d^{kl}}, s^{jkl} = \theta(\underset{jK}{\rightarrow}, \underset{kl}{\rightarrow})$
Where,
$\underset{\boldsymbol{jK}}{\rightarrow}\ and\ \underset{\boldsymbol{kl}}{\rightarrow} = Angle\ formed\ by\ two\ different\ vectors$
Gabor wavelets for the whole face or specific regions to extract feature vector
$\boldsymbol{\psi_\theta(b_x b_y, x, y, x_0 y_0)} = \frac{1}{\sqrt{b_x b_y}}\ \psi\ (\frac{x - x_0}{b_x} + \frac{y - y_0}{b_y})$
Where,
$\boldsymbol{\psi_\theta} = Mother\ wavelet\ of\ image\ plane$
Implementation procedure - continuation
$\boldsymbol{P_{opt}} = Optimal\ position$
$\boldsymbol{A'} = Global\ and\ local\ locations$
Distance-based metric, $DB\left(I_{target\_exp}(\lambda_k)\right) = e^{\vec{E}target\_exp(\lambda)}$
Muscle distribution-based model,
$\boldsymbol{IL^{MD}(u, v)} = M(u, v)IL(u, v)$
Where,
$\boldsymbol{IL(u, v)} = Illumination\ muscle\ details$
$\boldsymbol{M(u, v)} = Mask\ pixel\ details$

6. Result

6.1 Setup

The goal of this experiment is to calculate the accuracy and to evaluate the performance metric of the Convolutional neural network using the open-source data-set known as FER2013. The behavior of the system is highly dependent on

the available data, and this test was conducted using the publicly available database. The facial images were extracted from the Kaggle's and Karolinska Directed Emotional Faces (KDEF) and were cropped to a dimension of 48x48. The training was performed with five different methods – Tang 13, Devries et al. 14, Zhang et al. 15, Guo et al. 16, Kim et al. 16. The method was based on the convolution neural network (CNN) network type with a network size of 4-10. There were a total of 28,000 facial emotions extracted from the Kaggle's and Karolinska Directed Emotional Faces (KDEF) used from training, 3,500 facial emotions from validation sets, and 3,500 for the test set. Out of the five methods performed Zhang et al. 15, performed on the network type CNN, shown higher accuracy rates as compared to the other methods.

6.2 Accuracy rate

Table 3, shows the comparative analysis and the performance of the five methods using the CNN network type. The design model implemented uses FER2013 as the database. Out of all the facial expressions trained, the most prominent of them was the angry and the happy facial expressions with the angry expression showing a promising – 70% accuracy and the happy expression showing an accuracy of 100%.

The algorithm used to process the images is FaceEX and it is working on the base principle of Viola-Joneses. Viola-Jones is the first object detection algorithm to run in real-time and is widely used for face detection.

It is a reliable algorithm and comes with a ready-to-use image processing software package known as Open CV. Above mentioned fig.5, shows the facial expression analysis for accuracy and recognition rate in percentage mode. This accuracy value is calculated as follows along with the precision value. The main tools used for evaluation are Recognition rate, Precision, and Accuracy.

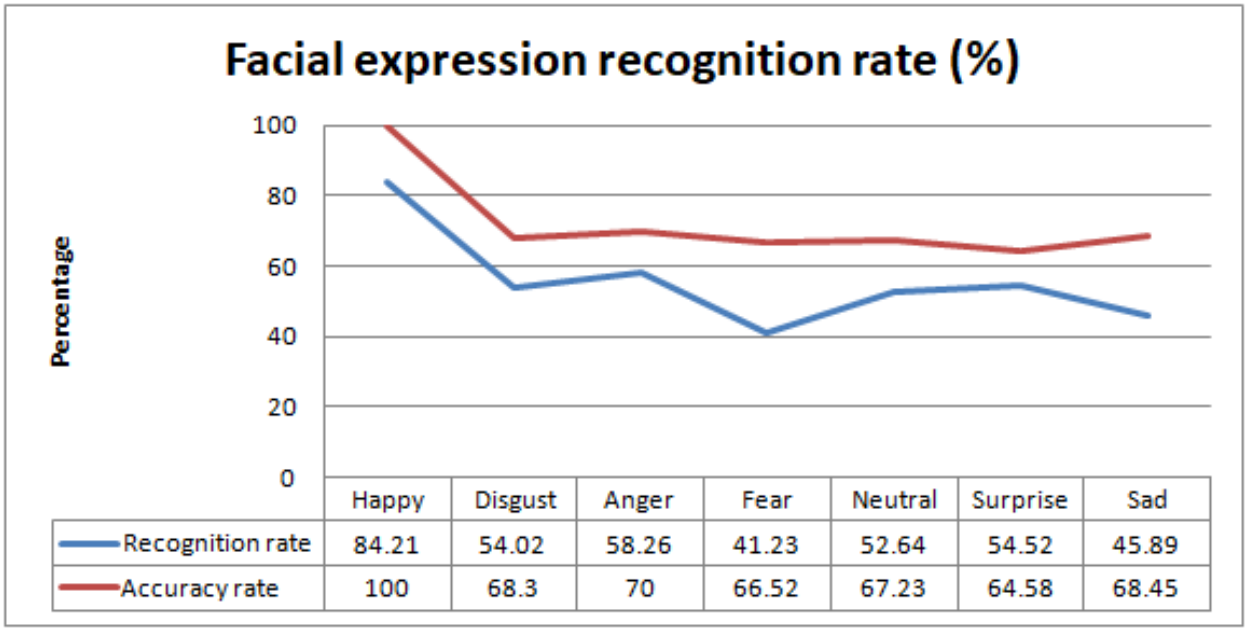

Fig. 6: Shows the accuracy rate and the recognition rate in determining the facial expressions.

6.2.1 Calculation

$$Precision = \frac{True\ positive}{True\ positive + True\ negative}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where,

TP *= True positive,* ***TN*** *= True negative,* ***FP*** *= False positive, and* ***FN*** *= False negative*

Table 3: Depicts the accuracy rate of the facial expressions in the existing model

Facial expression	**Recognition rate (%)**	**Accuracy rate (%)**
Happy	84.21	100
Disgust	54.02	68.3
Anger	58.26	70
Fear	41.23	66.52
Neutral	52.64	67.23
Surprise	54.52	64.58
Sad	45.89	68.45

Table 4: Algorithm used and its recognition rate

Algorithm	Recognition rate (%)
Face_EX	64.23

7. Conclusion

Facial expression recognition is a boon to mankind and we can see many enthusiasts experimenting and making the experience better, be it in medical treatment, or in contributing to the Global Happiness Index. In this experiment, various databases had been explored and at last with comprehensive analysis, Kaggle's and Karolinska Directed Emotional Faces (KDEF) were used as a database.

The challenges faced in this project were at a par level. The most interesting thing to explore was to learn that every human emotion is different from the rest. For example, person A can show him/her being happy in a different way than person B. There is a wide spectrum to it, and detecting each expression, and categorizing them can be hard.

In this experiment, the accuracy rate was satisfactory. The rate of detecting happiness was 100% and that was good, however as the facial expression got more complex, the accuracy rate reduced. At a later stage, the accuracy rate hovered at a rate of 70%, which is neither too bad nor too good.

As mentioned earlier, human expressions can vary a lot, and concretely categorizing them takes a lot of effort and intelligence.

As future enhancements, the facial images can be flattened before training and be converted to ASCII codes. The ASCII codes can then be mapped to the vector points so that facial expression categorization is comprehensive.

We also need to work on image clarity, alignment, illumination, facial rotation. All

of these factors can play a vital role in deviating the accuracy and recognition rate.

8. References

1. B.-F. Wu and C.-H. Lin, “Adaptive feature mapping for customizing deep learning-based facial expression recognition model,” IEEE Access, 2018.
2. B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, “Hierarchical committee of deep convolutional neural networks for robust facial expression recognition,” Journal on Multimodal User Interfaces, vol. 10, no. 2, pp. 173–189, 2016.
3. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning.” in AAAI, vol. 4, 2017, p. 12.
4. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
5. D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in Computer vision and pattern recognition (CVPR), 2012 IEEE conference on. IEEE, 2012, pp. 3642– 3649.
6. F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815– 823.
7. G. Pons and D. Masip, “Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition,” arXiv preprint arXiv:1802.06664, 2018.
8. G. Pons and D. Masip, “Supervised committee of convolutional neural networks in automated facial expression analysis,” IEEE Transactions on Affective Computing, 2017.

9. G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, “Hand-crafted feature guided deep learning for facial expression recognition,” in Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018, pp. 423–430.
10. J. Lu, V. E. Liong, and J. Zhou, “Cost-sensitive local binary feature learning for facial age estimation,” IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5356–5368, 2015.
11. J. Zeng, S. Shan, and X. Chen, “Facial expression recognition with inconsistently annotated datasets,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 222–237.
12. K. Liu, M. Zhang, and Z. Pan, “Facial expression recognition with cnn ensemble,” in Cyberworlds (CW), 2016 International Conference on. IEEE, 2016, pp. 163–166.
13. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multilevel network for saliency prediction,” in Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016, pp. 3488–3493.
14. P. Ekman and E. L. Rosenberg, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.
15. R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, “An all-in-one convolutional neural network for face analysis,” in Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE, 2017, pp. 17–24.
16. S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, “Feature selection mechanism in cnns for facial expression recognition,” in BMVC, 2018.
17. W. Shang, K. Sohn, D. Almeida, and H. Lee, “Understanding and improving convolutional neural networks via concatenated rectified linear units,” in International Conference on Machine Learning, 2016, pp. 2217–2225.

18. Y. Lv, Z. Feng, and C. Xu, “Facial expression recognition via deep learning,” in Smart Computing (SMARTCOMP), 2014 International Conference on. IEEE, 2014, pp. 303–308.
19. Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in European Conference on Computer Vision. Springer, 2016, pp. 499–515.