International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

Facial Expression Recognition using Deep Learning

Suhail Ahmed, S. Ponmaniraj

Abstract: ***Facial expression recognition (FER) is now getting extensively popular because of its ability to predict an unknown data-set, and to its extent with some accuracy. An average human being possesses or shows seven different expressions based on the situation, namely anger, sad, happy, surprise, disgust, neutral and scared. Each individual has a unique way to express the afore-mentioned expressions and hence the term "an unknown data-set". To identify human's present mindset through facial expressions, many data sets are prepared based on face components (such as lips, cheek, nose, eyes and eye brows etc.,) dislocations and elasticity of all the facial parts. Many facial recognition systems are functioning on muscle distribution analysis from the mother image set's pixel parameters. This research paper is going to present about image pre processing, facial expression learning methods, classification methods and implementation of FaceEx algorithm for facial expression analysis through FER2013 CNN data sets and Viola-Jones Principle.***

Keywords: ***Convolutional Neural Network, Deep Belief Network, Facial Expression Recognition, Face Normalization, Gabor Wavelet Form, Local Response Normalization.***

I. INTRODUCTION

Facial expression is a comprehensive tool that distinguishes an individual from another. Although facial expressions vary from person to person but still the underlying feelings that they showcase are the same. Significantamount of studies have been conducted on the topic 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 [6][2]. Same with the case in medical treatment. Altogether a human being 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 the faces based on the emotion shown [19][10].

II. RELATED WORK

There are many studies and research techniques are done for analysing facial expressions. Those techniques have been implemented through geometrical features and location based parameters distances. Geometrical feature points extractions and eigenface values are used to identify facial organs and it's movement analysis.

Revised Manuscript Received on April 21, 2020.
* Correspondence Author
Suhail Ahmed*, UG Student, School of Computing Science and Engineering, Galgotias University, Greeater Noida, bg.kenz@icloud.com
S. Ponmaniraj, Assistant Professor, School of Computing Science and Engineering, GalgotiasUniversity, Greater Noida, ponmaniraj@gmail.com

Roberto Brunelli and *et.al* had an idea about relative position and some other parameters of different features such as muscle distributions by eyes, eyes brows, mouth and chin [11]. They implemented same idea to identify facial organ movements from orign pixel values to the new positions for getting their expressions.

III. PRE-PROCESSING

Prior to the training of deep neural networks to train the model so that it can classify the facial images, we need to list important criteria that an image should fit. For example, facial learning and the results can comprehensively vary based on the image background, contrast, facial rotation (head poses), and image alignment. Pre-processing can help in normalizing the visual appearance and the semantics, which can then help neural networks to learn better.

A. Facealignment

It is a traditional normalizing process that helps eliminate visual aspects that are not required or is irrelevant in deep learning. Below are some well-known approaches and implementations that are widelyused.

a) Holistic [AAM] – Categorized by poor performance and slow speed[15].
b) Part-based [MoT, DRMF] – Better performance as compared to Holistic [9].
c) Cascaded regression [SDM, 3000fps] – Categorized by superior speed and real-time results. The performance too is significantly better as compared to the above-mentioned technology[10].
d) Deep learning [CNN, MTCNN] – Similar in performance to the cascaded regression and also consists of real-time results[18][19].

B. Data augmentation

Any machine learning requires a significant amount of data training and FER is no less. A good amount of training helps in generalizing a given recognition task and we too can be sure about the consistency in performance. Unfortunately, publicly available databases have an insufficient amountof images for training, as a result data-augmentation is an important step for deep Facial Expression Recognition.

C. Facenormalization

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 (facialalignment).

Contrast normalization**:** Contrast and brightness may vary in facial images due to unconstrained backgrounds. Numerous findings

have shown that histogram equalization in addition with contrast normalization results in superior facial recognition. Global Contrast Normalization (GCN), local normalization, histogram normalization has shown consistent results [4][19].

***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.

IV. FACIAL EXPRESSIONLEARNING METHODS

A. Support vector machine(SVM)

A support vector machine (SVM) is used to perform classification of the given images based on its extracted features. Statistical and geometrical/topological features are extracted from images to identify the components which are really used to make up a particular object for analysis purpose. This extractions process involves high tolerance of distortion and variations in 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 methodsperformnonlinear mapping functions [4][15].

B. Deep belief network (DBN)

It is especially beneficial if the facial images are perfectly aligned, by means of size and rotation. Fig. 1, shows that DBN functional structure. In the era of machine learning, DBN has many layers for an image classification kind of problems. Each layer is trained with the help of a greedy layer-wise strategy.

Fig.1 Layers of deep belief network

DBN learns probabilistically when it is trained on a set of examples without supervision. The layers then detectsthe features and the model canfurther be trained with supervision to perform classification. This deep belief networks does fine tuning of the image given as input in it's module. This DBN performs the operations for trim the features through visible layers and "N"number of perceptron layers to segregation of image features.

C. GoogLeNet

GoogLeNet is the award winning concept of ILSVRC 2014 (ImageNet Large Scale Visual Recognition Competition) presented by Prof. Yan LeGun'sLeNet for better image classification and finally it is adopted with google services. When compared with VGGNet, GoogLeNet gives minimum of error rate. It started to work on 1X1 to 5X5 layered structure for fully connected layer. Inception module concept is used in this technique where, different single input image is passing through number of 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 testing used as follows; Multi – Scale and Multi – Crop Testing.

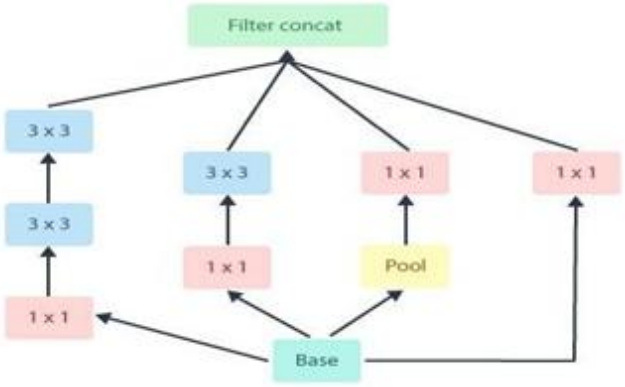

Fig.2 The inception structure of GoogleNet

V. CLASSIFICATION OFFACIAL EXPRESSIONS

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 human being shows seven basic emotions - anger, sad, happy, scared, surprise, disgust and neutral (see Fig. 3). In a broad spectrum, images are organized and classified as one of the above- mentioned categories.

Fig. 3 Categorization of emotions

The question that needs to be asked now is, how does a machine learn to differentiate the images. Unlike the older methods, newer ones focus on the three step approach. The first step is to load a face of a known person. The second step is to load the face of the same person and the last step is to load the face of a different person. The algorithm then looks at the measurements of the face which is completely dependent on the training model itself. It makes small tweaks and adjustments and makes sure that the first image and the second image fed is the same while making sure that the first and the third,

or the second and the third are kept apart.
This process is done millions of times until and unless the result is the same for other facial images as well. After successful classification of the persons, irrespective of the facial expressions, the next step is to classify a person's emotion. This step is harder and complicated as every individual has its own set of emotions. Person 1 may show a different facial expression when happy than Person 2 shows. The process remains the same as that of facial recognition.

VI. PROPOSED METHOD

Holistic [AAM] and Part-based [MoT, DRMF] mechanisms are much slower in speed as compared to Cascaded regression [SDM, 3000fps] and Deep learning [CNN, MTCNN]. These provide faster results without sacrificing speed. It is also characterized by real-time results [15][9].

A. Convolutional neural network (CNN)

Shift or space invariant in the artificial neural networks (SIANN) is the principle concept behind convolutional neural networks (CNN). Video image recognition, image processing, medical image analysis and financial analysis are naming the few applications of CNN. As discussed earlier in GoogLeNet, CNN also working with many layered approach and Convolutional layers, pooling layers and fully connected layers are the three main layers of the CNN for image feature segregation. The Convolutional layer passes three mandatory functions such as local connectivity (To learn neighbouring pixels relationship) for parameter optimization; Pooling layers are functioning to reduce computational cost along with spatial size of the maps [3]. Conversion of 2D feature to map with 1D pixel map feature is done through fully connected layer [7].

$$L_i = SUM_{j \neq yi}\, Max(0, w_j^T x_i - w_j^T x_i + \Delta) \quad (1)$$

The function $f\ (x_i\ w) = wx_i)$ is a model for linear score and in (1), $w_j^T\ and\ x_i$ are the softmax loss function parameters for a score of i-th element is depends on the j-th element. Δ is the classification boundary. His model equation is used to calculate the loss value for processing images.

B. Architecture – VGG-16

VGG-16 is one of the functional and comprehensive model architecture for 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 hyper parameters. 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 px 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 soft max layer. The setup of the Fully Connected layers is the equivalent in all networks. All concealed layers are furnished with the correction non-linearity [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 [7].

Implementation Procedure For Facial Expressions and Analysis
Read the input image
If (Non face image (F_{Nd}) == yes)
Read face(F_d) and non face(F_{Nd}) data sets to identify face regions from kernel lines
Trace the face from background situations
Remove non face regions using $\boldsymbol{f\{\theta\}}$
Calculate distance between (F_d) and (F_{Nd}),
$\boldsymbol{f\{\theta\} = \frac{\theta^T\{[M_x - M_y][M_x - M_y]^T + R_x + R_y\}\theta}{\theta^T R_x \theta}}$ (2)
Where,
$\boldsymbol{M_x}$ & $\boldsymbol{M_y = Mean\ values\ of\ F_d\ and\ F_{Nd}}$
$\boldsymbol{R_x}$ & $\boldsymbol{R_y = Covariance\ values\ of\ F_d\ and\ F_{Nd}}$
Apply coarse region detection to refine the resulting images,
$\boldsymbol{d_q(x,y)^2 = (x-y)^T A (x-y)}$ (3)
Where,
$\boldsymbol{A}$=Adjecency information
Analysis of face and head pose estimation
$\boldsymbol{P_t = \frac{\sum_{i=1}^{N} S_t^{(i)} \pi_t^{(i)} W_t^{(i)}}{\sum_{i=1}^{N} \pi_t^{(i)} W_t^{(i)}}}$ (4)
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)
$\boldsymbol{r^{jkl} = \frac{d^{jk}}{d^{kl}},\ s^{jkl} = \theta(\overrightarrow{jk}, \overrightarrow{kl})}$ (5)

Where,

$\overrightarrow{jk}$ and $\overrightarrow{kl}$ = Angle formed by two different vectors

Gabor wavelets for whole face or specific regions to extract feature vector

$$\psi_\theta(b_x, b_y, x, y, x_0, y_0) = \frac{1}{\sqrt{b_x b_y}} \psi\left(\frac{x - x_0}{b_x} + \frac{y - y_0}{b_y}\right) \quad (6)$$

Where,

ψ_θ = *Mother wavelet of image plane*

b_x, b_y, x, y, x_0, y_0 = *Shifting and scaling parameters at spatia shifting*

Facial expression changes - identified as facial action units or prototypic emotional expressions

Elastic model,

$$P_{opt} = \arg \min_{P, P \epsilon Win} \left\| \sum_i \overrightarrow{F_{Bi}}(P) + \overrightarrow{F_{A'}}(P) \right\| \quad (7)$$

Where,

Implementation Procedure - Continuation

P_{opt} = *Optimal Position*

A' = Global and Local locations

Distance based metric,

$$DB\left(I_{target\_exp}(\lambda_k)\right) = e^{\|\vec{E}_{target\_exp}(\lambda_k) - \vec{E}_{source\_exp}\|} + \omega_{db} \cdot e^{\|\vec{E}_{target\_exp}(\lambda_k) - \vec{E}_{target\_gt}\|} \quad (8)$$

Muscle distribution based model,

$$IL^{MD}(u, v) = M(u, v) IL(u, v) \quad (9)$$

Where,

$IL(u, v)$ = *Illumination muscle details*

$M(u, v)$ = *Mask pixel details*

VII. EXPERIMENTAL 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.

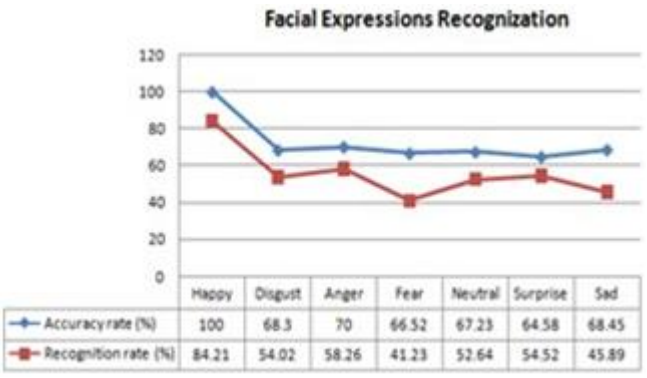

Fig. 4 Architecture Depiction Of VGG-16

Fig. 4 shows that architecture of VGG – 16, which explains pools of data set analysis and classification through multiplelayers.

A. Setup

The training data model was carried out on a NVIDIA GeForce GTX 1080 Ti with 4GB of memory. The estimated training time taken was approximately 15 minutes and a loss of 0.8% was seen. All fivemethods (insection6) were trained using the same hardware and software integration.

B. Result

Table – I, 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%.

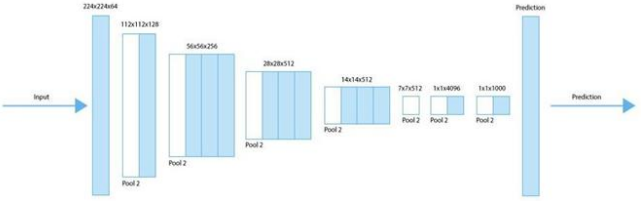

Fig. 5 Facial Expression Analysis

The algorithm used to process the images is FaceEX and it is working on the base principle of Viola – Jones..

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 asOpenCV.
Above mentioned fig.5, shows that facial expression analysis for Accuracy and recognition rate in percentage mode. This accuracy value is calculated as follows along with precision value. The main tools used for evaluation are Recognition rate, Precision, and Accuracy.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (10)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

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

Table – I: Depicts the accuracy rate of the facial expressions in 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 – II. Algorithm used and its recognition rate

Algorithm	Recognition rate (%)
Face_EX	64.23

VIII. CONCLUSION

Facial expression recognition is a boon to mankind and we can see many enthusiasts experimenting and make 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.

REFERENCES

1. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016, pp. 3488–3493.
2. B.-F. Wu and C.-H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," IEEE Access, 2018.
3. 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.
4. 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.
5. 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.
6. 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.
7. S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in cnns for facial expression recognition," in BMVC, 2018.
8. 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.
9. 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.
10. 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.
11. 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.
12. G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," IEEE Transactions on Affective Computing, 2017.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
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. 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.

AUTHOR'S PROFILE

Mr Suhail Ahmed pursuing his Bachelor of Technology in the stream of Computer Science and Engineering from Galgotia's Univeristy, Greater Noida. As a part time work, he spent his valuable time at Anjuman Islamia(Charitable society since 1879), Kalimpong since 2020 for exploring computer knowledge to the juniors.His interest research area focuses on image processing, segmentation and feature extraction. On the basis of his area of interest, he started to do research on facial expression analysis under his faculties guidance and supervision.

S. Ponmaniraj, is an Assistant Professor in School of Computing Science and Engineering at Galgotias University, Greater Noida. He got his Master Degree from Compuer Science and Enginerring department at Sri Muthukumaran Institute of Technology, affiliated to Anna University, Chennai. Currently he is perusing Ph.D. in the area of "Network Security" for the concept of "Intrusion Detection Mechanism Based on Web Images and Sub Links" . He has 12 years of academic experiences from various educational institutions.