Smart Facial Emotion Recognition along with Gender and Age Factor Estimation Through KNN, SVM, CNN and VGG – 16

Surya Teja Chavali[1], Charan Tej Kandavalli[1], Sugash T M[1], Subramani R[2]
[1]Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.
[2]Department of Mathematics, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.
bl.en.u4aie19014@bl.students.amrita.edu, bl.en.u4aie19013@bl.students.amrita.edu, bl.en.u4aie19062@bl.students.amrita.edu, r_subramani@blr.amrita.edu

***Abstract*— Human-Computer Interaction (HCI) in an intelligent way, which aims at creating scalable and flexible solutions. Big tech firms and businesses believe in the success of HCI as it allows them to profit from on-demand technology and infrastructure for information-centric applications without having to use public clouds. This paper proposes a system of recognizing the emotional condition, along with the age and gender factors of humans. The model can currently detect 7 emotions based on the facial data of a person - (Anger, Disgust, Happy, Fear, Sad, Surprise, and Neutral state). The proposed system is divided into three segments: a.) Gender Detection b.) Age Detection c.) Emotion Recognition. The initial model is created using 2 algorithms - KNN, and SVM. We have also utilized the architectures of some of the deep learning models such as CNN and VGG - 16 pre-trained models (Transfer Learning). The evaluation metrics show the model performance in regard to the accuracy of the Recognition system. Future enhancements of this work can include the deployment of the DL and ML model onto an android or a wearable device such as a smartphone or a watch.**

Keywords — HCI, Emotion recognition, Gender and Age recognition, SVM, KNN, CNN, VGG - 16

I. Introduction

Face recognition and Face differentiation have always been innate capacities in humans. Computers can now perform the same thing. This brings up a slew of human computers. Face detection and recognition can be used to improve access and security, process payments without actual cards, identify criminals, and provide individualized healthcare and other services. Face recognition has been the topic of extensive research in recent decades, owing to rising security needs and its potential in commercial and law enforcement applications. Face recognition has a variety of applications, including Human-Computer Interaction, Content-based coding of image video, Biometric Analysis, and so on. It has the potential to be a non-intrusive biometric identification method that does not require human collaboration. In this paper, well-known machine learning algorithms like K-nearest neighbours (KNN), Support Vector Machine (SVM) have been used, and also popular Depp Learning techniques like CNN and VGG-16 with Transfer Learning have also been used.

Age Detection, provide additional data about an individual's identification. Being able to distinguish people based on their age might be beneficial in product sales. You might personalise adverts in stores based on who is watching or exiting the store. It could also be used to reduce the number of elements studied in a database if they are first filtered by age, cutting a search time from minutes to seconds or even minutes. In the forensic sector, where a suspect's description is frequently all that is given, an age range is usually the only data that is always accessible, which would be very valuable in narrowing down the list of candidates to be compared with the sketch to get more exact answers. The same Machine Learning and Deep Learning techniques that were used for Facial recognition have also been used for Age Detection

Many studies on human-machine emotional communication have been identified as one of the most critical issues in human-computer interaction. As part of human-computer interactions, the emotional relationship between humans and machines has gotten a lot of attention. And, among humans, facial expressions are the most potent and natural form of communication. Significant head motion, temporal patterns, and partial occlusions are common features of spontaneous interactions' facial expressions. However, utilizing facial expression to characterize human emotional states is problematic since these facial expression traits are sensitive to lighting circumstances, dynamical head motion, and other factors. It's also difficult to categorize a specific feeling because they're so diverse. People can develop simulations of facial emotion that are characteristic of each unique emotion in reaction to similar facial expressions that occur in response to certain emotion eliciting circumstances. For emotion recognition also, previously considered Machine Learning and Deep Learning models were considered

There are a variety of real-life applications for Emotion recognition, some of them are:

- Security Systems
- Interactive Computer Simulations/designs
- Psychology and Computer Vision
- Driver Fatigue Monitoring
- Video Games

The structuring of this paper is as follows: The second section talks about the datasets used and the methodologies incorporated in the implementation part. Coming on to section three, we analyze different results obtained from the model executions. Following this would be the conclusion and references part.

II. Materials and System Design

A. Facial Emotion Dataset

There are multiple international datasets utilized in different research works on the topic of emotion recognition. We have used one such dataset known as the FER – 2013 dataset which can be found on the Kaggle repository. Briefing about it, it consists of 7 different folders namely – Angry, Disgust, Fear, Happy, Natural, Sad, and Surprise.

Each of these categories consists of nearly a thousand 48 x 48 images. These faces were automatically recorded such that they are almost centered in each picture and take up around the same amount of area.

Figure – II.a. A glimpse of the emotion dataset

B. *Age and Gender Dataset*

For this particular estimation part of the model, we have considered the CelebA dataset which can be found on the Kaggle Repository. Research scientists at MMLAB, The Chinese University of Hong Kong, originally contributed in the development of this dataset.

The CelebFaces Attributes Dataset (CelebA) is a large collection of good-quality celebrity faces. Here, the data is split into 900 male and 900 female photos as the training category. For the testing phase, a separate directory of 170 female and 170 male photos have been provided.

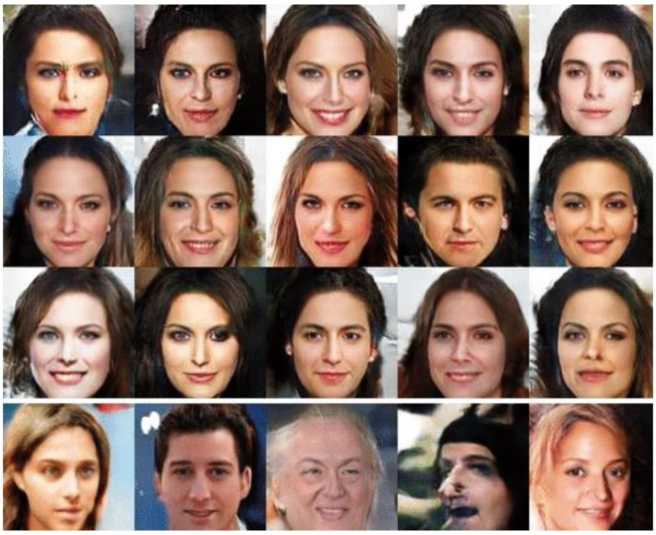

Figure – II.b. A sample of the CelebA dataset

C. *Design and Architecture*

a. *Support Vector Machine (SVM)*

SVM belongs to the group of non–linear classifiers, that uses a kernel transformation matrix to convert the input feature vectors into a generally higher dimensional feature space. The dividing plane seen between borders of 2 categories is optimally placed to achieve maximum classification. The support vectors cover the plane, reducing the number of references.

For example, we have a set of points x_1, x_2, $x_3, \dots \dots, x_n$, where $x_i \in R^n$. Our objective is to find the equation and model a hyperplane that separates the set points on the 3D plane. If our data points were linearly separable, then there exists a relationship:

$$a_i\ [(p.x_i) + b] \geq 1,$$

Therefore, $(p.x_i)$ will form our hyperplane. This plane is referred to as the diving hyperplane, which is then formulated using an optimal objective function.

In the implementation part of this work, we have incorporated the linear kernel in the SVM classifier while we train the model on our data points.

b. *K – Nearest Nwighbours (KNN)*

The k-NN method is a basic and intuitive classification approach used by researchers to distinguish feature vectors. This classifier relates a newly labelled sample (validation) to the training data and makes a decision based on the results. Class labels are included in the training data set itself and the algorithm discovers the k i.e. nearest neighbourhood in the learning data set for a given value from the data set.

The algorithm then allocates a class that is frequently seen in the area. The algorithm can be summariszed as follows:

- A training data set, which is a combination of input and class variables, is used to accomplish the k-nearest neighbour classification.
- Then assess test data that only comprises input variables to the reference set.
- By analyzing closest neighbour points, K-NN operates on k patterns, where the proximity of unknown k decides its category.
- A majority voting mechanism is employed, with each class receiving one point for each instance in the surrounding samples.

In this work, the value of k has been considered as 8 in case of Gender and Age estimation and 5 in the case of Emotion recognition and the optimal value of k has been decided based on the performance of the algorithm.

c. *Convolutional Neural Network (CNN)*

A convolutional neural network (CNN) is a type of artificial neural network (ANN) used in deep learning to evaluate visual information. CNN's (Shift Invariant or Space Invariant Artificial Neural Networks) are built on the shared-weight architecture of convolution kernels or filters that slide along input features and create translation-equivariant outputs known as feature maps. Surprisingly, most convolutional neural networks are only equivariant under translation, rather than invariant. Image and video recognition, recommender systems, picture classification, image segmentation, medical image analysis, natural language processing, brain-computer interfaces, and financial time series are only a few of the applications

An input, hidden, and output layer make up a convolutional neural network. Any middle layers in a feed-forward neural network are referred to as hidden since the activation function and final convolution mask their inputs and outputs. The hidden layers of the artificial neural network include convolutional layers.

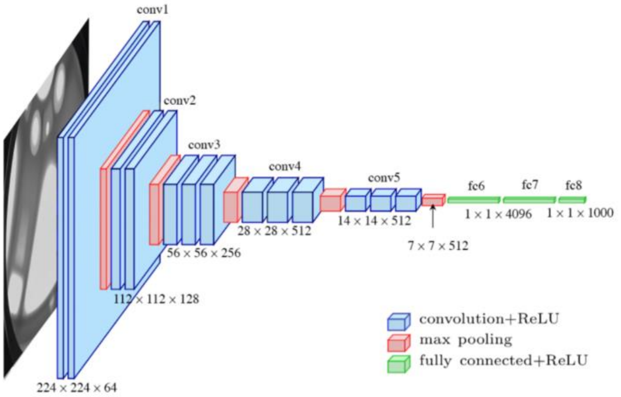

Figure – II.c. Architecture of a traditional CNN model

$$\text{(No. of inputs)} * \text{(Input Height)} * \text{(Input Width)} \\ * \text{(No . of inputs channels)}$$

A convolutional operation is usually carried out by laying the flipped filter over the image and then finding the sum of dot products of the corresponding elements of the filter and the image. A feature map is built of out the above process and the relu activation function is used to discard all of the negative values. The maximum value of each local cluster of neurons in the feature map is used in max pooling, while the average value is used in average pooling. In a neural network, each neuron computes an output value by applying a specified function to the input values received from the previous layer's receptive field. A vector of weights and a bias decide the function that is applied to the input values (typically real numbers). Iteratively modifying these biases and weights is what learning is all about.

d. VGG – 16 (Transfer Learning)

VGG16 is a CNN (Convolutional Neural Network) that is widely regarded as one of the best computer vision models available today. The creators of this model analyzed the networks and enhanced the depth using an architecture with very small (3, 3) convolution filters, which outperformed previous-art setups significantly. They increased the depth to 16–19 weight layers, resulting in 138 trainable parameters.

The 16 in VGG16 stands for 16 weighted layers. VGG16 comprises thirteen convolutional layers, five Max Pooling layers, and three Dense layers, for a total of twenty-one layers, but only sixteen weight layers, or learnable parameters layers. VGG16 uses a 224, 244 input tensor with three RGB channels. The most distinctive feature of VGG – 16 is that, rather than having a huge number of hyper-parameters, they focused on having 3x3 filter convolution layers with stride 1 and always used the same padding and max pool layer of 2 x 2 filter stride 2.

The convolution and maximum pool layers are organized similarly throughout the architecture. The Conv-1 Layer has 64 filters, the Conv-2 Layer has 128 filters, the Conv-3 Layer has 256 filters, and the Conv 4 and Conv 5 Layers have 512 filters. Following a stack of convolutional layers, three Fully-Connected (FC) layers are added: the first two have 4096 channels apiece, while the third performs 1000-way ILSVRC classification and so has 1000 channels (one for each class) and the soft-max layer is the final layer.

However, since the model has been trained on the dataset known as 'ImageNet', and the last layer where the classification has been done has many different neurons because of the number of classes the ImageNet dataset contains. Therefore, we have removed the last dense layer (ANN) part and added extra hidden layers for our dataset, and lastly added a dense layer with 4 neurons for our classification task. The above process is referred to as Transfer Learning.

III. Experimental Results and Analyses

We have performed the training part and validation part of the samples on Jupyter Notebook. Eight different notebooks were developed for Gender and Age recognition as one set and Facial Emotion recognition as another set, consisting of all the 4 algorithms in each set. Following are the results obtained after all the execution is completed.

TABLE I
ACCURACIES ACHIEVED BY THE MODELS

Method	Emotion	Gender	Age
SVM	36.074766 %	82.5 %	66.789 %
KNN	34.018691 %	76.25 %	60.086 %
CNN	84.63 %	90.000 %	91.17 %
VGG – 16	97.834 %	98.246 %	95.31 %

We can clearly observe the robustness of deep learning models over the traditional machine learning models in terms of their performance. Although VGG - 16 looks promising, SVM and KNN have shown good performance as they are one of the basic classifiers available.

Figure IV.a and IV.b show the Training vs Validation parameters across all epochs for the CNN model for Emotion Recognition. As we observe closely that although Training and Validation accuracies improve as the epochs keep increasing, the loss values for training and validation decrease and then remain constant for the rest of the training.

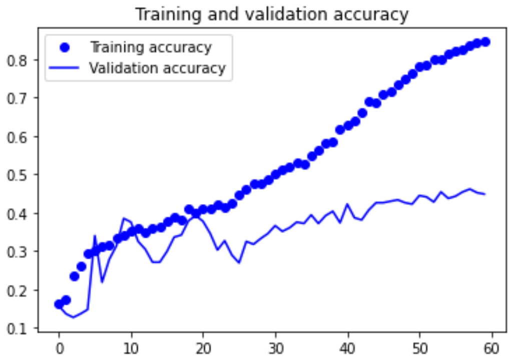

Figure – III.a. Training (vs) Validation Accuracies of the CNN model for Emotion Recognition

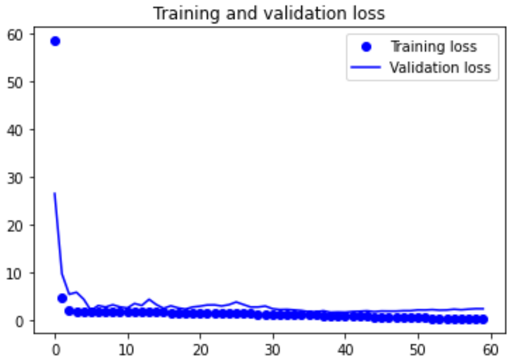

Figure – III.a. Training (vs) Validation Losses of the CNN model for Emotion Recognition

IV. Conclusion

Determining the age and gender of people we meet on a regular basis plays a significant function in our social lives as well. Many intelligent devices, such as visual surveillance, clinical diagnosis, and product marketing, now require persons to be assessed using their facial photos. We have successfully implemented Facial Emotion Recognition long with Gender and Age estimation, further improved it, and achieved excellent accuracy for VGG – 16 model. In conclusion, we hope that.

The current work is still under development, i.e., anyone who would like to extend this paper can implement several different pre-trained models such as AlexNet, ResNet - 50, GoogleNet, etc., and develop an application out of it.

References

[1] X. Zhao, S. Zhang and B. Lei, “Facial expression recognition based on local binary patterns and local fisher discriminant analysis,” WSEAS Transactions on Signal Processing, vol. 8, no. 1. pp.21-31, 2012

[2] M. Pourebadi and M. Pourebadi, “MLP neural network-based approach for facial expression analysis,” 2016.

[3] Bradski G., “The OpenCV Library,” Dr. Dobb’s Journal of Software Tools, 2010.

[4] Boureau, Y-Lan & Ponce, J & Lecun, Yann, “A Theoretical Analysis of Feature Pooling in Visual Recognition,” ICML 2010 - Proceedings, 27th International Conference on Machine Learning, 111-118, 2010..

[5] R. R. Atallah, A. Kamsin, M. A. Ismail, S. A. Abdelrahman and S. Zerdoumi, "Face Recognition and Age Estimation Implications of Changes in Facial Features: A Critical Review Study," in IEEE Access, vol. 6, pp. 28290-28304, 2018, doi: 10.1109/ACCESS.2018.2836924.

[6] A. Garain, B. Ray, P. K. Singh, A. Ahmadian, N. Senu and R. Sarkar, "GRA_Net: A Deep Learning Model for Classification of Age and Gender From Facial Images," in IEEE Access, vol. 9, pp. 85672-85689, 2021, doi: 10.1109/ACCESS.2021.3085971.

[7] G. Ozbulak, Y. Aytar and H. K. Ekenel, "How Transferable Are CNN-Based Features for Age and Gender Classification?," 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), 2016, pp. 1-6, doi: 10.1109/BIOSIG.2016.7736925.

[8] M. Uricár, R. Timofte, R. Rothe, J. Matas and L. Van Gool, "Structured Output SVM Prediction of Apparent Age, Gender and Smile from Deep Features," 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016, pp. 730-738, doi: 10.1109/CVPRW.2016.96.

[9] UKEssays. (November 2018). Age and Gender Prediction using Convolutional Neural Network. Retrieved from https://www.ukessays.com/assignments/predicting-age-and-gender-from-facial-images-4366.php?vref=1

[10] Muhammad Firdaus Mustapha, Nur Maisarah Mohamad, Ghazali Osman, & Siti Haslini Ab Hamid (2021). Age Group Classification using Convolutional Neural Network (CNN). Journal of Physics: Conference Series, 2084(1), 012028.

[11] Amit Dhomne, Ranjit Kumar, & Vijay Bhan (2018). Gender Recognition Through Face Using Deep Learning. Procedia Computer Science, 132, 2-10.

[12] P. Smith and C. Chen, "Transfer Learning with Deep CNNs for Gender Recognition and Age Estimation," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2564-2571, doi: 10.1109/BigData.2018.8621891.

[13] Linnan Zhu, Keze Wang, Liang Lin and Lei Zhang, "Learning a lightweight deep convolutional network for joint age and gender recognition," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3282-3287, doi: 10.1109/ICPR.2016.7900141.

[14] Gil Levi, & Tal Hassner (2015). Age and Gender Classification Using Convolutional Neural Networks. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops.

[15] https://doi.org/10.48550/arXiv.1811.07344