Facial Expression Recognition and Recommendations Using Deep Neural Network with Transfer Learning

Narayana Darapaneni
Director - AIML
Great Learning/Northwestern University
Illinois, USA
darapaneni@gmail.com

Rahul Choubey
Student - AIML
Great Learning
Pune, India
r.choubey37@gmail.com

Pratik Salvi
Student - AIML
Great Learning
Pune, India
pratiksalvi.diablo@gmail.com

Ankur Pathak
Student - AIML
Great Learning
Pune, India
ankur.x.pathak@gmail.com

Sajal Suryavanshi
Student - AIML
Great Learning
Pune, India
sajalsuryavanshi18@gmail.com

Anwesh Reddy Paduri
Research Assistant - AIML
Great Learning
Pune, India
anwesh@greatlearning.in

***Abstract*—This study is an attempt to understand and address the mental health issue, of working professionals through facial expression recognition. As a society, we are all currently talking about ways as to how a person who is suffering from any emotional issue can adopt certain ways to come out of a specific circumstance and how we as a society can support such people in these situations.**

Our endeavor is to work on a way where the identification of such persons who are going through a difficult phase in their life can be performed. It is not always evident that a person going through a tough phase may open up about their feelings to people around them and hence making use of AI/ML to identify a person's emotion through their facial expressions captured over a span of time thereby recommending them some activities, thoughts which can help them in getting over their emotions when they are sad, fearful or else will address the problem to some extent.

Keywords—Facial Expression Recognition, Deep Neural network, transfer learning, recommendations, [1] VGG16

I. INTRODUCTION

Face is one of the most important means of human communication [20]. It plays a central role in all social interactions [21]. Facial expressions are non-verbal clues to emotions. Indeed, some facial muscles are specifically associated with certain emotional states and allow, [22]according to Ekman the expression of primary emotions (Sadness, Anger, Fear, Joy, Disgust and Surprise). These external signals express the internal emotional state of an individual, and therefore his intentions. [18] In fact, 7% of the communication relies on verbal interaction, 38% represent tone and sound of voice, 55% are articulated around gestures and expressions of the face according to Meharabian. Automatic recognition of facial expressions is an interesting problem which finds its interest in several fields such as eLearning and [23] affective computing. When designing an automatic facial expression recognition system, three problems are considered: face detection, facial feature extraction, and classification of expressions. First, face acquisition is a processing stage to automatically locate the face region in the input images. The next step is to extract and represent facial changes caused by facial expressions. Finally, the classification task allows to infer the facial expressions. In this approach of ours, we have restricted our solution on facial feature extraction & classification of expressions.

According [2] to emotion theorist and psychologist, various emotion can be categorized into six different fundamental emotions to complicated emotions which are originated from different culture with. From several models that were developed in the field of emotion detection, two models have hold the command in this research domain: Ekman's fundamental set of emotions [3] and russell's circumflex representation of influence. Ekman's and Freisen in 1971 put forward six basic emotions like fear, disgust, happy anger, sad, surprise which are globally identified as facial expression.

II. RELATED WORK

Our proposed solution measures the reliability of the *[24]* Facial Expression Recognition (FER) result. Based on the research we performed on the existing work present on the web related to reliability estimation for helping FER, we found that work performed in this direction is less, although there certainly exists related work that estimates the reliability of classification results. In the existing works, it has been shown that the classification performance can be further improved by exploiting the reliability estimation of the classification probability.

As part of the solution that we implemented, the purpose is to improve the reliability of the classification performed by the

978-1-7281-9656-5/20/$31.00 ©2020 IEEE

model using deep CNN and transfer learning methodology, we have tried to improve the model performance such that the overall runtime involved in performing the classification is minimized so that the model can be easily integrated with the overall solution where the recommendation can also be performed using the classification given by the model.

III. DATA SELECTION

A. Data Approach process

In order to identify the expression on face a neural network model has to be provided with images which have all emotions identifiable clearly, because model will learn on the basis of these images only and perform its predictions on not seen images efficiently. To achieve this motive, a thorough research was performed on available datasets on open platforms, shortlisted datasets as below: eNTERFACE-05 [4] : audio-vision emotion database (42 subjects, 14 different nationalities, mix of gender, has images with glasses, beard, video/audio sequences, has 6 emotions) Emotiw [5] : contains 2 sub databases, acted facial expression in the wild (AFEW) and static facial expression in the wild (SFEW) Extended CK+ [6] : contains grey images with a good mix of age, ethnicity, contains labeling with 8 emotions JAFFE [7]: The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects

A. Finalization of dataset

To solve the purpose as mentioned in the previous section, a decision was required to pick the dataset which will be used to train the model. After much discussion and introspection, JAFFE image dataset was finalized to be used to train our model for emotion recognition. The decision was derived by taking into consideration following points:

- All the images are 256*256 Grayscale
- Exposure Level (Contrast, Brightness, Background etc.) is consistent and are face focused.
- Expression labeling of the images is clearly divided into separate folders, there are 6 expressions captured in the images which are anger, disgust, fear, happy, sadness, surprise.
- Total number of images in the dataset are around 213, these are segregated into train and test folders such that each folder has subset of 6 subfolders containing images specific to 6 expressions respectively (Surprise, Anger, Disgust, Happy, Sad, Fear)

B. Data Preprocessing and data augmentation

In order to increase the volume of training data for the model, image augmentation [16][17] technique was used via a number of random transformations, so that the model never get to see the same picture twice, this helps prevent over fitting and helps the model generalize better.

In tensorflow.keras this can be done via the tf.keras.preprocessing.image.ImageDataGenerator class. This class allows you to: configure random transformations and normalization operations to be done on your image data during training instantiate generators of augmented image batches (and their labels) via. flow (data, labels) or. flow_from_directory(directory). These generators can then be used with the tensorflow.keras model methods that accept data generators as inputs, fit_generator, evaluate_generator and predict_generator.

Below example shows the usage of these libraries:

```
datagen = ImageDataGenerator(rotation_range=20,
zoom_range=0.15, width_shift_range=0.2,
height_shift_range=0.2, shear_range=0.15,
horizontal_flip=True, fill_mode="nearest")
```

where:
-rotation_range is a value in degrees (0-180), a range within which to randomly rotate pictures
-width_shift and height_shift are ranges (as a fraction of total width or height) within which to randomly translate pictures vertically or horizontally
-rescale is a value by which we will multiply the data before any other processing. Our original images consist in RGB coefficients in the 0-255, but such values would be too high for our models to process (given a typical learning rate), so we target values between 0 and 1 instead by scaling with a 1/255. factor.
-shear_range is for randomly applying shearing transformations
-zoom_range is for randomly zooming inside pictures
-horizontal_flip is for randomly flipping half of the images horizontally --relevant when there are no assumptions of horizontal assymetry (e.g. real-world pictures).
-fill_mode is the strategy used for filling in newly created pixels, which can appear after a rotation or a width/height shift.

IV. METHOD

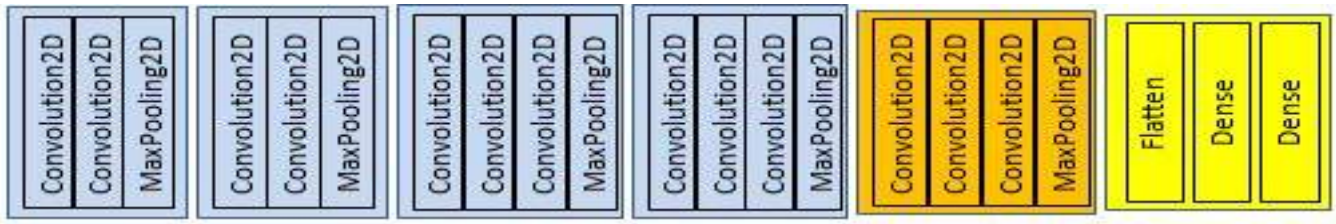
Summary: To solve the purpose, we employed a deep neural network with transfer learning approach to extract bottle features from the input images and saving these features. At later stage a network of fully connected layers is used where the bottle features are loaded back and images are then passed to the model for prediction.

Overview:

1.) Data pre-processing: As part of pre-processing, we performed dimension Fixing(rescaling) as well as to increase the data volume for our model to train on we employed image augmentation on the images before they are fed to the neural network for prediction

2.) Using neural network: We employed deep neural network with transfer learning to extract the bottleneck features from JAFFE images dataset and fed these features to a set of fully connected layers to predict the facial emotions of these images. Finally, complete content and organizational editing before formatting.

3.) Using Transfer Learning[8]: We have used transfer learning as follows:



1. Reused VGG16 Model.
2. Loaded with Imagenet[10] weights.
3. Extracted bottleneck feature from it for our testing scenario.
4. Tuned our model with bottleneck weights to achieve higher accuracy.

V. STEP-BY-STEP WALK THROUGH OF THE SOLUTION

Below steps draft a detail workflow of our model to predict emotions in the input images:

- Read images data using OpenCV [18]libraries for visualizing the dataset
- As per Fig 1. Applied image augmentation on images stored in structured training and validation folders
- Loaded VGG16 model with ImageNet weights
- Extracted bottleneck features of the images using VGG16
- Saved the bottleneck features from the VGG16 model
- Trained a small network using the saved bottleneck features for classification
- Performed model compilation using check point to extract the best accuracy and save the model
- Used VGG16 model and a sequential model of fully connected layers to make predictions as in Fig 2.
- Run the model to predict facial expression of the image both inside and outside of the dataset

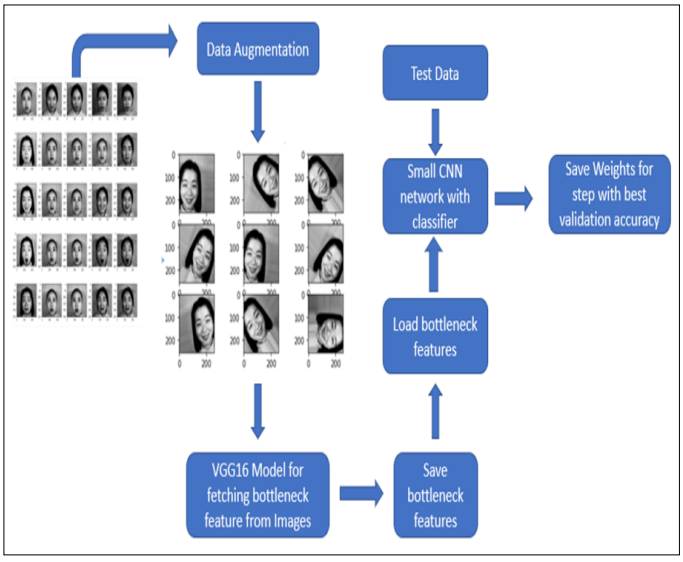


Fig 1. Solution flow chart

VI. IMPLEMENTATION AND RESULTS

Evaluation Criteria/Model Evaluation

1.) Performance of two prominent architecture VGG16 andInceptionV3[9] with different configuration of trainable and non-trainable layers, and multiple optimizers, were evaluated. Tuning the layers with different combination to analyze and evaluate for best result were performed. Below are the combinations of layer tuning done on VGG 16 architecture:

i.) Last 5 layers trained
ii.) Last 3 layers trained
iii.) Last 1 layer trained
iv.) All layers trained

2.) This was done to ascertain how many layers are actually being used on an input image to read the most prominent or specific class related feature so that it would be able to differentiate and classify emotion. Below are the 4 combinations of layer tuning done on inceptionV3 architecture:

i.) Last 5 layers trained
ii.) Last 3 layers trained
iii.) Last 1 layer trained
iv.) All layers trained

3.) Table 1. enumerates different model configuration, architecture trained on and its detailed evaluation metric (Precision, Sensitivity and F1 score):

Model Configuration	Trainable parameters	Optimizer	Precision	Recall	F1 Score
VGG16 : Feature Extractor	All Layers trainable	Adam	91	90	90
VGG16 : Feature Extractor	Freezed 5 layers from the last	Adam	94	93	93
	Freezed 4 layers from the last	Adam	90	88	88
	Freezed 3 layers from the last	Adam	88	86	86
	Freezed 2 layers from the last	Adam	92	90	90
	Freezed 1 layers from the last	Adam	88	86	86

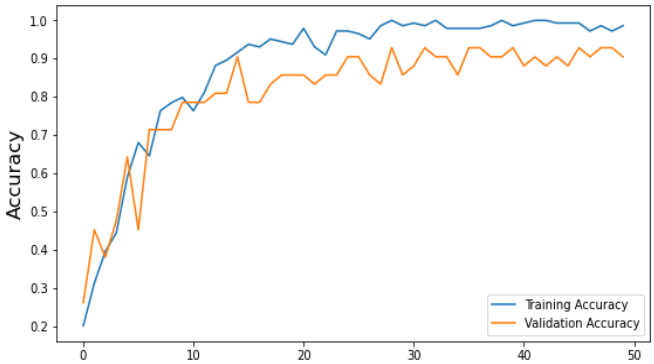
Table 1. Various model configurations and performance

4.) Among all the different parameter combinations of experiments across 2 different architectures, the best performing model with its results are listed in Table 2:

Model Configuration	Trainable parameters	Optimizer	Precision	recall	F1 Score	Accuracy
VGG16 : Feature Extractor	All Layers trainable	Adam	91	90	90	4s 159ms/step - loss: 0.1141 - accuracy: 0.9792 - val_loss: 0.3404 - val_accuracy: 0.9286
VGG16 : Feature Extractor	Freezed 5 layers from the last	Adam	94	93	93	4s 146ms/step - loss: 0.0881 - accuracy: 0.9931 - val_loss: 0.3599 - val_accuracy: 0.9286

Table 2. Best model configuration

5.) Having experimented with multiple optimizer functions [15] like [11]Adam, RMSprop (These are known as most prominent optimizer to be used in most cases), the study deduces that with RMSprop as an optimizer on VGG 16 architecture with last 3 layers trained the model accuracy is90% %(as shown in A. Fig2.). The model gives best accuracy with Adam as an optimizer.



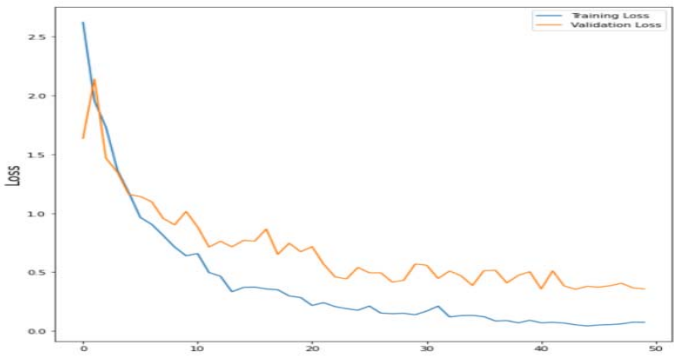
A. Fig2.Representation of Accuracy matrix

A. Evaluation Matrices

1.) Loss Function

The cost or loss function has an important job in that it must faithfully distill all aspects of the model down into a single number in such a way that improvements in that number are a sign of a better model. In calculating the error of the model during the optimization process, a loss function must be chosen. This can be a challenging problem as the function must capture the properties of the problem and be motivated by concerns that are important to the project and stakeholders.

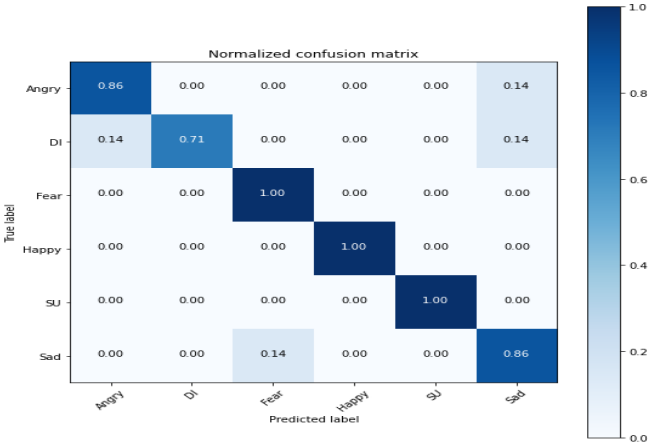
To establish the performance and how close the model is predicting outputs that are as close/same as the actual emotions, a loss function graph was prepared as shown in B.Fig.2:



B.Fig.1. Representation of loss function

2.) Confusion matrix

In order to evaluate the classification performance of the model another effective way is to plot confusion matrix. Using confusion matrix [14] helps to better understand the areas where a classification model is making the errors, more so when the model is built to perform classification on more multiple classes.



B.Fig.2. Heat map representation

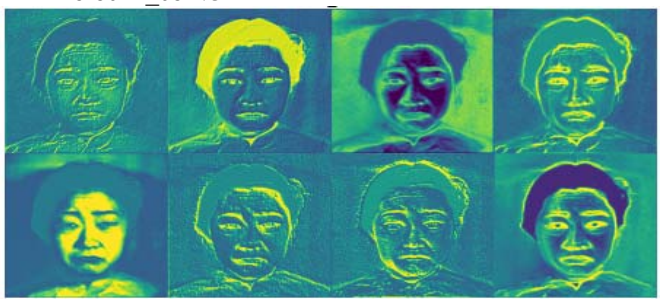
VII. ANALYSING FEATURE MAPS

Deep learning neural network models are generally uninterpretable models as one cannot explain its predictions, although they can make useful and skillful predictions. This is the reason all the deep learning models are considered as black box models where one cannot list any quantifiable metrics to validate model predictions. But what we know is how input features are read by a convnet filter operating on the input. Generally multiple filters do a convolutional operation on the input and as a result we get a feature map. The feature maps [13] of a CNN [12] capture the result of applying the filters to an input image i.e. at each layer, the feature map is the output of that layer.

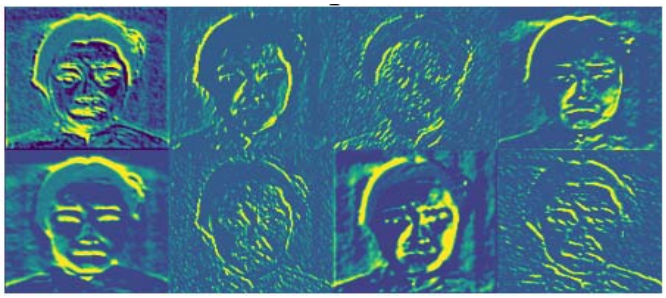
The reason for visualizing a feature map for a specific input image is to try to gain some understanding of what features our CNN detects. Let's try visualizing the feature maps by passing on a set of 6 images from the image data pool and visualize the feature maps extracted from different layers in VGG16 and try to correlate with the model performance. For our visualization purpose we can consider all the conv layers before max-pooling to understand what feature map has been read by each of those layers as in the max-pooling we decrease the spatialsize of feature maps for representation.

Blocks/layers considered to visualize the feature maps are as follows:

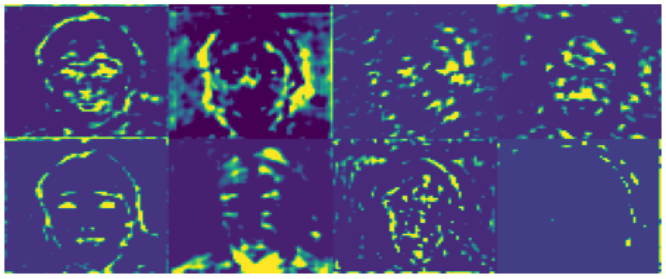
- block1_conv2
- block2_conv2
- block3_conv3
- block4_conv3



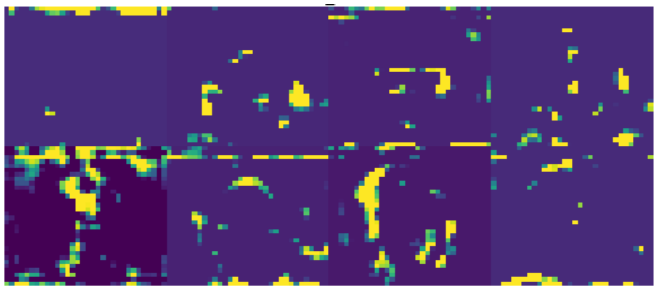
VII.Fig.1 –Features learnt during after block1 completion



VII.Fig.2 –Features learnt during after block2 completion



VII.Fig.3 –Features learnt during after block3 completion



VII.Fig.4 –Features learnt during after block4s completion

As you can see from the above image the first layer feature maps retain the edge and spatial features of the image as we go deeper more granular features are extracted. The block2_conv2 image has the spatial features a lot more visible, whereby we can detect this as a chest x-ray. In block2_conv3 the expression is partially visible but the features are unrecognizable. The reason for that is as we go deeper the filters try to extract deeper features. The observations from these block images are that most of spatial and granular features are detected in first 12 layers (block1_conv2, block2, conv2, block3_conv3) which can give us better predictability in our case of classification of expressions but it was not holding good in predicting all type of images. Since we have most granular features detected only in the last 4 layers our VGG16 model was giving best accuracy when tuned with the last 4 layers as trained.

VIII. Model Validation and Benchmark

Identifying a benchmark is a necessary and mandatory step to know your target model performance, before starting the work on facial expression recognition a study was mandated to be performed on work done in the field of facial expression recognition which could give us a fair and scalable stat that we could aim to achieve during our work in this topic.

A study we performed on articles published under Paper on [19] PERFORMANCE COMPARISONS OF FACIAL EXPRESSION RECOGNITION IN JAFFE DATABASE by Frank Y. Shih and Chao-Fa Chuang it uses two strategies i.e. 2D-LDA and SVM. Recognition rate of this method was around 95% and around 94% by using Cross Validation strategy.

Based on above study, the benchmark laid out for our model was to achieve a test accuracy of around 90%. However major challenge in this work was human emotion detection in the data base we finalized for the same i.e. JAFFE. This dataset has around 200 image and in this paper, we present a systematic comparison of feature extraction and classification methods to the problem of fully automatic recognition of facial expressions and to find the optimal solution to it, for facial emotion detection model this number appeared very less.

As a known fact that training a model on a large number of images could have led us to over fitting of the model thereby getting very less accuracy. We used Transfer learning approach with VGG16 model and the lead to the test accuracy of around 88.26%.

IX. Conclusion and Recommendation

Mental health problems and well-being of employees are a few things which if gone unnoticed, will not only lead to personal downfall but are also directly linked with the efficiency and energy that an employee can give to his/her work. To make this model more effective, the model can be trained with images taken on varied conditions like angles, light conditions, colors, etc. which will enhance the performance of the model so that in future it can work on video clippings which will predict expressions in real time and prepare an integrated system such that the predictions made by the model would be displayed on a screen with best recommendations.

In this research addressed the problem of automatic facial expression recognition to know the mental state of the person based on their facial emotions. We have developed a computer vision system that automatically recognizes facial expressions with subtle differences. Image augmentation was performed on the Jaffe dataset on which a VGG16 model was applied to extract facial motion features. Finally, a neural network system is compiled along with saved extracted weights to make facial expression predictions. Facial expressions of different types, intensities, and durations from a large number of untrained subjects have been tested. The results show that our system has high accuracy in facial expression recognition.

With all the proposed aspects of facial expression recommendation discussed in this paper, we plan to design an end system such that when this system captures images of the

employees over a predefined period of time then based on the classifications the system could propose some fruitful activities to the employees so which are proven to help and get over any critical mood condition. This will help employee to get over that critical mood quickly and will also act as an accompany that understand employee's emotions confidentially and suggests activities which will be confidential to the employee without exposing the data to any third person, this way a reliable and effective system could come into existence over a long period of time. The recommendation could be design on a web page which can be integrated to our model and based on model's output a set of activities come show up to the employee on the web page itself.

REFERENCES

[1] Keras Team, "Keras documentation: VGG16 and VGG19," Keras.io. [Online]. Available: https://keras.io/api/applications/vgg/.

[2] D. Dagar, A. Hudait, H. K. Tripathy, and M. N. Das, "Automatic emotion detection model from facial expression," in 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2016, pp. 77–85.

[3] A. De and A. Saha, "A comparative study on different approaches of real time human emotion recognition based on facial expression detection," in 2015 International Conference on Advances in Computer Engineering and Applications, 2015, pp. 483–487.

[4] O. Martin, I. Kotsia, B. Macq and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006, pp. 8-8, doi: 10.1109/ICDEW.2006.145.

[5] "EmotiW2020," Google.com. [Online]. Available: https://sites.google.com/view/emotiw2020. .

[6] Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/224165246_The_Extended_Cohn-Kanade_Dataset_CK_A_complete_dataset_for_action_unit_and_emotion-specified_expression. .

[7] M. Lyons, M. Kamachi, and J. Gyoba, The Japanese female facial expression (JAFFE) database. 1998.

[8] D. F. T. Orozco, C. Lee, Y. Arabadzhi, and V. K. Gupta, "Transfer learning for Facial Expression Recognition," 2018.

[9] Keras Team, "Keras documentation: InceptionV3," Keras.io. [Online]. Available: https://keras.io/api/applications/inceptionv3/. .

[10] "ImageNet," Image-net.org. [Online]. Available: http://www.image-net.org/. .

[11] Keras Team, "Keras documentation: Adam," Keras.io. [Online]. Available: https://keras.io/api/optimizers/adam/. .

[12] Keras Team, "Keras documentation: Convolution layers," Keras.io. [Online]. Available: https://keras.io/api/layers/convolution_layers/. .

[13] "Confusion matrix — scikit-learn 0.23.2 documentation," Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html. .

[14] E. M. Dogo, O. J. Afolabi, N. I. Nwulu, B. Twala and C. O. Aigbavboa, "A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 92-99, doi: 10.1109/CTEMS.2018.8769211.

[15] P. Pai, "Data Augmentation Techniques in CNN using Tensorflow," Medium, 25-Oct-2017. [Online]. Available: https://medium.com/@prasad.pai/data-augmentation-techniques-in-cnn-using-tensorflow-371ae43d5be9. .

[16] "Data augmentation | TensorFlow Core," Tensorflow.org. [Online]. Available: https://www.tensorflow.org/tutorials/images/data_augmentation.

[17] V. Agarwal, "Complete Image Augmentation in OpenCV | Towards Data Science," Towards Data Science, 13-May-2020. [Online]. Available: https://towardsdatascience.com/complete-image-augmentation-in-opencv-31a6b02694f5.

[18] "'Silent Messages' -- Description and Ordering Information," Kaaj.com. [Online]. Available: http://www.kaaj.com/psych/smorder.html.

[19] F. Y. Shih, C.-F. Chuang, and P. S. P. Wang, "Performance comparisons of facial expression recognition in Jaffe database," Intern. J. Pattern Recognit. Artif. Intell., vol. 22, no. 03, pp. 445–459, 2008.

[20] "Exploring Nonverbal Communication," Ucsc.edu. [Online]. Available: https://nonverbal.ucsc.edu/facerev.html.

[21] L. M. Mayo, J. Lindé, H. Olausson, M. Heilig, and I. Morrison, "Putting a good face on touch: Facial expression reflects the affective valence of caress-like touch across modalities," Biol. Psychol., vol. 137, pp. 83–90, 2018.

[22] B. Neel, "What Are Basic Emotions?," Psychology Today, 07-Jan-2016.

[23] "MIT Media Lab: Affective Computing Group," Mit.edu. [Online]. Available: https://affect.media.mit.edu/. [Accessed: 14-Oct-2020].

[24] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial Expression Recognition: A survey," Symmetry (Basel), vol. 11, no. 10, p. 1189, 2019.