Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 11, June 2020: 823-830

Research Article

Image Processing: Human Facial Expression Identification using Convolutional Neural Networks

Anusha Padala[a], Dr. D. Shravani[b]

[a,b]Department of computer science,
Stanley College of Engineering and Technology for women, Abids, Hyderabad, Telangana, India

Corresponding author: [a]anushapadala4@gmail.com

Abstract

Facial expressions identification is the growing area of interest as it attracts the advancement in technology by focusing on human computer interaction. Many researchers has got their hands on various approaches for automatically generating the expressions of human faces. Emotions and expressions are inter relatable where expressions are facial movement for expressing the emotions of human being. Image processing is the technology which helps in identifying expressions by including factors such as the face detection, the feature extraction and the expressions classification. Two datasets FER-2013 and CK+48 are used in the process of identifying the expressions like sad, fear, happy, angry, surprise, neutral, disgust, contempt. HAAR features based adaboost cascades are used to identify the features of a face which helps in detecting facial points and makes it easy for further process of expression detection. The deep learning technique, convolutional neural network (CNN) is implemented for classification of expressions for prediction

Keywords: Facial expression identification; FER2013; CK+48; HAAR features based adaboost cascades; CNN.

1. Introduction

Emotions are psychological representation of a facial expression. An expression is a representation of the emotional state of human and therefore helps in predicting the human state of mind and further behaviour. The ability to identify facial expressions automatically is the important aspect to non-verbal communication and helps in human-computer interaction using application which enables in the study of Facial expression identification. The expressions which are the most common emotions of human beings at certain situations are like Fear, Happy, Sadness, Anger, Neutral, Disgust and Surprised.

1.1 Background study

FER system predicts the expressions in both static images and also in the real time video. In both the cases it is important to give the image for testing while in video streaming the image is taken from the frame and feed to the model to get the predicted output. Extracting features from input is the first task to be performed in the FER system for detecting emotions. But this is the challenging part as features extraction from an image purely depends on the angle, positions and directions of the face in the image. A good feature representation helps in efficient and effective recognition process. The feature vectors help in to train the classifier and is used to predict and assign the expression labels to the input face.

However, facial expression analysis is a much challenging because of complex and noisy backgrounds and major challenge is due to subject dependence and head-pose which can effect the performance of the FER system

1.2 Motivation and scope

As the time started passing the rapid emerging of technologies like pattern identification and Image processing drew a line of hope for researchers and significantly helped with research works on automatically detecting the facial expressions. According to the research, many efforts were made to identify in images using many optimized techniques. The major motivation was to keep up with the human state of mind and track their reactions and emotions. The main aim is to come up with the solution to the problem of face emotion identification by the means of classification method of trained models**.**

The scope is to design a FER system using deep neural network to classify the basic expressions with a possible maximum accuracy. Facial Expression Identification is applied in various research areas, such as in detecting physiological interaction or mental diseases diagnosis. The common applications are like in various fields like working on Security, Counselling Systems, Lie detection in interrogation, Analysis on attempting the crime etc. A facial expression identification system is an automated system which analyses the facial features from both static input and video.

1.3 Applications

The face expressions identification system can be used widely in many sector in day to day life. The major applications are to keep up with the human emotional state and track their psychological tendencies. Some of the very most common applications are as follows:

- Teaching sector as teaching assistant and make business meeting more effectively
- Airports and Railway stations, to detect the faces and recognise any suspected expression and set an alert
- Educational institutions to check with the student’s expressions while listening to the class.
- Shopping malls and marts, to check with customer’s satisfaction and also security check based on expressions.
- Defence agencies to interrogate the criminal and check with lie detector.

2. Facial Expression Identification system

According to the extensive usage of FEI systems in various fields like crime analysis, deep-fake, Medical analysis, etc., the researchers have considered them from a variety of views. An efficient FEI system plays a crucial role in different fields to accurately characterize the facial geometry in featuring out the facial detection and expression identification.

2.1 Proposed system approach

The major challenges in face and expression identification as mentioned earlier are the variations, subject dependence, head-pose. Every technique is reviewed with respect to the rate of accuracy of FER system with future changes and future works. In order to encounter these challenges an idea to build a correct and reliable FEI system is considered by detecting the face and identify the facial expressions with the help of the trained images using FER-2013 and CK+48. The proposed system focuses on facial detections and classifying the expressions according to the extracted features using the algorithm. The input image is turned to gray scale to avoid noise, then face detection by localizing parts of face like eyes, eyebrows, mouth, cheeks, chin, etc and cropped by focusing on center of image normalizing to a size of 48x48. There after features are been extracted from the image which is pre-processed by the trained convolutional kernels using HAAR features based cascade classifier.

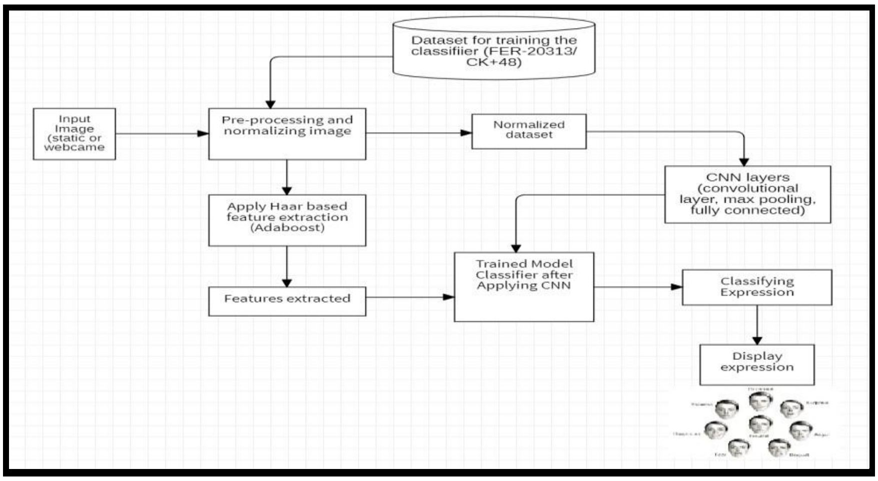

Dataset for training the classifiier (FER-20313/ CK+48)
Input Image (static or webcame
Pre-processing and normalizing image
Normalized dataset
CNN layers (convolutional layer, max pooling, fully connected)
Apply Haar based feature extraction (Adaboost)
Features extracted
Trained Model Classifier after Applying CNN
Classifying Expression
Display expression

Fig. 1. System Architecture of facial expression Identification system

2.2 Data collection

A good collection of data can surely help in getting the efficient model be designed and implemented. According to the requirement of the proposed system of FEI, two different datasets which are already defined with some attributes and constraints are preferred. The two datasets are: FER 2013 and CK+48 (Extended Cohn-Kanade).

FER2013: This dataset was made by Pierre-Luc Carrier and Aaron Courville, to help in a research project. This dataset is a .csv file which consists data of images with faces in grayscale with 48x48 pixel. The faces are registered automatically as the faces are centered and occupies the same restricted space in each image. The data is categorized with seven (7) different expressions of different human faces based on their emotions. The seven face expressions are categorized with values as 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

The file fer2013.csv (dataset file) consists of three columns, 'emotions', 'pixels', 'usage'. The column 'emotions' has the emotion values of each of those seven expressions. The column 'pixel' consists the pixel values of the images with expressions. The column 'usage' has the three categories as sets as the images are divided as Training, PrivateTest, PublicTest sets respectively. The training set consists of 28,709 examples, private test-set consists of 3,589 examples and public test-set consists of 3,589 examples.

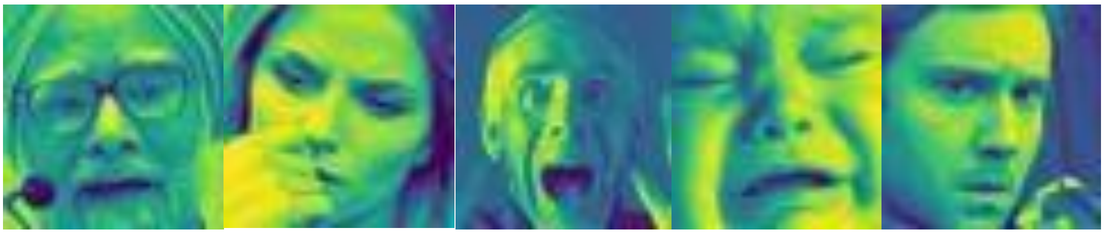


Fig. 2. Examples of FER2013 dataset

CK+48: Cohn kanade dataset's main purpose is to detect individual facial expressions automatically. The CK+ dataset is used prominently for testing the algorithm development and evaluation. The CK+ dataset has been used for emotion detection where the facial database of 210 adults were recorded under different action units and emotion labels like posed and Non-posed expressions. The expressions are categorised into different emotions like Sadness, Fear, Happy, Surprise, Disgust, Contempt, Anger. Each emotion have their set of captured images of the people relative to the concerned expression. The images with respect to each expression are as: Anger = 135, Disgust = 177, Contempt = 54, Fear = 75, Happy = 207, Surprise = 249, Sadness = 84. All these images are 48x48 pixels and are segregated according to their respective expression in a folder.

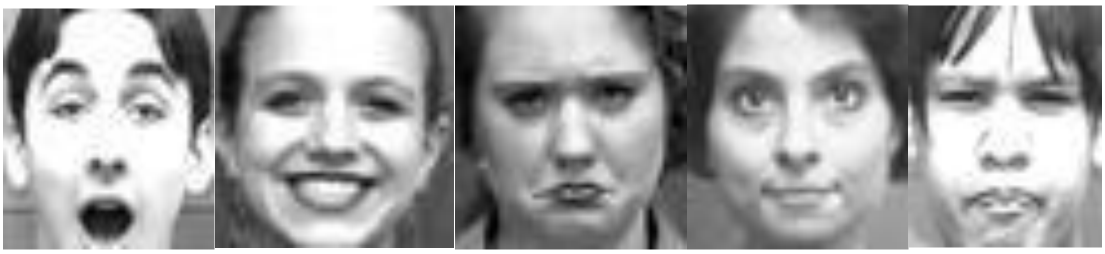


Fig. 3 Examples of CK+48 dataset

3. Working Methodology

The working of the FEI system is given below in step wise manner on how the input is taken and until the output predictions are made. Every step is important and explained in detailed way to depict the functionality of the application. The steps are as follows:

Step 1: Input Image (static or webcam):

The very first basic step is to provide an input image either using webcam or any static or still image. The input image undergoes many further processing steps to get a clear and clean data in order to get required predictions done effectively.

Step 2: Pre-processing and Normalising image:

Pre-processing is a procedure applied on data to transform it according to the model requirement before feeding it to the algorithm. Generally, pre-processing of data is converting the raw data to the clean and noise free data so as to increase the level of accuracy and the avoid errors. Normalising image is a pre-processing technique where given data is refined into a format and that data is used to intensify the results. The main aim of pre-processing is to improve the data by supressing the disturbances and distortions. This enhances the features and helps to process the image further.

General pre-processing steps that are implemented in the initial stages of the development of a model. These steps include: Noise reduction; Converting image to gray scale; Brightness of pixels transformation; Geometric transformation.

Step 3: Applying Haar based feature extraction Adaboost algorithm:

Features extraction is an important step in the process of detecting expressions as it is the process of locating particular regions, facial points, landmarks on a face in the images. In this step, a feature vector which is numerical is generated from a registered image. The very most common features extracted are: lips, eyes, eyebrows, jawline, nose tip.

HAAR features extraction Adaboost algorithm is one of the effective machine learning approach to detect objects in image processing. This is proposed by Paul Viola and Michael Jones in 2001. The cascade function is trained set from a large number of positive and negative images. This algorithm is used on other images during testing for detection.

In this case, Cascade classifier relies on Adaboost algorithm for detecting faces from the images. This is called as Viola-Jones object detection framework. This includes different steps for face detection both on static and live images. They are:

- **Haar Feature selection**

Haar features are same as those of convolution kernels which enables detecting those particular features in the given image. There are some very common features that are found commonly on human faces, such as: Eye region being dark compared to cheeks, the nose bridge region and some particular location of eyes, nose and mouth. All these regions are the commonly called as Haar features. These features are Haar wavelets rescaled sequence of square/rectangle shaped functions which are same as convolutional kernels.

The feature selection includes features extraction in the basic level and this is the process of figuring out black region and white region on the face. Each feature results are calculated by subtracting the sum of pixels under white region from sum of pixels under black region. The features calculation is done using different sizes of each kernel and are applied on the given input image to detect a face where the value of black region is ‘+1’ and value of white region is ‘-1’ relatable to convolutional kernel having one row, two columns or two rows and one column or two rows and columns.

Equation Number 1:

Feature Value (Δ) = ∑(pixels of black region) - ∑(pixels of white region)

Viola-Jones uses 24x24 sub window and calculates all features from an image. For example, considering the image Fig. 4. which showcases the process of applying haar based features on an image explains the point that haar features move from the very first pixel till the last bottom pixel in a given image as to calculate feature value throughout the window. This procedure follows similarly by increasing the shapes and size of the features on pixel by pixel to calculate the value. Considering all the variations and possible parameters of features the final output would be around 160,000+ features in a window.

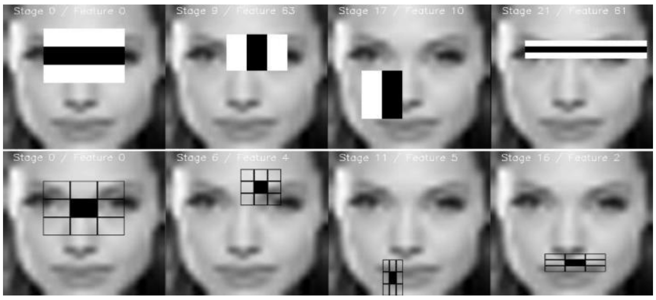


Fig. 4. Applying Haar features on image

Real values detected from image most likely to have value closer to ‘1’

White region = ‘0’; Black region = ‘1’

Example: Considering values from the above image, we know,

Feature Value (Δ) = ∑(pixels of black region) - ∑(pixels of white region)

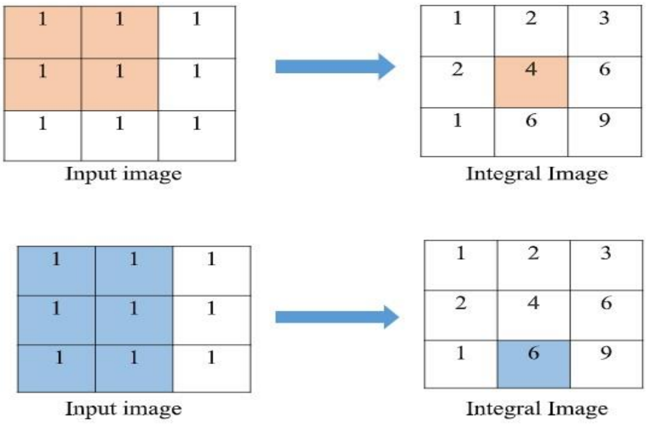
Feature Value **(Δ) = 0.74 - 0.18 = 0.56** From the equation to calculate feature value (Δ), for ideal feature it is 1. The closer the value to 1, the most likely a haar feature is found. Similarly, the final result would be the input values for the pixels in

an image with the closest value to ‘1’ throughout. The time complexity is O(n2) [1, 2].

- **Create integral image**

Integral image is defined as the representation of every pixel in image with the sum of a corresponding input pixel with pixels above and to the left of the input pixel. Integral image can simply be said as image with new pixel values. There are so many operations as different Haar features are used with possible sizes and locations and therefore more than 160,000 features are calculated.

Table 1: Conversion of input image to Integral image



1	1	1
1	1	1
1	1	1

Input image

1	2	3
2	4	6
1	6	9

Integral Image

1	1	1
1	1	1
1	1	1

Input image

1	2	3
2	4	6
1	6	9

Integral Image

Integral image is used to calculate the sum of pixel values fast and effectively in the rectangular subset. The average intensity is calculated in a given image. Integral image calculates haar features really very fast. Doing this, the time complexity is reduced from O(n2) to O(1) and hence time is very much saved in procedure of calculating features.

- **Adaboost Training**

As stated in beginning, while moving the Haar features from the very first pixel to bottom pixel in 24x24 window gives out more than 6000 feature values. Among them only few set of features are used to detect or identify faces and remaining are irrelevant for detection.

In order to find out the most relevant and irrelevant features in an image, Adaboost algorithm is used.

Adaboost determines relevance and irrelevance. It selects few features which are relevant to us. It will

identify certain number of features and gives weight to those selected features and linear combination of all 6000 features are used to determine whether it is a face or non-face. Adaboost finds the best features, called as 'WEAK classifers'. Weak classifiers are the best features or the relevant features and performs better than random guessing extracted by Adaboost. We apply relevant feature and find corresponding weight of that and combine all those relevant features with corresponding weights and form a strong classifier.

Equation Number 2:
Formula for STRONG classifier: Fx=1f1x+2f2x+ 3f3x+…

- **Cascading classifiers**

Generally 6000 features are used to form a strong classifier which is the sum of weak classifiers. The basic principal of this algorithm is to slide the image with different sizes and shapes of detector for many times. So, in every 24x24 window evaluating all the 6000 features throughout the image resolution and see whether it exceeds certain threshold to detect a face. This process would probably spend time on face regions and discarding non-face regions. Hence, a single strong classifier formed by linear combinations of features would not be effective to evaluate on each window. This mean out of 6000 features first 10 features are kept in one classifier, then next set of 20 or 30 features in another classifier and so on. Therefore a cascade classifier is a set where all the features are hierarchically segregated and face detection is performed.

Step 4: Applying CNN on dataset to get trained model:

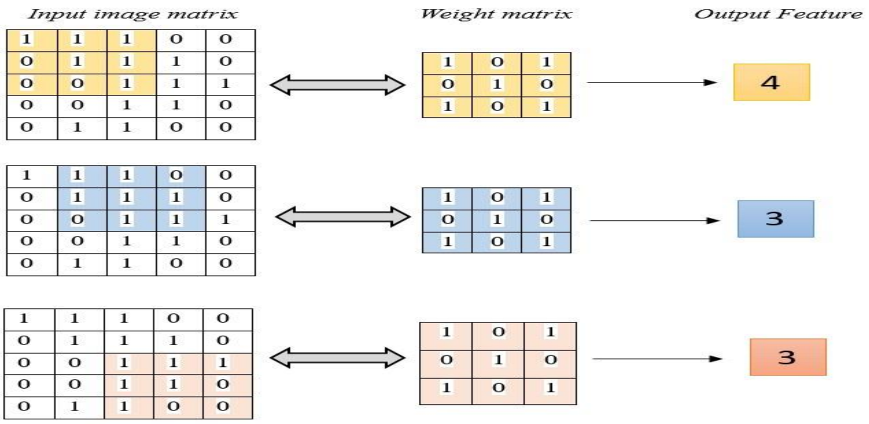
CNN is feed forward neural network that is generally used to analyse visual images by processing data with grid like topology. So it is called as ConvNet. In CNN, every image is represented in the form of arrays of pixel values. The ConvNet layers are as follows:

- **Convolutional Layer:**

Convolutional layer has number of filters to perform convolutional operations. Every input image is processed as matrix of pixel values. In order to preserve the spatial arrangement in both horizontal and vertical directions, weights are taken as 2D matrix which take pixels together in both horizontal and vertical directions. According to both horizontal and vertical movement of weights, the output is one pixel lower in both horizontal and vertical direction.

Example: Suppose an image size of 5x5 and define a weight matrix to extract certain features from the images. The weight matrix is slided over the image pixel matrix both horizontally and vertically simultaneously computing the dot product.

Table 2: Sliding the weight matrix on image pixels to compute and generate features



1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

1	0	1
0	1	0
1	0	1

Accordingly the size of an image would reduce and to avoid that, padding the input image with zeros would solve the problem and size of an output image will be same as input image and is called as *same padding.* The output from the each filter the ***activation map*** or ***feature map*** is the output of the convolution layer.

- **ReLU Layer:**

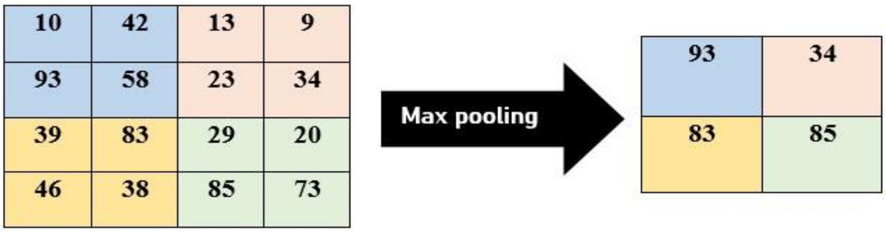
ReLU is Rectified Linear Unit which locates the features. After the feature map is generated or extracted from the convolutional layer, is a feeding to ReLU, which means the output of convolutional layer is the input to ReLU layer. It performs element wise operations and sets all the negative pixels to zero. This introduces non-linearity to the network, where linearity has some value like edges, features, points, etc. Then from this the output is called rectified feature map. **R(z) = max(0,z)**

- **Max Pooling Layer:**

Sometimes when the image size is too large, we need to reduce the number of parameters which are trainable. It is important to have pooling layers between every subsequent convolution layers. Pooling is done on image to reduce the spatial size of that image. The depth of the image remains unchanged as pooling is done irrespective of dependency. Therefore, the pooling layer generally applied is called the ***max pooling***. Pooling is a down sampling [5, 6] operation which reduces the dimensionality of feature map from ReLU. Pooling layer again uses different layers to identify parts of the image like eyes, nose, mouth, etc. Pooling is nothing but gathering or collecting the maximum values from the input feature map.

Example: Suppose a feature map of 4x4 matrix and output would be 2x2 matrix as pooled feature map.

Table 3: Max pooling the values from the acquired featured map



10	42	13	9
93	58	23	34
39	83	29	20
46	38	85	73

93	34
83	85

- **Fully Connected Layer – Output Layer**

The convolution and pooling layers would only be able to extract features and reduce the number of parameters from the original images. A flattened matrix is then from the pooling layer is the input to fully connected layer to classify image.

Step 5: Classify, Predict and display expression

From the final stage of the application, it depicts the output as the predicted expression from the classified expressions based on the trained model from the input image. The predicted expression is the result of the facial expression identification system.

4 Analysis

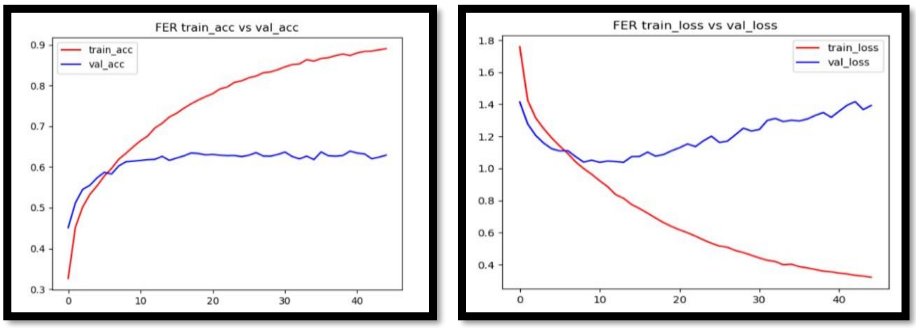


Fig.5. Loss and accuracy of FER2013 datas

Analysis of datasets FER2013 and CK+48 based on the training and validation graphs gives the clarity on working of the model. In comparison with both datasets, analysis on accuracy of prediction is explained. The validation loss and validation accuracy graph lines explains how the model is learning and how normally the model gets trained and works on prediction. It is very important that a model get trained well in order to use it for real time applications in predictions.

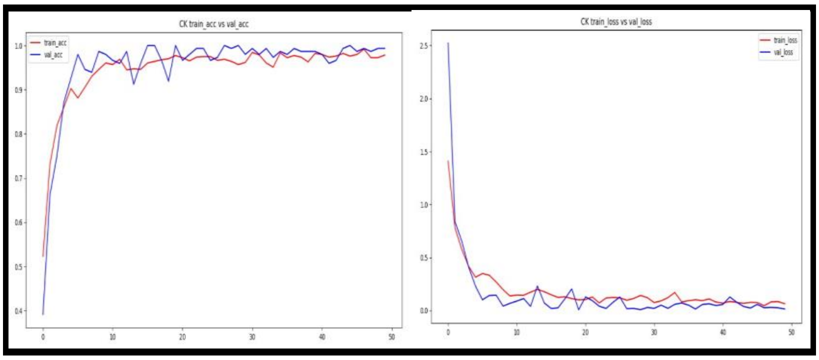


Fig. Loss and accuracy graph of CK+48

While comparing validation loss and validation accuracy,

1. If val_loss < val_acc; Model runs normally.

2. If val_loss is greater and val_acc is lesser; then Model isn't learning.

3. If val_loss is greater and val_acc is also greater with respect to train_loss and train_acc; then it is overfitting.

So, it is important that a model learns correctly and runs normally in predicting the outputs.

5. Conclusion

FER2013 (model accuracy: 63.8%) and CK+48 (model accuracy: 97.9%) which almost predicts same but missed with few of the expressions. Anyways, the system projects on classifying different expressions is the prior issue to be noted and another observation on working with datasets is having good and huge collection of data which is always good and gives better efficiency and results.

Future work can include the developed system of human face expression recognition can be made much more effective by still using much more elements related for expression detection. It can be extended by integrating a database and improving interface by specifying more features like face identification and include security based interactions and make it real time usage in places like educational institutions, crime investigations and large organisations during interviews, etc.

References

[1] Lei Xu, Minrui Fei, Wenju Zhou, Aolei Yang, "Face Expression Recognition Based on Convolutional Neural Network", 2018 Australian & New Zealand Control Conference (ANZCC)-Swinburne University of Technology, 2018.

[2] Imane Lasri ; Anouar Riad Solh ; Mourad El Belkacemi, "Facial Emotion Recognition of Students using Convolutional Neural Network", 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), doi:10.1109/icds47004.2019.8942386.

[3] S.Nithya Roopa, "Research on Face Expression Recognition", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, July 2019, Volume-8, Issue- 9S2.

[4] Burhanudin Ramdhani, Esmeralda C. Djamal*, Ridwan Ilyas, "Convolutional Neural Networks Models for Facial Expression Recognition", International Symposium on Advanced Intelligent Informatics (SAIN), 2018.

[5] Kewen Yan, Shaohui Huang, Yaoxian Song, Wei Liu1, Neng Fan, "Face Recognition Based on Convolution Neural Network", Proceedings of the 36th Chinese Control Conference, July 26-28, 2017.

[6] Matthew Turk; Gang Hua, 'Book-Vision-Based Interaction-Synthesis lectures on Computer Vision', published by Morgan and Claypool, 2013.