Implementation and Comparison of Facial Expression Detection and Classification Techniques

Anupam Tripathi
Department of Computer Engineering
K. J. Somaiya College of Engineering
Mumbai, India

Nikhil Thakurdesai
Department of Computer Engineering
K. J. Somaiya College of Engineering
Mumbai, India

ABSTRACT
Facial expressions are one of the most important behavioral measures for emotion recognition. Expressions can tell a lot about the person, his behavior, what he is thinking and this data is vital in making various predictions which can have a variety of applications. In this paper we have implemented and compared three types of facial expression recognition and classification techniques. The first one is a state-of-the-art convolutional neural network, the second one is a transfer learning approach using the InceptionV3 model and in the last one, we have extracted the 68 facial points which have been identified as important for recognizing the expression of a person and passed it to a deep neural network. All these techniques have given accuracies over 90%, so comes the need to compare them in detail and determine which one of them would give results more accurately and efficiently.

General Terms
Facial Expression Recognition, Deep Neural Network

Keywords
Facial Expression Recognition, CNN, Transfer learning, Haar Cascades

1. INTRODUCTION
Facial expression are an important part of non verbal communication among human beings, using which one can understand the mood and the mental state of another person. It is reported that 7% of any message is conveyed through words, 38% through certain vocal elements, and 55% through nonverbal elements (facial expressions, gestures, posture, etc). This is only possible because humans are able to recognize emotions quite accurately and efficiently. This makes it very important for machines to be able to recognize human expressions to facilitate a smooth conversation.

Emotions more often are communicated by subtle changes in one or a few discrete facial features, such as tightening of the lips in anger or obliquely lowering the lip corners in sadness. To capture such subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed. The facial action coding system (FACS) is a system based on facial muscle changes and can characterize facial actions to express individual human emotions as defined by Ekman and Friesen in 1978. FACS encodes the movements of specific facial muscles called action units (AU). They code the fundamental actions (46 AUs) of individual or groups of muscles typically seen when producing the facial expressions of a particular detected and the system classify facial category according to

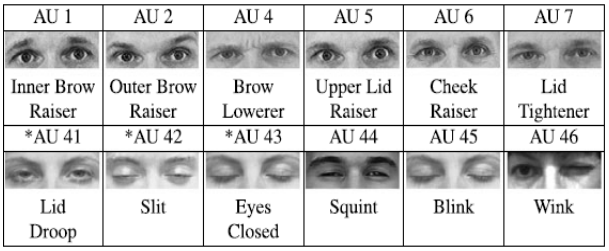

Fig 1. Upper Face Action Units

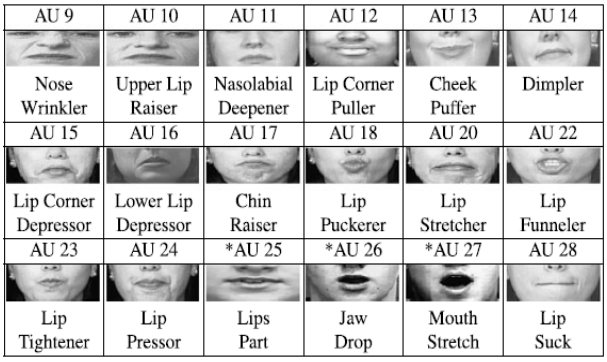

Fig 2. Lower Face Action Units

the combination of AU's. For example, if an image has been annotated as having 1, 2, 25, and 26 AUs using an algorithm, the system will classify it as expressing an emotion of the "surprised" category. Figure 1 and Figure 2 shows some examples of combinations of FACS action units.

2. LITERATURE REVIEW
Several methods have been reported in the literature to automatically recognize facial expressions. In 1971, the American psychologist Ekman and Friesen defined seven categories of basic facial expressions, which are Happy, Sad, Angry, Fear, Surprise, Disgust and Neutral [2]. For automatic facial expression analysis, Suwa et al. [3] presented an early attempt in 1978 to analyze facial expressions by tracking the motion of 20 identified spots on an image sequence. In the 1980s, Yaan Lecun et al. [4] presented a paper about Convolutional Neural Network (CNN). Considerable progress has been made since then. The most popular feature extraction techniques are Gabor filters [5][6], Local Binary Patterns (LBP) [7][8], Principal Component Analysis (PCA) [9][10], Independent Component Analysis (ICA) [11][12], Linear Discriminant Analysis (LDA) [13], Local Gradient Code (LGC) [14], Local Directional Pattern (LDP) [15]. Very few of them crossed the 90% accuracy mark, which is not quite good considering the high accuracy demanding applications which we implement today.

3. EXPERIMENTAL SETUP

3.1 Database

The two databases used by us are the Extended Cohn-Kanade database (CK+) and the Japanese female facial expression (JAFFE) database. The CK+ contained facial images taken from 97 subjects with age ranging from 18 to 30 years. The database had 65 percent female subjects. Fifteen percent of the subjects were African-American and three percent were Asian or Latino. The camera was located directly in front of the subject. The subjects performed different facial displays (single action units and combinations of action units) starting with a neutral face and ending with the target emotion. The displays were based on descriptions of prototypic emotions (i.e., neutral, happy, surprise, anger, fear, contempt, disgust, and sad). Figure 3 shows examples of CK+ database. The images on the top level are taken from the CK database and those on the bottom are representative of the extended data. Examples of the emotion and AU label are: (a) Disgust – AU 1+4+15+17, (b) Happy – AU 6+12+25, (c) Surprise – AU 1+2+5+25+27, (d) Fear – AU 1+4+7+20, (e) Angry – AU 4+5+15+17, (f) Contempt – AU 14, (g) Sadness – AU 1+2+4+15+17, and (h) Neutral - AU0 This has been specified in more detail in CK+ paper [1].

3.2 Data Preprocessing

Data preprocessing plays a key role in overall process since it enhances the quality of input image and locates data of interest by removing noise and smoothing the image. The preprocessing techniques used by us are face cropping and histogram equalization.

Fig 3. CK+ Database

Fig 4. JAFFE Database

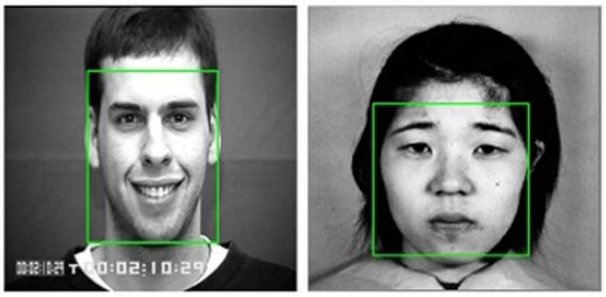

Fig 5. Face crops taken from CK+ dataset (left) and JAFFE dataset (right)

3.2.1 Face cropping

All the images from the dataset were cropped to the facial part only. This was done using the dlib library which helped us extract coordinates of the four corners of the bounding box of the face as shown in Figure 5. All the part outside the bounding box was cropped away. This removes most of the redundant portion of the image and helps us in extracting the useful features, thus saving a lot of computational time.

3.2.2 Histogram equitation

Histogram equalization is a technique for adjusting image intensities to enhance contrast. The mathematical procedure for mapping the image pixels to new pixel values that will increase the overall contrast in the image is quite straightforward. The intensity of a pixel is represented by i and the available intensities are $0 \leq i \leq L\text{-}1$ where L for an 8-bit image would be 256. The probability of finding a pixel with intensity k can be given by

$$p_k = \frac{number\ of\ pixels\ with\ intensity\ k}{total\ number\ of\ pixels} \qquad (1)$$

Consider a function T, which assigns a new intensity value to any given pixel. So for all pixels with intensity i, the new intensity value will be given by T(i)

$$T(i) = \left\lceil (L-1) * \sum_{k=0}^{i} p_k \right\rceil \qquad (2)$$

4. METHODOLOGY

We have implemented three techniques for facial expression recognition. All the three methods can be generalized to have the stages depicted in Figure 6. The input image is first preprocessed. We have used two preprocessing techniques: face cropping and histogram equalization. After this, the facial features are extracted and fed to a deep neural network to finally get the identified facial expression as the output. The three techniques implemented by us are:

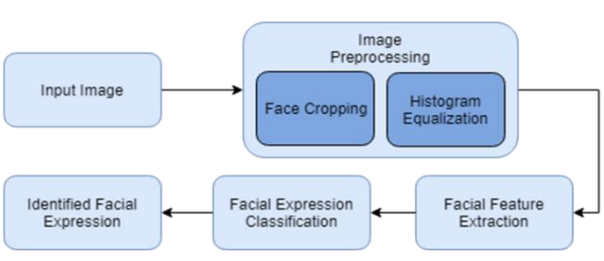

Fig 6. Stages for classification

4.1 CNN

Among the several deep-learning models available, the convolutional neural network (CNN) is the most popular network model. In CNN based approaches, the input image is convolved through a filter collection in the convolution layers to produce a feature map. Each feature map is then combined to fully connected networks, and the facial expression is recognized as belonging to a particular class based on the output of the network.

The main advantage of a CNN is to completely remove or highly reduce the redundant portions of an image which we don't need to concentrate on. The number of features which

are sent as input into the fully connected neural network are exponentially decreased when using this technique. For this reason, CNN has achieved state-of-the-art results in various fields, including Object Recognition, Face Recognition, Scene Understanding, and Facial Expression Recognition.

We have implemented a deep convolutional neural network which consists of 4 convolutional layers. This is followed by a fully connected network in which the output layer consists of 8 emotions. The model is shown in Figure 7.

The activation functions used for the convolutional layers and the first 4 fully connected layers is the relu activation function given by (3) and for the output layer is the softmax activation function given by (4).

$$f(x) = \max(0, x) \tag{3}$$

$$f(x_i) = \frac{Exp(x_i)}{\sum_{j=0}^{k} Exp(x_j)} \quad i = \ 0, 1, 2, \ldots . k \tag{4}$$

4.2 TRANSFER LEARNING

Transfer learning or inductive transfer is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks. Thus, such pre-trained models can be used to increase your system's accuracy and it is also useful if you do not have the hardware to train big data sets.

For this technique, we used the InceptionV3 [12] model which was trained on the ImageNet [13] data set. The InceptionV3 was made by Google. Their first inception model was the GoogLeNet followed by BN-Inception-V2. The ImageNet

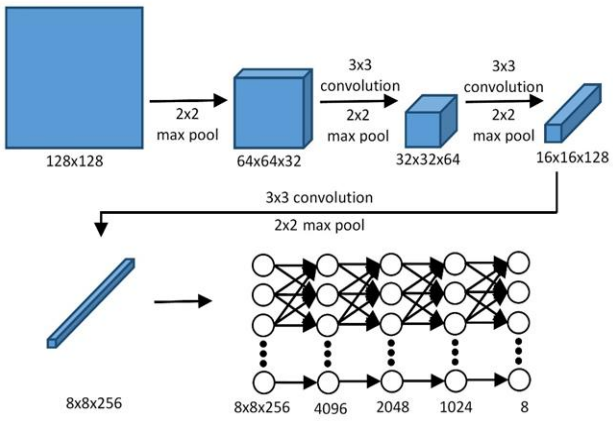

Fig 7. CNN Architecture

dataset is a 1000 image classifier which is considered as an academic benchmark for computer vision. The InceptionV3 achieved an as low as 3.46 % error rate on the validation set.

Figure 8 shows the architecture of the InceptionV3 model. The transfer layer of the Inception model outputs a 2048 feature vector. The 2048 feature vector was then taken as input for our own deep neural network which we trained using back propagation. We used a 3 layer deep neural network shown in Figure 9. The output layer had 8 nodes which outputted values for 8 facial expressions including neutral face. The last layer of the inception model was replaced by three fully connected layers of our own.

Figure 9 shows the inception model giving a feature vector of size 2048 for a RGB image. This is connected to two more hidden layers. The last layer is the output layer with 8 classes for the 8 emotions.

4.3 FEATURE EXTRACTION USING HAAR CASCADES

68 points have been identified on a person's face which are necessary in recognizing the expression or the emotion of a person. So, one way to classify the expression of a person is to plot the 68 points on a person's face and take those coordinates as feature inputs to our network. A haar cascade file is used to plot these 68 points on a person's face. Figure 10 shows the facial landmarks marked on an image from the CK+ database (left) and a JAFFE database (right). The way in which each person smiles might be different and the facial features might vary, like someone may have a larger distance between the eyebrow and the eye than some other person but for everyone. We can use this common relativity to classify a

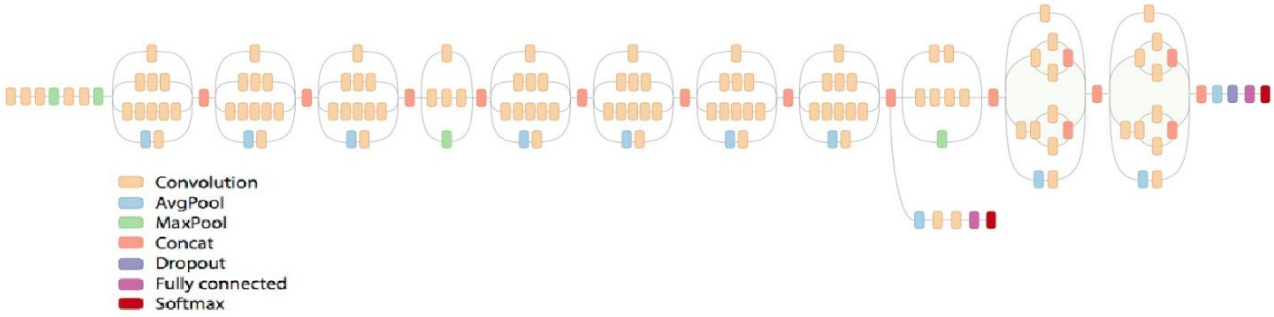

Fig 8. Inception Architecture

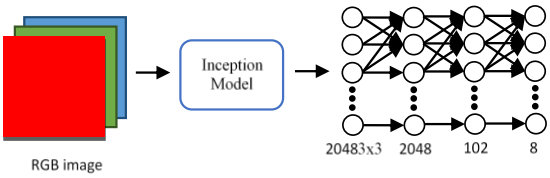

Fig 9. Transfer Learning Architecture

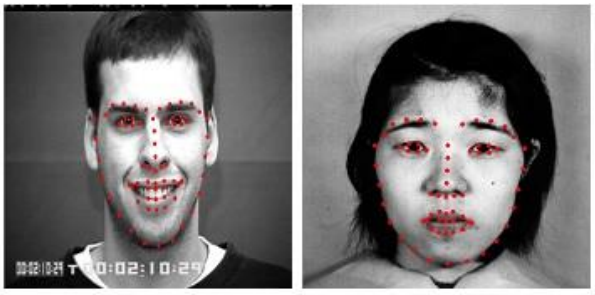

Fig 10. 68 face points mapped onto the CK+ dataset (left) and JAFFE dataset (right)

person's face into one of the 8 emotions. When trained over sufficient number of images, the neural network should be able to map this relativity, e.g. - in Figure 11, the angle made by point 49 with point 58 will closely be the same for every person with a happy face and the network on sufficient training; will be able to map that angle to a happy face and classify the image as happy.

The dlib library is a general purpose cross-platform software library written in the programming language C++. The facial landmark detector implemented using dlib produces 68 (x, y)-coordinates that map to specific facial structures. The facial landmark detector takes a haar cascade file as input. The 68 feature points which our dlib model detects include the Jaw line of the face, left and right eyes, left and right eyebrows, the nose, and the mouth. These 68 (x, y)-coordinates were fit into a 136 feature vector which was passed as input into the neural network.

The deep neural network consisted of four hidden layers as

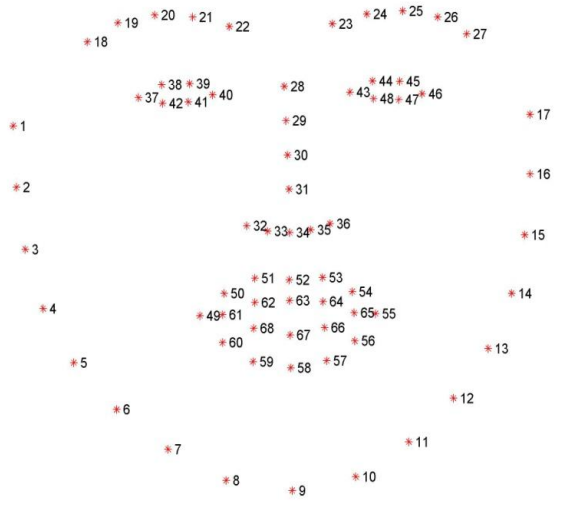

Fig 11. 68-Point Markup

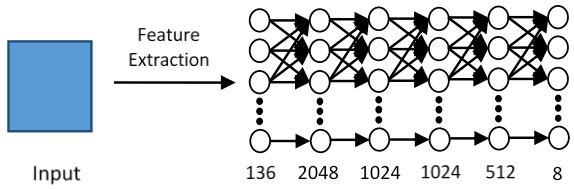

Fig 12. Transfer Learning Architecture

shown in Figure 12 and one output layer with the number of classes for the output layer being the 8 basic emotions i.e. neutral, happy, surprise, anger, fear, contempt, disgust, and sad. The learning rate chosen was 0.00001.

This method is computationally less expensive than the two preceding methods. The reason being that we have only used 136 features for every input image instead of taking every pixel as an input feature as in the convolutional approaches which amounted to 8 x 8 x 256 features in the CNN approach and 2048 features in the Transfer Learning approach before passing to the fully connected network. This decreased the features from 8 x 8 x 256 or 2048 to just 136 which is an exponential decrease in the number of features. Thus, this process only takes in the useful features further removing all the redundant background information which does not contribute to classifying the emotion of a person.

5. RESULTS

Table I shows the comparison between the best accuracies for the three methods. And Figure 13, 14 and 15 show the relative validation graphs drawn using Tensorboard. The number of epochs for which theses graphs were made are 50 for CNN, 50 for Transfer Learning and 1200 for Feature extraction using haar cascades. As compared to the CNN and the transfer learning techniques, feature extraction using haar cascades took the least processing time. This is because in this method, we pass only the co-ordinates of the 68 facial points, which sums up to only 136 features per image, whereas for the other two techniques, each pixel is treated as a feature at start and multiple layers are required to extract the real facial features. But the haar cascade approach took a lot more epochs than the other two since it was slow in learning those features.

Table 1. Validation Accuracies

Model	Accuracy
CNN	94.34
Transfer Learning	94.62
Feature Extraction using Haar Cascades	96.39

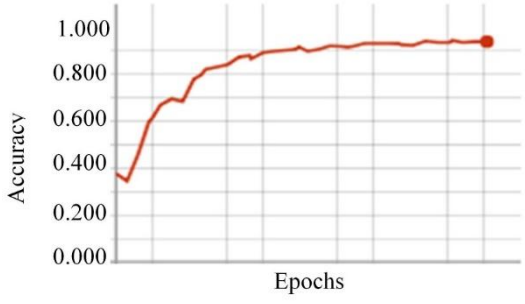

Fig 13. 68-Point Markup

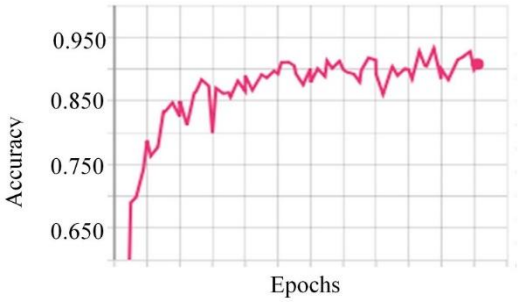

Fig 14. Transfer Learning

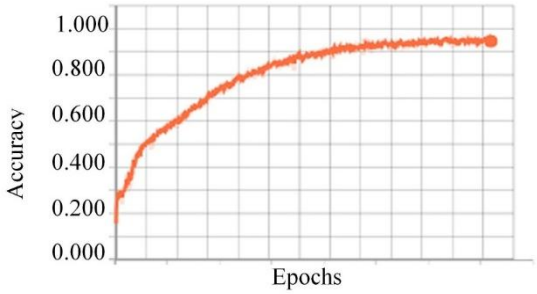

Fig 15. Feature Extraction using Haar Cascades

Feature extraction using haar cascades has given the highest accuracy. It also required the least computing power and time. Both the other models gave accuracies over 90 percent but were more resource consuming and complex as compared to this method. The pre-trained inception model gave an accuracy slightly better than the CNN model used by us but the haar technique is more suitable for expression detection and classification. Thus, we can say it with confidence that the 68 points extracted from the face provide far more valuable information than analyzing the entire face as far as expression detection is concerned.

6. FUTURE WORK

Automated emotion recognition has many applications including human behavior understanding, detection of mental disorders and improving human-machine interactions (HMI). Using facial expressions as a communication channel in HMI is one way of making use of the inherent human ability to express themselves via changes in their facial appearance. Thus, the robustness and genuineness of HMI is greatly improved. We can even get the feedback of a seminar or any event using just some pictures of the audience, by analyzing everyone's emotions.

7. REFERENCES

[1] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[2] Ekman, Paul, and Wallace V. Friesen. "Constants across cultures in the face and emotion." Journal of personality and social psychology 17.2 (1971): 124.

[3] Suwa, M., Sugie, N., Fujimora, K.: A preliminary note on pattern recognition of human emotional expression. In: International Joint Conference on Pattern Recognition, pp. 408– 410 (1978).

[4] Lecun, Y. "Generalization and Network Design Strategies." Connectionism in Perspective 1989.

[5] Lyons, Michael J., Julien Budynek, and Shigeru Akamatsu. "Automatic classification of single facial images." IEEE Transactions on pattern analysis and machine intelligence 21.12 (1999): 1357-1362.

[6] Kulkarni, Ketki R., and Sahebrao B. Bagal. "Facial expression recognition." India Conference (INDICON), 2015 Annual IEEE. IEEE, 2015.

[7] Ojala, Timo, Matti Pietikäinen, and David Harwood. "A comparative study of texture measures with classification based on featured distributions." Pattern recognition 29.1 (1996): 51-59.

[8] Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. "Robust facial expression recognition using local binary patterns." Image Processing, 2005. ICIP 2005. IEEE International Conference on. Vol. 2. IEEE, 2005.

[9] Turk, Matthew A., and Alex P. Pentland. "Face recognition using eigenfaces." Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. IEEE, 1991.

[10] Gosavi, Ajit P., and S. R. Khot. "Facial expression recognition using principal component analysis." International Journal of Soft Computing and Engineering (IJSCE) 3.4 (2013): 2231-2307.

[11] Bartlett, Marian Stewart, Javier R. Movellan, and Terrence J. Sejnowski. "Face recognition by independent component analysis." IEEE Transactions on neural networks 13.6 (2002): 1450-1464.

[12] Guo, XiaoHui, et al. "Facial Expression Recognition based on Independent Component Analysis." Journal of Multimedia 8.4 (2013).

[13] Belhumeur, Peter N., João P. Hespanha, and David J. Kriegman. "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection." IEEE Transactions on pattern analysis and machine intelligence 19.7 (1997): 711-720.

[14] Tong, Ying, Rui Chen, and Yong Cheng. "Facial expression recognition algorithm using LGC based on horizontal and diagonal prior principle." Optik-International Journal for Light and Electron Optics 125.16 (2014): 4186-4189.

[15] Jabid, Taskeed, Md Hasanul Kabir, and Oksam Chae. "Facial expression recognition using local directional pattern (LDP)." Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010.

[16] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[17] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.