International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Illumination Invariant Facial Expression Recognition using Convolutional Neural Networks

K. Prasada Rao, M. V. P. Chandra Sekhara Rao

Abstract: *— In this work, we propose a prospective novel method to address illumination invariant system for facial expression recognition. Facial expressions are used to convey nonverbal visual information among humans. This also plays a vital role in human-machine interface modules that have invoked attention of many researchers. Earlier machine learning algorithms require complex feature extraction algorithms and are relying on the size and uniqueness of features related to the subjects. In this paper, a deep convolutional neural network is proposed for facial expression recognition and it is trained on two publicly available datasets such as JAFFE and Yale databases under different illumination conditions. Furthermore, transfer learning is used with pre-trained networks such as AlexNet and ResNet-101 trained on ImageNet database. Experimental results show that the designed network could recognize up to 30% variation in the illumination and it achieves an accuracy of 92%.*

Keywords : Classification, Convolutional Neural Network, Facial expression recognition, Illumination.

I. INTRODUCTION

Facial expressions play a major role in human communication. They serve as one among the most of information channels in social communications. Facial expression recognition (FER) has been a vigorous area of research in computer vision with application areas including social robots, personalized banking, animation, e-learning etc.

The accurate analysis and interpretation of the emotional content of human facial expressions is essential for the deeper understanding of human behavior. Facial expressions significantly assist in direct communication and it has been indicated that during face-to-face human communication, 7% of the information is communicated by the linguistic part, such as the spoken words, 38% is communicated by paralinguistic, such as the vocal part, and 55% is communicated by the facial expressions [3].

Manuscript published on November 30, 2019.
*Correspondence Author
K. Prasada Rao* , pursuing Ph.D in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, AP, India.
Dr. M. V. P. Chandra Sekhara Rao, professor at Dept of CSE, RVR&JC College of Engineering, Guntur, AP, India.

Therefore, the main aim of facial expression recognition methods and approaches is to enable machines to automatically estimate the emotional content of human face. In intelligent tutoring systems, emotions and learning are inextricably bound together; so recognizing learners' emotional states could significantly improve the efficiency of the learning procedures delivered to them [9]–[11].

Facial expression recognition is the process of identifying human emotions and state of mind based on their facial expressions. In human beings from generation immemorial, this recognition ability has been treated as one of the most important social skills in understanding the other persons. In general, there is universality in the facial expressions of human beings in expressing certain emotions. It is observed that these universalities can efficiently be used for human-machine interaction. So far, many researchers have developed different machine learning algorithms to implement this application and the concept of deep learning is considered as the most successful approach in delivering promising results for human-machine interactions [2].

Developing an effective facial representation from the original face image is a vital step for successful facial expression recognition. Even the best classifier may fail to achieve better performance when features are inadequate. Two types of approaches have been proposed to extract facial features for expression recognition: geometric feature-based method and appearance-based method. In the geometric feature-based method, the shape and location of facial components are considered. Geometric relationships between these components are used to form a feature vector. The appearance-based method extracts features by applying an image filter or filter banks on the whole face or some specific regions of the face such as principal component analysis (PCA) [15].

In most of the facial expression recognition approaches, it is assumed that the facial images consider for training and test dataset are captured under similar or same environmental conditions but in real-time these conditions are uncontrollable and illumination changes are unavoidable [1]. In this scenario, the performance of the recognition system decreases and sometimes may give incorrect predictions. This paper concentrates on providing a solution for expression recognition under varying illumination conditions. The work aims to design a deep convolutional network that is smaller, faster and simpler that could meet the objective of recognizing the expression under changing illumination.

II. RELATED WORK

In recent past, there has been a boom in the use of convolutional neural networks (CNN) for image classification and object detection. A CNN can be viewed as a framework combining a feature extractor and a classifier. The convolutional layers work as feature extractors that learn the representations automatically from the input data. Features such as shapes, edges, and color blobs are learned in the earlier layers of the CNN. While the following layers learn features more specific to the original dataset. The learned features are fed to the last fully connected layers to classify the data into one of the classes.

Deep neural networks are applied to many computer vision applications like classification, detection, recognition, and others. Several authors and researchers have designed the networks for the specific applications and some of them used transfer learning with existing trained networks. For image classification, many of the researchers employed transfer learning with famous CNN (convolutional neural networks) like AlexNet [2], the network is deeper and bigger in which the convolutional layers are stacked on each other. Another most cited network is VGGNet [3] which has multiple convolutional layers with a small filter size of 3x3 which are stacked on top of each other.

As in [4], Alizadeh and others developed CNNs with variable depths for facial expression recognition and evaluate the performance of these models using different post-processing and visualization techniques. The model has attained more accuracy than the hybrid feature sets, here they employed raw pixel data to train the network in PyTorch. In [5] PrudhviRaj introduced RAUs (Representational Auto-Encoder Units) based on two independent methods for emotion detection. The first one uses representational auto-encoders to construct a unique representation of any given emotion. Auto-encoders are a different class of neural networks that can rebuild their input in some lower-dimensional space. The second method is based on convolutional neural network.

The effect of CNN parameters like kernel size and several filters on the classification accuracy is investigated by Abhinav and others in [6]. In this work, they used the FER-2013 dataset as a benchmark and analyzed the influence of parameters on accuracy. They also proposed two models that are suitable for hardware deployment that sustainable to the change in the kernel size. Li and others in [7], designed a faster region-based convolutional neural network in which the dimensions of the extracted features are reduced using maximum pooling. This network includes the region proportional networks which help to detect the region with which the expression can be recognized accurately.

In [8] Garimella et.al trained a network for detecting the driver fatigue condition, however, the network was tested on 6 emotions that were taken from the Jaffe database. From the experiments, they attained an average accuracy of 68 %.

In all the above papers most of the researchers have concentrated on recognizing the expression with customized networks and with transfer based learning considering the same environmental condition that was present during training the network. In this work, a customized deep network is developed for recognizing the emotions under varying illumination conditions. The classification performance might be inadequate due to the poor lighting conditions even with an optimal classifier and a great training set.

The images captured in low-light condition usually suffer from both low contrast and more noise. Various low-light image enhancement methods were proposed [21],[22], most of them focused on contrast enhancement. However, these methods usually imposed a uniform enhancement on the whole image which tended to cause over enhancement for very bright regions or under enhancement for very dark regions.

Ruiz-Garcia et.al [11] proposed a deep Stacked Convolutional Autoencoder (SCAE) method. It is trained to reconstruct the images with different illumination to the mean luminance. Jiang et al [13] developed a local feature hierarchy network (LFHN) model, unlike the conventional ConvNet, where the classification is performed based on the features extracted from the last layer, In LFHN, different local features extracted from various layers are concatenated and fed them as input to the final layer. Li, Chongyi et al [14] proposed the LightenNet method that learns a mapping between weakly illuminated image and the corresponding illumination map which is subsequently used to obtain the enhanced image.

Our approach to the problem has been to focus on simple convolutional neural networks on individual images. The reason for this stems from two arguments: (1) we want to create a neural network model which does not require too much of computational power, since the area of use might be to embed the software on a network camera, and (2) when trying to analyze facial expressions in a video stream it is hard to distinguish when a facial expression starts and when it ends [12].

III. PROPOSED APPROACH

In this work, the deep convolutional neural network architecture is designed for facial expression recognition. The network is made of convolutional layers, pooling layers, ReLu layer with activation function and fully connected layer. The output layer consists of 7 neurons justifying 7 emotions. In this work, seven emotions like happiness, sadness, disgust, fear, normal/ neutral, surprise and anger were considered for recognition.

Fig.1 depicts the proposed architecture for emotion recognition, where the input layer is provided with a grayscale image of size 256x256 as shown in the first block of the figure. Fig.2 shows the generalized convolutional neural network architecture for facial expression recognition. The detail specifications of the architecture are shown in Table- I and Table- II.

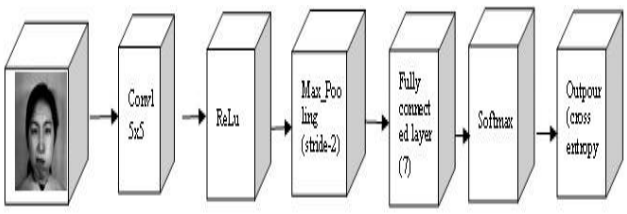

Fig. 1.Block diagram of the proposed network architecture

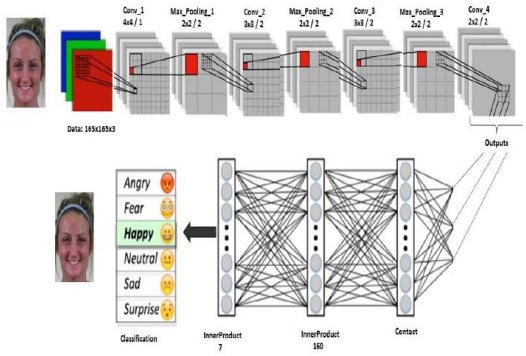

Fig. 2. Architecture of CNN.

Table- I: Proposed architecture details model 1

Input : 256x256
Conv2d: 5x5x1 (20) stride: 1 + ReLu
Max_pooling_1 : 2x2 , stride : 2
Conv2d: 5x5x20 (20) , stride : 1+ ReLu
Max _pooling_2: 2x2, stride: 2
Fully connected: 7 layers, (74420)
Softmax , output: cross entropy

Table- II: Proposed architecture details model 2

Input :243x320
Conv2d: 5x5x1 (20) stride: 1 + ReLu
Max_pooling_1 : 2x2 , stride : 2
Conv2d: 5x5x20 (20) , stride : 1+ ReLu
Max _pooling_2: 2x2, stride: 2
Fully connected: 5 layers, (87780)
Softmax , output: cross entropy

The proposed architecture consists of 2 convolutional layers with (size, number of filters) is (5 × 5, 1), (5 × 5, 20) respectively and the stride is 1. The first pooling layer is a max-pooling layer with a filter size of 2x23 and the stride is 2. The next two pooling layers are also max-pooling layers with a filter size of 2 × 2 and the stride is 2. The next one is a fully connected layer with 7 layers for model 1 and 5 layers for model 2. Passing this layer, from each input image we obtain an N-dimensional vector, which is a high-level feature descriptor representing the input image [20].

Finally, this vector is passed to the last fully connected layer of neurons, which in turn, outputs a 7-d vector s = (s1, s2, .., s7)representing class scores towards 7 kinds of emotions. The connectivity between the layers is depicted in Fig. 3.

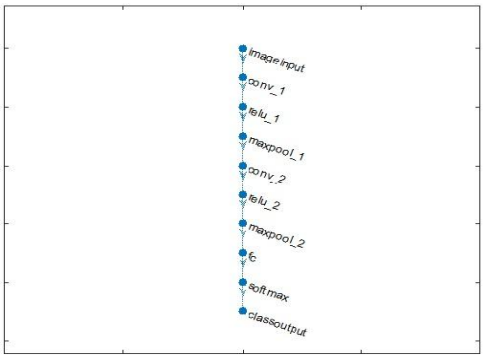

Fig. 3.Connections of layers in the network

IV. DEEP LEARNING ARCHITECTURES

AlexNet [17]: AlexNet, named by its creator Alex Krizhevsky, is an architecture originally submitted in the ILSVRC competition in 2012. The network consists of 8 layers: five convolutional layers and three fully-connected layers. The output of the last fully-connected layer is fed to a 1000-way softmax function, which produces a distribution over 1000 class labels. Starting with the first convolutional layer shown in Fig. 4.

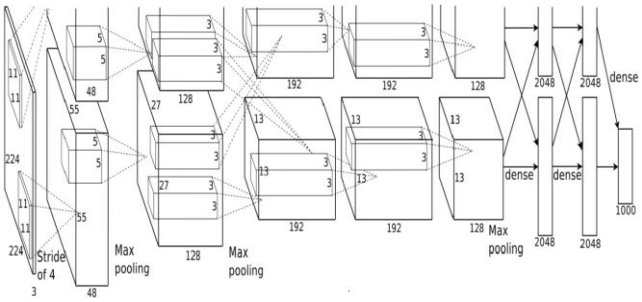

Fig. 4. Illustration of AlexNet's architecture [17].

ResNet-101: A degradation problem has been uncovered when deeper networks start converging. As the network depth starts increasing, accuracy gets saturated and then degrades rapidly. This is not caused by overfitting (which has been confirmed through experiments on the accuracy of training data) [12, 19], which means that the deeper architecture seems unable to find additional features in an image.

A proposed solution to this problem was introduced with a deep residual learning framework (ResNets) [18]. This is implemented by having shortcut connections in the network, which means that the input signal takes a shortcut around several layers as shown in Fig. 5. These shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers.

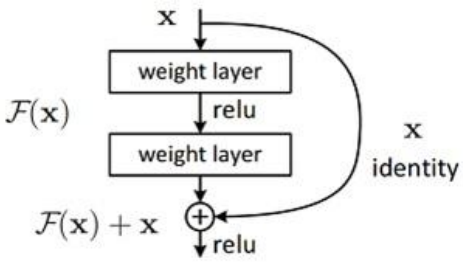

Fig. 5. Residual learning; the idea of shortcut connections [18].

V. EXPERIMENTAL RESULTS

The proposed network has tested with two datasets JAFFE [9] and Yale [10] with seven emotions of both males and females.

A. Illumination Invariant Reconstructions

Training a deep learning model to deal with illumination is a challenging task due to several factors, such as limited multi-illumination training data, or the large distribution of data containing different illumination variations which cause the search space to grow.

B. Databases

JAFFE [16]: The Japanese Female Facial Expression (JAFFE) database contains 213 images of female facial expression expressed by 10 subjects. Each image has a resolution of 256×256 pixels with almost the same number of images for each category of expression. Each person has seven types of facial expressions: anger, disgust, fear, happiness, neutral, sadness, and surprise.

Yale [9]: The Yale facial expression database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject. Each subject exhibited one of the six facial expressions; ha, ne, sa, sleepy (sl), surprise (su) and wink (wi).

C. Data Augmentation

It is important to consider overfitting in deep convolutional neural network. Overfitting causes the model performs execessively well on the training data but leads to poor performance in validation and test data. Data augmentation and dropout are two primary ways to prevent overfitting. For data augmentation, the easiest and the most common way are artificially reproduce new training instances from the existing ones. This increases the dataset size. In dropout, randomly selected neurons are ignored during training. In this paper, considering the images in the two datasets are grayscale, we do not manipulate the contrast, brightness or color. First, we take a random rotation on the images to capture different-angle invariance. Second, we flip the images horizontally to capture the reflection invariance. Finally, we randomly shift the images to capture the translation invariance. For each image, the following operations are performed. Every image is scaled and tilted the resultant images look as shown in Fig. 6.

Fig. 6. (a1-a2) Original images (b1-b2) rotated images (c1-c2) scaled images.

D. Comparison between different models

The proposed models are compared with the pre-trained models on ImageNet such as AlexNet and ResNet-101using by applying transfer learning. It is evident from Table- III that the proposed models have given significantly high accuracy on JAFFE and Yale datasets.

Table- III. Accuracy of the Pre-trained and proposed network architectures.

Network	Database	
	JAFFE	Yale
AlexNet	78.25	76.5
Resnet101	80.79	78.6
Proposed model 1	92.8	-NA-
Proposed model 2	-NA-	91.3

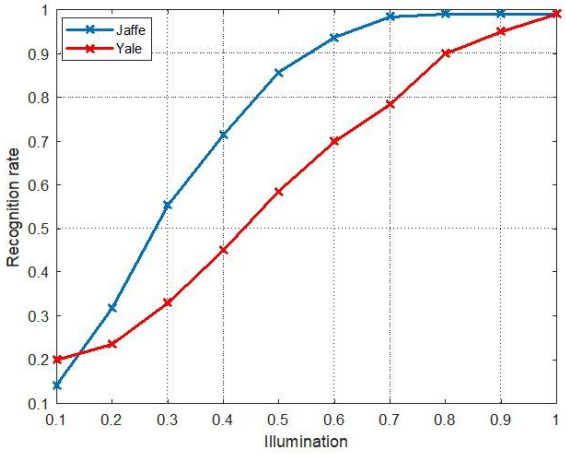

Fig. 7. Recognition rate versus change in the illumination for the trained images

In Table-IV, the performance of the proposed model has shown. Each row shows how often the labels in each column has been predicted for that given ground truth. The correct predictions are found along the diagonal in highlighted bold text.

Table-IV. Confusion matrix of proposed network.

	Anger	Disgust	Fear	Happy	Normal	Sad	Surprise
Anger	**1**	9.46e-17	2.48e-12	1.66e-18	2.31e-23	1.04e-28	4.39e-39
Disgust	8.49e-6	**0.8**	1.91e-6	2.71e-19	3.2e-26	6.52e-23	1.35e-23
Fear	1.1e-14	4.52e-21	**0.35**	1.67e-23	5.00e-17	2.43e-23	9.21e-20
Happy	6.9e-20	7.14e-20	7.24e-20	**0.25**	1.16e-13	6.91e-20	9.18e-20
Normal	7.4e-20	7.34e-20	4.01e-18	6.95e-11	**0.2**	1.09e-16	4.28e-18
Sad	6.3e-18	4.2e-15	9.08e-12	7.23e-10	3.2e-10	**0.16**	6.06e-15
Surprise	5.1e-15	5.19e-15	1.31e-10	8.19e-15	8.06e-12	1.05e-13	**0.142**

The training set is constructed after the data augmentation process. So a total of 642 images for the JAFFE dataset and 225 for Yale database were generated. Out of these samples, 459 from JAFFE and 157 from Yale were used for training the network. The proposed architecture is compared with AlexNet and ResNet 101 [10] and it is observed when simulated on MATLAB 2018a version, that the accuracy of the proposed architecture is yielding 11~12% more than these earlier networks. Fig. 7 depicts the variation of recognition rate under varying illumination and it can be observed that the network is sustainable for 30% variation in the illumination. From the graph, it is observed that when the illumination is decreased by 70% still it acheives the recognition rate of 98% for the trained samples. However, when tested with unknown samples this recognition rate is decreased slightly to the tune of 92% with JAFFE and 91.3% with Yale dataset.

VI. CONCLUSION

This paper presents a customized convolutional neural network that is designed to recognize facial expression.

The network is trained with the CPU and tested with two datasets and found that the proposed models attaining 92.8% and 91.3% recognition rates on JAFFE, Yale datasets respectively. And this rate is higher than those of the earlier networks used here with transfer learning methodology. The network is tested with varying illumination and it is observed that this model is yielding a good recognition rate even at a decrease of illumination by 70%. This work can be further extended with gender recognition too and also the model is designed for grayscale images and this has to be applied for true color images and for the images that are acquired in different environmental conditions.

REFERENCES

1. Haibin Yan, "Transfer subspace learning for cross-dataset facial expression recognition," Neuro-computing, Vol. 208, 2016, pp. 165-173.
2. A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Image-net classification with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097–1105, 2012.
3. K. Simonyan and A. Zisserman "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
4. ShimaAlizadeh, Azar Fazel, "Convolutional Neural Networks for Facial Expression Recognition," CoRR, abs/1704.06756 .
5. Dachapally, Prudhvi Raj. "Facial emotion detection using convolutional neural networks and representational autoencoder units," arXiv preprint arXiv:1706.01509 (2017).
6. Agrawal, Abhinav, and Namita Mittal. "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," The Visual Computer (2019):1-8.
7. Li, Jiaxing, et al. "Facial expression recognition with faster R-CNN." Procedia Computer Science 107 (2017): 135-140.
8. Bairaju, Siva Prasad Raju, A. Sowmya, and Rama Murthy Garimella. "Facial Emotion Detection using Different CNN Architectures: Hybrid Vehicle Driving," (2017).
9. http://vision.ucsd.edu/content/yale-face-database
10. Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks," arXiv preprint arXiv:1605.07146 (2016).
11. Ruiz-Garcia, Ariel, et al. "Deep learning for illumination invariant facial expression recognition," 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018.
12. Söderberg, Erik, and Andreas Jönsson. "Recognizing Spontaneous Facial Expressions using Deep Convolutional Neural Networks," Master's Theses in Mathematical Sciences (2018).
13. Jiang, Xiaoyue, Dong Zhang, and Xiaoyi Feng. "Local feature hierarchy for face recognition across pose and illumination," 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2016.
14. Li, Chongyi, et al. "Lightennet: A convolutional neural network for weakly illuminated image enhancement," Pattern Recognition Letters 104 (2018): 15-22.
15. Tian, Ying-Li, Takeo Kanade, and Jeffrey F. Cohn. "Facial expression analysis," Handbook of face recognition. Springer, New York, NY, 2005. 247-275.
16. Lyons, Michael J., et al. "The Japanese female facial expression (JAFFE) database." Proceedings of third international conference on automatic face and gesture recognition. 1998.
17. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems. 2012.
18. He, Kaiming, et al. "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
19. R. K. Srivastava, K. Gre_, and J. Schmidhuber. Highway Networks. url: https://arxiv.org/abs/1505.00387.
20. Sang, Dinh Viet, and Nguyen Van Dat. "Facial expression recognition using deep convolutional neural networks," 2017 9th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2017.
21. Dong, Xuan, et al. "Fast efficient algorithm for enhancement of low lighting video." 2011 IEEE International Conference on Multimedia and Expo. IEEE, 2011.
22. Abdullah-Al-Wadud, Mohammad, et al. "A dynamic histogram equalization for image contrast enhancement," IEEE Transactions on Consumer Electronics 53.2 (2007): 593-600.

AUTHORS PROFILE

K. Prasada Rao , has received M.Tech from JNTUK, Kakinada, AP, India in 2009. He obtained B.Tech degree in CSE, from JNTU, Hyderabad in 2004. Now he is from pursuing Ph.D in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, AP, India. His areas of research includes Computer vision, Image processing and Pattern

Dr. M. V. P. Chandra Sekhara Rao, has received Ph.D from JNTU, Hyderabad in 2012. He received M.Tech from JNTUK, and MS (Software systems) from BITS, Pilani. His area of research includes Data Mining, Pattern recognition, and Image processing. He is currently working as a professor at Dept of CSE, RVR&JC College of Engineering, Guntur, AP, India. He has published about 30 research articles in National and International.