Recognition of Facial Expression using Landmark Detection in Deep Learning Model

Palak Girdhar[1], Vishu Madaan[2], Tanuj Ahuja[1] and Shubham Rawat[1]

[1]Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, India
[2] School of Computer Science Engineering, Lovely Professional University, Punjab, India

Abstract

With the advent of Deep Learning algorithms and especially Convolutional Neural Network (CNN), is used to extract the important features from face. However, most of the discriminative features come from the mouth region, nose and eyes. Whereas the other regions such as forehead, hair, ears hold a very small role in the analysis. In this paper, we present a deep learning-based approach for facial expression recognition using Landmark detection in CNN, that has the ability to focus on the sensitive area of the face and ignores the less sensitive information. This method of extracting information is known as landmark detection. The proposed work uses CNN with Landmark detection having a learning rate of 0.001 and on 50 epochs. The proposed model concentrates on the chief areas of the face, instead of less important regions of face. The methodology is tested and validated on Jaffe dataset using 10-fold cross validation of 141 images. The empirical results of the proposed methodology show that accuracy of recognition of facial expression increases by 87.5% as compared with state-of-the-art methods of classical CNN with 78.1% using Adam optimizer. The methodology can be further utilized as a base for emotional and behavior analysis using soft computing techniques.

Keywords

Human Computer Interaction, Facial Expression Recognition, landmark detection Convolutional Neural Network, Deep Learning

1. Introduction

Expressions on our face plays a vital role in daily human-to-human communication. Automatic detection of these facial expressions has long been studied due to its potential applications in various domains such as service robots, driver drowsiness monitoring, and intelligent teaching systems. It is also gaining popularity in the field of Human Computer Interaction (HCI). It refers to the interaction of humans and computer technology. It has almost impacted on every area of our daily lives.

It is gaining strength in the area of visual interactive gaming, data driven animations, robotics, surveillance systems and many more. Verbal communication (speech and textual data) and non-verbal communication (face expression, gestures, eye movement, body movement) are the two categories through which human emotions can be expressed. Emotions are the representation of the human nervous system towards the external situations. Brain first sends the instructions for the corresponding feedback which may reflect through human facial expression, pitch of the voice, body movement, gestures also influence human organs like heart rate and brain etc. Face Expressions can be studied for various reasons like: they hold numerous useful features for expression recognition, they are visible, and their datasets are readily available as compared to other expression recognition features. Expressions of the face can be grouped into six principal classes: anger, surprise, sadness, disgust, fear and happiness.

Emotions are a critical part of our communication with another party. It gives clues of our current state of mind even without saying something. Facial expression recognition (FER) has become an active area research due to its applications in medicine, e

ISIC'21: International Semantic Intelligence Conference, February 25-27, 2021, Delhi, India. *https://orcid.org/0000-0002-4042-6001 EMAIL: palakgirdhar@bpitindia.com (P. Girdhar)*

learning, monitoring, entertainment, marketing, human-computer interaction etc. Therefore, there is a need to develop a mechanism to detect emotions. Traditionally, handcrafted features were used along with machine learning algorithms to address this problem. But with the recent triumph of deep learning and especially convolution neural networks (CNN) in tasks such as object recognition, face recognition, and object detection researchers have started to explore these methods in the area of expression recognition [1][2].

Despite achieving excellent results, robust facial expression recognition is still a tough task for existing these deep learning methods because the images in the wild vary a lot in pose, background, occlusion etc. Moreover, deep learning requires a lot of data and relatively small datasets are available for emotion recognition tasks, making the training for deep networks difficult. Therefore, there is a need to develop a system that can accurately identify the emotional state of a person, along with the given constraints. In this research, we aim to artificially increase the size of the dataset using data augmentation methods [4].

Furthermore, motivated by the fact that human observers pay close attention where the expressions are most prevalent, we decide to focus on only the essential parts of the image and ignore other irrelevant parts such as background details as they contribute little or no information [3]. In this proposed method, landmark detection is used to ignore irrelevant features of the image using a tool 'Open Face' developed by Zadeh et. al [9].

Outline: The paper is organized as follows: Section II discusses the state-of-the-art methods used in literature survey. Section III presents the materials and methods used in developing the proposed methodology. Section IV and V show the experimental results and conclusion with future scope respectively.

2 Related Work

Ekman [6] et. at. had performed one of the earlier works in the domain of emotion recognition. They identified the six basic principle emotions namely anger, happiness, surprise, fear, sadness and disgust. This work set the foundation for all the other future work done in the field of emotion recognition. Earlier works on facial expressions mainly involved a two-step process. The first step involves manual extraction of features from the faces by a human or automatically by computer software, then in the second step popular classification algorithms like as support vector machines, or k - nearest neighbours are used to classify emotions. A few of the traditional, well-known, approaches to extract features from images are Gabor wavelets [10], histogram of oriented gradients [11], Haar features [12] etc. These methods worked well on limited sized datasets but didn't generalize well on larger datasets and the datasets, which had more variations. In recent times, deep learning-based approaches specifically CNN have gained a lot of popularity because of their ability to extricate features automatically from the images. Deep learning has been found to perform very well for object recognition and other vision related problems, as a result several researchers have proposed deep learning driven facial expression recognition (FER) models. Recent works have concentrated on building a compounded network and training that network on the input images, the mixture of multiple structures makes the model extremely powerful. Mayya et al. [7] proposed a deep convolutional neural network (DCNN) based approach to identify the facial expressions. They used ImageNet, which is a famous DCNN architecture to extract the facial features. The last layer of the network gave them a dimensional vector, which they plugged into a support vector machine (SVM) classifier to recognize emotion on the faces. They obtained an accuracy of 96.02% and 98.12% for 7 classes of emotions on two separate databases namely CK+ and JAFFE, respectively. Despite achieving competitive results their approach has three major downsides, firstly is it difficult to understand, secondly it is not an end-to-end approach, and lastly it takes a lot of time to train. Zhang et al. [8] proposed a unique well-architecture of CNN, while training the network has the ability to maximize different-emotion differences and at same time minimize same-emotion variations. They used a two-way soft-max activation function which requires a high level of expertise and skill set. However, their model is for smile detection, and the size of the dataset is way larger than any FER dataset, close to 4000 images for a single expression.

Lucy et al. [13] proposed a system based on DCNN using facial parts, it used a mixture of algorithms for feature extraction, face detection and classification. It used a two channel CNN. For the first channel the input was the extracted eyes and for the second channel the input was the extracted mouth. Although most of the previous works obtained notable improvements over the orthodox works on facial emotion recognition, they haven't focused on the principal regions of the face. In this work, we aim to tackle this problem and focus on salient face regions

3 PROPOSED METHODOLOGY

In this section, we discussed the materials and methods used in developing the proposed methodology. The section presents the details of dataset used (JAFFE), use of CNN architecture with landmark detection

3.1 Dataset used

We used JAFFE dataset, a widely used dataset for FER to train our model. The JAFFE dataset contains 7 facial expressions consisting of 10 Japanese female models. In total there are 213 images. The size of each image 128 x 128. The expressions in the dataset include happiness, anger, disgust, fear, surprise, sadness, and neutral with around 30 images each. Figure 1 shows a few images from the Jaffe dataset.

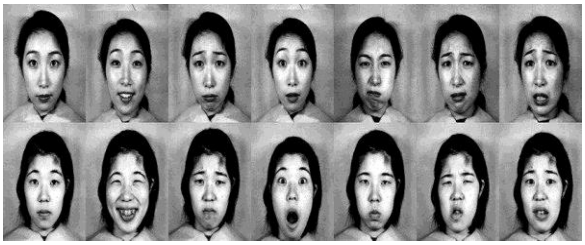


Figure 1: JAFFE dataset sample images

3.2 Convolutional Neural Networks (CNN)

We implemented CNN to build the proposed model. CNNs are known to emulate the human brain when working with visuals. In this research we have used a CNN to extricate facial features and detect the emotions from the dataset. Figure 2 shows the network architecture used for FER. We used three convolution layers, three max pooling layers, one flatten layer and one output layer in our proposed network. The first layer is a convolution layer with 6 kernels and the third layer is also a convolution layer with 16 kernels. Both the kernels have a size of 5 × 5. The second and fourth layers are max pooling layers, with a size of 2 × 2. We used rectified linear unit (ReLU) as an activation function for the CNN and flattened the max pooling layer to get a 14400-dimensional vector. The vector then acts as an input to the output layer which has a soft-max activation function and dropout of 0.5. The CNN uses a glorot uniform initializer, adam optimizer and a cross entropy loss function.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 128, 128, 6)	456
max_pooling2d_1 (MaxPooling2	(None, 64, 64, 6)	0
conv2d_2 (Conv2D)	(None, 64, 64, 16)	2416
activation_1 (Activation)	(None, 64, 64, 16)	0
max_pooling2d_2 (MaxPooling2	(None, 32, 32, 16)	0
conv2d_3 (Conv2D)	(None, 30, 30, 64)	9280
max_pooling2d_3 (MaxPooling2	(None, 15, 15, 64)	0
flatten_1 (Flatten)	(None, 14400)	0
dense_1 (Dense)	(None, 128)	1843328
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 7)	903

Total params: 1,856,383
Trainable params: 1,856,383
Non-trainable params: 0

Figure 2: CNN Network Architecture

3.3 Landmark Detection on JAFFE dataset

Motivated by the fact that human observers pay close attention where the expressions are most prevalent, we decide to focus on only the essential parts of the image and ignore other irrelevant parts such as background details as they contribute little or no information. We have used image cropping to get important features from the images and ignore irrelevant features like hair, ears, neck, etc., which contribute little or no information. [9] Zadeh et. al. has performed landmark detection using modified

Constrained Local Models (CLMs). They developed an end-to-end framework that combines the benefits of mixtures of experts and neural architectures. Their proposed algorithm outperformed state-of-the-art models by a large margin.

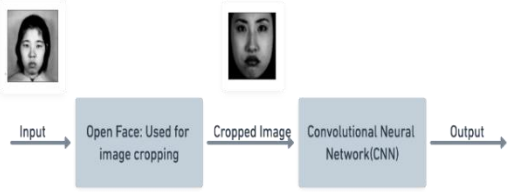
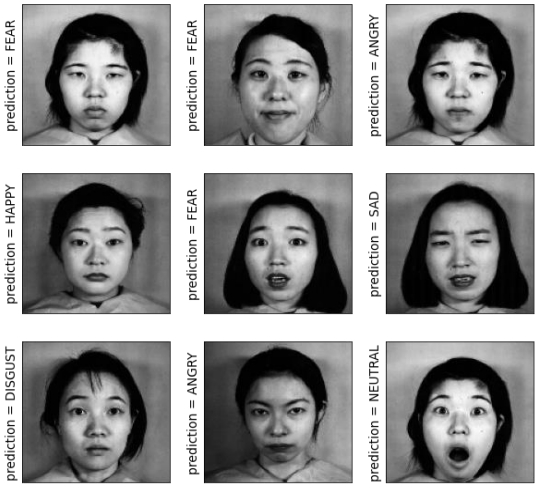


Figure 3: Proposed Approach

The designer of the tool has made the resources available for external use through an open-source software OpenFace. Machine learning researchers, groups keen on working on interactive applications involving facial behaviour analysis and organizations researching in the area of affective computing are the tools main consumers. The tools offer the following capabilities: head pose estimation, eye-gaze estimation, landmark detection and facial action unit recognition. Also, it's an open-source project with available source code for both training the network and the models. The advantages of this tool are two-fold, firstly it comes with real-time performance and secondly it does not require any special hardware.



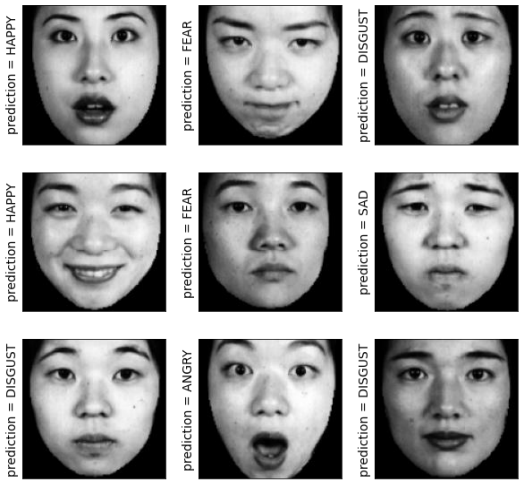


Figure 4 Original Images (top three rows), Cropped Images (bottom three rows) for the JAFFE dataset

In Figure 4 there are nine original images from the JAFFE dataset (in first three rows) and the set of images obtained after cropping (last three rows). It is visible that the cropped images do not contain irrelevant features like hair, neck, ears and background.

4 RESULTS AND ANALYSIS

We tested the proposed model performance on the JAFFE dataset. The proposed model is trained on the subset of the available dataset and validated on the validation set. To measure the accuracy, test dataset is used. Architecture and hyper parameters are kept the same for all the experiments in the training procedure. For the comparison purpose, each model is trained on 50 epochs. The JAFFE dataset is trained on the CPU.

We initialized network weights with a glorot uniform initializer and for optimization purposes and used adam optimiser with a learning rate of 0.001. As in the JAFFE dataset, there are a limited number of images, so it took very less time to train the model.

For training and testing purposes, 10-cross validation has been performed. It is taken care of that every set of data has a balanced distribution of classes.

For the experiment, 141 images are used for the training, 29 images for the validation and 43 images are used to perform testing. And the proposed model

is found to work better with an increased accuracy of almost 9%.

The proposed model is tested on the JAFFE dataset and is validated on 50 epochs with accuracy 87.5%.

Table 1 Accuracies obtained for different models and different datasets.

Method	Optimizer Used	Learning Rate	Dataset	Accuracy
CNN	Adam	0.001	JAFFE	78.1
CNN + Landmark Detection	Adam	0.001	JAFFE	**87.5**

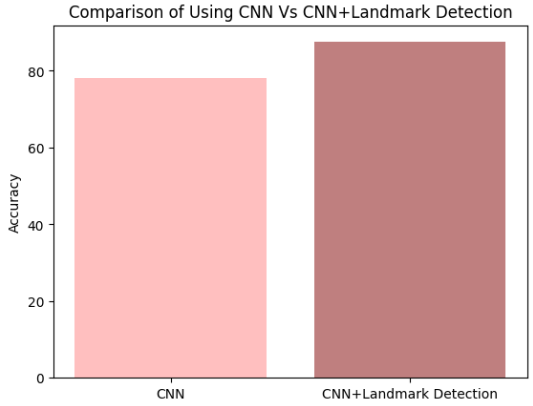


Figure 5 Comparison of the two approaches

Successfully applied landmark detection on JAFFE dataset and the result is tabulated in Table 1. Figure 5 shows the comparison of the two approaches that is using CNN and CNN in addition with landmark detection.

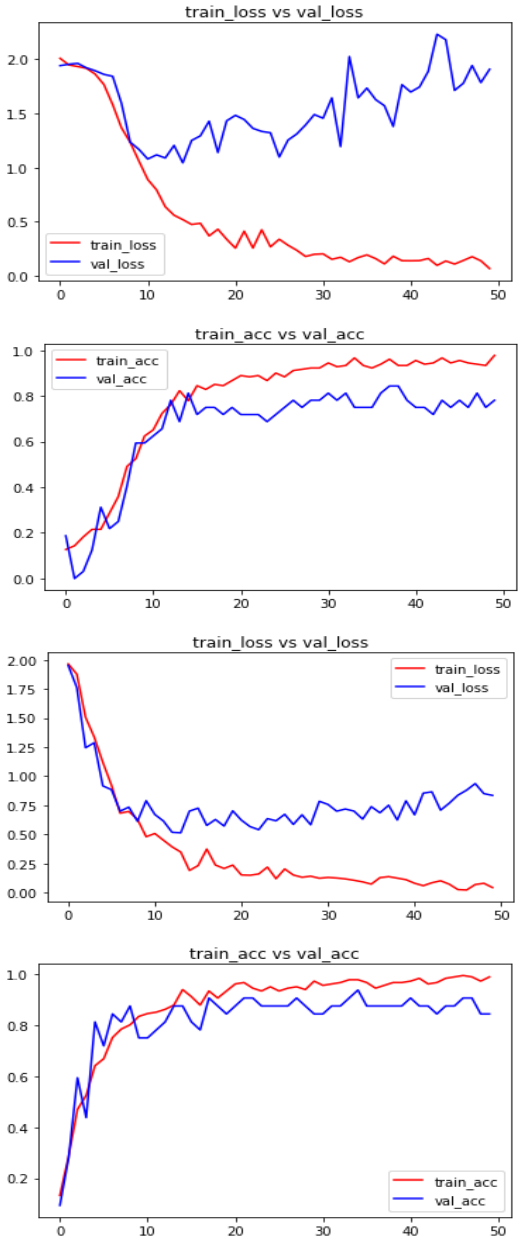


Figure 6: Model loss and model accuracy for experiment I (first two graphs) and experiment II (second two graphs) respectively.

The four graphs in Figure 6 show the model loss and model accuracy with the increase in epoch count. The figure on the top left corner shows that the validation

loss after roughly around 10 epochs does not improve and the model starts overfitting which is clearly visible as the training loss starts to decrease and the validation loss remains the same or increases further. However, the figure on the bottom left corner shows that the gap between the validation loss and training loss is much less as compared to figure on the top left corner. Also, the loss steeply decreases in the first 10 epochs showing the network certainly benefits from the image cropping methodology followed in this research.

5 Conclusion and Future Scope

Proposed an efficient approach for the Facial Expression Recognition. The proposed approach for FER uses Convolutional Neural Networks and facial landmark detection. The use of CNN makes the model more accurate towards the feature extraction and classification process [14-17].

In our proposed work, we applied CNN with landmark detection approach to extract important image features and to remove the irrelevant image features like ear, hair, neck, background, etc. The idea of using landmark detection is to remove irrelevant features from the face that does not contribute or contains very less cue for the analysis. By applying landmark detection, the original face images are cropped. Now, the cropped images are clear enough to read the expression from the human face. The approach is validated on JAFFE dataset. The accuracy achieved with the CNN model is 78.1% and with the proposed method, accuracy is raised to 87.5%. Researchers could certainly use more complex models and achieve higher accuracy but the time would also increase correspondingly. Whereas, in this study we used a simple model which can be trained quickly and have concentrated in making the dataset efficient in terms of the feature set.

This work can be extended to real world applications like in driver's drowsiness detection, pain assessment, lie detection etc.

References

[1] Khorrami, P., Thomas P., and Thomas H., (2015), "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?", *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, pp. 19-27, doi: 10.1109/ICCVW.2015.12.

[2] Han, S., Meng, Z., Khan, A. S. and Tong, Y., "Incremental boosting convolutional neural network for facial action unit recognition", International Conference on Neural Information Processing Systems (NeurIPS'2016), pp. 109-117, 2016.

[3] Minaee, S. and Amirali A., "Deep-emotion: Facial expression recognition using attentional convolutional network", arXiv preprint, arXiv abs/1902.01019, 2019.

[4] Li, K., Yi, J., Akram, M. W., Han, R. and Chen, J., "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy." The Visual Computer 36(2), pp. 391-404, 2020.

[5] Mehrabian, A.: Communication without words. Communication Theory, 2nd Edition, pp. 193-200, Taylor & Francis, 2008

[6] Paul, E. and Friesen, W. V., "Constants across cultures in the face and emotion", Journal of personality and social psychology, 17(2), pp. 124-126, 1971.

[7] Mayya, V., Radhika, M. P., and Manohara, M. M. P., "Automatic facial expression recognition using DCNN", Procedia Computer Science 93 (1), pp. 453-461, 2016.

[8] Kaihao, Z., Huang, Y., Wu, H. and Wang, L., "Facial smile detection based on deep learning features." 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 534-538. IEEE, 2015.

[9] Zadeh, A., Yao, C. L., Baltrusaitis, T. and Morency, L. P., "Convolutional experts constrained local model for 3d facial landmark detection", International Conference on Computer Vision Workshops, pp. 2519-2528, 2017.

[10] Lee, T. S., "Image representation using 2D Gabor wavelets", IEEE Transactions on pattern analysis and machine intelligence, 18(10), pp. 959-971, 1996.

[11] Nigam, S., Singh, R., and Misra, A. K., "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain",

Multimedia tools and applications, 77(21), pp. 28725-28747, 2018.

[12] Wilson, P. I., and Fernandez, J., “Facial feature detection using Haar classifiers”, Journal of Computing Sciences in Colleges 21(4), pp. 127-133, 2016.

[13] Lucy, N., Wang, H., Lu, J., Unwala, I., Yang, X. and Zhang, T., “Deep convolutional neural network for facial expression recognition using facial parts”, International Conference on Pervasive Intelligence and Computing, pp. 1318-1321. IEEE, 2017.

[14] Girdhar, P., Virmani, D. and Kumar, S. S., “A hybrid fuzzy framework for face detection and recognition using behavioral traits”, Journal of Statistics and Management Systems 22(2), pp. 271-287, 2019.

[15] Agrawal, P., Madaan, V., Kundu, N., Sethi, D. and Singh, S. K., "X-HuBIS: A fuzzy rule based human behavior identification system based on body gestures", Indian Journal of Science and Technology, 9(44), pp. 1-6, 2016.

[16] Kaur, G., Agrawal, P., “Optimization of image fusion using feature matching based on SIFT and RANSAC”, Indian Journal of Science and Technology, 9(47), pp. 1-7, 2016.

[17] Agrawal, P., Chaudhary, D., Madaan, V., Zabrovskiy, A., Prodan, R., Kimovski, D. and Timmerer, C., “Automated bank cheque verification using image processing and deep learning methods”, Multimedia Tools and Applications, pp. 1-32, 2020. https://doi.org/10.1007/s11042-020-09818-1