A Self Learning Emotion Recognition Approach using Deep Learning

**1Anisha Shaikh, 2Bhumika Singh, 1Geetanjali Agarwal and 4Lokpriya Kumawat**

**1Department of Computer Science and Engineering, Geetanjali Institute of Technical Studies, Udaipur**

**2,4Department of Electrical Engineering, College of Technology and Engineering, Udaipur**

## ABSTRACT

The facial expression of a person varied one to another, while it makes great sense of humour to predict the expression of a person which helps to get the idea how to interact with them. Detection of Human Facial Expression make the task autonomously that can help for the financial institute, education sector to grow and interact with person they are dealing with them. As far back as PCs were created, researchers and engineers thought of artificially insightful frameworks that are rationally as well as physically proportionate to people. In the previous decades, the expansion of, for the most part, accessible computational power gave assistance for growing quick learning machines, though the web provided a huge measure of information for preparing. In this paper, we are providing a research-based technique that can be implemented through Convolutional Neural Network. That shifts the traditional based approach to new generation machine learning technique that connects with facial dataset which run over Tensorflow to get result. The Facial Expression Recognition Challenge (FERC-2013) Dataset containing 32000 wild images has been used with the CNN to training and testing purpose. The Alex Net Architecture is applied for the whole processing to make the training faster in the minimum system requirement.

1. **Introduction**

As far back as PCs were created, researchers and engineers thought of artificially insightful frameworks that are rationally as well as physically proportionate to people. In the previous decades, the expansion of, for the most part, accessible computational power gave assistance for growing quick learning machines, though the web provided a huge measure of information for preparing. These two advancements helped the exploration of shrewd self-learning frameworks, with neural systems among the most encouraging procedures.

## BACKGROUNDS

One of the present best uses of artificial intelligence utilizing neural systems is the acknowledgement of appearances in photographs and recordings. Generally, strategies process visual information and look for general examples present in human countenances. Face acknowledgement can be utilized for reconnaissance purposes by law masters as well as in group the board. Other present-day applications include programmed obscuring of appearances on Google Streetview film and programmed acknowledgement of Facebook companions in photographs.

A much further developed advancement in this field is feeling acknowledgement. Notwithstanding just recognizing faces, the PC utilizes the course of action and state of for example eyebrows and lips to decide the outward appearance and henceforth the feeling of a each child. One conceivable application for this lies in the territory of observation and social investigation by law enforcement. Moreover such methods are utilized in advanced cameras to naturally take pictures when the client grins. That’s why emotion recognition plays an important role in improving human behaviour with machine interaction.

## RESEARCH OBJECTIVE

In this paper we for the most part centre around the neural system based artificially wise frameworks equipped for determining the feeling of an individual through photos of his or on the other hand her face. Distinctive methodologies from existing writing will be tried different things with and the consequences of different decisions in the planned procedure will be assessed. After the final model processed performs the final assessment that shown in the result section. It might be noticed that the point of our work isn't to structure a feeling recognizer sans preparation but instead to audit plan decisions and upgrade existing procedures with a few new thoughts.

# LITERATURE REVIEW

For the improvement of a framework that can recognize feelings through outward appearances, past look into in transit people uncover feelings also as the hypothesis of programmed picture arrangement is seen again. In the first part of this area, the job of deciphering outward appearances in feeling acknowledgement will be talked about. The last part overviews past contemplate on programmed picture classification. The regular ordinary methodology for the assignment of facial include acknowledgement basically pursues the pipeline. Lion's share of the customary and business facial investigation techniques depend on the Facial Action Coding Framework (FACS), which includes distinguishing different facial muscles that can cause changes in physical facial appearance. utilizes a model based methodology called the Active Appearance Model to group feeling while at the same time constructing a 3-D model of the face that encodes more than 500 facial tourist spots from which facial strong developments (Action Units, characterized by the FACS) can be inferred. The Active Appearance Model is created utilizing PCA specifically on the pre-handled pixels and is encoded as the deviation of a face from the normal face. This model is then used to order the feelings communicated by the face utilizing a solitary layered neural system.

In the assignment of feeling acknowledgement from faces, Tang's sets the best in class on the Facial Expression Recognition Challenge (FERC) dataset. This is accomplished by executing a two-arrange organize: a convolutional arrange prepared in an administered way on the primary stage and a Support Vector Machine as the second stage-prepared on the yield of the main stage. Late work by

Kahou et al. effectively exhibits a multi-modular profound learning based structure for feeling acknowledgement in recordings.

# PROPOSED WORK & EXPERIMENT ISETUP

To assess the three approaches mentioned previously on their capability of emotion recognition, we developed three networks based on the concepts. This section describes the data used for training and testing, explains the details of each network, and evaluates the results obtained with all three models.

## DATASET

Neural networks, and deep networks in particular, are known for their need for large amounts of training data. Moreover, the choice of images used for training are responsible for a big part of the performance of the eventual model. This implies the need for a both high qualitative and quantitative dataset. For emotion recognition, several datasets are available for research, varying from a few hundred high resolution photos to tens of thousands smaller images. The three we will discuss are the Facial Expression Recognition Challenge (FERC-2013) [8], Extended CohnKanade, and Radboud Faces Databas (RaFD).

The datasets differ mainly on quantity, quality and 'cleanness' of the images. The FERC-2013 set for example has about 32000 low resolution images, where the RaFD provides 8000 high resolution photos. Furthermore it can be noticed that the facial expressions in the CK+ and RaFD are posed (i.e.'clean'), while the FERC-2013 set shows emotions 'in the wild'. This makes the pictures from the FERC-2013 set harder to interpret, but given the large size of the dataset, the diversity can be beneficial for the robustness of a model. We reason that, once trained upon the FERC-2013 set, images from 'clean' datasets can easily be classified, but not vice versa. Hence for the three networks under consideration, training will be done using 9000 samples from the FER-2013 data with another 1000 new samples for validation. Subsequently testing will be done with 1000 images from the RaFD set to get an indication of performance on clean high quality data. This latter set has an even distribution over all emotions. Please note that non-frontal faces and pictures with the label contemptuous are taken out of the RaFD data, since these are not represented in the FERC-2013 training set. Furthermore, with use of the Haar Feature-Based Cascaded Classifier inside Emotion Recognition using Deep Convolutional Neural Networks TU Delft IN4015 the OpenCV framework, all data is preprocessed. For every image, only the square part containing the face is taken, rescaled, and converted to an array with 48x48 grey-scale values.



Figure 1: Number of images per emotion in the training set

## NETWORKS

The networks are programmed with use of the TFLearn library on top of TensorFlow, running on Python. This environment lowers the complexity of the code, since only the neuron layers have to be created, instead of every neuron. The program also provides real-time feedback on training progress and accuracy, and makes it easy to save and reuse the model after training. More details on this framework can be found in reference.

* + 1. The system comprises of three convolutional layers and two completely associated layers, consolidated with max-pooling layers for decreasing the picture estimate and a dropout layer to diminish the opportunity of over fitting. The hyperparameters are picked with the end goal that the quantity of counts in each convolutional layer remains generally the equivalent. This guarantees that data is protected all through the system. Preparing is performed utilizing contrasting quantities of convolutional filters to assess their impact on the execution.
		2. In 2012, the AlexNet convolutional organize was produced for grouping pictures in more than 1000 unique classes, utilizing 1.2 million example pictures from the ImageNet dataset. Due to the certainty that in this exploration the model just needs to recognize seven feelings, and due to our limited figuring assets, the measure of the first organize is viewed as excessively vast. Thus, rather than 5 convolutional layers, we employed 3, and in the ensuing 3 completely associated layers the number of hubs of each completely connected was decreased from 4096 to 1024. While the first system was separated for parallel preparing, it was seen that was redundant for the littler variant. The system additionally makes utilization of nearby standardization to accelerate the training and dropout layers so as to decrease the overfitting.

The keep going tests are performed on a network-based crafted by Gudi. Since this explore likewise pointed on perceiving 7 feelings utilizing the FERC-2013 dataset, the design bought to be a decent beginning stage for our examination. The first system begins with an info layer of 48 by 48, coordinating the measure of the info information. This layer is trailed by one convolutional layer, a neighbourhood differentiates standardization layer, and a maximum pooling layer individually. The system is finished with

two more convolutional layers and one completely associated layer, associated with a delicate max yield layer. Dropout was connected to the completely associated layer and all layer contain ReLu units. For our exploration, a second max-pooling layer is connected to lessen the number of parameters. This brings down the computational force of the arrange, while the decrease in execution is professed to be just 1-2%. Besides the learning rate is balanced. Rather than directly decreasing the learning rate as done by Gudi, we trust a learning rate which makes utilization of momentum would meet quicker, as the momentum expands the learning rate when the gradient props up a similar way.

## EVALUATION

All networks are trained for 60 epochs with the data mentioned. Various details of the training process and the final model. For network A, the final accuracy on the validation data is around 63%. Already after 1 epochs, the accuracy raised above 60%, indicating quick learning capabilities. Furthermore it is noteworthy that adjusting the alter dimension did not have a big influence on the accuracy, though it has 5 on the processing time. This means that fast models can be made with very reasonable performance. Surprisingly, the second, much larger, network learns quickly as well, but converges to an accuracy of about 54%. Apparently reducing the network size breaks down the promising performance of the original network more than expected. Together with the much higher computational intensity, and therefore slower live performance, this model is not a worthy challenger of the other two architectures. Network C shows a somewhat slower learning curve, but the final accuracy on the validation set is similar to that of network

A. The processing demands are in between that of the other networks, so based on this fact, network A seems to be the most promising approach for our emotion recognition task. However, the performance of network C on the extra RaFD testset is signi\_cantly better (60%) than that of network A (50%). This indicates better

generalizing capabilities, which is very important for future applications. Hence, in the next chapter, the model from network C will be further investigated and tested.

Figure 2: Overview of the network architecture of the model

1. **Final Models**

The last described network from section 3.2 was observed to have the most promising performance for practical applications. An overview of its architecture. The source used for this network, as well as other scripts used for this project can be found on.

As can be seen from the figure the accuracy seems to still increase in the last epochs. We therefore will train the network for 100 epochs in the final run, to make sure the accuracy converges to the optimum. In an attempt to improve the final model even more, the network will be trained on a larger set than the one described previously. Instead of 9000 pictures, training will be done with 20000 pictures from the FERC-2013 dataset. The ratios of the emotions present in this set are given. Newly composed validation (2000 images) and test sets (1000 images) from the FERC-2013 dataset are used as well, together with the well-balanced RaFD test set from the previous experiment.



Figure 3: Number of images per emotion in the final training set

# RESULTS

The exactness rates of the last model. On all approval and test sets the exactness was higher than amid past runs, underlining that more information and longer preparing can improve the execution of a system.

Table 1: Details of Trained Data Accuracy of the network

| Network | FERC-2013 | RAFD |
| --- | --- | --- |
|  | Validation | Test |  |
| Final | 67% | 63% | 71% |

# CONCLUSION

The final output of the research module conclude that the emotion recognition through Deep learning can gently overtake the tradition module over wild images that mostly come into our real life scenario than clear & vivid images. Hence satisfactory output can be generated by collecting more and more samples images over time. it is hard to objectively assess, the live application shows promising performance. Though, it encounters problems when shadows are present on the face of the subject. All emotions are easily recognized when acted by the user, and when pointing the camera on the television, most emotions

in the wild can be classified. This once again emphasizes the power of using neural network based models for future applications in emotion recognition.

The project can be forecast on CCTV module through IOT implementation where the data can be used to evaluative the personality behaviour of a person which ultimately profit the organization for the development of the business. While mobile application can be further developed through the cloud implementation which can track the emotion behaviour as a track record.

# REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012, pp. 1097–1105.
2. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.
3. R. Girshick, "Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.
4. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems 28 (NIPS 2015), 2015, pp. 91–99.
5. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779– 788.
6. W. Liu et al., "SSD: Single shot multibox detector," in European Conference on Computer Vision (ECCV), 2016, pp. 21–37.
7. T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
8. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961–2969.
9. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
10. A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.
11. M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10781–10790.
12. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4203–4212.
13. T. Y. Lin et al., "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.
14. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, 2010.
15. A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
16. J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in Advances in Neural Information Processing Systems 29 (NIPS 2016), 2016, pp. 379–387.
17. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
18. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
19. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7263–7271.
20. X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint arXiv:1904.07850, 2019.
21. K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in European Conference on Computer Vision (ECCV), 2014, pp. 346–361.
22. S. Liu et al., "Path aggregation network for instance segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8759–8768.