IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Review of Computer Engineering Studies
Vol. 6, No. 3, September, 2019, pp. 64-68
Journal homepage: http://iieta.org/journals/rces

3D Facial Emotion Recognition Using Deep Learning Technique

Pankaj Rao, Ashish Choudhary, Vijay Kumar*

Department of Computer Science and Engineering, National Institute of Technology Hamirpur, Himachal Pradesh 177005, India

Corresponding Author Email: vijaykumarchahar@nith.ac.in

https://doi.org/10.18280/rces.060303

Received: 15 May 2019
Accepted: 20 August 2019

Keywords:
face recognition, computational intelligence techniques, convolutional neural networks, depth map, multi view

ABSTRACT

In this paper, a 3D facial emotion recognition model using deep learning technique is proposed. In the deep learning architecture, two convolution layers and a pooling layer is used. Pooling is performed after convolution operation. The sigmoid activation function is used to obtain the probabilities for different classes of human faces. In order to validate the performance of deep learning based face recognition model, Kaggle dataset is used. The accuracy of the model is approximately 65% which is less than the other techniques used for facial emotion recognition. Despite dramatic improvements in representation precision attributable to the non-linearity of profound image representations.

1. INTRODUCTION

2D facial recognition has been actively studied since the 17th century. These methods need statistical models to perform the tasks of classification. Three common 2D models of face detection and identification are the Active Look, Active Form Model, and Constrained Local System. In many restricted and unconstrained settings, the techniques have proved their usefulness and identification reliability differs with various illuminations and orientations. All methods rely primarily on 2D entity characteristics and shape statistical models to carry out classification tasks. Three specific techniques for face detection and recognition based on 2D are the Active Face method, the Active Form framework, and the restricted local design. While these 2D methods have proved effective in many restricted and limited situations, their detection quality varies with specific illuminations. They capture more information in order to prevent the adverse effects caused by external factors.

Cohen et al. [1] suggested a system for analyzing facial expression using 3D laser scanners based on human-computer communication data. Blanz and Vetter [2] came up with method based on 3D face recognition for the decomposition of 3D image data from the key property research. Newcombe et al. [3] which can also be used in facial mapping, introduced recent advances in RGBD cameras to make the 3D test faster, less costly and more precise. Chen et al. [4] displayed a method for recording facial exposure. Thies et al. [5] revealed a continuous-looking procedure for facial reconstruction. Since a depth camera offers 3D information despite the usual 2D images, it increases position and identification precision and consequently facilitates further applications.3D information can also be generated for 3D deep learning methods. Sinha et al. [6] developed a CNN- dependent 3D surface modeling strategy by generating a graphical object from the 3D shapes info. Su et al. [7] used multi-view techniques to transform 3D structures to some 2D objects and using them to configure a multi-view CNN for shape training. In summary, as a contribution to the CNN framework, these techniques always move 3D shapes into 2D images. Wu et al. [8] used 3D voxel channels to handle voxelization depth information instead of doing 2D operations in CNN. Our system has far exceeded current methodologies for assignments of type identification. In the case of an exterior image identification project, such tests can be done outside the 3D facial template reproduced.

To achieve a consistent examination frame on the 3D faces, 3D facial models are replicated by applying a deformable facial design to the surfaces being inspected in which a dense vertex match will normally be achieved. A visual inspection method is used for dealing with the influential points. These are critical for modifying and optimizing the execution of better networks through an intuitive description of the informed strengths. The proposed system will display focus groupers dependent on their activation habits by giving customers an intuition. It helps to calibrate the trained framework and streamline for over-learning issues.

The remaining section of this paper is as follows. Section 2 describes related works done in the field of face recognition. Section 3 describes proposed approach. Experimental results and discussions are presented in Section 4. The concluding remarks are given in Section 6.

2. RELATED WORK

This field incorporates the most related work that has occurred to date, the template, for face. In the important research assistance, the study relevant to this field is tabulated and distributed in a chronological fashion to demonstrate some of the significant disputes that have been consigned. There are many deep learning methods that have been used to date to overcome some of the most significant difficult and regressive challenges and progress the research, such as Convolutional Neural Network, Recurrent Neural Networks, and Strong Short-Term Memory. CNN has the ability to assess objects and evaluate any research related to Computer Vision, including interpretation, comprehension and identification.

Jiang et al. [9] came up with a technique commonly known as Multiple Graph Embedding, a mechanism for analyzing various patterns from the sample of the disorder. Using a noisy several graph embedding, they use this method for the crawl facial recognition of solitude- insulated IoT applications. Necessarily three service givers were taken into account in this arrangement for security and the goal of assessment. The repository is same to the record of Japanese women face emotions, directory of MUG expressions, and database of Cohn-Kanade [10]. The authors [11] used the fixed wavelet transform to interpret facial expression to take out the facial attribute. Both spectral and spatial realms were regarded. For increasing the dimension of the function, a distinct cosine transform is used. For process the leading aural chain branch, the backpropagation steps are used as a classifier. The three JAFFE, Ck and MS-Kinect servers were used to obtain reliability of roughly 98.83%, 96.61%, and 94.28%. Du et al. [11] came out with the recognition mixture of facial expressions and the properties of the 22 separate categories of emotions. All he did is use the study of "Facial Activity Coding System" to demonstrate the output of these 22 separate categories of emotions and also publish the server that is used in this research after doing this. The file includes the 5000 Photos mark with the 7 normal human emotions and the 15 different basic human emotions variations. This method suggested that by using just geometric specifics, 73.61 percent accuracy was achieved and by using aspect details, 76.91 percent accuracy was achieved by using the database created. Benitez-Quiroz et al. [12] suggested a model that could accommodate the study of facial expression, the model dubbed this Continuous model. Throughout their study, a diverse array of emotions can be measured by linear frit disconnected continuous facial spaces. The author also showed how the established template can be used to detect recognition of the face and also give an advanced direction in the area of machine learning and computer vision that will enable the societies concerned to step forward in this area. The data set called EmotionNet was generated by this proposed model. Throughout their function, both mathematical and adumbration were captured. Symmetrical attributes are in sync with deviation and height of different metrics of the head. Adumbration characteristics were implemented using the Gabor tub, in which essentially a template was created to catch improvements due to skin domain variability [13-15]. Majumder et al. [16] implemented a deep facial expression recognition neural network that subsists as the classifier focused on Autoencoders and Self-Organizing Graph. Usage of the auto-ciphers, assembling the pictorial presentation with the regional organization presentation have a better estimation of facial expression. The outcome of this proposed method was 97.55 percent accuracy for the MMI database and 98.95 percent accuracy for the CK+ database. Zhao et al. [13] suggested a system for understanding micro-expression by using frames near to the image apex. They used the Eulerian method to increase the intensity or magnify the subtle changes to improve the understanding of the phrase. Their proposed method has done great work on the approaches based on the apex frame. This new method's key idea is to reduce the materialistic confrontation and maximize the visual activity. Architecture of the ELRCN. ELRCN used CNN to derive spatial characteristics and LSTM to understand how these spatial characteristics apply temporally. In the CASME II index, they finally got 50.0 percent F1-score. Berretti et al. [17] suggested a two-stream 3D CNN pre-trained model focused on datasets for macro-expression. ELRCN design was developed by Alyuz et al. [18]. For the spatial presence and temporal interaction between the spatial presentation, CNN and ELRCN were used. On the CASME II list, they hit around 50.0 percent of F-1 rating.

Zhang et al. [15] used a tool to derive temporal characteristics and spatial characteristics to integrate softmax and model emotions. Alyuz et al. [18] proposed a fully automatic and effective 3D face recognition method, which is robust to face occlusion. Colombo et al. [19] proposed a brand-new recovery strategy that can effectively recognize 3D faces even when faces are partially occluded by unforeseen and unrelated objects (such as scarves, hats, glasses, etc.). The occlusion region is detected by considering their influence on the face projection in suitable face space. Tang et al. [20] proposed a local binary model (LBP) based on a 3D facial segmentation scheme.

3. PROPOSED APPROACH

3.1 Motivation

There is a need to maintain information security in today's networked world. The crime rate is increasing day by day in state like Assam. There are no automated systems that can monitor the actions of a human. If we can constantly monitor people's facial expressions, then we can easily find the suspect as facial expressions change with different activities [21-23]. Therefore, a system is required to recognize facial expressions. The above facts motivate us to develop a system that recognizes facial expression and tracks the activity of one person.

3.2 Proposed algorithm

The input layer has defined dimensions that are pre-determined, so the image must be pre-processed before it can be fed into the layer. For practice, validation and evaluation, structured grayscale images of 48 X 48 pixels from the Kaggle dataset is used. Images from the system are used for testing purposes, and image pixel values are read using open CV.

Batch processing is used to render convolution and pooling. Each batch has N images and these batches are modified with CNN filter weights. An object batch input of four dimensions N X Color-Channel X width X height is taken from each convolution surface. There are also four-dimensional feature maps or converting filters (number of feature maps in, number of feature maps out, filter width, height of the filter). Four-dimensional convolution is determined between the batch of images and the feature maps in each convolution layer. Only the parameter that changes after convolution is the width and height of the object.

There are two popular methods of pooling Max and Average Pooling. Max pooling is done in this project after convolution. The pool size of (2x2) is 12, splitting the image into the block grid of 2x2 size each and taking up to 4 pixels.

Only height and width will be affected after pooling. In the architecture, two convolution layers and a pooling layer is used. The picture batch is N X 1 X 48 X 48 at the input's first convolution layer width. The image batch size here is N, the number of color channels is 1 and the height and width of the image are 48 pixels. Convolution is a N x 20 x 44 x 44 scale image batch with a function map of 1x20x5x5 performance.

Pooling is performed after convolution with a pool size of 2x2, resulting in the N x 20 x 22 x 22 scale object array. This is followed by a second convolution layer with a 20 x 20 x 5 x 5 x 5 feature map resulting in the Nx20x18x18 size object pool. This is accompanied by a pooling surface of pool length 2 X 2, resulting in a batch of images of size N X 20 X 9 X 9. This way neurons send messages through the brain inspires this surface. It requires a large number of input features and transforms features by linking layers with trainable weights. the fully connected network, two hidden layers of size 500 and 300 units are used. These layers' weights are learned by forwarding training data propagation and then reverse propagation of their errors. Backpropagation starts from determining the difference between prediction and true-value and back-calculating beforehand the weight adjustment needed for each surface. By tuning the hyper-parameters, such as learning rate and network, we can monitor the training speed and complexity of the architecture. The output from the second pooling layer is Nx20x9x9 and the input from the first secret layer is N X 500. The output of the pooling layer is flattened to the size of N X 1620 and fed to the first layer that is hidden.

The production is fed to the second hidden layer from the first hidden layer. The second secret layer is Nx300 in size and its output is supplied to the size output layer equivalent to several categories of facial expression. Output from the second hidden layer is connected to the output layer having seven distinct classes. Using the sigmoid activation function, the output is obtained using the probabilities for each of the seven classes. The class with the highest percentage is the predicted class. Algorithm 1 depicts the proposed approach. Figure 1 shows the flow diagram of proposed approach.

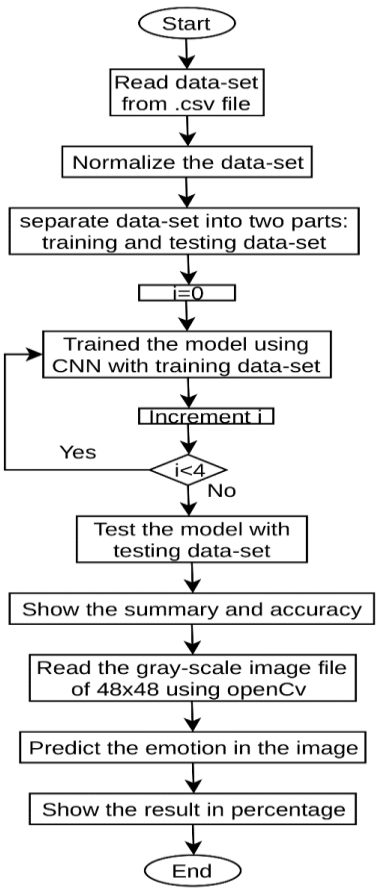

Figure 1. Flowchart of proposed approach

4. EXPERIMENTAL RESULTS

4.1 Database used

This project uses the Kaggle dataset. The data consists of facial images with a gray scale of 48x48 pixels. The faces are automatically registered so that the face is more or less centered and in each photo occupies about the same amount of space. The goal is to categorize the face (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral) into one of seven categories.

The emotion column contains a numerical code for the emotion present in the picture, ranging from 0 to 6 inclusive. The column of pixels contains a sequence of quotes for each image.

Algorithm 1: Facial Emotion Recognition Algorithm (FERA)

Input:
48X48 grey-scale image file,
Output:
Summary of testing phase including accuracy of the model.
Predicted emotion in the image given for prediction(% of each emotion)

begin
label_emotion=['Anger','Disgust','Sad','Happiness','Surprise','Fear','Neutral']
dataset = read(dataset.csv)
***for** each pixel in dataset:*
Divide each pixel with 255.0
Split the dataset into two parts (i.e., training and testing)
***for** layer from 1 to 4:*
Train the dataset with CNN model.
Predicted the values from training dataset.
Print the model summary.
end for
Read the image for testing.
Divide each pixel with 255.0
Predicted the values of testing dataset.
end for
return the result of each emotion in percentage.
end

4.2 Performance metrics

The evaluation metrics of the facial emotion recognition approaches are crucial because they provide a standard for a quantitative comparison. The evaluation metrices are classified into four methods using different attributes such as precision, recall, score, and accuracy [5].

4.3 Results and discussions

Four convolutional layers and two fully-connected layers are used with the Kaggle dataset. And the model was trained for 150 epochs with a batch size of 128. Batch Normalization is also applied to the dataset to get the best classification results. The accuracy of the model is approximately 65% which is less than the other techniques used for facial emotion recognition. The performance of the proposed model is given in Table 1.

Table 1. Performance obtained from the proposed model

Total Params	**4,478,727**
Trainable Params	**4,474,759**
Non-Trainable Params	**3,968**
Accuracy on test set	**0.6544998606854276**

This model is tested for the different types of images that are taken from the internet or camera. This model takes the input of a 1x48x48x1 gray-scale image and then predicts the percentage of every emotion and gives the result. One constraint in this model is that it can take only a 48x48 image. Table 2 shows the results obtained from testing data.

Table 2. Accuracy obtained from different emotions using proposed model

Emotion	Accurancy (in %)
Neutral	99.9047762547
Happy	0.0496994878
Sad	0.0265211639
Anger	0.0121861008
Fear	0.0059911908
Disgust	0.0007975856
Surprise	0.0000311941

5. CONCLUSIONS

In this paper, a facial emotion recognition system is proposed. The proposed system ultizes the concepts of deep face for emotion detection. The convolutional neural network is used. The proposed model is tested on Kaggle dataset. Experimental results reveal that the proposed model is able to capture the emotion and identify efficiently.

The proposed model can be extended to detect facial emotion from video sequences. The emotion oriented deep learning architecture can be developed. Internet –of -Things sensors can be in conjection with deep learning for further improvement in the accuracy. 3D models can be used to pose variations, which greatly affect the performance of system. An appropriate fusion method can be used to tackle the illumination and background variability issues.

REFERENCES

[1] Cohen, I., Garg, A., Thomas, T.S. (2000). Emotion Recognition from facial expressions using multilevel HMM. In Proceedings of Neural Information Processing Systems.

[2] Blanz, V., Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9): 1063-1074. http://dx.doi.org/10.1109/TPAMI.2003.1227983

[3] Newcombe, R.A., Fox, D., Seitz, S.M. (2015). DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 343-352. http://dx.doi.org/10.1109/CVPR.2015.7298631

[4] Chen, Y.L., Wu, H.T., Shi, F., Tong, X., Chai, J. (2013). Accurate and robust 3D facial capture using a single RGBD camera. In Proceedings of IEEE International Conference on Computer Vision, pp. 3615-3622. http://dx.doi.org/ 10.1109/ICCV.2013.449

[5] Thies, J., Zollhofer, M., Niebner, M., Valgaerts, L., Stamminger, M., Theobalt, C. (2015). Real time expression transfer for facial reenactment. ACM Transactions on Graphics, 34(6): 183(1-14). http://dx.doi.org/10.1145/2816795.2818056

[6] Sinha, A., Bai, J., Ramani, K. (2016). Deep learning 3D shape surfaces using geometry images. In Proceedings of European Conference on Computer Vision, pp. 223-240. http://dx.doi.org/10.1007/978-3-319-46466-4_14

[7] Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E. (2015). Multi-View convolutional neural networks for 3D shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile. http://dx.doi.org/10.1109/ICCV.2015.114

[8] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA. http://dx.doi.org/10.1109/CVPR.2015.7298801

[9] Jiang, R., Al-Maadeed, S., Bouridane, A., Crookes, D., Celebi, E.M. (2005). Face recognition in the scrambled domain via salience-aware ensembles of many kernels. IEEE Transactions on Information Forensics and Security, 11(8): 1807-1817. http://dx.doi.org/10.1109/TIFS.2016.2555792

[10] Kanade, T., Cohn, J.F., Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp. 46-53. http://dx.doi.org/10.1109/AFGR.2000.840611

[11] Du, S., Tao, Y., Martinez, A.M. (2005). Compound facial expressions of emotion. Proceedings of the National Academy of Sciences, 111(15): E1454-E1462. http://dx.doi.org/10.1073/pnas.1322355111

[12] Benitez-Quiroz, C.F., Srinivasan, R., Martínez, A.M. (2019). Discriminant functional learning of color features for the recognition of facial action units and their intensities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(12): 2835-2845. http://dx.doi.org/10.1109/TPAMI.2018.2868952

[13] Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S. (2016). Peak-piloted deep network for facial expression recognition. In European Conference on computer Vision, Springer, pp. 425-442. http://dx.doi.org/10.1007/978-3-319-46475-6_27

[14] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. http://dx.doi.org/10.1109/CVPR.2016.90

[15] Zhang, Z., Luo, P., Loy, C.C., Tang, X. (2014). Facial landmark detection by deep multi-task learning. In European Conference on Computer Vision, pp. 94-108. http://dx.doi.org/10.1007/978-3-319-10599-4_7

[16] Majumder, A., Behera, L., Subramanian, V.K. (2011). Automatic and robust detection of facial features in frontal face images. In UKSim 13th International Conference on Modeling and Simulation, pp. 331-336. http://dx.doi.org/10.1109/UKSIM.2011.69

[17] Berretti, S., Bimbo, A.D., Pala, P. (2013). Sparse matching of salient facial curves for recognition of 3-D faces with missing parts. IEEE Transactions on Information Forensics and Security, 8: 374-389

http://dx.doi.org/10.1109/TIFS.2012.2235833
[18] Alyuz, N., Gokberk, B., Akarun, L. (2013) 3-D face recognition under occlusion using masked projection. IEEE Transactions on Information Forensics and Security, 8(5): 789-802. http://dx.doi.org/10.1109/TIFS.2013.2256130
[19] Colombo, A., Cusano, C., Schettini, R. (2006). Detection and restoration of occlusions for 3D face recognition. In International Conference on Multimedia and Expo, pp 1541-1544. http://dx.doi.org/10.1109/ICME.2006.262837
[20] Tang, H., Yin, B., Sun, Y., Hu, Y. (2013). 3D face recognition using local binary patterns. Sign Processing, 93(8): 2190-2198. http://dx.doi.org/10.1016/j.sigpro.2012.04.002
[21] Sharma, S., Kumar, V. (2018). Performance evaluation of 2D face recognition techniques under image processing attacks. Modern Physics Letters B, 32(19): 1850212(1-9). http://dx.doi.org/10.1142/S0217984918502123
[22] Sharma, S., Kumar, V. (2019) Transfer learning in 2.5D Face image for occlusion presence and gender classification. Handbook of Research on Deep Learning Innovations and Trends, pp. 97-113. http://dx.doi.org/10.4018/978-1-5225-7862-8.ch006
[23] Meng, W., Mao, C., Zhang, J., Wen, J., Wu, D. (2019). A fast recognition algorithm of online social network images based on deep learning. Traitement du Signal, 56(3): 575-580. http://dx.doi.org/10.18280/ts.360102