2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)

Automated Student Review System with Computer Vision and Convolutional Neural Network

Sadman Chowdhury Siam
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
sadman.siam@northsouth.edu

Abrar Faisal
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
abrar.faisal@northsouth.edu

Niazi Mahrab
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
niazi.mahrab@northsouth.edu

Akm Bahalul Haque
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
bahalul.haque@northsouth.edu

Md. Naimul Islam Suvon
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
naimul.suvon@northsouth.edu

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397164

***Abstract*— Detecting Emotion from a face is a prevalent topic, and computer vision researchers are continuously working on perfecting this challenging task. Facial emotion is widely used in applications like Snapchat, Face apps, Cameras, etc., to predict emotionsfrom faces,to detect smiles, and many more. The task is very challenging, as a person's facial features vary from one to another.As convolutionalneuralnetworkscan detect such complicated features so, the idea of deep learning has been used to tackle this problem.We have tried to predict facial emotion from images and generate reviews from that. Since there are already many works on emotion detection, we focused mainly on its application. Our main aim is to design a system that can generate automated reviews from human emotion. We have labeled different facial expressions with some scores and used that score to predict student review in the classroom.**

Keywords—computervision, convolutional neural network,deeplearning, data normalization, features extraction

I. INTRODUCTION

Facial expressions convey a person's visible emotion that apprises us about the intuitive state and intention of a person and develops interpersonal communication. Emotion can be classified by other factors such as voice, text, or video,but the facial expression is widely used and acceptable. Automatic recognition of facial expressions has drawn considerable attention because of its wide applications. It helps to develop various potential applications in so many different areas such as medical, commercial uses, human-computer interaction, psychology, robotics, and artificial intelligence. Facial expression carries a lot of information about humans. In our project, the system will classify seven different human emotions, including anger, disgust, fear, surprise, happiness, sadness, and neutrality.

Previously, many applications were made using FER (facial emotion recognition), likethe interaction between human-computer, the detection of pain and mental disorders, and understanding psychological behavior,etc. An overview of the early works in FER analysis can be found in this paper [1]. We have come up with a new idea that is entirely different from previous applications. Our main goal is to make a student review system where the system will generate a review based on the mental conditions of students from a class. There are many manual software systems [2] that are used for generating reviews, but our approach is based on computer vision and deep learning.

We have arranged the paper in the following sections: We have shownourassociated works inSection II. Section III of the paper contains the proposed methodologyof our work. Section IV includesan assessment of our system's performance,and Section V brings the paper to an end with a summary of our work.Lastly, Section VIshows some possible approaches to extend our work and future research possibilities.

II. RELATED WORK

In this section, we have discussed some papers that are related to our work. Firstly,the author here [3] worked on an approach based on two types of CNN to classify an image. The individual facial emotion recognition system using CNN from faces, and the universal image-based CNN's depended on general images. They made an average of all the CNN scores from all faces and the images to show the categories of groups of emotion. They used a kind of edge SoftMax classifier for discriminative learning and also for overfitting problems. They won theEmotion Recognition prize in the Wild Challenge 2017,and their accuracy was 83.9% and 80.9% on the validation set and testing set accordingly.

In another paper [4], the authors dispensed a new deep neural network architecture to classify human facial expressions using HOG with CNN. They trained their model to recognize emotion from facial images using deep learning andconvolutional neural network model. They usedthe FER2013 dataset. This paper showed approachesto enhance image processing in the FER system. The approach of this paper can also extract the key features from images by joining Local Binary Patternand HOG operators using their designed models.

Moreover, Liyanage C. DE SILVA et al. [5], in their work,wanted to give a facial emotion detection outcome by using visual information and auditory together. They pickedtwo peoplewho could speak two different languages (Spanish and Sinhala Language). They made thirty-six different kinds of emotional sentences in front of a camera, and their facial expression had been recorded.Mostly recorded image sequences had the same length. They had

calculated both video and audio supremacy for each emotion. From the effect of this assessment research,theytried to predict emotions through their visible appearance and voice solely.

In the paper [6], by Mao Xu et al., the author wanted to show an efficient way to classify facial expressionswithtransfer learning from convolution networks. They took 2062 sample images from four facial expressions related dataset like CK+, JAFFE, KDEF, and Pain expressions form PICS. They achievedan accuracy of 80.49% with theSVM modelfrom theirdesigned dataset. They used the Viola-Jones face detection algorithm and artificial selection to process the dataset images.They constructed the validation set with different ratiosfrom range 0 to 0.5 to experiment with varying transfer featuresgenerated from a deep convolution network.

The authors of [7] estimated different kernel sizes and several filters and proposed two unique Convolutional Neural Network (CNN) architectures to recognize facial expression for image classification using the FER2013 dataset. The second model was derived from the first model, so both models are almost similar except for the number of filters. To improve accuracy, both the models used dropout layers and had a smaller number of parameters. As a result, they attained an accuracy of 65%, which can be further improved by experimenting with hyper-parameters to improve their designed models.

Lastly, In the paper [8], H. Jung et al.wanted to showhow deep learning techniques can be applied todetect facial expressions.They first detected input images of faces by using Haar-like features. They compared two types of deep learning networks and wanted to show a comparison between them.Finally, they said that CNNis betterthan a deep neural network. They used the CK+, FER 2013 database. They also used 71774 training images with an image size of 48×48. The recognition result was good except the disgust expression, which the model was predicting wrong for most test images.

III. Proposed System Methodology

We tried various deep learning approaches to train images of faces to generate emotion. For generating review, we first used the computer vision technique to extract faces. There are many approaches for detecting faces like Haar Cascade, MTCNN, DNN face detector of the Caffe model, etc. [9].For our project, we have used the DNN Face detector of OpenCV. It is a Caffe model that is designed based on SSD and uses ResNet-10 as its model. After this, we used the extracted image to detect the facial expression and generate a review from that expression.We have broken our project into five stages for completing the task.

Fig. 1 represents a flowchart of our proposed model. In our proposed method, we aim to design a system that captures images through the front camera of a laptop during online class or from CCTV cameras to capture images in the classroom for offline environment.Next the system will process the captured images to detect students' emotions and describe overall students' feelings in real-time. The system is designedbased on deep learning techniques to build a facial emotion recognition (FER) model. Then we have designed an algorithm to generate reviews from emotion. The entire system will execute a real-time emotion detection process, which will help us to evaluate faculty reviews from an automated student review system.

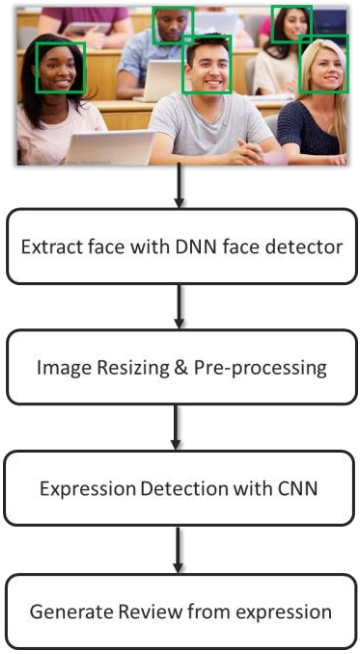

Fig. 1. Proposed System Flowchart

A. Dataset

We have used the Facial Expression Recognition 2013 (FER-2013) dataset for our work [10]. It is a free and open-source dataset that was published in International Conference on Machine Learning (ICML). Pierre-Luc Carrier and Aaron Courville generated the dataset by gathering search results from google of different human emotionsbased onfeelings. Then the dataset was openly shared before ICML 2013 for a Kaggle competition. The dataset consists of 35,887 grayscale images, from which 28,709 labeled for training, 3589 images for validation, and 3,589 images for testing. The dataset contains 48-by-48-pixel grayscale images of faces, where each image is labeled as one of the seven facial emotions.

In Fig. 2. some sample images from our dataset have been shown with all the labeled expression thatwe have used in our work.The labels of categories indexed from (0 – 6) where 0: Angry (4593 images), 1: Disgust (547 images) , 2: Fear (5121 images) , 3: Happy (8989 images), 4: Sad (6077 images) , 5: Surprise (4002 images) , 6: Neutral (6198 images) .The dataset contains 48-by-48-pixel grayscale images of faces.

B. Data Pre-Processing

Fig. 2. Image samples of facial expressions from our dataset with labels

The dataset comes in CSV format with three columns; emotion, pixels, and usage. The pixel column contains the pixel values of the images in 1D. So, we converted the 1D list of pixels into 2D to get the images' width and height, which is 48 by 48 according to the dataset.

	emotion	pixels	Usage
0	0	70 80 82 72 58 58 60 63 54 58 60 48 89 115 121...	Training
1	0	151 150 147 155 148 133 111 140 170 174 182 15...	Training
2	2	231 212 156 164 174 138 161 173 182 200 106 38...	Training
3	4	24 32 36 30 32 23 19 20 30 41 21 22 32 34 21 1...	Training
4	6	4 0 0 0 0 0 0 0 0 0 0 0 3 15 23 28 48 50 58 84...	Training

Fig. 3. Five random data according to pixels

Fig. 3 shows five random data from our dataset,which represents the pixel values of our images.Our dataset comes with pictures of faces in size 48 by 48 pixels. Since we will use VGG architecture as one of our training approaches, which expects the input images to be of size 224 by 224, we resized the images to fit in proper measure. We applied random horizontal flip and random rotation by 30 degrees for image augmentation. As each image has pixel values from 0 to 255, so we performed normalization on the color channels to rescale them to be in range 0 to 1.We divided the dataset into three groups as the dataset was provided with three different sets, i.e., the training, testing, and validation set. So, we extracted them from the CSV file to be in the proper set.

C. Modeling

For model training, we have tried two approaches. First, we have tried using transfer learning. We have used VGG-16 architecture as our CNN model. It contains about 138 million [11] learning parameters. The architecture is simple and elegant, but the network is very deep. So, training from scratch would have taken a lot of time and required a huge dataset. For this reason, we have chosen to use transfer learning for training. Since VGG is a pre-trained architecture so, the CNN layers work very well as a feature extractor. So, we decided to freeze thelearning parametersof the CNN layers and used them as the feature extractor for the images. Then wemodified the fully connected layers of the architecture and added the final output layer with seven nodes. But this approach failed to give a good result. The validation loss was getting stuck. So, we had to reject this approach.

Next, we build a new CNN architecture from scratch to train the model. The architecture expects an input of size 48 by 48 greyscale images. So, it was an advantage as our dataset also had images of the same size, unlike VGG which expects to have input with image size of 224 by 224 pixels and three different color channels.

In Fig. 4, We have shown our CNN model architecture with layers and filter size.We have used filters of size 3 for all the convolutional layers. We have usedstride size of 2 and applied same padding over the layers. The same padding allows toadd extra padding to the image so that the whole image gets covered. We added batch normalization and pooling layers and used a dropout of 0.5 for most of the layers.

.

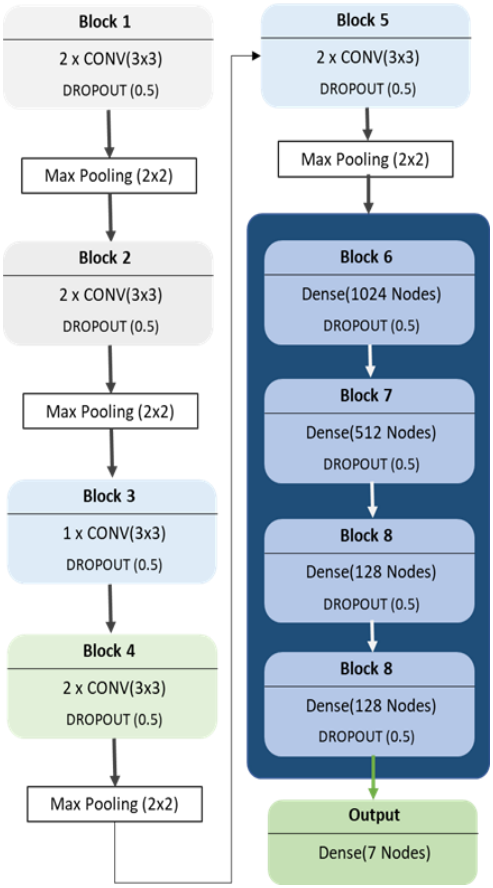

Fig. 4. Different CNN Layers and parameters

The total number of learning parameters in our model is over 13 million. To train our model, we used batch size of 32 and trained the model for about 40 epochs. We have used the following function for calculating errors.

$$CCE = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot \log(\hat{y}) + (1 - y_j) \cdot \log(1 - \hat{y}_j)$$

Equation 1

We have used equation (1),which isthecategorical cross-entropy loss function,to calculating error.We have also used Adam optimizer to update weights with an initial learning rate of 0.003.

D. Generating Review

Our model can predict the facial expression from an image of a single face. But for generating reviews from an image with multiple faces, we have used computer vision. DNN Face Detector from OpenCV's Caffe model [12] is one of the most popular methodsused to detect faces with high accuracy. So, we have used this DNN classifier to extract faces from images. Then for each face, we have used our model to generate expression. As we have seven facial expressions, we have assigned specific weights to each of them. The weights are shown in Table 1.

Table 1 shows the corresponding scores for each class of emotions. The weights here are being usedto generate the overallscore after the model predicts. These scores are assigned based on a general survey of student's behavior. In the survey, we have compared student's results and their response during the class. Students with excellent results tend to be very excited during lectures as they enjoy what they learn. And generally, attentive students are found to have neutral expressions during their class. So, happy and neutral expressions are assigned high scores.

On the other hand, some students remain unfocused during the class and focus on mobile phones or other topics. So, they often get surprised or remain in fear when asked questions, and many students who give negative reviews are found to have sad or dull expression during their class. Based on these general student behaviors, we have assigned the scores.

TABLE 1: EMOTIONSCORE

Emotion	Score
Happy	1
Neutral	0.7
Surprise	0.6
Fear	0.5
Sad	0.4
Angry	0.25
Disgust	0

To generate the review, we havegenerated result based on an average calculation. We have calculated the sum from the scores generated from all the faces. Then we havetaken the average of the scores. We have classified the scores intothree labels. If the average review is over 0.6, we say it is positive, and if the score is below 0.4, we say the review is negative. If the review is above 0.4, and below 0.6, we have labeled the review as neutral. But during the class,a facial expression of an individual may change every minute. So, taking review from just a single image for each individual may cause a biased result. So, we took more than one image of the same scene with a specific time interval and then extracted the same individual faces for review. We have used the following approach.

Algorithm 1. Generate Average Review Pseudocode

```
review_count ← 0
repeatfor every 5 munities up to 1 hour:
        temp_review ← 0
        Capture images and extract faces
        temp_review ← Score(extracted_face)
        review_count ← review_count + temp_review
end loop
final_review = Average(review_count)
```

In Algorithm1, wehave designed and approach that takes several images for every time frame at a specific gap during the class. Then it extracts faces from the images and generatesa review score for each of the detected faces. For face extraction, we have usedCaffe model's DNN face detector that comes with the OpenCV library. Then the algorithm repeats the same step several times and calculates the average score. This is howitgenerates the final score from the average of the calculated scores.

IV. PERFORMANCE EVALUATION

We have trained our model on two different architectures with the FER2013 dataset. We have used 270 images from the dataset for testing our models. We kept the dataset separate for the fair evolution of the model. First, we have used VGG-16 architecture and trained for seven epochs. It gave us a test accuracy of 54%. Then we have trained for about 40 epochs with the second model, and found 186 out of 270 test images were predicted correctly. So, we found almost 69% in our test accuracy.

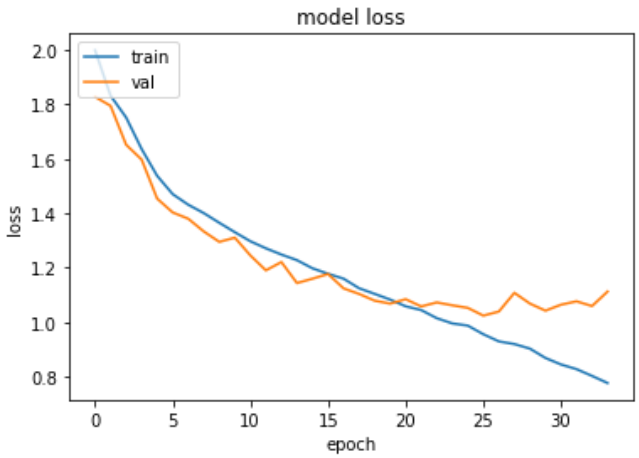

Fig. 5. Training loss Graph

Fig. 5shows the training loss and the validation loss graph. For preventing overfitting, we used early stopping by monitoring the validation loss.

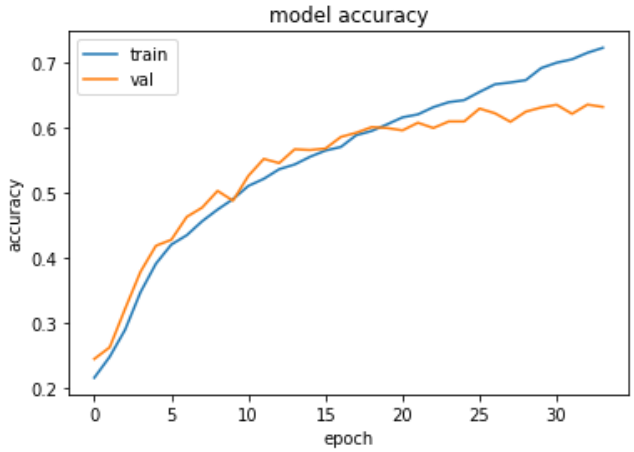

Fig. 6. Model Accuracy track

Fig. 6 shows the model accuracy graph generated while training the model. We have shown the training and validation accuracy in the graph.

Finally, to check how our review system works, we have used a different approach. We first collected 60 text reviews from students for three different faculties of the North South

University. Then we went through the review and labeled each of them as positive, negative, and neutral and extracted the average review for those faculty members. We then collected images of students doing class during the online session under the same faculties and used our algorithm to generate an average review. From our text review, we labeled 3 of the faculties with Positive review. Our model got two of the reviews as neutral and one of the reviews positive. So, we can conclude that the model is doing quite well. With a better emotion detector model, we believe our algorithm will perform as well as other methods that are generally used as faculty evaluation systems [13].

V. Conclusions

In ourwork, we have experimented with two models to predict facial expression from images of the face. We have assigned scores to the expression and designed an algorithm to generate reviews from that score. The main goal of our work wasto design a system that can generate reviews. We focused less on tuning the accuracy of the models as there are already pre-trained architectures with high accuracy for predicting facial expressions. The system will help the faculties understand how students are responding during the class and help authorities get an idea about the interactivity of students during class.

VI. Limitations& Future reSEARCH

In this work, we focused on designing an algorithm thatuses a trained model to extract facial expressions and generate reviews. With a better dataset of larger image size like 224 by 224,it is possibleto train the data on pre-trained architectures like inception, ResNet, VGG etc. to improve the accuracy of the model. It is possible to experiment with other methods like Histogram of Oriented Gradients (HOG) for feature extractions. These might result in better accuracy and improve review prediction.This system can be incorporatedwithIoT devices that will be able to provide real-time reviews of the students during the class. Another approach that can also beexperimented is to work with an image as a whole. Instead of working with an image of a single face separately, we could use the whole classroom imagewhich will have faces of multiple students and use that image as the input to a Convolutional Neural Network to directly predict a review from that image. This will open many opportunities to experiment with more complicated architectures and use other methods like single-shot detectors to build better models. We could not collect enough data for testing these options. But with the power of CNN, we believethere are lots of scope for further research on this topic to improve and design better models to improve accuracy and generate reviews.

References

[1] Y. l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity," in Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 229-234, 2002.

[2] "How does a faculty evaluation system keep things fair? - Interfolio", Interfolio, [Online]. Available: https://www.interfolio.com/resources/blog/how-does-a-faculty-evaluation-system-keep-things-fair/. [Accessed: 19- Dec- 2020].

[3] Tan, Lianzhi & Zhang, Kaipeng & Wang, Kai & Zeng, Xiaoxing & Peng, Xiaojiang & Qiao, Yu. (2017). Group emotion recognition with individual facial emotion CNNs and global image based CNNs. 549-552. 10.1145/3136755.3143008.

[4] Zafar, Sahar & Ali, Fayyaz & Guriro, Subhash & Ali, Irfan & Khan, Asif & Zaidi, Adnan. (2019). Facial Expression Recognition with Histogram of Oriented Gradients using CNN. Indian Journal of Science and Technology. 12. 10.17485/ijst/2019/v12i24/145093.

[5] L. C. De Silva, T. Miyasato and R. Nakatsu, "Facial emotion recognition using multi-modal information," Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat., Singapore, 1997, pp. 397-401 vol.1, doi: 10.1109/ICICS.1997.647126.

[6] Mao Xu, Wei Cheng, Qian Zhao, Li Ma and Fang Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," 2015 11th International Conference on Natural Computation (ICNC), Zhangjiajie, 2015, pp. 702-708, doi: 10.1109/ICNC.2015.7378076

[7] Agrawal, Abhinav & Mittal, Namita. (2019). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer. 36. 10.1007/s00371-019-01630-9.

[8] H. Jung et al., "Development of deep learning-based facial expression recognition system," 2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), Mokpo, 2015, pp. 1-4, doi: 10.1109/FCV.2015.7103729.

[9] V. Agarwal, "Face Detection Models: Which to Use and Why?", Medium, [Online]. Available: https://towardsdatascience.com/face-detection-models-which-to-use-and-why-d263e82c302c?gi=1d1c9add7285. [Accessed: 02- Dec- 2020].

[10] "fer2013", Kaggle.com, [Online]. Available: https://www.kaggle.com/deadskull7/fer2013. [Accessed: 06- Jan- 2021].

[11] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition*arXiv e-prints*, arXiv:1409.1556.

[12] "OpenCV: Deep Neural Network module", Docs.opencv.org, [Online]. Available: https://docs.opencv.org/4.0.0/d6/d0f/group__dnn.html. [Accessed: 11- Jan- 2021].

[13] Kumar, Alok & Jain, Renu. (2018). Faculty Evaluation System. Procedia Computer Science. 125. 533-541. 10.1016/j.procs.2017.12.069.