2022 International Conference on Artificial Intelligence in Everything (AIE) | 978-1-6654-7400-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/AIE57029.2022.00025

FACIAL EXPRESSION RECOGNITION USING CNN

1st and 2nd Fahreddin M. SADIKOĞLU
1st *Department of Computer Engineering*
Near East University
99138 Nicosia, North Cyprus, Mersin 10, Turkey
fahreddin.sadikoglu@neu.edu.tr
2nd *Department of Information Technology*
Odlar Yurdu University, Azerbaijan
fahredin.sadikoglu@oyu.edu.az

3rd
Mohamed Idle Mohamed
Deptartment of Electrical and Electronic Engineering
Near East University
99138 Nicosia, North Cyprus, Mersin 10, Turkey
Neylka9@gmail.com

***Abstract*—Facial is the most dynamic part of the human body that conveys information about emotions. The level of diversity in facial geometry and facial look makes it possible to detect various human expressions. To be able to differentiate among numerous facial expressions of emotion, it is crucial to identify the classes of facial expressions. The methodology used in this article is based on convolutional neural networks (CNN). In this paper Deep Learning CNN is used to examine Alex net architectures. Improvements were achieved by applying the transfer learning approach and modifying the fully connected layer with the Support Vector Machine(SVM) classifier. The system succeeded by achieving satisfactory results on icv-the MEFED dataset. Improved models achieved around 64.29 %of recognition rates for the classification of the selected expressions. The results obtained are acceptable and comparable to the relevant systems in the literature provide ideas a background for further improvements.**

***Index Terms*—Facial expressions, facial expression analysis, facial expression recognition, icv-MEFED, deep learning, Convolutional Neural Networks,**

I. Introduction

Research has shown that the face recognizes as the organ that conveys emotions. It is also a major "channel" of nonverbal communication. A facial expression is a collection of the many different positions and movements of muscles beneath the face. Humans have an inborn capacity to use a wide range of facial expressions independently and involuntarily. May deduce a person's mental state by looking at his face and not by hearing his words. Human emotions have been found to date back to at least the 19th century when Charles Darwin stated that universal emotions exist [1]. Although numerous unique methods have been presented in face expression identification, this has been investigated for several years. Finally, CNN-based techniques have reached an advanced degree of development recently. However, as the existing camera equipment is more usually capable of high resolution, the model's proposed ability to differentiate additional subtleties about emotion would face some resistance. A significant advance will be made in human-computer interaction if computers can identify more subtle signs of human involvement. Since Emotions use different facial expressions, researchers Iiris et al [1]. The iCV Multi-Emotion Facial Expression Dataset (iCV-MEFED) was created to record and classify facial expressions with a variety of emotional states. The most difficult part of FER is having complete pre-processing, feature extraction, and classification, especially when working with varied inputs such as face posture, context, and lighting. Even after applying deep learning techniques to FER, there are still problems. Additionally, deep neural networks need large quantities of training data that are free from overfitting. Even yet, databases that provide face expressions do not provide enough training data for the general. other than the great degree of diversity in face shape and facial appearance, Facial Expression Recognition (FER) has proven to be a difficult issue in computer vision for many decades. Neural networks use deep learning. Deep learning is a kind of machine learning in which a model is designed to train to perform categorization functions that are carried out directly from the textual, visual, or audio input. A deep learning algorithm carries out a certain job repeatedly, changing it slightly each time to improve the outcome, and therefore, for a train's computer system, it means finding ways to be more human-like: learning from instances. Driverless vehicles are mostly powered by deep learning, which helps them identify and react to people and road signs. Recently, for a justifiable reason, deep learning has begun to get much attention, and therefore, it is generating outcomes that could not be obtained before. AI models that use deep learning may achieve an accurate state-of-the-art from time to time, even if it exceeds human capacity. Data has been trained using many different pieces of information in conjunction with neural network frameworks with varying numbers of layers. The huge datasets and neural network frameworks needed to

Identify applicable funding agency here. If none, delete this.

978-1-6654-7400-9/22/$31.00 ©2022 IEEE
DOI 10.1109/AIE57029.2022.00025

train deep learning models do the heavy lifting in this case, whereas hand-selecting features are not required [2].

This Article will focus on set of problems which is disable people and there is computational limitations and low feature extraction and recognition rate for training and testing the IcvMEFED and misclassifications during testing and training labels

II. Methodology

This section discusses the approach proposed in this paper. Which is divided into three distinct phases: Steps for detecting facial expressions include data pre-processing, feature extraction, and classification and it was implemented for MATLAB simulation Deep Learning Toolbox [2].

A. Data and preprocessing

We compared the proposed approach on the dominant and complementary multi-emotional facial expression identification challenge using the iCV-MEFED dataset, which is accessible on the competition page the figure below shows several samples from this dataset. This dataset contains 31250

Fig. 1. Face emotion examples from the iCV-MEFED dataset .

facial faces, all of which have distinct emotional expressions and were recorded from individuals who portrayed 50 different emotions, however for each expression, only five samples were captured using a Canon 60D camera under constant lighting. Psychologists and the participants are all taught to play out these emotions, contributing to the dataset's validity. Complementary dominating and each kind include 7 options, including "angry", "disgust", "fear", "happy", "sad", "surprise", "neutral". The emotion expressed by the answer or result is identified with a number or letter N. We trained using the label transformation rule mentioned in the previous part [1].

B. Resizing of the Data Set Images

The training dataset included pictures with varying dimensions, necessitating that images be scaled to serve as training inputs. It resized the Square pictures to the specified shape (5184 ×3456 pixels), as illustrated in Figures below. The picture was reduced to its smallest dimension (227 pixels), after which the center 227×227 square was clipped. In order to produce a 227×227 picture after training augmentation, the network expects to receive input images that are 227 pixels by 227 pixels. When processing images of faces, it is common practice to crop and align faces to remove any distortions due to posture. First, we extract each face's landmarks using MATLAB script code, then we utilize the eyes' two points, and the upper lip, to do a comparable transformation that resizes and centers the face.

Fig. 2. Original Image

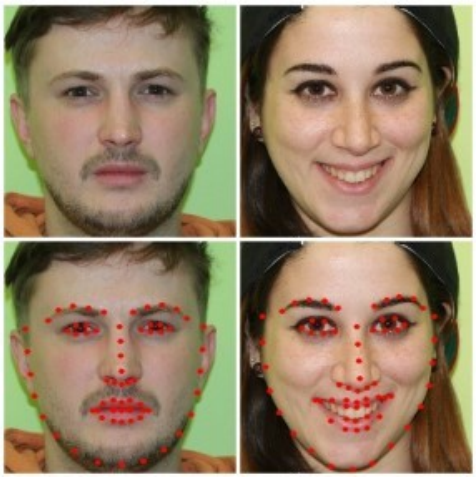

Fig. 3. Cropped and aligned pictures of size 224 × 224 are the first row

C. Data Augmentation

Label-preserving modifications may help to prevent overfitting on picture data by artificially enlarging the dataset. We use two different types of data augmentation: The data transformation process that we use enables us to generate changed pictures from the original photos so that the converted images do not need to be kept on disk, and we employ other methods to increase the overall data diversity. Pictures produced in MATLAB on the CPU, and GPU training on the previous batch of images, are used in our solution. So, the computational freedom of both data augmentation strategies is the same. The initial type of data augmentation, known as image translation and horizontal reflections, is generated by producing image translations and horizontal reflections. To produce random 227x227 patches (and their horizontal reflections), we take a random 227x227 picture and extract 227 random patches from it. We then use the network trained on these 227 patches to produce the 227x227 patches. By providing an increased training set of over four thousand times the size, while resultant training examples are, of course,

highly interdependent, this method effectively doubles the size of our training set. We would have had to utilize smaller networks because of overfitting, and we did not want to do that. When the network's SoftMax layer is applied to the five 227× 227 patches (the four corner patches and the center patch) and their horizontal reflections, the network produces a prediction (the network's total error is five 227× 227 patches plus their horizontal reflections). To improve the performance of a convolutional neural network, you need to put in many data. The dataset's size enables it to retrieve even more characteristics from the unlabeled data and match it. However, if gathering data proves to be difficult, data augmentation may enhance the model's performance. Additional pictures are generated by performing image manipulation operations such as random rotation, shifts, shear, and flips on the current dataset. In this study, we scale the facial pictures to 227 × 227 pixels to do further analysis. Now, a higher-quality photograph may be obtained, and lowering the picture resolution lowers their visual ability. However, the less resolution a system has, the more rapidly it learns.Here, just zero-mean normalization is required for data preparation. We use random cropping to extract the core areas of the original pictures, and then we rotate the images by flipping them horizontally. Overfitting is thought to be lessened due to the increased variety of training samples [2].

D. ALEXNET TRAINING MODELS

In 2012, Alex's network completed the ImageNet large-scale visual recognition challenge. In the article published in ImageNet Classification with Deep Convolution Neural Network in 2012, researcher Alex Krizhevsky and his colleagues developed the model. The Alex network contains eight levels, each of which may be controlled through a learnable parameter. The model comprises five layers, each of which has a ix of max pooling and Relu activation except for the last layer. To prevent their model from overfitting, they utilized the dropout layers. The model is trained on the ImageNet dataset, which has more than 16 million images. This database of nearly 14 million pictures spanning 1,000 classifications is known as the ImageNet dataset.

E. Alexnet Architecture

A good point to keep in mind since Alex net is a complex deep architecture, padding was included to ensure feature maps do not dramatically reduce in size. This model's input is 227x227x3-sized pictures. This thesis was implemented in Alex's net convolutional neural networks by using facial expression recognition. In MATLAB deep learning toolbox, the Data set used ICV-MEFED as trained in 70% and 30% for testing with data augmentations [3]. As shown in the below tables, we have used the architecture described in this article to summarize our findings which is called Alex net architecture contains twenty-five layers, with each layer learnable. The Model has an RGB input. It contains five convolution layers, each with a different maximum pooling operation and Each layer has three direct connections. All layers in the network

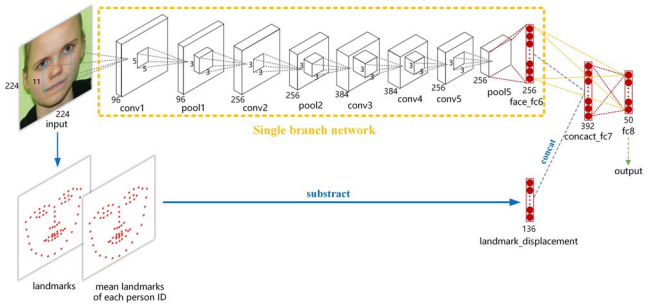

Fig. 4. ALEXNET ARCHITECTURE 1 [3]

Layer	# filters / neurons	Filter size	Stride	Padding	Size of feature map
Input	-	-	-	-	227 x 227 x 3
Conv 1	96	11 x 11	4	-	55 x 55 x 96
Max Pool 1	-	3 x 3	2	-	27 x 27 x 96
Conv 2	256	5 x 5	1	2	27 x 27 x 256
Max Pool 2	-	3 x 3	2	-	13 x 13 x 256
Conv 3	384	3 x 3	1	1	13 x 13 x 384
Conv 4	384	3 x 3	1	1	13 x 13 x 384
Conv 5	256	3 x 3	1	1	13 x 13 x 256
Max Pool 3	-	3 x 3	2	-	6 x 6 x 256
Dropout 1	rate = 0.5	-	-	-	6 x 6 x 256

Fig. 5. ALEXNET ARCHITECTURE 1 [3]

```
25x1 Layer array with layers:

   1   'data'     Image Input                   227x227x3 images with 'zerocenter' normalization
   2   'conv1'    Convolution                   96 11x11x3 convolutions with stride [4  4] and padding [0  0  0  0]
   3   'relu1'    ReLU                          ReLU
   4   'norm1'    Cross Channel Normalization   cross channel normalization with 5 channels per element
   5   'pool1'    Max Pooling                   3x3 max pooling with stride [2  2] and padding [0  0  0  0]
   6   'conv2'    Grouped Convolution           2 groups of 128 5x5x48 convolutions with stride [1  1] and padding [2  2  2  2]
   7   'relu2'    ReLU                          ReLU
   8   'norm2'    Cross Channel Normalization   cross channel normalization with 5 channels per element
   9   'pool2'    Max Pooling                   3x3 max pooling with stride [2  2] and padding [0  0  0  0]
  10   'conv3'    Convolution                   384 3x3x256 convolutions with stride [1  1] and padding [1  1  1  1]
  11   'relu3'    ReLU                          ReLU
  12   'conv4'    Grouped Convolution           2 groups of 192 3x3x192 convolutions with stride [1  1] and padding [1  1  1  1]
  13   'relu4'    ReLU                          ReLU
  14   'conv5'    Grouped Convolution           2 groups of 128 3x3x192 convolutions with stride [1  1] and padding [1  1  1  1]
  15   'relu5'    ReLU                          ReLU
  16   'pool5'    Max Pooling                   3x3 max pooling with stride [2  2] and padding [0  0  0  0]
  17   'fc6'      Fully Connected               4096 fully connected layer
  18   'relu6'    ReLU                          ReLU
  19   'drop6'    Dropout                       50% dropout
  20   'fc7'      Fully Connected               4096 fully connected layer
  21   'relu7'    ReLU                          ReLU
  22   'drop7'    Dropout                       50% dropout
  23   'fc8'      Fully Connected               1000 fully connected layer
  24   'prob'     Softmax                       softmax
  25   'output'   Classification Output         crossentropyex with 'tench' and 999 other classes
```

Fig. 6. SoftMax layer Alex net architecture implementation 1 [3].

utilize Relu as their activation function. Two Dropout layers were utilized. The activation function in the output layer is SoftMax. A total of 62.3 million parameters may be found in this design.

III. RESULT AND DISCUSSION

A. Experimental Implementation

The suggested system is powered by an Intel Core i5 7th Gen processor running at 2.7 GHz with 4GB of RAM. The MATLAB R2020b device was used to evaluate the approach and execute the classification and feature selection tasks. The training subset of 70% was used to train the network

for classification, while the testing subset of 20% and 10% was utilized to determine the likelihood that a face picture corresponds to a certain class of facial expression.

B. Experimental Results

Images scaled to 227 × 227 are used in our tests: transfer learning and data augmentation combined with a technique employing mini-batch sizes of 20. For the learning rate, the process begins at 0.0001, and it is completed for 300 iterations. In addition to a decay rate of 0.0001, we also have a momentum of 0.9 and a frequency of 2 iterations. In order to conduct our tests, we utilize the Alex net model with SoftMax classifier. We have 140 pictures that include the ground truth for 20 images per label, and the remainder is for test images. Approximately ten percent of all pictures were selected for the validation set, including approximately 140 images from the previous nine individuals.

C. REPORT OF SOFTMAX CLASSIFIER

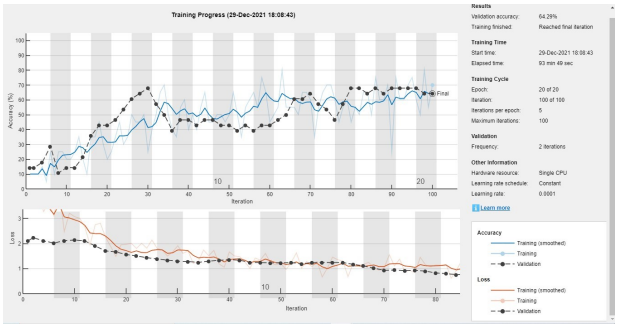

Fig. 7. Final training validation accuracy of SoftMax classifier iteration 1

In this figure illustrates that we trained and tested 140 sample images from the dataset of icvMEFED using Alex net model with SoftMax classifier and in training cycle there was 2 iterations. In the above figure is the iteration 1 and the training cycle was 300 of 300, the Epoch cycle was 20 of 20, every iteration per Epoch cycle was 15 and during the training time was 29 December 2021 at 18:08:43 and the elapsed time was 93 minutes and 48 seconds that shows us when the class of expression are more, the classification images are more and the iteration progress are more than one the system training takes time because of making more classifying the image emotions and the classifying the classes. Our system iteration achieved in the final iteration for the validation correct accuracy of 64.29% for seven classification images in SoftMax classifier and the second part of the result which is loss accuracy illustrates that the final training misclassified images of SoftMax classifier in iteration 1. The X axis shows the loss accuracy and Y-axis shows the iteration training program and the loss validation accuracy of 35.71% of seven classes' emotions.

This figure illustrates the confusion matrix and training validations of the recognition accuracies for seven classes of

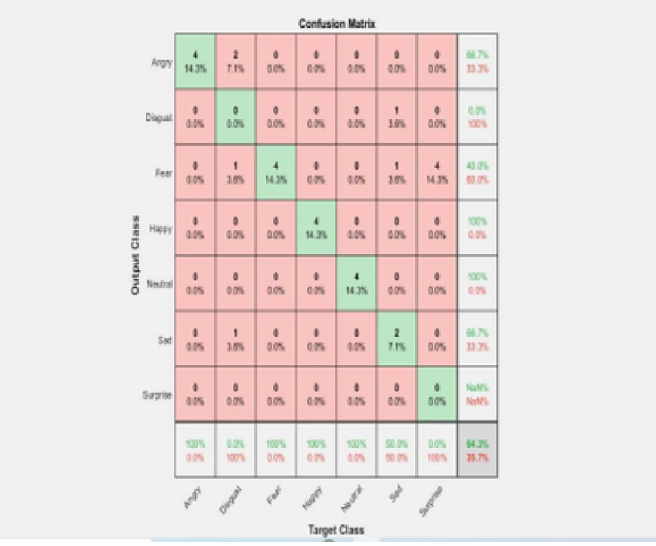

Fig. 8. Alex net SoftMax classifier of confusion matrix

FER. The accuracy recognition rate provided our system when implemented Alex net with SoftMax classifier was 64.29% because of the classes are seven and the images was 20 images per class and the total number of images trained and tested in the simulation was 140 images that's why the accuracy is 64.29% note that if the classes are more and the images is low the system will make miss classification and the recognition accuracy will be good around 60% above and our system shown when we used SoftMax classifier obtained 64.29% which is satisfactory result.

D. Validation Accuracy of Proposed Methods

Sensitivity	***Specificity***	***Accuracy rate***	***Error rate***
90%	80%	64.29%	35.71%

TABLE I
COMPARISON OF RESULTS WITH APPROACHES IN METRIC MEASUREMENTS

The accuracy recognition rate provided our system when implemented Alex net with SoftMax classifier was 64.29% for seven classification images with the proposed method are shown in Table

E. OVERALL RECOGNITION RATE

The overall recognition rate in the below figures shows the overall recognition rates. It is observed that the overall recognition rates under unknown expressions of the known faces are reported in satisfactory range after applying convolutional neural networks in Alex net model with SoftMax layer but SVM has achieved the best overall recognition rate of 64.29% and the error rate of 35.71% for twenty compound classes.

F. EXPRESSION RECOGNITION RATE APPLIED WITH SOFTMAX

As shown in figure below illustrates the recognition rate of averaging the training set and comparing it with the compound dataset images with applied SoftMax classifier reports that the recognition rate of 64.29%

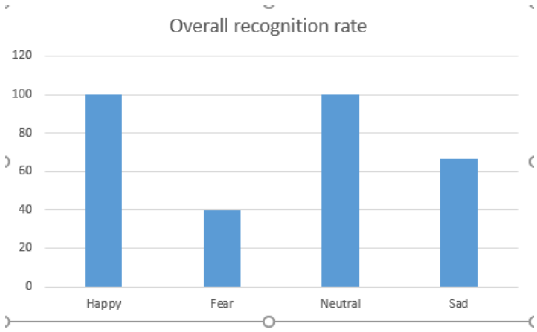

Fig. 9. overall expression recognition rate applied with SoftMax

CONCLUSION

This paper proposed facial expression recognition using convolutional neural networks. Several experiments were carried out to obtain the results presented in the previous chapter. The facial expression database was first acquired from the icv-MEFED database, the images were then organized into the different facial expressions, normalized and Alex net models applied on them. The icv-MEFED database on which this experiment was tested includes 140 subjects, female images and male images. For each of the subjects, there are seven facial-expressions which are arranged as Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise expressional faces. In our experiments the result proves that applying SoftMax classifier, enhances the results of the recognition rate enhances the overall recognition rate it even greater. During the experimental stage, it was discovered that including the male subject from both the test and training dataset affects the overall recognition result due to the presence of facial hair and other variation on the male face. The experiments conducted during this research also confirms that the when the classes are more, and the expression dataset images are low can affect the overall recognition rate and the processing time will take more time. The comparison results with approaches in metric measurements. We used deep learning convolutional neural networks to examine the Alex Net architectures for facial emotion recognition. As shown by the findings, we succeeded in obtaining acceptable results in the ICV-MEFED dataset. We further refined these models, with the Alex Net model and SVM classifiers achieving the greatest recognition accuracy 64.29%, followed by the Alex Net model with SoftMax classifiers. Finally, the findings of this paper may aid in increasing the rate of facial expression. The comparison of the system performance shows that softmax has achieved higher accuracy then SoftMax in the final iteration for the validation correct accuracy for seven facial classification images with the proposed method are shown in the previous chapters.

REFERENCES

[1] Guo, J., Zhou, S., Wu, J., Wan, J., Zhu, X., Lei, Z., & Li, S. Z. (2017). Multi-modality Network with Visual and Geometrical Information for Micro Emotion Recognition. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 814–819. https://doi.org/10.1109/FG.2017.103

[2] Liu, K., Zhang, M., & Pan, Z. (2016). Facial Expression Recognition with CNN Ensemble. Proceedings - 2016 International Conference on Cyberworlds, CW 2016, 163–166. https://doi.org/10.1109/CW.2016.34

[3] Nour, N., Elhebir, M., & Viriri, S. (2020). Face Expression Recognition using Convolution Neural Network (CNN) Models. International Journal of Grid Computing & Applications, 11(4), 1–11. https://doi.org/10.5121/ijgca.2020.11401

[4] Liu, Y., Yuan, X., Gong, X., Xie, Z., Fang, F., & Luo, Z. (2018). Conditional convolution neural network enhanced random forest for facial expression recognition. Pattern Recognition, 84, 251–261. https://doi.org/10.1016/j.patcog.2018.07.016

[5] Rajendra Kurup, A., Ajith, M., & Martínez Ramón, M. (2019). Semi-supervised facial expression recognition using reduced spatial features and Deep Belief Networks. Neurocomputing, 367, 188–197. https://doi.org/10.1016/j.neucom.2019.08.029

[6] Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing, 27(6), 803–816. https://doi.org/10.1016/j.imavis.2008.08.005

[7] Ryu, B., Rivera, A. R., Kim, J., & Chae, O. (2017). Local Directional Ternary Pattern for Facial Expression Recognition. IEEE Transactions on Image Processing, 26(12), 6006–6018. https://doi.org/10.1109/TIP.2017.2726010

[8] Chen, X., Yang, X., Wang, M., & Zou, J. (2017). Convolution neural network for automatic facial expression recognition. Proceedings of the 2017 IEEE International Conference on Applied System Innovation: Applied System Innovation for Modern Technology, ICASI 2017, 814–817. https://doi.org/10.1109/ICASI.2017.7988558

[9] Ryu, B., Rivera, A. R., Kim, J., & Chae, O. (2017). Local Directional Ternary Pattern for Facial Expression Recognition. IEEE Transactions on Image Processing, 26(12), 6006–6018. https://doi.org/10.1109/TIP.2017.2726010