Paper ID: 125

Real Time Facial Expression Analysis Using VGGNet Architecture

Anonno Singha Ray[1] , Mohsiul Mumit Alik[2] , Laboni Mostofa Labonno[3] , F.M. Rezwan Parvez[4],
[1-4]Department of Mechatronics Engineering,
Rajshahi University of Engineering & Technology, Rajshahi.
E-mail: singharayanonno@gmail.com[1]; www.mohsiul@gmail.com [2] ;labonimostofalabonno@gmai.com[3] ; rezwanparvez67@gmail.com[4]

Abstract

Facial expression recognition using Convolutional Neural Networks (CNNs) is a rapidly evolving field within machine learning. It is a fundamental task in computer vision with numerous applications, including human-computer interaction, emotion detection, and psychological research. The emotions that have grown in the face have a significant impact on judgments and debates on a variety of topics. Facial emotion recognition can also be implemented in zoom platform to detect attentiveness of the attendances. This study proposes the development of a system capable of predicting and recognizing face emotions based on feature extraction using the Convolution Neural Network (CNN) method in real time using the OpenCV, TensorFlow and Keras library. The study design includes three primary processes: face recognition with eye and mouth, facial extracted features, and facial emotion categorization. The model uses convolutional, batch normalization, activation, and pooling layers to capture local spatial patterns and hierarchical representations, which improves its capacity to recognize emotions including happy, sadness, rage, surprise, disgust, fear, and neutrality. The model is trained with 28821 images and validated using 7066 images and has an accuracy of 66.907%.

Keywords: Facial emotion, Convolutional Neural Networks, computer vision, OpenCV, deep learning.

1. Introduction

Artificial intelligence (AI), which has been working towards the objective of replicating human cognitive functions, has made incredible progress over the previous decade, while facing significant challenges. Recognizing subtle facial emotions is a challenging task that many individuals deal with [1]. It is crucial to note that the way a certain emotion is portrayed can change greatly from person to person, depending on aspects such as ethnicity, age, and gender. Additionally, the perception of an individual's emotional state is based on several contextual factors, such as lighting conditions, posture, and background situations.

The investigation of how facial expressions may be used to identify feelings is an intriguing area of research that has a wide range of applications, including those in the fields of healthcare, psychological analysis, lie-detection, security and studying effectiveness of advertisement and many more. In the field of robotics, the robot's capability, enable to more intuitive and human-like interactions can significantly enhance by the use of automated facial emotion recognition [11]. One of the basic goals of sentiment analysis is to determine the perspective of a person about a certain subject matter or the general emotional tone of a document altogether. Analysis of sentiment plays a significant part in the field of nonverbal communication, namely in the process of interpreting the feelings that are communicated via facial expressions. Researchers working in this field are committed to improving models and approaches, as well as extracting a wide variety of variables, in order to allow computer- generated sentiment predictions that are made with greater precision. Positive, negative, or neutral are the standard classifications that are used to describe the polarities of sentiment. There are two main approaches can be recognized as Facial expression classification. The first involves support vector machine (SVM) and k-Nearest Neighbor (KNN) as distance-based techniques, and The others are Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) as statistical learning method. Although there are many different techniques to sentiment analysis, but the availability of enormous-scale of data with enhancement in computational capabilities has been highly highlighted deep learning architectures stand out [2] [12].

The most recent developments in facial expression recognition have placed a greater emphasis on spontaneous facial expressions in order to achieve greater generalization capacity. It is still challenging to acquire a discriminative representation of facial expressions, despite the fact that deep learning has achieved exceptional performance on a variety of face recognition tasks (such as face detection and face verification). On the one hand, the explanation for this difficulty is that the training data for facial expression recognition is very limited in comparison to the data for the other tasks. For instance, the amount of data used for training purposes in the face verification job is often millions, but the amount used for expression recognition is merely thousands. It is, on the other hand, difficult to categorize when there are significant variances within the same class. Images belonging to

a single expression class in a dataset include a wide variety of individuals from different backgrounds. Because of differences in appearance, gender, and ethnicity, there will be significant variances within the same class [3]. Generalizing models is difficult in face emotion recognition. Across persons, circumstances, and civilizations, facial expressions may transmit diverse emotions via changes in intensity, duration, and spatial arrangement. Another problem with FER is the uneven distribution of data across categories. Positive emotions like happiness and surprise are more prevalent than negative ones like disgust and disdain. Biased models may perform badly on underrepresented emotions due to this imbalance [4]. More robust and accurate models are needed for face and emotion identification due to the problems outlined above. The use of the FER model in virtual settings, especially online meeting platforms, is an additional technological contribution. The VGGNet model can improve communication dynamics, increase user experience, and encourage participation in virtual meetings by providing real-time insights into participants' emotions [13].

In order to overcome these obstacles, we provide a method for real-time FER based on the VGGNet architecture, emphasizing the extraction of discriminative features that improve classification accuracy across emotional categories. VGGNet is appropriate for FER because of its deep layers, which allow it to collect fine-grained characteristics necessary for identifying nuanced expressions. To strike a compromise between accuracy and computational economy, our method involves improving the network using varying numbers of convolutional layers, filter sizes, and fully connected layers [14]. The majority of expression databases lack enough data to train the complete network from the beginning. There is a lot of promise in using VGGNET for classification when face recognition is used as a biometric approach. It provides high recognition rates while remaining easy to use. Several convolution layers with tiny filters and a few pooling layers are added by the VGG network model. Convolutional layers work better than completely connected layers for classifying images. This benefit makes it possible for the VGG network model to more effectively handle the difficulties brought on by the complexity and diversity of facial emotions [15].

2. Related works

In case of nonverbal communication, facial expressions are the most popular forms as it is the best way to transmit emotions. For this reason, this paid attentions of many researchers in the field of computer vision.

Pranav et al. [5] proposed a Deep Convolutional Neural Network (DCNN) model for facial emotion recognition that classifies facial emotions of 5 different humans. By using an emotion image dataset, the model is being trained, as an optimizer they used Adam and for loss function, cross-entropy is utilized. The system offers good accuracy. In [6] asresa et al. developed a model for 7 facial emotion detection and recognition utilizing CNN for training data. As database FERC-2013 is used which consists of about 35889, 48 × 48- pixel gray-scale images of faces. Epoch 10 and dropout 0.3 is used to quantify the model performance. In [7] Usman et al. proposed a technique based on a stage classifier, Neural Architecture Search Network by using the FER 2013 dataset utilizing RNN and child networks, for emotion identification and classification. To evaluate the performance of the model, CK+ data is used.. Lasri et al. [8] developed a system consisting of three phases (face detection, normalization, and emotion recognition) for recognizing students 7 types of facial emotion expressions using CNN on FER 2013 database. The accuracy rates the system is quite compatible and the efficiency of the model is evaluated by confusion matrix, precision, recall, and F1-score. The author of [9] proposed a system for facial expression recognition on real- time mobile applications using a set of support vector machines (SVM). Six fundamental emotions and neutral expressions have been classified. An approach is developed by the author of [10], based on two-part convolutional neural networks (FERC) for facial emotion recognition. To identify the five distinct categories of facial expressions, expression vectors (EVs) are employed in the proposed model. Accuracy goes down by increasing the number of images as well as decreasing the number of training images.

3. Proposed model

Our proposed Convolutional Neural Network (CNN) is designed for facial emotion recognition, capable of identifying seven emotions—happiness, sadness, anger, surprise, disgust, fear, and neutrality—from grayscale facial images resized to 48x48 pixels. The model architecture begins with four convolutional layers that extract hierarchical features such as edges, textures, and patterns from the input images. These layers are followed by batch normalization, which stabilizes and accelerates the training process, and ReLU activation functions to introduce non-linearity, allowing the network to learn complex relationships within the data. Max-pooling layers are incorporated after each convolutional layer to reduce spatial dimensions while preserving critical information, enhancing computational efficiency and tolerance to spatial variations in the input. Dropout regularization is applied during training to randomly deactivate neurons, reducing overfitting and improving generalization. After the convolution and pooling operations, the features are flattened and passed through two fully connected layers to learn high-level representations essential for classification. The final output layer employs a softmax activation function, generating a probability distribution over the predefined emotion categories and predicting the most likely emotion for a given input. The model is trained using the Adam optimizer, which dynamically adjusts learning rates for faster and more stable convergence, with a learning rate set at 0.001. Categorical cross-entropy is used as the loss function to measure the divergence between predicted

and true probabilities. Training was conducted over 75 epochs with a batch size of 128, resulting in a peak validation accuracy of 66.90% achieved at epoch 74. Preprocessing steps such as resizing, grayscale conversion, and normalization ensure consistent input scaling, while potential data augmentation techniques, including rotations and flips, enhance the model's robustness by simulating variations in real-world scenarios. This architecture effectively addresses challenges like overfitting, data variability, and computational limitations, making it a robust and scalable solution for emotion recognition. The model's design and performance highlight its applicability in diverse fields such as human-computer interaction, psychological assessments, and behavioral analysis, offering a reliable framework for future enhancements and real-world applications.

```
Model: Sequential
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================

 batch_normalization_4 (Bat  (None, 256)               1024
 chNormalization)

 activation_4 (Activation)   (None, 256)               0

 dropout_4 (Dropout)         (None, 256)               0

 dense_1 (Dense)             (None, 512)               131584

 batch_normalization_5 (Bat  (None, 512)               2048
 chNormalization)

 activation_5 (Activation)   (None, 512)               0

 dropout_5 (Dropout)         (None, 512)               0

 dense_2 (Dense)             (None, 7)                 3591

=================================================================
Total params: 4478727 (17.08 MB)
Trainable params: 4474759 (17.07 MB)
Non-trainable params: 3968 (15.50 KB)
_________________________________________________________________
```

Fig 1: The Part of the construction of our CNN model with parameters.

4. Datasheet used

Jonathan Oheix's facial expression detection dataset on Kaggle is a comprehensive collection designed to facilitate study and development in the field of emotion recognition from facial photos. This dataset contains 35,887 photos that depict a variety of emotions, including happiness, sadness, rage, surprise, contempt, fear, and neutral expressions.



Fig. 2. Greyscale pictures of various emotions of the dataset

The collection is divided into seven folders, each representing a distinct facial expression category. Each folder contains a different amount of photos, reflecting the diversity and distribution of facial emotions in the dataset. This diversity enables complete training and evaluation of facial emotion recognition algorithms. The dataset is separated into two parts: 80% (28821 images) are testing data and 20% (7066 images) are for validation.

The image expressions in our training dataset are balanced, except for the 'disgust' category. The dataset have total 4103 fear images, 436 disgust images, 4982 neutral images, 7164 happy images, 3993 angry images, 3205 surprise images and 4938 sad images for training. For validation the dataset contains 960 angry images,111 disgust images,1018 happy images, 1825 neutral images, 1216 sad 1139 and surprise 797 images.

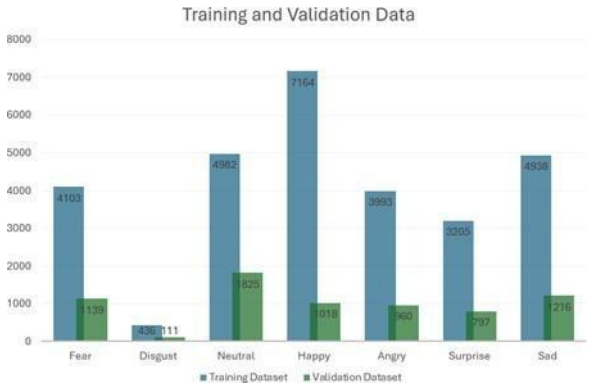


Fig. 3. Visualization of training and validation data

5. Experimental setup

The model is trained online in Google Collaboratory. The code was written in Python 3 and was run on the hosted runtime using the hardware accelerator T4 GPU. The model was executed with a system having Intel core i5-12500H CPU, NVIDIA RTX 3050 4GB GPU, and 16GB of RAM. The model was executed in real-time on the Spyder compiler with Python 3.9. The TensorFlow framework was used to create and execute the model. The ReLu activation function is applied. SoftMax converts a vector of integers into a vector of probabilities. The Adam optimization method is also utilized.

6. Results

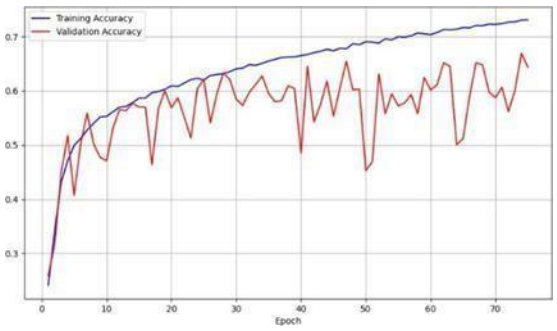


Fig. 4. Training & validation accuracy

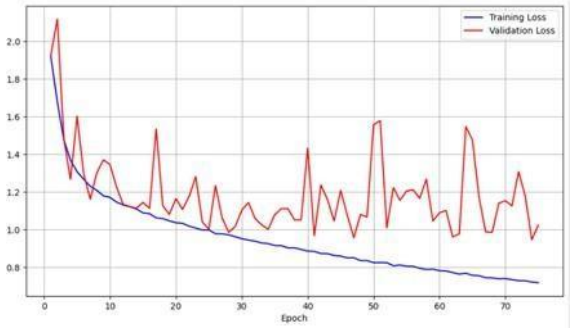


Fig. 5. Training & validation loss

Here, In Fig. 4. the blue line indicates training accuracy and the red line indicates validation accuracy. Training accuracy gradually increases as the number of epochs increases, but validation accuracy shows a distinct number of point fluctuations due to the number of images or frames are not adequate for validating 7 classes. After nearly 40 epochs validation accuracy increases very slowly, probably it is because of overfitting. The best validation accuracy was found 66.90 after 74 epochs.

The Fig. 5. plots both training and validation loss vs number of epochs. The validation loss actually gives an estimation of the expected error. The training loss decreases whether the model learns signal or noise. However, the validation loss decreases only once the model learns the signal. The size of the gap between training and validation loss tells us how much noise the model has learned.

We assessed our laptop camera through the OpenCV library using Python and detected real-time facial emotion. Frontal face-eye and mouth are classified using haarcasade classifier. A Haar cascade classifier is a machine learning algorithm that employs Haar characteristics to identify objects in photos. The method operates by categorizing positive data points as part of the identified item and negative data points as not containing the object. Some of the captured emotions are shown below.

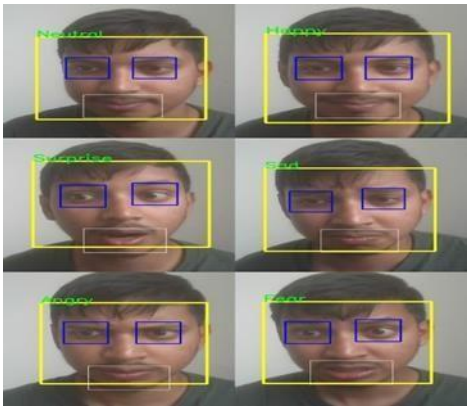


Fig. 6. Real-time emotion detection using OpenCV.

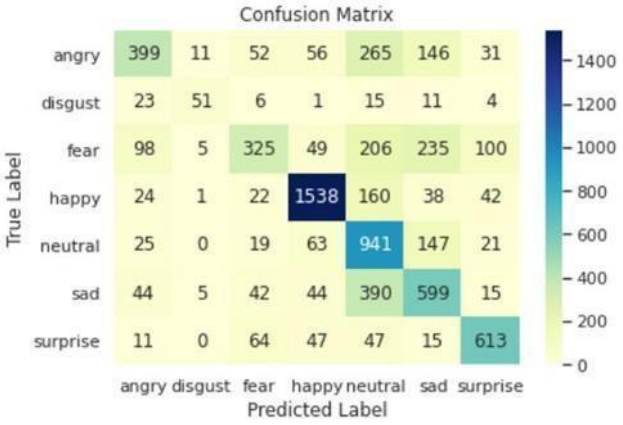


Fig. 7. Confusion matrix of classified emotions

A confusion matrix is a table used to describe the performance of a classification method. It depicts and summarizes the performance of a classification algorithm. Fig. 7 illustrates a confusion matrix, which is a tabular visualization used in machine learning to classify emotions. The rows depict the real emotions, while the columns represent what the algorithm anticipated. In the confusion matrix we can see that happiness and neutrality are predicted almost accurately, while emotions like disgust and sadness show misclassification as there are less training images available for these two classes. The model struggles to differentiate between similar emotions, highlighting areas for improvement in emotion recognition accuracy.

7. Discussion and conclusion

CNN-based activity recognition has significantly improved over manual representation techniques, especially on difficult datasets, and the computer science community has taken note of these impressive advances. CNN-based action recognition is thoroughly studied in this research, which synthesizes the abundance of information available on the subject. Despite the remarkable progress achieved in facial emotion recognition using CNNs, several challenges remain. These include handling large variations in facial expressions, addressing class imbalance in datasets, mitigating biases in training data, and ensuring model interpretability and transparency. Deploying facial emotion recognition systems in real-world scenarios presents additional challenges related to privacy, ethical considerations, and algorithmic fairness. Future research directions may involve exploring multimodal approaches that combine facial and contextual information, developing more robust and interpretable CNN architectures, and investigating techniques for mitigating biases and improving generalization across diverse demographics.

The proposed Convolutional Neural Network (CNN) architecture achieved a validation accuracy of 0.66903 after 74 epochs, using a batch size of 128. The training accuracy gradually increased with the number of epochs, but the validation accuracy showed significant point fluctuations due to the inadequate number of images or frames available for validating the seven classes, suggesting the possibility of overfitting. The model's highest validation accuracy score suggests that it has the potential to accurately recognize human facial emotions

8. References

[1] Haq, H., et al. "Enhanced Real-Time Facial Expression Recognition Using Deep Learning," in Acadlore Trans. Mach. Learn, vol. 3, no. 1, pp. 24–35, 2024.

[2] Meena, G., et al. "Identifying emotions from facial expressions using a deep convolutional neural network-based approach," in Multimedia Tools and Applications, vol. 83, no. 6, pp. 15711–15732, 2024.

[3] J. Wang, C. Yuan, "Facial expression recognition with multi-scale convolution neural network," in Advances in Multimedia Information Processing-PCM 2016: 17th Pacific-Rim Conference on Multimedia, Xi an, China, September 15-16, 2016, Proceedings, Part I, 2016, pp. 376–385.

[4] W. Liu, M. Xia, "The application of deep learning in computer vision: Facial emotion recognition based on convolutional neural network," in Journal of Physics: Conference Series, 2023, pp. 012020.

[5] Pranav, E., et al, "Facial emotion recognition using deep convolutional neural network," in 2020 6th International conference on advanced computing and communication Systems (ICACCS), 2020, pp. 317–320.

[6] T. Asresa, G. Tigistu, M. Bayih. "Convolutional Neural Network Driven Computer Vision Based Facial Emotion Detection and Recognition," in Indian Journal of Computer Graphics and Multimedia (IJCGM), vol. 3, no. 2, pp. 8–11, 2023.

[7] Islam, U., et al. "Facial Emotions Detection using an Efficient Neural Architecture Search Network," in Research Square, 2023.

[8] I. Lasri, A. Solh, M. El Belkacemi, "Facial emotion recognition of students using convolutional neural network," in 2019 third international conference on intelligent computing in data sciences (ICDS), 2019, pp. 1–6.

[9] M. Suk, B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014,pp.132–137.

[10] N. Mehendale. "Facial emotion recognition using convolutional neural networks (FERC)," in SN Applied Sciences, vol. 2, no. 3, pp. 446,2020.

[11] Sang, D. V., Van Dat, N., & Thuan, D. P. (2017). Facial expression recognition using deep convolutional neural networks. IEEE. https://doi.org/10.1109/kse.2017.8119447

[12] Jun, H., Shuai, L., Jinming, S., Yue, L., Jingwei, W., & Peng, J. (2018). Facial Expression Recognition Based on VGGNet Convolutional Neural Network. IEEE. https://doi.org/10.1109/cac.2018.8623238

[13] "Prediction of Human Misbehavior Detection From CCTV Video Recording Using Convolutional Neural Network Compared With VGG-16 Method," IEEE Conference Publication | IEEE Xplore, Sep. 2023, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10398039

[14] "Real-time face recognition using an enhanced VGGNET Convolutional neural network," IEEE Conference Publication | IEEE Xplore, Oct. 06, 2023. https://ieeexplore.ieee.org/abstract/document/10353479

[15] Cheng, S., & Zhou, G. (2019). Facial Expression Recognition Method Based on Improved VGG Convolutional Neural Network. International Journal of Pattern Recognition and Artificial Intelligence, 34(07), 2056003. https://doi.org/10.1142/s0218001420560030