UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

A COMPREHENSIVE STUDY OF CNN-BASED FACIAL EMOTION RECOGNITION IN IMAGES AND VIDEOS

Bhadra Sai Tarun Mediboina
Department of Computer science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh , India
2301050093cse@gmail.com

Basant Sah
Department of Computer science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh , India
basantbitmtech2008@gmail.com

Abstract— Human emotional states are categorized using facial emotion recognition. Sorting each face image into the seven classes of facial emotions is the goal. Convolutional Neural Networks (CNNs) are employed in the emotion classification process. Real-time videos and A range of grayscale images from the dataset are captured for input. Next, the CNN's sequence of convolution and purpose of pooling layers is information abstraction, and the SoftMax layer is used for classification. A few methods are used to address the model's overfitting issue, including dropout, cluster standardization, and L2 regularization. The facial expression dataset from the Picture Folders (fer2013) collection is applied to experiments, and our model performs more accurately in predicting individual emotions than previous research has. Furthermore, the model exhibits good performance in predicting the mood of every picture in the live video stream.

Keywords— Emotional States, Seven Classes, Surprising, Convolutional Neural Networks (CNNs), Information abstraction, SoftMax Layer, Realtime Videos

INTRODUCTION

The best results from emotion recognition are probably going to come from combining many modalities text, sound, image, and video to comprehend the actions and emotions of a particular person. Acknowledging emotions benefits many organizations and aspects of daily life. In terms of safety and healthcare services, it's important and beneficial. Furthermore, it is necessary for easy and uncomplicated identification of human physiology in a specific moment without actually requesting it [1]. Interaction and cooperation with others involve the communication of emotions. Emotions can be broadly categorized as follows: fear, fear, surprise, happiness, sadness, and anger. Artificial intelligence systems are used in Automatic Emotion Recognition to identify a person's emotion without interfering. It is acknowledged that each emotion has a unique Action Unit (AU) [2].

This implies that a particular development of facial muscles is associated with a specific propensity. For instance, in the event that the AI notices. The "upside-down smile" and the "wrinkled forehead" of the AU at the same time, it may determine that the subject of the investigation is unhappy. These fundamental groupings are combined by advanced feeling locators to identify complex sentiments. To ascertain a person's emotional state, a technology known as Face Emotion Recognition analyses both still and moving images [3]. Images and videos can be used as well as other sources to analyse emotions thanks to a technology called identification of facial emotions. It is a component of the advances in the multidisciplinary field of "emotion processing," which examines how well computers are able to identify and comprehend human emotions. In this field, advances in artificial intelligence are often built upon [4].

Facial Recognition verifies whether or not the two appearances at the event are the same. Facial recognition is widely used in security, biometrics, entertainment, personal health, and other areas [5]. Valence and arousal, two crucial components of a fervent perception of appearances, have been been identified by the systems involved in face recognition. Valence, which measures a face's pleasantness, is measured on a linear scale that positive or strong inclinations on one end and negative or hurtful feelings for the other. Temptation demonstrates extent face elevates an observer to an increased level of notable preparedness.

I. LITERATURE SURVEY

The majority among these discreet methods use cams on video to take pictures at regular intervals, and some of them use stereo vision to take pictures of the human face in three dimensions. The accuracy and efficacy of these techniques may be impacted by variations in lighting conditions and pose due to the camera's fixed position. In order to categorize thirty facial actions using artificial classifiers (such as fluttering and yawning movements, among ther jerks) derived from the system of facial action coding, machine learning was employed [6]. Even so, mouth features are also employed in the evaluation of tiredness and sleepiness. Qiang and associates Qiang and colleagues suggested employing an using an active infrared sensor identify head and student movements in

dimly lit environments. The movement of facial features is smoothed out by the facial feature-recording Kalman filter. The Gabor wavelets are used to rapidly identify features. [7, 8]. Once the face and eyes were identified by the Adaboost classifier, The Gabor filter of Esra et al. was used. The output is normalized before being sent to a support vector machine, a classifier that is data-driven. (SVM) [9]. Picture analysis is implemented using top-down architecture in a number of current works. First, the human face is identified using a technique founded on an enhanced Haar wavelet cascade look for the eyes in the face. The optical fluxes around the eyes are studied to be able to measure any noticeable eye blinks.

The Audio/Visual Emotion Challenge and Workshop (AVEC 2019) presented algorithms to detect depression or state-of-mind using artificial intelligence and cross-cultural affect. [10]. There are three smaller challenges in all. 1. State of mind (SOMS), which uses personal narratives to forecast mood

2. An artificial intelligence agent conducts virtual interviews to detect depression and predict mood. 3. Cross-Cultural Emotion Recognition: In this bilingualism in the natural environment, German and Hungarian are used as instruction and evaluation languages, while Chinese is only used for testing. The SOMS mood level was the most reliably predicted, even though both were used to train the system two types of evaluation: dynamic and static evaluation. Because of variations in dimension calculation in a noisy environment, alternative approaches to depression level prediction felt more difficult. Combined audio and visual characteristics used with the VGG and ResNet architectures for emotion classification

Innovative hybrid algorithm developed by Smys et al. demonstrated successful emotion recognition. Furthermore, There has been an increase in forecast accuracy. using multiple classifiers instead of just one [11]. According to this research, one can utilize can use The dataset from Twitter to assess people's depression by looking at their internet footprint, including their tweets. With increased sensitivity and accuracy, the analytical output on the selected dataset showed an early identification of the melancholic phenomenon.

To recognize the facial emotion, Pandey and associates. suggested a system utilizing deep learning and fuzzy inference combination. The index value for the emotion classification with the anticipated class is generated using VGG16. The index of classes and related The corresponding Fuzzy inference system receives images for the purpose of estimating the degree of intensity attained emotion. This system can identify the feelings of happiness, sadness, surprise, and anger [12]. This paper uses the FER 2013, CK +, and KDEF merged datasets for testing. An essential part of technical interaction with computing devices is Facial Emotion Recognition (FER). Despite the task's stimulation of precise results, CNN's VGGNet architecture makes it feasible [13]. The unprocessed image data can be used to train the CNN model, or additional auxiliary data can be added. We use the LBP classifier to identify the emotion on the face, and we can also remove the HOG attributes to prepare the image for model training [14].

II. Dataset

Image folders for the facial expression dataset (fer2013) 48 x 48 pixel grayscale images of appearances [12,13] make up the dataset [15]. The faces have subsequently been enlisted with the intention of essentially engaging the face. Assemble each face according to one of seven orders, taking into account the tendency revealed in the investigation (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The dataset is divided in the ratio 8:1:1 into train, value, and test. We combined value and test images to create a test set for our model. Finally, there are 28,709 photos in the data set. As stated in Table 1 below, the train set consists of 22968 models, and the test set consists of (test + value) 5741 models.

TABLE I. TRAIN AND TEST DATASET: EMOTIONS MATTER

Emotions	Train	Test
Angry	3995	962
Disgust	436	111
Fear	4097	1024
Happy	7215	1774
Neutral	4830	1247
Sad	3171	831
Surprise	4965	1233

The classification algorithm uses the characteristics of the face that are taken from this dataset and used as a source for determine the image's corresponding emotion. To train the model, a convolution neural network is employed. The purpose of the tests is intended to support the model and predict the related emotion. This includes the display of different emotions in test and train data.in Fig 1.

A. Video Dataset

The video focuses on how people react in a public place when they witness something intriguing on screen. The video, which lasts for about 29 seconds, shows various people's fast- changing reactions while also showing activities that move in relation to the screen. Two are present men and three women among its five members. They are articulating what is happening in front of them.

This is how the input will be used to advance the batch normalization and max pooling processes. The activation function will take part in this, going back to the earlier pooling and reshaping it before receiving the input as a result.

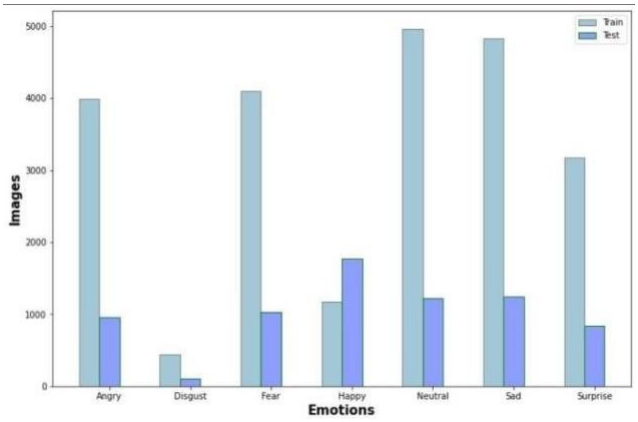

Fig. 1Number of Distinct Emotions Employed in Training and Testing Datasets

III. METHODOLOGY

A kind of artificial neural network with deep learning is a convolutional neural network (CNN) is employed for the classification and recognition of images.. The salient features are automatically identified. CNN is primarily utilized for computer vision applications like segmentation, localization, object recognition, video analysis, and image processing. The class of convolutional neutral networks is responsible for managing image extractions and associated symbolism information. In order to address this face affirmation problem, we therefore chose to incorporate a neural network with convolutions. The neural network type in question is of Neural Network is widely utilized for the picture evaluation duties like picture portrayal and is unquestionably very good at removing image aspects. Convolutions have been used to analyze expressions on faces. A particular kind of learning framework made up of artificial neurons arranged in multiple layers is called a neural network. Weighted data is received by each source node, which then passes it through an initial limit and receives the delayed outcome of the original limit. a data layer that will absorb the information. The dimensionality of the data layer is determined by the amount of data. The NN can discover intricate interactions in the data by utilizing a few enigmatic layers.

In the CNN architecture, every layer of artificial neurons generates multiple activation mechanisms that are transmitted to the layer above it. The second layer receives the first layer's output, which extracts the greatest basic features like diagonal or horizontal edges, and uses it to extract the finer details, such as combinational edges or corners. It can reach farther into the network. identify more complex elements such as people, objects, etc.

A neural network with numerous concealed layers are referred to as a deep neural network. [3]. The composition as it stands now revolves around the seven basic look classes that have been identified: fear, sadness, disgust, anger, neutrality, surprise, and happiness. The CNN computation included in the first version of this document concentrates on expressional evaluation and categorizes the provided image among these seven core emotion classes [7]. The most popular method for identifying human emotion is emotion recognition [4]. People generally vary in how accurately they pick up on other people's emotions. One area of relatively early investigation is the use of innovation to help people recognize emotions.

Many strategies have been implemented forth, and many more are emerging. A basic neural network can be efficiently set up to carry out classification. Regardless, it might not function properly with images. CNNs, in contrast, are fully integrated feed-forward neural networks.

CNNs are very good at minimizing boundaries without sacrificing the idea of the model. Since every pixel in a picture is regarded as a component, pictures have a high dimensionality, which is equivalent to the previously stated CNN limits. Although this review is still in its early stages, it offers a basic intuition. The CNN workflow that we used for our study is shown in Fig 2.

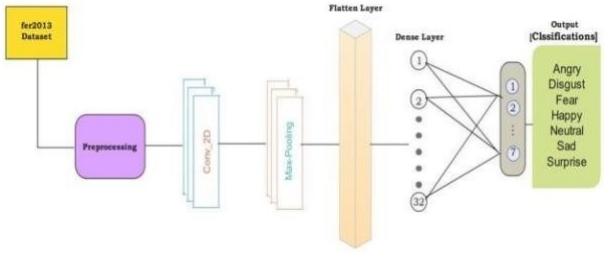

Fig. 2 Convolutional Neural Network Workflow for Classifying Facial Emotions

The convolution technique establishes the relationship depending on the kernel between the pixels. In this piece of work

, the image's features are retrieved utilizing the 3X3 kernel. Using stride value 1, this convolution operation calculates the convoluted value for each 3X3 pixel in the image. Thus, as seen in Fig. 3, the image's dimensionality will be decreased. The image will shrink to 4X4 in size after four layers of convolution. Next, use flattening to arrange the resulting convert a picture to a vector. exhibited in the convolutional neural network layer Fig 3.

CNN	Number of convolutional kernels of the layer 1	32
	Number of convolutional kernels of the layer 2	64
	Number of convolutional kernels of the layer 3	128
	Number of convolutional kernels of the layer 4	256
	Number of convolutional kernels of the layer 5	1024
	Size of convolutional kernels	3×3
	Strides size	2
	Number of convolutional layers	5
	Number of fully connected layers	2
	Activation function for convolutional and first fully connected layer	ReLU
	Activation function of the last layer	Softmax
	Learning rate	0.001
	Loss function	Binary Cross Entropy (BCE)
	Number of neurons of the first (fully connected) layer	32
	Number of neurons of output (fully connected) layer (CICIDS 2018)	Multi-class- 7
	Drop-out probability	0.5
	Number of epochs	60
	Batch size	3000

Fig. 3. Hyperparameter details for proposed models

Dropout: reduces overfitting by indiscriminately holding off on refreshing some hubs' loads. This prevents the NN from unduly depending on one hub in the layer. The SoftMax activation mechanism, which is typically used for multi-label classification, is applied to classification tasks in this work.

We built models for this project using CNN architecture. First, we imported every library needed to build the model. Next, we set the initial values for the generators for validation

and training. This involved rescaling and converting all of the images to grayscale that we required for model training. We developed the model using the fer2013 dataset. In this instance, we used the entire fer2013 dataset to train our network before storing the model's weights for later predictions. In the end, we used OpenCV harassed xml to predict the emotions and identify the face's bounding boxes in the webcam.

IV. Eperimental Results

An output layer, like a class assumption, that will offer a potential outcome. The quantity of courses that are actually desired determines the output layer's dimensions. Convolutional layers in Convolutional Neural Networks collect neighboring pixels. Consequently, those advancements have led to a prevalent comprehension of models found in images. For each layer in this work, I employed a few standard techniques. Examination of fer2013 Data.

A. Analysis of Static Image Results

During this cycle, the data will flow through two fully associated layers and four convolutional layers. The suggested work uses CNN's soft max and Relu activation functions for the flattening procedure. The training and testing processes' accuracy and loss are shown graphically in Fig. 4.

Training and validation loss is shown in the first graph, and training and testing accuracy is shown in the second. Test results for the different emotions show that several of the emotions have a 75% accuracy rate.

The correctness of each sensation is addressed in Fig. 5 Confusion Matrix by building a library of images expressing different emotions. 1. Anger; 2. Disgust; 3. Fear; 4. Happiness; 5. Neutral; 6. Sadness; and 7. Surprise are among the emotions mentioned. Every image used in the approval process is analyzed to identify emotions, and the resulting network is displayed here.

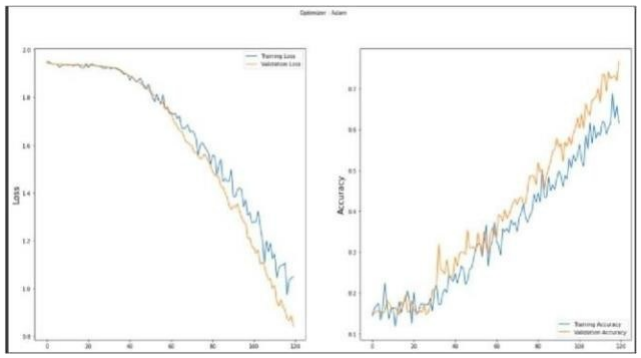

Fig. 5. Confusion Matrix Dataset Representation for Specific Emotions

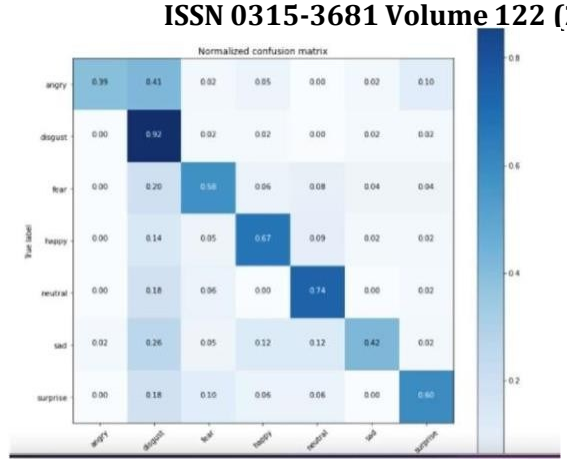

Fig. 6. Different emotions represented during model train

A model of a convolutional neural network and the aforementioned method are used to create a visual chart that illustrates the accuracy of different emotions, as seen in Fig. 7. This displays each feeling expectation rate as it passes through the aforementioned model. Initially, the model was evaluated with pictures tested using images of feelings found above the precision rate using a dataset that included a variety of happy pictures.

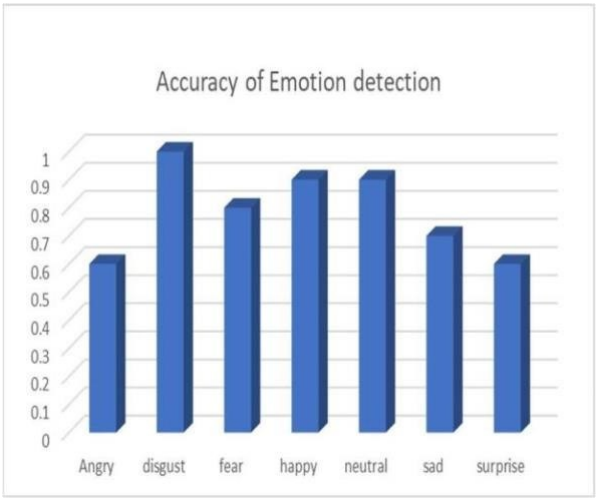

Fig. 7. Accuracy of Each Emotion with a Bar Graph

B. Analysis of Video Results

Expressions in the picture are recognized using the CNN model. Start by setting up the model using an image dataset in the preparation area. The information was thrown through the layer that was constructed using two completely associated layers and two convolutional layers. Finally, the SoftMax enactment layer is being used to level this. The Haar cascade is then used, which has the potential to yield impressive leads to the recognition of expressions [9].

Then it went on to a multi-person video, where it treated the entire video as a collection of frames and recognized the emotions of many persons in each frame. The result that is displayed is the identification of emotion in that specific video frame.

Fig. 8. Identifying Feelings While Watching the Video

The outcomes of a system that uses video input to recognize emotions in videos are shown in Fig 8. It's the process of using continuous video to identify people's emotions. It is possible to identify emotions by taking a frame every second. All of the characters in the video have their on-screen activities captured on video. Our framework is able to distinguish between the emotions of each character in the video.

V. Conclusion

Using the fer2013 dataset, a CNN-based method has shown promising results for facial emotion recognition, classifying expressions into seven distinct groups: happy, sad, neutral, startled, disgusted, fearful, and angry. The system's 90% accuracy rate in real-time video and image analysis is impressive. In order to maximize the accuracy of the CNN model, future research may investigate the use of pre-processing methods that are customized to the particular needs of the dataset. This would make it easier to implement a more durable and trustworthy facial emotion recognition system in real-world settings.

REFERENCE

[1] Proceedings of the Sixth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022) IEEE Xplore Part Number: CFP22J88ART; ISBN: 978-1-6654-8271-4.

[2] Lonare,Ashish,andShwetaV.Jain(2013).ASur- vey on Facial Expression Analysis for Emotion .Recognition. International Journal of Advanced Research in Computer and Communication Engi- neering 2.12.

[3]] N. Mehendale, “Facial emotion recognition using convolutional neural networks (FERC),” SN Appl. Sci., vol. 2, no. 3, 2020, doi: 10.1007/s42452-020-2234-1

[4] F. Abdat, C. Maaoui, and A. Pruski, “Human-computer interaction using emotion recognition from facial expression,” in Proceedings - UKSim 5th European Modelling Symposium on Computer Modelling and Simulation, EMS 2011, 2011, doi: 10.1109/EMS.2011.20

[5] Divya Meena Ravisharan, “An approach to face detection and recognition.” 2016 IEEE international Emotion Recognition Using Convolutional Neural Network (CNN) July 2021Journal of Physics Conference Series (https://www.researchgate.net/journal/JournalofPhysic s-Conference-Series-1742-6596).

[6] G. Bhardwaj, S. V. Singh and V. Kumar, "An Empirical Study of Artificial Intelligence and its Impact on Human Resource Functions," 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 2020, pp. 47-51, doi: 10.1109/ICCAKM46823.2020.9051544. https://developers.google.com/machine-learning/crash-course/firststepswith-tensorflow/toolkit .

[7] Kukla E., Nowak P. (2015) Facial Emotion Recognition Based on Cascade of Neural Neworks. In: Zgrzywa A., Choroś

K., Siemiński A. (eds) New Research in Multimedia and Internet Systems. Advances in Intelligent Systems and Computing, vol314. Springer, Cham. doi.org/10.1007/978- 3-319-10383- 9_7.

[8] P. Qiang Ji; Zhiwei Zhu; Lan, "Real Time Non- intrusive Monitoring and Prediction of Driver Fatigue," Vehicular Technology, IEEE Transactions, vol. 53, pp. 1052 - 1068, 2004.

[9] M. C. Esra Vural, Aytul Ercil, Gwen Littlewort, Marian Bartlett and Javier Movellan, "Drowsy Driver Detection through Facial Movement Analysis" in Human– Computer Interaction. vol. 4796, ed: Springer Berlin / Heidelberg, 2007, pp. 6-18 .

[10] Akamatsu, S., Sasaki, T., Fukamachi, H., Masui, N., and Suenaga, Y. 1992. An accurate and robust face identification scheme. In Proceedings, International Conference on Pattern Recognition. 217--220.

[11] Dr. S. Smys, Dr. Jennifer S. Raj, “Analysis of Deep Learning Techniques for Early Detection of Depression on Social Media Network - A Comparative Study”, Journal of trends in Computer Science and Smart technology (TCSST) (2021) Vol.03/ No. 01 Pages: 24- 39 [12] Pandey, A., Kumar, A. Facial Emotion Intensity: A Fusion Way. SN COMPUT. SCI. 3, 162 (2022).

[12] Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588.

[13] Amal, V. S., Suresh, S., & Deepa, G. (2022). Real-time emotion recognition from facial expressions using convolutional neural network with Fer2013 dataset. In Ubiquitous Intelligent Systems (pp. 541-551). Springer, Singapore.

[14] ULADZISLAU ASTRASHAB, Facial expression dataset image folders (fer2013) - Photos of faces expressing different emotions "https://www.kaggle.com/datasets/astraszab/facialexpressi on- dataset-image-folders-fer2013" , accessed on Feb 2022.