Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765805

NOVEL FRAMEWORK FOR CNN BASED FACE PHYSIOGNOMY

Shraddha Habbu
Dept. Of E&Tc, Vishwakarma Institute of Information Technology
Maharashtra, Pune

Gauri Ghule
Dept. Of E&Tc, Vishwakarma Institute of Information Technology
Maharashtra, Pune

Archana Ratnaparkhi
Dept. Of E&Tc VIIT, Maharashtra, Pune

Pallavi Deshpande
Dept. Of E&Tc VIIT, Maharashtra, Pune

Ketaki Kshirsagar
Dept. Of E&Tc VIIT, Maharashtra, Pune

Arti Bang
Dept. Of E&Tc VIIT, Maharashtra, Pune

Abstract—Humans have a tendency to express a standard and cardinal range of emotions through their facial expressions. There are seven basic raw emotions: happiness, anger, sadness, surprise, fear, neutral, contempt, and disgust. However, it is clear that most research focuses on the five primary emotions: happiness, surprise, anger, sadness, and neutrality. With the advancement of technology, it is now possible to automatically recognise a per- son's emotions through videos and images, thanks to algorithms that detect, extract, and evaluate these facial expressions. The algorithm we propose and present is a hybrid face detection and feature extraction model that can detect the previously mentioned emotions in real time. The former is accomplished by means of the Haar cascades object detector, while the latter is accomplished by means of Deep Convolutional Neural Networks (DCNN). The hybrid method aids in the rapid initial face detection step, followed by the extraction of features (starting with generalised [low frequency] features and subsequently more specific [high frequency] features). On the FER 2013 dataset, the proposed model achieved an accountable training accuracy of 84.40% and a validation accuracy of 74.74%.

Index Terms—Convolutional Neural Networks, Facial Expres- sions, Real-Time Detection, Feature Extraction

I. INTRODUCTION

Significant, intricate, and ever-expanding in the fields of psychiatry, neuroscience, biomedical science, and health [2] is the study of emotions. The primary topics of study and research in

these fields are predicting human emotions and computer-assisted psychological problem diagnosis. Analysis of voice and facial gestures, Galvanic Skin Response (GSR), Electroencephalography (EEG), and multisensory scanning behaviour are methods used to detect emotional states [1]. Humans have an innate and largely unconscious tendency to read and interpret the emotions of others based on a variety of verbal and nonverbal cues, including their words, voices, and body language. As previously stated, this analytical skill or ability likely stems from the fact that all beings share the same set of emotions. In spite of language and cultural barriers, these emotions are manifested through facial char- acteristics that are universally consistent. Positive, negative, angry, surprised, and neutral are the five most fundamental emotions. These emotions can be recognised by machines for a variety of applications, including content analysis, en- tertainment industries, personalised automobiles, etc. With the rise of deep learning in recent years, significant progress has been made in image classification. Convolutional Neural Networks (CNN), invented in 1988 by Yann LeCun, resemble an artificial neural network. One of the most popular deep learning architectures for image recognition, classification, and segmentation, these networks are well-known. This is because CNN has a significant advantage over other algorithms in terms of preprocessing. CNN's neuron architecture resembles the patterns that human brain cells use to communicate with one another. Keeping these facts in mind, we implemented Face Physiognomy, an emotion recognizer for the face. This AI-powered technology analyses facial expressions from both static images and video feed to determine the emotional state of the subject. In addition, we utilised some of Python's most potent libraries to create a web application that identifies human facial expressions in real time. TensorFlow, Keras, Open CV, Matplotlib, and Flask, to name a few, are included.

II. LITERATURE SURVEY

The science of machine learning has made significant advancements in recent years. Deep learning is a technique that use deep networked architectures comprised of many linear/nonlinear transformations to model high-level data ab- stractions. Deep learning systems are intelligent systems that represent complex facts from real-world events in a manner similar to the human brain and aid in making intelligent decisions. Deep learning, also known as deep structured learn- ing or hierarchical learning, is a subset of machine learning techniques focused on data representation comprehension. It has had a dramatic impact on the performance of computer vision for various tasks, such as image categorization and object recognition, that were previously impossible. Research involving graphical modelling, pattern recognition, and signal processing [1], computer vision [2], speech recognition [3], language recognition [4,5], audio recognition [6], and face recognition (FR) [7] utilise deep learning. Numerous feature extraction techniques have been proposed for use in biometric systems, such as principal component analysis (PCA) [9], independent component analysis (ICA) [10], local binary patterns (LBP) [11], and the histogram method [12]. Deep learning, specifically the convolution neural network (CNN), has recently been the standard technique for feature extraction in FR [13], demonstrating amazing benefits.

A. Proposed Model

The system we've proposed is a hybrid strategy with two phases. Using the Haar Cascades object detector, faces in the given image are identified in the first stage. The second stage

employs a Deep Convolutional Neural Network (DCNN) to extract facial features for the classification of the subject's emotion. Through this simple architecture and hyperparameter tuning, we were able to achieve significant results and a level of accuracy comparable to other complex models. Below is a comprehensive comparison between the proposed model and other models.

B. Database

For training a face emotion recognizer, there are numerous open-source datasets available. We utilised the FER 2013 [10] data 1, an additional publicly accessible dataset on Kaggle. The images have a resolution of 48x48x1, preserving the context of the faces. Faces of the subject in the images are roughly centred so that they occupy the majority of the image's area, thereby facilitating the training of the neural network model. The dataset contains 30,219 images, which are then divided into 24,282 images for training and 5,937 for testing.

Fig. 1. Database

C. Object Detection

As shown in the figure 2Two phases comprise the hybrid model: detection and classification. Using the Haar cascades object detector, the first phase detects faces from uploaded photos or in real-time via the video feed. Edge or line detection features introduced by Viola and Jones in their research work

[5] are utilised by the method. The method entails sliding attribute boxes across a picture and computing the difference in cumulative grey levels between adjacent spots. The difference is then compared to a threshold value to determine if an object has been discovered. This involves the application of cutoff values that have been trained on specific attribute boxes and characteristics.

D. Model Architecture

For the suggested model, we employed a neural network architecture that is solely convolutional. In the CNN archi- tecture [7, 8], we utilised four convolutional layers, each with batch normalisation for improved efficiency, followed by max-pooling and dropout regularisation to prevent overfitting. These tiers included 64, 128, 512, and 512 kernels, respectively. Each layer's percentage of neurons dropping randomly was set to 25%. This was followed by a flattened layer and 2 layers with 256 and 512 neurons, respectively, each with

full connectivity. The completely connected layers also contained a 25% dropout layer. The activation function utilised throughout the network was relu, with the exception of the final layer, which employed a softmax activation function to produce probability of the image class membership.

E. Model Training Phase

In order to consolidate and optimise the network parameters, the batch size was set to 64 and the epochs were determined to be 50 following several network training experiments and adjustments. The rate of learning is set to 0.0005 All kernel sizes were set as 3x3. The kernel size of only the second con- volutional layer was 5x5. During model training, we employed ReduceLRonPlateau, model checkpoint, and plot live losses as callbacks. Adam [9] was employed as the optimizer, and the loss function was categorical cross-entropy.

III. APPLICATION AND IMPLEMENTATION

As a real-world application and to evaluate the effectiveness of our model, we constructed a fully functional website capable of identifying the emotions of the user, where the user has the choice of uploading an image or using the device's webcam. According to our use case, we modified and adapted a bootstrap template for website creation. HTML, CSS, and JavaScript were the primary technologies employed for this project. As depicted in Figure 3, the model was given real-time photos and videos via a representational state transfer application interface (REST API) developed with the flask module in Python. To validate and demonstrate that the trained model was sufficiently generalizable, the included faces had any indicated limitations (were fully random) and exhibited substantial variance in age, gender, ethnicity, image

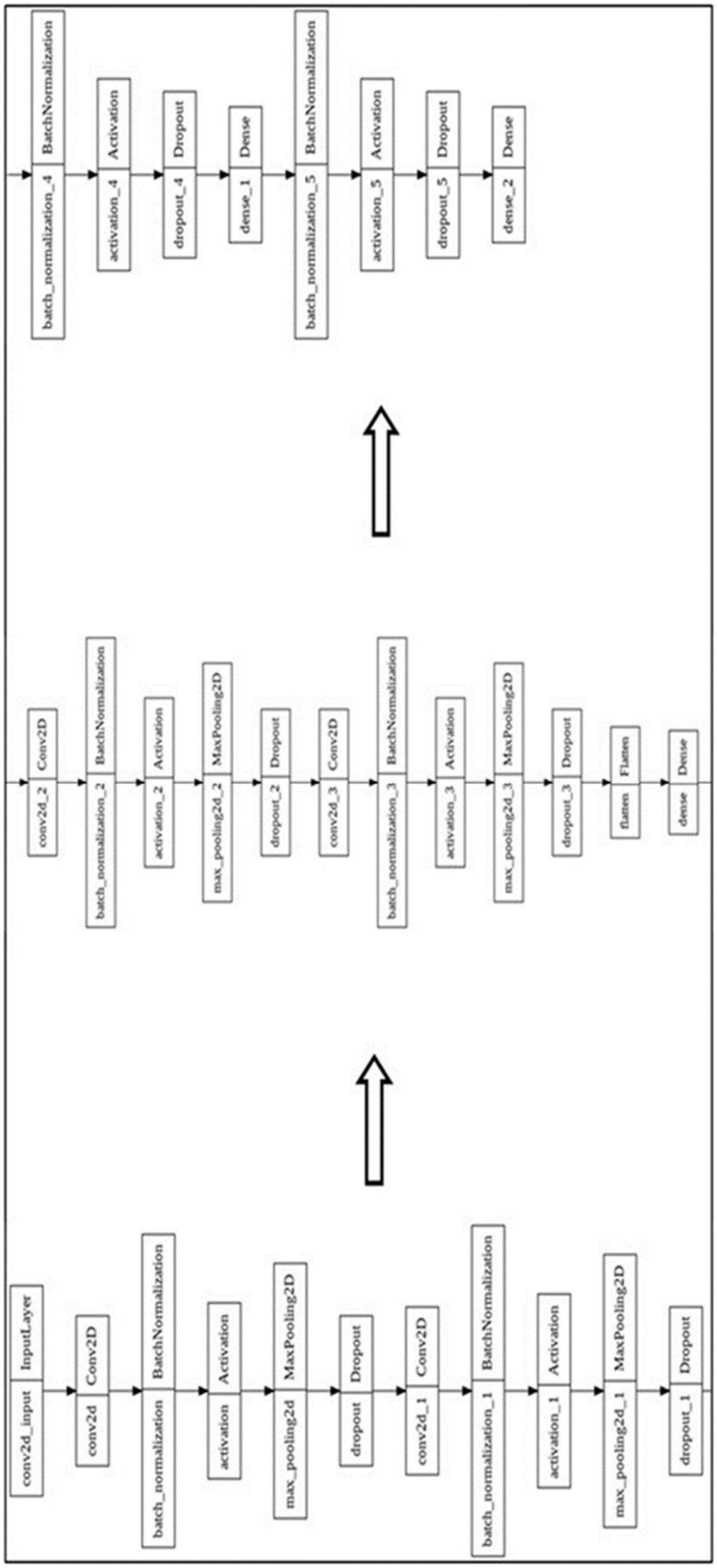

Fig. 2. Block Diagram

background, and lighting conditions. The model accurately identified and predicted the majority of emotions, with an accuracy of well over 90%, demonstrating its robustness and efficacy.

Fig. 3. Database

IV. RESULTS AND DISCUSSIONS

The complete implementation starts with accepting the input (image or video) from the user. These images/frames are then converted into grayscale images using the OpenCV library in python. Next, we use Viola-Jones' Haar cascades to detect faces. Keeping the detected face in context, we crop and resize the image to a target of 48x48x1, so that we can serve it as input to our trained classifier. Following that, the Convolutional Neural Network captures features from the input image. The feature extraction takes place such that in the initial layers, the low-frequency features, such as smoothening, and in the subsequent layers, more specific, high-frequency features like edges are learned. This serves as a good learning technique that makes CNN a popular network for dealing with images. Lastly, the emotion of the subject in the image is classified in its respective class. With the proposed model, we achieved a reasonable training accuracy of 85.40% and a validation accuracy of 74.74%. Figure3 and Fig4 shows the training and validation accuracy and loss against the number of epochs plot. Here, we can clearly comprehend that the loss saturates after a certain number of epochs.

A. Object Detection

A Confusion matrix is a N x N matrix that is used to evaluate the efficacy of a classification model, where N is the number of target classes. The matrix compares the ground truth to the predictions of the deep learning model. This offers us a comprehensive picture of how well our classification model is working and what kinds of errors it is making. Another way to assess the classifier's performance is through a classification report, which is used to evaluate the predictive accuracy of the algorithm. The classification report enlists metrics like precision, recall, and F1-Score for each class. How many of the model's predictions are true, and how many are wrong are analyzed in this report. More specifically, as depicted in Figure 5, True Positives, False Positives, True Negatives, and False

TABLE I
CLASSIFICATION REPORT

	Precision	Recall	F1-Score	Support
Angry	0.70	0.61	0.65	960
Happy	0.87	0.87	0.87	1825
Neutral	0.65	0.69	0.67	1216
Sad	0.63	0.64	0.63	1139
Surprise	0.85	0.85	0.85	797
Accuracy			0.74	5937
Macro Avg	0.74	0.73	0.73	5937
Weighted Avg	0.75	0.74	0.74	5937

TABLE II
COMPARATIVE ANALYSIS

	Precision	Recall	F1-Score	Support
Angry	0.70	0.61	0.65	960
Happy	0.87	0.87	0.87	1825
Neutral	0.65	0.69	0.67	1216
Sad	0.63	0.64	0.63	1139
Surprise	0.85	0.85	0.85	797
Accuracy			0.74	5937
Macro Avg	0.74	0.73	0.73	5937
Weighted Avg	0.75	0.74	0.74	5937

Negatives are utilized to predict the metrics of a classification report. Through the confusion matrix and classification report (figure 5 and Table II respectively), it was observed that while happy, neutral and surprise were “easy-to-distinguish” emotions, the model was somewhat confused

V. CONCLUSION

A hybrid face detection and classification method has been implemented in which face images are used to train a clas- sification model that predicts the five rudimentary emotions given an image or a video feed. Through the proposed model, we achieved results that are almost homogeneous to previous work in this domain. We accomplished the same with a smaller, completely convolutional, but definite and concrete neural network design, as well as hyperparameters such as the number of neurons in each hidden layer, kernel size, batch size, number of epochs, and learning rate. The model is reasonably accurate at predicting positive, neutral, negative, and surprise emotions. However, it is not very good at detecting sad expressions. Possible explanation for this performance is because the model confuses the classes for sad and furious. Another possibility for this performance is an imbalance in the data for the sad class. Nonetheless, the experimental results suggest that the proposed model is effective at recognising the subject’s emotions. Future research should focus on enhancing the robustness of the classifier by incorporating more training images from diverse datasets,

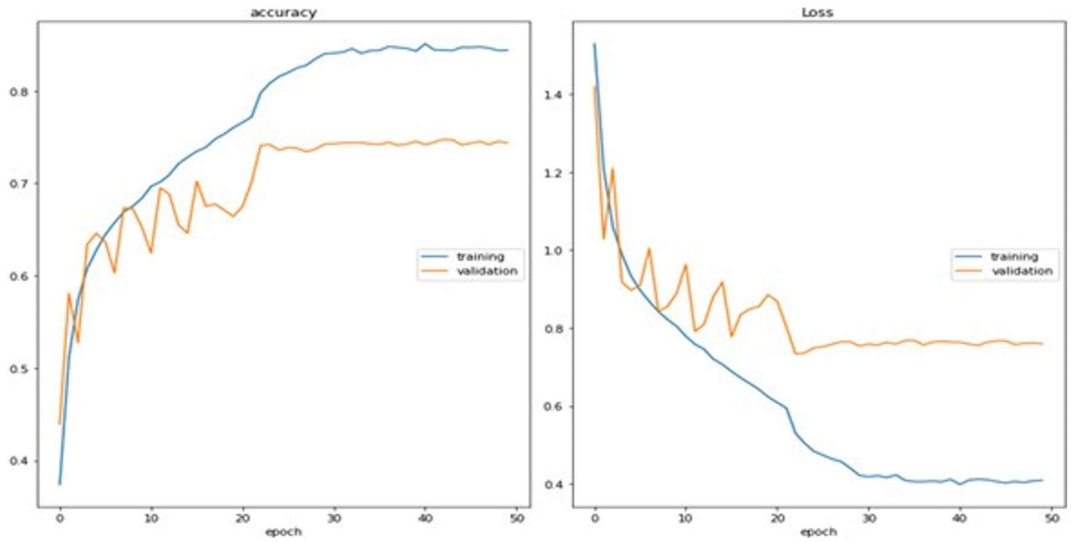

Fig. 4. Graphs for Training and Validation Accuracy and Loss

Fig. 5. Confusion Matrix

exploring more accurate detection methods, and considering the classification of more nuanced and complex expressions, all while maintaining computational efficiency. Depending on the use case and extent of applicability of the suggested model, we may even consider generating a custom dataset for a certain group of persons. Alternatively, one might utilise the power of transfer learning by employing robust CNN models trained on potentially enormous data sets. These pre-trained networks can then be fine-tuned according to our use case. This model could be enhanced with additional research and resources, resulting in greater accuracy.

ACKNOWLEDGMENT

We are extremely thankful to Research and Development wing and administration section for their support and encour- agement.

REFERENCES

[1] M. Cabanac, ”What is emotion?,” Behavioral processes, vol. 60, pp. 69-83, 2002.

[2] R. Roberts, ”What an Emotion Is: a Sketch,” The Philosophical Review, vol.97, 1988

[3] Cohen, I., Garg, A., & Huang, T. S. (2000, November). Emotion recognition from facial expressions using multilevel HMM. In Neural information processing systems (Vol. 2). State College, PA, USA: Citeseer.

[4] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046.

[5] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.

[6] Srivastava, Saransh. (2020). Human Facial Expression Recognition Us- ing TensorFlow And OpenCV. 10.13140/RG.2.2.19218.89288.

[7] Tensorflow. TensorFlow. (n.d.). Retrieved February 2, 2023, from https://www.tensorflow.org/

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Handwritten digit recognition with a back- propagation network,” in NIPS’89

[9] Kingma, Diederik P., and Jimmy Ba. ”Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980 (2014).

[10] Wolfram Research, ”FER-2013” from the Wolfram Data Reposi- tory (2018). [Online] Available:https://www.kaggle.com/c/challenges-in- representation-learning-facial-expression-recognition-challenge/data

[11] Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 971–987. [CrossRef]

[12] Liu, Y.; Lin, M.; Huang, W.; Liang, J. A physiognomy based method for facial feature extraction and recognition. J. Vis. Lang. Comput. 2017, 43, 103–109. [CrossRef]

[13] Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in faceverification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus,OH, USA, 23–28 June 2014; pp. 1701–1708.