Identifying Human Emotions from Facial Expressions with Deep Learning

Phavish Babajee, Geerish Suddul, Sandhya Armoogum, Ravi Foogooa
University of Technology, Mauritius
Pointes-Aux-Sable, Mauritius
g.suddul@umail.utm.ac.mu

***Abstract*—The identification of facial expressions that reveal human emotions can help computers to better assess the human state of mind, so as to provide a more customized interaction. We explore the recognition of human facial expressions through a deep learning approach using a Convolutional Neural Network (CNN) algorithm. The system uses a labelled data set containing around 32,298 images with multiple facial expressions for training and testing. The pre-training phase involves a face detection subsystem with noise removal, including feature extraction. The generated classification model used for prediction can identify seven emotions of the Facial Action Coding System (FACS). Results of our work in progress demonstrate an accuracy of 79.8% for the recognition of all basic seven human emotions, without the application of optimization techniques.**

Keywords—Convolutional Neural Network, Human Emotions, Facial Action Coding System, Face Detection, Feature Extraction

I. Introduction

Human facial expressions are linked to emotions, and learning to depict them is an important step for creating seamless human to machine communication. To understand human emotions, facial expressions play a major role along with both speech annotations and non-verbal communications such as hand gestures and head motions [1]. The 7% rule which dictates that 93% of human communication is nonverbal [2]. As at today, there are approximately 6,909 spoken languages in the world [3]. To convey 7% of the communication, it is estimated that over 90 thousand languages is required to understand the 93% other form of communication. The Facial Action Coding System (FACS) Ekman & Friesen [4] introduced in 1978, defines seven major facial expressions which are conveyed by human beings without the use of words, namely fear, anger, surprise, disgust, happiness, sadness and neutral. This system is considered as the threshold when it comes to Facial Expression Recognition (FER) [5]. Our human vision system can detect those emotions almost impeccably as the average human can accurately guess up to around 150 frames per second (fps). Face to face communication is a real-time process, that requires fast thinking to simultaneously detect those expression and process them. Our minds are not always capable to achieve this. Affective Computing [6], is an integrative field across data science, psychology, engineering and cognitive science tasked to study and develop devices and systems that can identify, interpret and process human emotions. This particular field requires systems, using modern artificial intelligence algorithms, in particular neural networks, to identify the basic human emotions with the help of the Facial Action Code System (FACS).

Gajarla and al. proposes a system based on VGG Image Net [7] to perform sentiment analysis on images obtained from Twitter and Flickr APIs in accordance to the FACS system. The underlying architecture is a Support Vector Machine (SVM) classifier which yields results only for the happy emotion. The network uses a deep layer approach (50 layers) reaching an accuracy of 73%. Revina and al. proposes an approach which is based on the Gaussian Filter [9] to resize input images, Viola/Jones algorithm [10] to detect faces and Scale Invariant Feature Transform (SIFT) for face alignment. A combination of the Jaffe database [11] and Cohn Kanade database [12] have been used to test the classifier built using a Support Vector Machine alongside a Convolution Neural Network (CNN). Their results demonstrate high accuracy and identification of 7 expressions. The work of Tarnowski and al. is based on the Kinect Device [13] which measures the displacement and distance of the 121 datapoints identified by the FACS system. The classifier is based on a Mutli-Layer Perceptron (MLP) with two hidden layers trained and successfully tested using the KDEF [14] database.

II. Research Methodology

The research work is based on an experimental approach, which consists of incremental improvements on the developed prototype of the system. A traditional Facial Expression Recognition (FER) system constitutes of three key steps: facial detection, feature extraction, and classification of the facial expression. Face detection refers to the pre-processing stage where face regions are located [15]. Facial feature extraction aims to find the most fitting representation of facial images for recognition. Our approach is based on a geometric feature-based method which extracts shapes and locations of facial components information using an edge detection framework [9]. As for the last step, facial expression classification, a probabilistic classifier is used to identify different expressions relative to the extracted facial features. A Convolutional Neural Network is used, with the neurons in the output later defined by the amount of facial expressions targeted by our scope. The results will consequently be evaluated using the accuracy rates of similar previous work in the area of Facial Emotion Recognition.

III. Proposed Work

Figure 1 represents the two-fold approach of our System. It consists firstly of training a model based on dataset of images containing a mixed set of facial expressions. The images require preprocessing, including face detection, followed by feature extraction. Once the model is trained, it can be used for making predictions based on a new input facial image.

978-1-7281-8259-9/20/$31.00 ©2020 IEEE

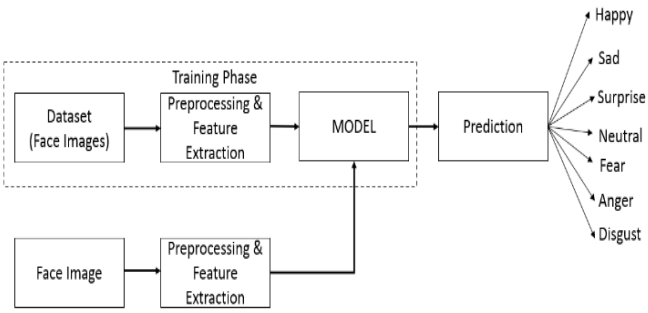

Figure 1: Proposed Architecture

A. *Face Detection*

The first step of our proposed system is to extract the useful facial areas of a human being that will facilitate the feature extraction process and the classification process. The pre-processing is a two-fold approach using the Viola-Jones algorithm [16]. The photograph is first cropped using a Haar Feature Selection [17] to choose only the face area and to eliminate unimportant surrounding data. The cropped image is then converted to greyscale.

B. *Feature Extraction*

The next step after detecting a face from a given image is the feature extraction process that helps isolate the important facial areas. Two methods can be used for feature extraction: Analytic approach, Holistic approach [18]. The holistic approach uses raw facial image as input, while the analytic approach focuses on some of the important facial features are been detected and extracted from face. The analytic approach is being used in this paper, where we the extracted selected features that were obtained using edge detection from the image are sent as input to a classifier. The holistic approach takes the global properties of the patterns being obtained into consideration while the analytic approach we used computes a set of geometrical features based on feature vectors.

C. *Neural Network Classifier*

Neural networks are at their core made up of various layers which are in turn made up of various neurons [8]. The basic idea of any neural network is to mirror the synapses in our human brains. Unlike traditional programming concepts, a neural network can learn to recognize patterns, make decisions, and most importantly make predictions without being programmed explicitly. Similar to our human minds, a well-constructed neural network can learn independently. The distinction between an actual human mind and a Neural Network however is an obvious one as the latter is at its core just a software. A traditional neural network ranges from tens to hundreds, or even millions of artificial neurons arranged sequentially in a layered series, each of which connects to the layers on either side. There are different types of units, namely; the Input Unit which are designed to receive data from outside the constructed network. They are found in the left most layer of a neural network. The second type of units are the Output Units that show what the network is 'learning' and are usually situated on the right most layers. And the third, Hidden units, are usually made up of different layers, hidden units interconnect the Input and Output unit in the middle. The connections between units are represented by numbers called weights that are either positive or negative. High weights are relative to the amount of influence, as such a unit with a high weight will have an equally high influence rate on other. Our proposed system uses a 3 layered Convolutional network with n input layers, 7 output layers for each of the emotions in the FACS and a hidden layer in the middle.

CNN yields a superior accuracy rate than other neural network-based classifiers [9]. CNNs generally consists of two layers; a convolutional layer and a subsampling layer wherein the two-dimensional images are taken as input. In first aforementioned layer, the feature maps are produced by complex convolution kernels with the two-dimensional images whereas in the subsampling layer, pooling and redeployment are performed [19]. Furthermore, two important perceptions sharing weight and sparse connectivity are also present in the CNN [20]. The diagram below depicts a similar 5-layer neural network.

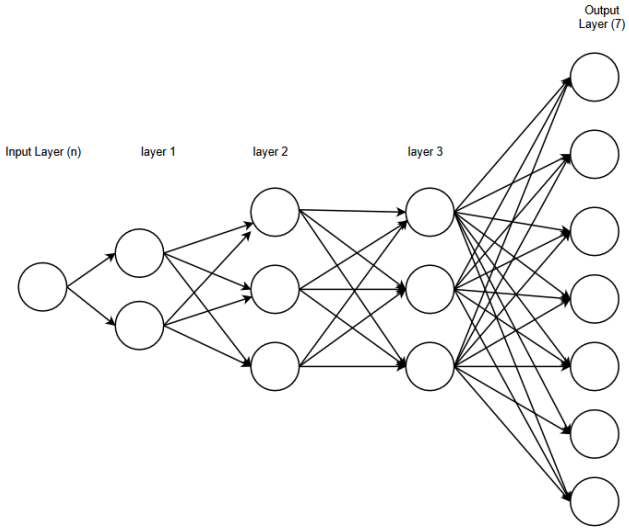

Figure 2: Layered Neural Network Model

IV. Experiments & Results

A. *Dataset*

We are using the Face Expression Recognition (FER2013) dataset [21] as it is among the biggest open-sourced datasets for emotion recognition. It consists of around 32,298 images, each individually labelled. The FER 2013 dataset is preprocessed, consisting of cropped grayscale images in the size of 48 by 48 pixels. The dataset takes the form of a CSV file with three columns. The first two represents "emotion" and "pixels" and the last its usage. The "pixels" column contains the pixelized value of each image. The "emotion" column contains a numeric code from 0 to 6 representing the emotion emitted in the image which is represented in the Table 1.

Table 1: Dataset Key

Number	Emotion
0	Angry
1	Disgust
2	Fear
3	Happy
4	Sad
5	Surprise
6	Neutral

B. Model Training & Testing

70% of the FER2013 dataset is used as input for the training phase, to which feature extraction is applied. Key attributes of the image are computed and stored as feature vectors which represent the essential properties detected in the faces. This preprocessing step allows for reduced data sizes as just a handful important features from an entire image are selected and, more concise information is obtained from feature selection. The classification rate is traditionally determined in the testing phase. The testing and training phases both follow the same steps of feature extractions, and classification. Classification is however different for the testing phase as the features are tested against the model constructed in the training phase. The outcome of this step yields a score which indicates the emotion predicted by the model.

In an attempt to determine the accuracy of the model, the Convolutional Neural Network ran for three different set of iterations: 10, 100 and 500. All our experiments have been conducted on a device powered by a Nvidia GeForce GTX 1050Ti GPU with 8 GB of RAM.

Table 2: Iterations table

Epochs (Training Iterations)	Training Time (hours)	Accuracy Rate (Testing)
10	1	71
100	16	79.8
500	98	79.98

The optimal training phase lasted for 16 hours using the full dataset containing 28 709 designated images. This model was trained using 100 epochs on a 256-batch size and learned 122 images on each pass. Extensive tests were also carried out by increasing the number of epochs to 500 and it took approximately 4 days to complete, and yielded only a better accuracy rate by only 0.18%. It is assumed to have occurred due to oversaturation of the selected dataset in the model's training phase.

While testing the model, with images it has not seen before, subsequent training on 100 epochs, results in an accuracy of 79.8%. The comparison in Table 3 shows that our result is somewhere between the accuracy achieved from the work of Noh et al. and Zhang et al. at this stage. No optimization techniques have yet been applied on the CNN, which might provide improvements on accuracy but have implications on performance (e.g. increase in training time and use of more powerful hardware resources)

Table 3: Results Comparison

Author Name, Year	Accuracy (%)	No of Emotions Detected
Noh et al. [22]	75	6
Zhang et al. [23]	82.5	6
Our Model	**79.8**	**7**

V. CONCLUSION

Face to face communication occurring in real time is present in both work and personal life environments. As such, Facial Expressions Recognition systems have a wide spectrum of applications ranging from basic security systems to real time Facial emotion detection in interviews, as non-verbal cues are an essential form of communication. Being a work in progress, we presented a deep learning neural-network based method for facial expression classification in this paper limited to images containing a single human face. We plan to extend it for multiple faces and using live camera streaming. The future scope is also centered towards improving the accuracy of the results and encapsulating the model into a consumable service that will in turn be available for use by various cooperating systems.

REFERENCES

[1] Mehdi, G., 2015 A Review of Multimodal Biometric Systems Fusion Methods and Its Applications ICIS, USA.

[2] A. Mehrabian, Communication Without Words, Psychology Today 2 (4) (1968) 53–56.

[3] Ethnologuecom. 2016. Ethnologue. [Online]. [15 October 2019]. Available from: https://www.ethnologue.com/guides/how-many-languages

[4] Ekman, P. and Friesen, W. 1978. "Facial Action Coding System: Investigator's Guide.", Consulting Psychologists Press, Palo Alto, CA.

[5] Ekman, P. and Friesen, W. 2002. Facial Action Coding System: Investigator's Guide. Volume 1, pp. 3-5

[6] Jen-Chun Lin, Chung-Hsien Wu, Wen-Li Wei, Semi-coupled hidden Markov model with state-based alignment strategy for audio-visual emotion recognition, Proceedings of the 4th international conference on Affective computing and intelligent interaction, 2011, Memphis, TN

[7] Gajarla, V and Gupta, A. 2016. Emotion Detection and Sentiment Analysis of Images. Georgia Institute of Technology

[8] Gurney, K. 1997. An Introduction to Neural Networks. Taylor & Francis, Inc., Bristol, PA, USA.

[9] Revina, M and Emmanuel, W. 2018. A Survey on Human Face Expression Recognition Techniques. Dept of Computer Science, Christian College, Sunadaranar University, Tirunelveli, Tamil Nadu, India

[10] Biswas, S., 2015. An Efficient Expression Recognition Method using Contourlet Transform. Int. Conf. Percept. Mach. Intell. pp. 167–174.

[11] Lyons, Michael, Kamachi, Miyuki, & Gyoba, Jiro. (1998). The Japanese Female Facial Expression (JAFFE) Database [Dataset]. Zenodo. http://doi.org/10.5281/zenodo.3451524Cohn-Kanade AU-

Coded Expression Database, ‘‘http://www.pitt.edu/~emotion/ckspread.htm”, accessed on 2017.

[12] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.

[13] Tarnowski, P., Kołodziej, M., Majkowski, A. and Rak, J. (2017). Emotion Recognition using facial expressions. ICCS Zurich.

[14] Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces – KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

[15] Cootes, T., Edwards, G. and Taylor, C, “Active appearance models”, IEEE Trans. Patternnal.Mach. Intell., vol. 23, no. 6, Jun. 2001, pp. 681-685.

[16] Happy, S.L., Member, S., Routray, A., 2015. Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. 6, 1–12.

[17] Jayalakshmi, J and Mathew. T 2017. Facial Expression Recognition and Emotion Classification System for Sentiment Analysis Mar Baselios College of Engineering and Technology Trivandrum, India

[18] Pantic, M., & Rothkrantz, L. J. M. 2000. Automatic analysis of facial expressions: The state of the art. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(12), 1424-1445.

[19] Shan, K., Guo, J., You, W., Lu, D., Bie, R., 2017. Automatic Facial Expression Recognition Based on Deep Convolutional-Neural-Network Structure. IEEE 15th Int. Conf. Softw. Eng. Res. Manag. Appl. 123–128.

[20] Rashid, T.A., 2016. Convolutional neural networks-based method for improving facial expression recognition. Intell. Syst. Technol. Appl. 73–84. https://doi.org/10.1007/978-3-319-47952-1.

[21] Goodfellow, I., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Lee, D., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L. Xu, B., Chuang, Z., and Y. Bengio. (2013) Challenges in Representation Learning: A Report on Three Machine Learning Contests. In: Lee M., Hirose A., Hou ZG., Kil R.M. (eds) Neural Information Processing. ICONIP 2013. Lecture Notes in Computer Science, vol 8228. Springer, Berlin, Heidelberg

[22] Noh, S., Park, H., Jin, Y., Park, J., 2007. Feature-adaptive motion energy analysis for facial expression recognition. Int. Symp. Vis. Comput., 452–463

[23] Zhang, L., Tjondronegoro, D., Chandran, V., 2014. Random Gabor based templates for facial expression recognition in images with facial occlusion. Neurocomputing 145, 451–464. https://doi.org/10.1016/j.neucom.2014.05.008.