Improving the Performance of Deep Learning in Facial Emotion Recognition with Image Sharpening

Ksheeraj Sai Vepuri, Nada Attar

***Abstract*—**We as humans use words with accompanying visual and facial cues to communicate effectively. Classifying facial emotion using computer vision methodologies has been an active research area in the computer vision field. In this paper, we propose a simple method for facial expression recognition that enhances accuracy. We tested our method on the FER-2013 dataset that contains static images. Instead of using Histogram equalization to preprocess the dataset, we used Unsharp Mask to emphasize texture and details and sharpened the edges. We also used ImageDataGenerator from Keras library for data augmentation. Then we used Convolutional Neural Networks (CNN) model to classify the images into 7 different facial expressions, yielding an accuracy of 69.46% on the test set. Our results show that using image preprocessing such as the sharpening technique for a CNN model can improve the performance, even when the CNN model is relatively simple.

***Keywords*—**Facial expression recognition, image pre-processing, deep learning, CNN.

I. Introduction

FACIAL expression recognition (FER) plays an important role in many fields such as human computer interaction, security, marketing, new analysis and so on [1], [15], [20], [21]. However, it is still a challenge to process the data and extract the features required for the analysis. To classify an expression into a number of finite expression categories with a high accuracy, computers need to learn various features for each particular expression. To achieve this goal, the database for expression feature learning should contain large set of images with many expression features.

The common databases used for FER analysis are divided into two categories. One type depends on recognizing basic facial expressions (e.g. happiness, sadness, surprise, anger, disgust, fear, and neutral) from human facial expression in RGB or greyscales images [15], [16]. The other type focuses on extracting fine-grained descriptions for facial expressions [2]. Both types have two issues, which is either the limited number of images in the dataset, or the acted expression instead of a spontaneous one in highly controlled environments. Both issues make it difficult for learning models to effectively classify expressions.

The main approach that is used for FER is based on CNN [23] which is fast to train and allows for real-time FER even when using standard computers. Models that use CNN is able to predict the facial expression label based on the categories of emotions and CNN usually provides a simple solution for FER when it is combined with image pre-processing steps. Xie et al. [23] conducted an experiment to evaluate their CNN model on a Few Training Samples using public databases (CK+, JAFFE). They found that CNN and specific image pre-processing steps can achieve competitive results when compared with other FER methods.

In this paper, we developed a method using the FER-2013 database for FER. We enhanced the images clarity by using image sharpening technique, which enhanced the performance of the CNN model while keeping it simple. Our model did not need to use any other preprocessing filters or supervised machine learning model such as SVM [17], [25]. While our approach adds some noise edges, the overall effect emphasized the prominent edges in facial images that resulted in higher accuracy than previous studies [24], [5], [6], [27].

II. Related Work

Previous studies have developed different methods with increasing progress in facial emotion recognition performance [4], [7]-[9], [19]. Conventional classification has shown its robustness when it is preceded by image preprocessing techniques [11]. Rani et al. [10] used edge detection algorithms for prefiltering the raw images for FER. Other studies used Gaussian edge detectors [13], Colored edge detectors [12] for multi-view face detection, and Canny edge detection [14] for feature extraction. A study by Yu et al. [24] used standard histogram equalization for preprocessing facial images data. In unconstrained environments, [22] obtained a sparse representation of faces for person-specific verification.

CNN have been extensively used for image classification tasks, especially FER, due to their ability to extract image features [4]-[6], [23], [25]-[27]. Though CNNs perform well on their own, performing image preprocessing and feeding the preprocessed image as input to the CNN has shown significant improvement in accuracy as opposed to feeding a raw input image [27]. Earlier implementation of CNN architectures did not employ image data augmentation and preprocessing techniques [5], [6], [26] making them less robust to rotated and deviated facial images. Tang [25] implemented CNN with a linear Support Vector Machine (SVM). Wang et al. [27] employed the same technique, but that they used SVM to stack the result of the "softmax" activation function. They also used data augmentation with histogram equalization, which resulted in better performance. Varying preprocessing techniques and feature extraction parameters can also boost the performance of an ensemble of classifiers [3]. Nanni et al. [18] showed that it is possible to further boost performance by designing an ensemble of classifiers based on different preprocessing

Ksheeraj Sai Vepuri is with the San Jose State University, United States (e-mail: ksheerajsai.vepuri@sjsu.edu).

techniques and feature extraction parameters.

Our goal is to combine the best of both methods, using an image preprocessing technique with data augmentation to enhance each image features and feed it to the CNN model. We found that Image Sharpening technique enhances the prominent features of the input facial images. Our CNN architecture is tantamount to [27] with some additional changes in the convolutions and normalization, and without the need to stack SVM with the CNNs output.

III. Data Preparation

A. Dataset

FER-2013 is the standard dataset used for FER tasks. The images in FER-2013 were generated using the Google Image Search API. The dataset consists of 28, 709 training images, 3589 validation images, and 3589 test images. The "emotion" column from the FER-2013 is the target attribute which consists of seven categories labeled from 0 to 6, representing human emotions such as "Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral". The "pixel" column consists of 48*48 grayscale facial images stored as a 1-dimensional string. The facial images are not all frontal face photos and most of them consist of deviations. Human accuracy to detect facial expressions in the FER-2013 dataset is recorded to be around 65%. Most CNN recorded to have an accuracy of 64%-67% on the FER-2013 dataset.

B. Image Preprocessing

Background, illumination, and posture deviations are all factors which affect the accuracy of experimental results. Applying preprocessing techniques to clean the images and enhance the features that play an important role to detect the facial emotions can result in increased accuracy. The FER-2013 dataset consists of a few noisy unknown/comic images which have not been excluded in our experiments. There are mainly two preprocessing techniques that we employed in our experiments. The first is data augmentation, and the second is image sharpening. For data augmentation, we used ImageDataGenerator from Keras library. The "ImageDataGenerator" generates 32 different images from one image by rotating, flipping, and applying other methods on the original image. Since the FER-2013 is relatively small dataset, this data augmentation technique is beneficial to train the model with additional data. For image sharpening, we used "Unsharp Mask" filter from PIL (Python Imaging library) to improve the contrast and density changes as shown in Fig. 1.

Unsharp Mask uses a blurred or negative image to create a mask of the original image. The final positive "less blurred" image is obtained by combining the original image with the blurred image. The advantage of using Unsharp Mask over other sharpening filters like Gaussian High Pass is the ability to control the sharpening process. Unsharp Mask provides adjustable parameters which can be modified. By observing and understanding the images in FER-2013, we found that the images is slightly blurry and applying sharpening to these images helped to define the edges of prominent features, such as eyes and mouth which are relevant for detecting human emotions as shown in Fig. 2.

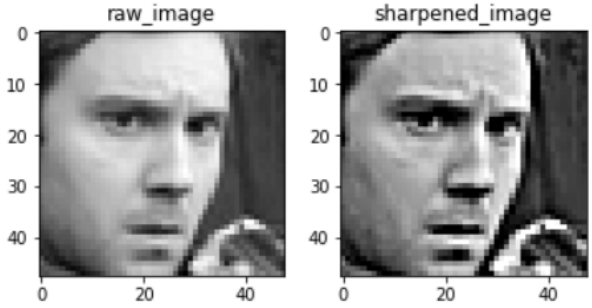

Fig. 1 Applying the Unsharp Mask filter on a raw image (a), The result (b) has high contrast and density

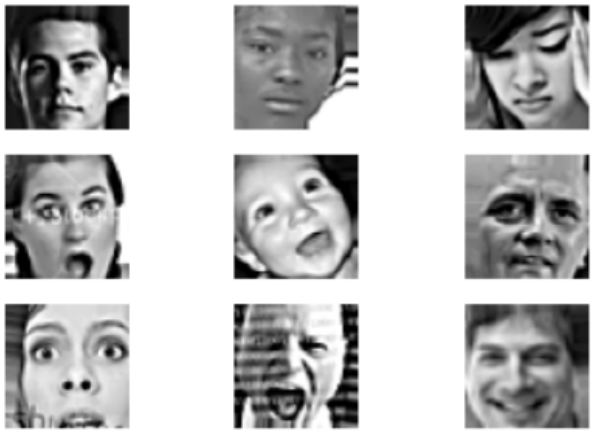

Fig. 2 Sample of images from FER-2013 dataset after applying the Unsharp Mask

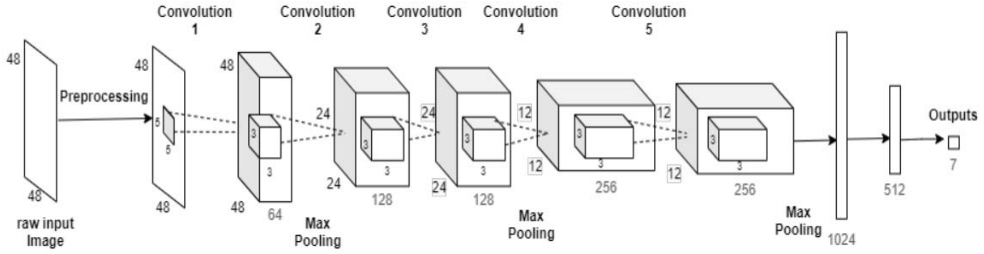

Fig. 3 CNN model architecture: 5 convolutional layers, 3 max pooling layers, and 3 fully connected layers

IV. Proposed Model

CNN is a class of Deep Neural Network which performs well for image related tasks as they tend to capture spatial information. The input to the CNN is a 48*48 grayscale image. These input images are the output of the ImageDataGenerator which produces rescaled, rotated, flipped, and sharpened images. The output of the CNN is a probability of seven categories of facial expressions. It is composed of five convolutional layers, three max pooling layers, and three fully connected layers as shown in Fig. 3. All the convolutional layers use "same" padding and "ReLU" activation function. The first two fully connected layers use "ReLU" as the activation function and the last fully connected layer uses "Softmax" activation function. The first, third, and fifth convolutional layers are followed by batch normalization, max pooling of size 2x2 and stride 2x2, and dropout of 20%. The first convolutional layer is composed of a 5x5 kernel with 64 filters. The second and third convolutional layers are composed of 3x3 kernels with 128 filters. The fourth and fifth convolutional layers are composed of 3x3 kernels with 256 filters. The first fully connected layer consists of 1024 neurons and the second fully connected layer consists of 512 neurons followed by a drop out 20%. The final fully connected layer, which produces the output, consists of 7 neurons representing the 7 categories. The model uses the "categorical_crossentropy" loss function and "adam" optimizer. The total trainable parameters in the model are 11, 075, 847 and non-trainable parameters are 896. Keras callback ReduceLROnPlateau has been used to tweak the learning rate if the validation accuracy does not increase by a certain amount for 10 epochs. The model was trained for 100 epochs using a batch size of 128.

V. Experimental Results

The model trained on the raw dataset without any preprocessing results in a training accuracy of 84.52% and validation accuracy of 59.76%. There is a significant improvement in the validation accuracy when the model is trained on the preprocessed dataset, and in comparison, to previous methods such as [24], [5], [6], [27]. Table I shows the improvement in the model performance with our image preprocessing techniques applied. After the dataset is preprocessed using ImageDataGenerator and Unsharp Mask, the model results in a training accuracy of 90.85% and validation accuracy of 67.32%.

TABLE I
Preprocessing Raw Images Using Our Method

Method	Accuracy (%)	Validation (%)
False	84.52	59.76
True	90.85	67.32

When comparing our classification accuracy on the test set to other algorithms, Liu et al. [5] achieved an accuracy of 65.03% by using CNN ensemble model to identify facial expressions. They use multiple different CNNs and average out the results. Shin et al. [6] achieved an accuracy of 68.79% and Wang et al. [27] achieved an accuracy of 68.79%. With the preprocessing technique we employed, the classification accuracy we have achieved on the test set is 69.46%. While Yu et al. [24] used Multiple Deep Networks, an ensemble of classifiers, to achieve a higher accuracy of 72.1%, the model used in our paper is substantively simple and takes lesser time to train. Table II shows the prediction accuracies of the previous methods and our method on the FER-2013 test set.

TABLE II
Classification Accuracies

Method	Recognition accuracy (%)
Raw-tang model	62.20
Ensemble CNN	65.03
Hist-eq-tang model	66.67
Histogram equalization with (SVM+CNN)	68.79
Our method	69.46
Multiple deep network	72.1

VI. Conclusion

In this study, we were able to demonstrate that the use of sharpening technique to preprocess data for a CNN model boosted performance even though the CNN model is relatively simple. This is improved performance vis a vis previously published methods such as Wang et al. [27] and Shin et al. [6]. Our image preprocessing technique led to emphasizing the prominent edges in facial images resulting in higher accuracy. Our baseline model has resulted in an accuracy of 69.46%. Using pre-trained models such as VGG16, Resnet-50, Inception v3, and SeNet-50 and applying transfer learning with the preprocessing techniques employed in our study on the FER-2013 dataset, the classification accuracy can be improved further. Future work can be extended to detect noise edge in the preprocessing and compose the CNN with an ensemble method. Better detection of human emotions can help children with autism, blind people to read facial expressions, robots to better interact with humans, and ensure driver safety by monitoring attention while driving. FER can also enhance the emotional intelligence of applications and improve customer experience by using emotion recognition.

References

[1] A. Lonare, and S. V. Jain. "A Survey on Facial Expression Analysis for Emotion Recognition". International Journal of Advanced Research in Computer and Communication Engineering 2.12

[2] C. Shan, S. Gong, and P. Meowan, "Facial expression recognition based on local binary patterns: A comprehensive study", Image & Vision Computing, vol. 27, no. 6, pp. 803-816, 2009.

[3] C. Padgett, and G. Cottrell. 1996. "Representing face images for emotion classification". In Proceedings of the 9th International Conference on Neural Information Processing Systems (NIPS'96). MIT Press, Cambridge, MA, USA, 894–900.

[4] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In Proc. CVPR, 2012. 4

[5] 24.K. Liu, M. Zhang, and Z. Pan. 2016. "Facial Expression Recognition with CNN Ensemble". In International Conference on Cyberworlds. 163–166.

[6] 25.M. Shin, M. Kim and D. Kwon, "Baseline CNN structure analysis for facial expression recognition," *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, New York, NY, 2016, pp. 724-729, doi: 10.1109/ROMAN.2016.7745199.

[7] M. Dantone, J. Gall, G. Fanelli, and L. J. V. Gool. “Real-time facial feature detection using conditional regression forests”. In Proc. CVPR, 2012. 1, 2

[8] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. “Metaanalysis of the first facial expression recognition challenge”. IEEE TSMC-B, 42(4):966–979, 2012.

[9] S. Jain, C. Hu, and J. K. Aggarwal. “Facial expression recognition with temporal modeling of shapes”. In ICCV Workshops, pages 1642–1649, 2011.

[10] S. Rani, V. Tejaswi, B. Rohitha, and B. Akhil,”Pre filtering techniques for face recognition based on edge detection algorithm. J. Eng. Technol. 13–218 (2017)

[11] J. Kaur, and A. Sharma, “Review Paper 0n Edge Detecti0n Techniques in Digital lmage Pr0cessing” lnternati0nal J0urnal 0f lnn0vati0ns & Advancement in C0mputer Science ljiacs lssn 2347 – 86l6, V0lume 5, lssue ll, N0vember 2ol6.

[12] J. Prasad, and G. P. Chourasiya, and N.S. Chauhan, "Face detection using color based segmentation and edge detection," International Journal of Computer Applications (0975-8887), voL72, no.16, pp.49-54, June 2013.

[13] M. Abo-Zahhad, R. Gharieb, S. Ahmed, and A. Donko.. Edge Detection with a Preprocessing Approach. Journal of Signal and Information Processing. (2014) 5. 123-134. 10.4236/jsip.2014.54015.

[14] M. Ali, and D. Clausi, "Using the Canny edge detector for feature extraction and enhancement of remote sensing images," IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217), Sydney, NSW, Australia, 2001, pp. 2298-2300 vol.5, doi: 10.1109/IGARSS.2001.977981.

[15] M. Pantic, M. Valstar, R. Rademaker and L. Maat, "Web-based database for facial expression analysis", IEEE International Conference on Multimedia and Expo (ICME), pp. 1-5, 2005.

[16] M. Kamachi, M. Lyons, and J. Gyoba, The japanese female facial expression (jaffe) database, 1998.

[17] T. Kanade, J. F. Cohn and Y. Tian, "Comprehensive database for facial expression analysis", Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46-53, 2000.

[18] L. Nanni, A. Lumini, and S. Brahnam.. “Ensemble of texture descriptors for face recognition obtained by varying feature transforms and preprocessing approaches”. Applied Soft Computing. 61. 10.1016/j.asoc.2017.07.057. (2017)

[19] R. Cui, M. Liu, M. Liu. Facial Expression Recognition Based on Ensemble of Mulitple CNNs[C] Chinese Conference on Biometric Recognition. Springer International Publishing, 2016:511-518.

[20] S. Li, and A. Jain, “Handbook of Face Recognition”, Springer Publishing Company Incorporated, 2011

[21] V. Bettadapura.” Face expression recognition and analysis: the state of the art”. arXiv preprint arXiv:1203.6722(2012)

[22] Y. Liang, S. Liao, L. Wang, and B. Zou, "Exploring regularized feature selection for person specific face verification," *2011 International Conference on Computer Vision*, Barcelona, 2011, pp. 1676-1683, doi: 10.1109/ICCV.2011.6126430.

[23] Z. Xie, Y. Li, X. Wang, W. Cai, J. Rao, and Z. Liu, "Convolutional Neural Networks for Facial Expression Recognition with Few Training Samples," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 9540-9544, doi: 10.23919/ChiCC.2018.8483159.

[24] Z. Yu, and C. Zhang.”Image based Static Facial Expression Recognition with Multiple Deep Network Learning”. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). Association for Computing Machinery, New York, NY, USA, 435–442.

[25] Y. Tang. 2013. Deep Learning using Linear Support Vector Machines. *Computer Science* (2013).

[26] S. Zhou, Y. Liang, J. Wan. 2016. Facial Expression Recognition Based on Multi-scale CNNs. In *Biometric Recognition*. Springer International Publishing, 128–135.

[27] X. Wang, J. Huang, J. Zhu, M. Yang, and F. Yang. “Facial expression recognition with deep learning”. In: Proceedings of the 10th International Conference on Internet Multimedia Computing and Service. New York, NY, USA: ACM, 2018. (ICIMCS ‘18), p: 10:1 10:4.