Emotion Classification of Facial Images

Jia-Hung Wu, Chun-Jung Chien, Shao-En Chen, Jeff Chang and Chu-Ya Yang
Department of Computer Science and Engineering
Texas A&M University
{jiahungwu, amy88328, seanchen47, jeff_chang, graceyang406}@tamu.edu

***Abstract*—Emotion classification is a task that classifies images into different emotional categories. In this project, we work based on an existing model, VGG19, and experiment methodologies from several aspects, including input processing and model optimization. The process can be separated into two main phases. First, we apply image processing techniques to the dataset, with which we suggest might help the existing model better learn the features of images and hence increase the prediction accuracy. Second, we try to optimize the learning process of the model by introducing a regularization term to make the model more general and robust. We conduct a series of experiments to verify our proposed solutions and conclude our findings accordingly.**

I. Introduction

Emotions, which affect both human physiological and psychological status, play an important role in human life. Understanding the unspoken words from facial and body cues is a fundamental human trait, and such aptitude is vital in our daily communications and social interactions. Additionally, the facial expression is one of the most natural, direct and universal signals for humans to convey the emotional states. However, due to the complexity of mutual interaction of physiology and psychology in emotions, recognizing human emotions timely and precisely is still limited to our knowledge and remains the target of relevant scientific research and industry, although a large number of efforts have been made by researchers in different interdisciplinary fields. In the fields of computer vision and machine learning, numerous Facial Emotion Recognition (FER) studies have been conducted to extract emotion information from facial representations. From previous studies, we realized that two main challenges are the overfitting due to the lack of sufficient training data, and the emotion-unrelated bias in the images such as head position and illumination. Therefore, we aim to compare different methodologies and develop a successful combination of preprocessing methods to outperform existing models by providing better experimental results on FER2013 dataset [1], which is one of the largest released FER datasets.

In our work, we utilize a convolutional neural network (CNN) as the baseline model to predict emotion classifications of the facial images. We then apply different concepts and tricks, including image preprocessing and a regularization term to the loss function to improve our baseline model.

The rest of this article is organized as follows. Previous researches of emotional recognition from facial images are reviewed in Chapter II. In Chapter III, the core concepts and steps of the proposed methods are introduced in detail. The experimental results and comparison to other work are shown in Chapter IV. And the conclusion and future work will be summarized in Chapter V.

II. Previous and Related Work

In 1978, Ekman and Friesen defined facial expressions as a rapid signal that varies with contraction of facial features such as eyebrows, lips, eyes, cheeks, etc., thereby affecting the recognition accuracy. They also identified happy, sad, fear, disgust, anger and surprise as the six basic expressions.

Since then numerous studies have been conducted on automatic facial expression analysis because of its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, advanced driver assistant systems (ADASs), health care, virtual reality (VR), games, e-learning and many other human-computer interaction (HCI) systems.

The early stage of FER(facial emotion recognition) researches methods had been organized and compared in Sanjay Kumar and Ayushi Gupta [2]. These methods were such as using Bezier-curve, k-means(Banu, Danciu, Boboc, Moga, Balan), Gabor filter (Wang Zhen, Ying Zilu), and DCT(Deepti, Archana, Dr. Jagathy).

Wei-Long Zheng and BaoLiang Lu [3] had developed an approach of EEG-based affective models without labeled target data using transfer learning techniques. Zixing Zhang, Fabien Ringeval, Fabien Ringeval, Eduardo Coutinho, Erik Marchi and Bjrn Schller [4] used Semi-Supervised Learning (SSL) techniques which delivered a strong performance in the classification of high/low emotional arousal and significantly outperformed traditional SSL methods.

In more recent work, Yu and Zhang utilized an ensemble of CNNs in the 2013 ICML competition on dataset FER2013. They employed data augmentation at both training and test time in order to improve the performance. Instead of performing ensemble voting via uniform averaging, ensemble predictions were integrated via weighted averaging with learned weights. This method ranked second in the recent EmotiW 2015 challenge. The winner of this challenge employed a large committee of CNNs. Certain properties of the individual networks (e.g. input preprocessing and the receptive eld size) varied in order to obtain more diverse models. The ensemble predictions were integrated with a hierarchical fashion, with network weights assigned according to validation set performance. Goodfellow [1] found that human accuracy on FER2013 was 65$\pm$5% on average, while the best recognition accuracy by machines so far was approximately 71% in 2013.

In this project, we also consider some of the most prevailing image preprocessing methodologies. Various image processing techniques can be used to enhance the captured image and then increase the recognition rate. Image normalization, de-noising, filtering, histogram equalization, image resizing, cropping and accurate face detection [5] are certain techniques to enhance image quality and improve the recognition rate.

Regarding CNN models, VGGNet was invented by VGG (Visual Geometry Group) from the University of Oxford. Though VGGNet was the runner-up rather than the winner of the ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2014, it had significant improvement in the classification task over ZFNet (The winner in 2013) and AlexNet (The winner in 2012). Simonyan and Zisserman introduced this state-of-the-art DNN model in their 2014 paper [6]. Their contribution was a thorough evaluation of networks by increasing the depth using an architecture with very small (3x3) convolution filters, which showed that a significant improvement on the prior-art configurations could be achieved by pushing the depth to 16-19 weighted layers.

Nitin Bansal, Xiaohan Chen, Zhangyang Wang [7] developed novel orthogonality regularizations on training deep CNNs, utilizing various advanced analytical tools such as mutual coherence and restricted isometry properties. They achieved consistent performance gains after applying those proposed regularizations on CIFAR-10, CIFAR-100, SVHN, and ImageNet.

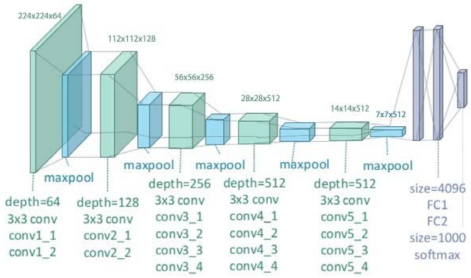

Fig. 1. VGG19 Model Architecture Source: Clifford K. Yang

III. Proposed Method

A. Problem Formulation

Algorithms for automated FER usually involve two main steps: feature extraction and classification. In the feature extraction step, a numerical feature vector is generated from the input image. However, by using neural networks, which is an end-to-end training approach, we do not have to worry about feature selection - as neural networks have the capacity to learn features that statistically allow the network to make correct classifications from the input data. In the second step, the algorithm attempts to classify the given face as portraying one of the seven labeled emotions.

B. Proposed Solution

We begin with the original images from the FER2013 dataset to propose a baseline, using VGG19 in almost all experiments as the origin. Performing different image processing operations, including blurring, sharpening, and illumination adjustments such as histogram equalization and gamma correction, each experiment we switch a set of input images to train the model, with the same structure, to see if any of the operations improves the performance.

Multiplying widths and heights of the original images by four times, we then use the enlarged images with higher resolution as another set of inputs to train the model. Provided with such benchmark, we carry out facial landmark and tilt correction to the upscaled images by locating facial features and generating the images accordingly. Our methodologies aim to see how human recognition and computer vision perform emotion classification similarly and differently.

To further improve the performance of our classification model, we also adopt other machine learning techniques, the most significant being adding an orthogonality regularization to the loss function. Details of our implementation are explained in the following sections.

1) Model: We developed our model based on VGG19 network. To have the model better fit the data propriety of this specific task (i.e., low resolution, grayscale), we removed the last two fully connected layers from the original VGG19 network. Fig. 2 shows the model structure after the modification.

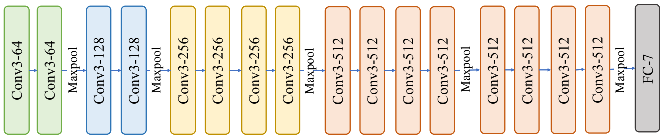

Fig. 2. VGG19 with last two FC layers removed

2) Box Blurring: Image blurring is achieved by convolving the 2D images with a low-pass filter kernel. The kernel removes high-frequency contents (e.g., noise, edges, etc.) and thus blurs the images. We use an averaging blurring technique to perform the blur operation, with an API *cv2.blur()* provided by OpenCV. The images are convolved with a normalized box filter, which takes the average of all the elements under the kernel area and updates the central element. The default kernel in our method is a 3x3 normalized filter constructed as below:

$$k = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

3) Sharpening: Image sharpening is also accomplished by convolving the images with a filter, this time with a high-pass filter kernel. The kernel emphasizes the high spatial frequency components in the images, augmenting the contrast between light and dark areas. The kernel we use is a 3x3 filter constructed as below:

$$k = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

4) Histogram Equalization: Histogram equalization is a technique to adjust image intensities and enhance contrast using the image's histogram. This method allows for areas of lower local contrast to gain a higher contrast. On grayscale images, this method first defines the accumulated histogram as the cumulative distribution function (CDF), in terms of the number of occurrence of pixels ranging from 0 to 255. Then it maps the normalized values in the histogram back into the original range, 0 to 255. Making the intensities better distributed on the histogram, it improves the visual quality of the images. A visualized example is shown in Figure 3.

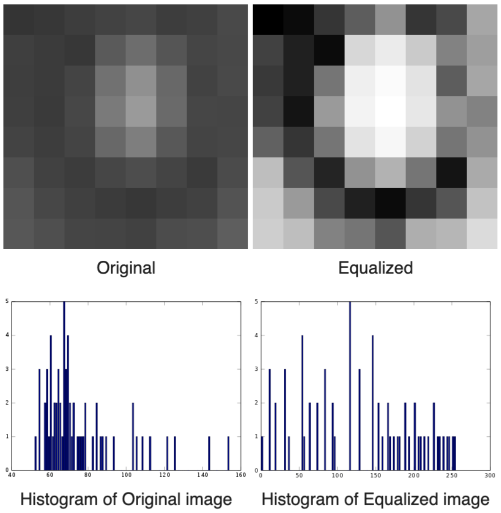

Fig. 3. An example of histogram equalization from Wikipedia.

5) Gamma Correction: Gamma correction is a nonlinear operation used to encode and decode the luminance in images and videos. It is defined by a power-law expression. It takes advantage of how humans perceive light and color, since our perception of brightness follows a nonlinear power function, with greater sensitivity to relative differences between darker tones than between lighter ones. Although we set a default value of gamma to *g = 0.5*, we haven't experimented and tuned different values of gamma. The correction for each pixel p in the images follows the power-law equation:

$$p = 255 \cdot (p/255)^g$$

The two methods proposed below, tilt correction and facial landmark labeling, are built on top of the upscaled images to provide higher resolution and perception from a human perspective. Although the processing methods have already been implemented, we fail to test our model with these corrected images due to the limited time and powerful computing resources, and report this work as one of our future goals.

6) Tilt Correction: Intuitively, we assume that machines can recognize facial features and emotions better when the faces are aligned vertically rather than rotated randomly. Although most of the faces are almost at right angles to the horizontal plane, we manage to set them up by labeling the eyes with OpenCV Haar Cascades face detection methods and rotating to the right angle according to the angle of the line connecting the two eyes detected.

7) Facial Landmark Labeling: We suggest that human emotion might strongly correlate to location information of human facial features and expressions. Based on that, we use OpenCV to extract facial features and plot facial landmarks on the upscaled images. An example of the results is shown in Figure 4.

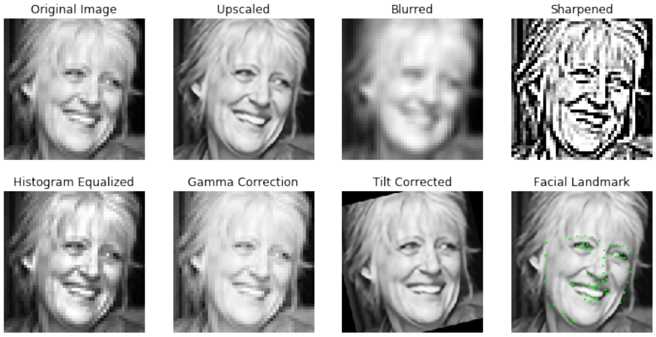

Fig. 4. An image from FER2013 processed in different methodologies: upscaling, blurring, sharpening, histogram equalization, gamma correction, tilt correction, and facial landmark labeling.

8) Image Upscaling: In the original FER2013 dataset, the resolution of the images is pretty low (48×48). We suppose that by upscaling the image resolution, it would be easier for a human to differentiate the images between different categories. Based on that we further suggest that training with the enhanced images would boost the model accuracy. We make use of EDSR [8], which is the model proposed for Single-Image-Super-Resolution task in NTIRE2017 Super-Resolution Challenge, to upscale images for four times in both widths and lengths. A sample result is shown in Figure 4.

9) Model Orthogonality Regularization: In order to optimize the learning process of the model, we introduce the orthogonality regularization term SRIP proposed by [7] into the CNN model to enforce orthogonality to deep network. The regularization term is formulated as below:

$$\lambda \cdot (W^T W - I)$$

where W represents the fully-connected layers $W \in^{m \times n}$ by default. For a convolutional layer $C \in^{S \times H \times C \times M}$ where S, H, C, M represent filter width, filter height, numbers of input channel and output channel, respectively, C can then be reshaped into a matrix $W' \in m' \times n'$ where $m_0 = S \times H \times C$ and $n' = M$.

IV. Experimental Results

We plan to explore and generate different features and demonstrate the outcomes dependency on these features and we hope to achieve state-of-the-art emotion classification accuracy on facial images. Experiments are conducted to compare our approach with the baseline approaches.

A. Data Formulation

FER2013 [1] is a large-scale dataset collected automatically by the Google image search API. It contains 28,709 training images 3,589 validation images and 3,589 test images with seven expression labels including anger, disgust, fear, happiness, sadness, surprise and neutral. Each sample consists of a grayscale 48 * 48 pixels facial image converted into a string of pixels. The dataset is originally prepared by Pierre-Luc Carrier and Aaron Courville, as part of their research project, and introduced during the ICML 2013 Challenges in Representation Learning. Some of the sample images are shown in Figure. 5.

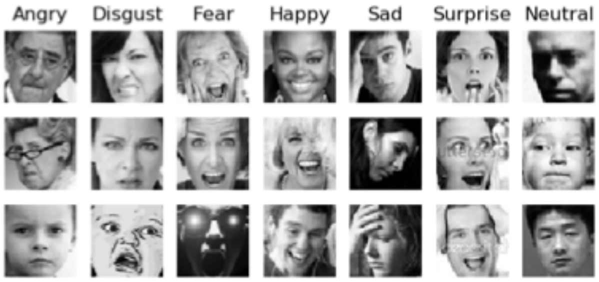

Fig. 5. Sample images from FER2013.

B. Results

From Figure 6 and Figure 7, we can see that experiments of VGG19 with the preprocessed inputs such as the upscaled, blurred, sharpened, histogram-equalized or gamma corrected images all fail to outperform our baseline performance. We think the reason why they fail is that neural network (i.e., VGG19) already has the power to process the images and detect certain patterns in the first few layers. When we try to strengthen some features in the images during preprocessing, some other features in the images are weakened. However, if we put the original images into the neural network, the neural network will do the preprocessing by itself and obtain the most important features. Besides, we can tell that VGG19 + SRIP achieves better performance than VGG19, which implies regularization really works. By using VGG19 and SRIP, we have the best validation accuracy of 71.440% and test accuracy of 72.945%.

By Figure 8, we can see that the correctly classified images usually have strong facial features. For example, there is often a smile on a happy face or an opened mouth on a surprise face. For the misclassified images, the facial features may be ambiguous. Even humans may not classify these images correctly. For instance, a sad face without a mouth or a neutral face with a smile may mislead the neural network.

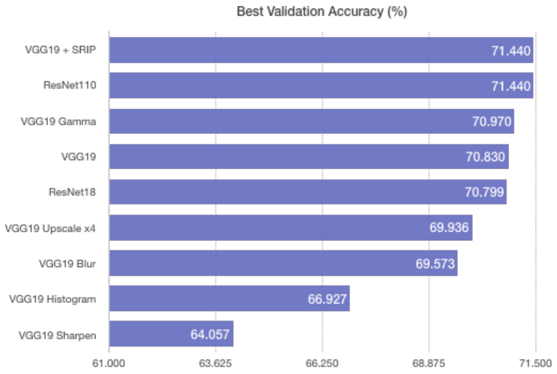

Fig. 6. Best Validation Accuracy

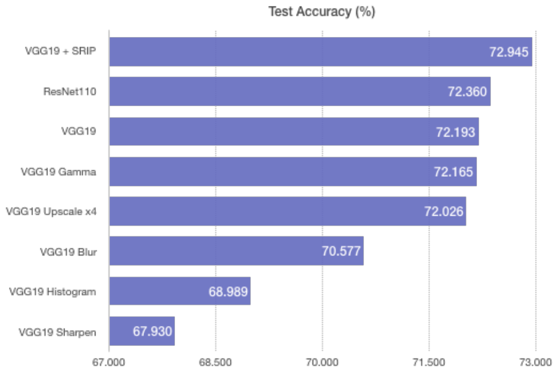

Fig. 7. Test Accuracy

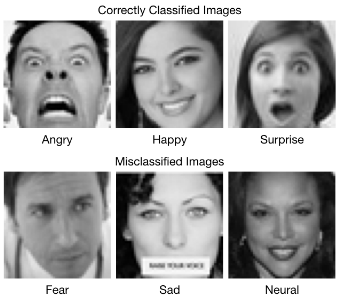

Fig. 8. Correctly Classified and Misclassified Images

V. Conclusion

Our work presents a deep neural network architecture for automated facial expression recognition. The network is based on effective VGG19 plus an orthogonal regularization term in the loss function. The proposed approach is a single component architecture that takes facial images as input and classifies them into one of the seven basic expressions. We evaluated our proposed architecture with different preprocessing methods on the FER2013 dataset. Our results confirmed that the regularization term works well with the expectation of consistent performance gains. However, not all of the preprocessing methods worked well due to the quantity and quality of the dataset. Hence, seeing may not be believing. We have found that better illumination, higher resolution or more features may be beneficial for humans to classify the emotion images, but may not be suitable in terms of DNN classification.

Another problem in facial expression for the FER2013 is

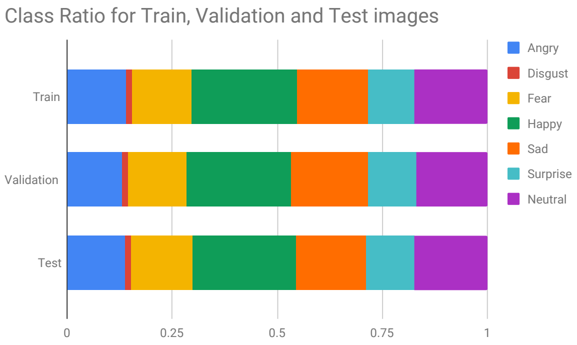

Fig. 9. Class Distribution for FER2013.

the class imbalance, which is a result of the practicalities of data acquisition: eliciting and annotating a smile is easy; however, capturing information for disgust, anger, and other less common expressions can be very challenging. There is a significant difference between the number of samples of different emotions in the dataset (i.e., disgust against the others) as plotted in Figure. 9. Probably one solution is to collect more images to balance the class distribution by using data augmentation and synthesis techniques.

References

[1] I. Goodfellow, D. Erhan, P.-L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," 2013. [Online]. Available: http://arxiv.org/abs/1307.0414

[2] S. Kumar and A. Gupta, "Facial expression recognition: A review," in *National Conference on Cloud Computing and Big Data*, 2015.

[3] W.-L. Zheng and B.-L. Lu, "Personalizing eeg-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, pp. 2732–2738. [Online]. Available: http://dl.acm.org/citation.cfm?id=3060832.3061003

[4] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5185–5189.

[5] F. A. T. Krishna Dharavath and R. H. Laskar, "Improving face recognition rate with image preprocessing," in *Indian Journal of Science and Technology*, 2014.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[7] N. Bansal, X. Chen, and Z. Wang, "Can We Gain More from Orthogonality Regularizations in Training Deep CNNs?" *ArXiv e-prints*, Oct. 2018.

[8] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.