Human Facial Emotion Recognition

(Using Convolution Neural Network)

Harsh Artwani
The University of Texas - Arlington
(UTA)
Arlington, TX, USA
hxa3350@mavs.uta.edu

Gaurav Sanjeev Taneja
The University of Texas - Arlington
(UTA)
Arlington, TX, USA
gst5801@mavs.uta.edu

***Abstract*—Facial expression recognition has been a hot topic in study for decades, and it remains difficult due to considerable intra-class variation. Hand-crafted features like SIFT, HOG, and LBP are used in traditional methods to this problem, followed by a classifier trained on a database of images or videos. The majority of these works do reasonably well on datasets of photographs recorded in a controlled environment, but they fall short on more difficult datasets with more image variance and partial faces. Several researchers have proposed an end-to-end system for facial expression identification using deep learning models in recent years. Despite these works' improved performance, there appears to be more potential for development. In this paper, we propose a deep learning strategy based on an attentional convolutional network that can focus on critical regions of the face. We also employ a visualization technique based on the classifier's output to locate critical face regions for detecting distinct emotions. We show that different emotions appear to be sensitive to distinct areas of the face using experimental results.**

I. INTRODUCTION

Emotions are an unavoidable part of any interpersonal interaction. They can manifest themselves in a variety of ways that may or may not be visible to the human eye. As a result, any indicators preceding or succeeding them can be detected and recognized with the correct instruments. In recent years, there has been an increase in the necessity to identify a person's emotions. Human emotion identification has piqued attention in a variety of sectors, including but not limited to human-computer interfaces, animation [2], medicine, and security.

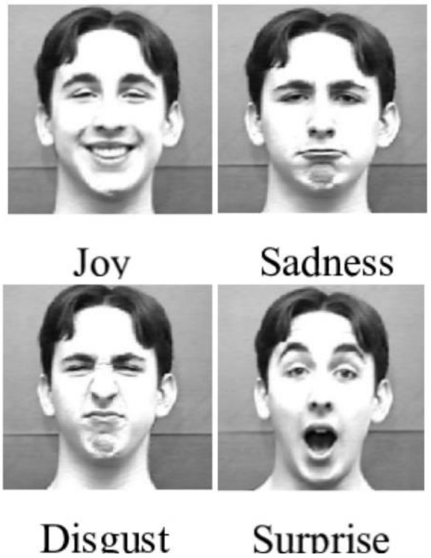


Figure 1: Types of Emotions

Face [2], speech [8], EEG [9], and even text [10] are all examples of features that can be used to recognize emotions. Face expressions are one of, if not the most common, of these features for a variety of reasons: they are visible, they contain many relevant features for emotion detection, and it is easier to collect a big dataset of faces (than other methods for human recognition) [2], [11].

Many features can be retrieved and trained for a competent facial expression detection system using deep learning and notably convolutional neural networks (CNNs) [32]. However, it's worth noting that in the case of facial expressions, a few portions of the face, such as the lips and eyes, provide the majority of the information, whilst other parts, such as the ears and hair, play a little role.

This means that the machine learning framework should ideally focus only on the most significant portions of the face and be less sensitive to the rest of the face.

In this paper, we offer a deep learning-based framework for facial emotion identification that takes into consideration the above observations and employs attention mechanisms to focus on the most important parts of the face. We show that employing an attentional convolutional network, even a network with only a few layers (less than ten layers) may obtain a high accuracy rate. More specifically, the following contributions are presented in this paper:

We propose a technique based on an attentional convolutional network that can focus on feature-rich areas of the face while still outperforming notable recent work in terms of accuracy.

We also apply the visualization technique presented in to emphasize the most important sections of the face image, i.e. the parts of the image that have the most impact on the classifier's output. Figure 1 shows examples of prominent regions for various emotions.

In Section II of the following parts, we first present an overview of relevant works. Section III explains the suggested framework and model architecture. In Section IV, we will present the experimental results, an overview of the databases used in this study, and model visualization. Section V will contain the analysis and Section VI brings the paper to a close.

II. RELATED WORKS

Paul Ekman listed happiness, sadness, anger, surprise, fear, and disgust as the six primary emotions in one of his most famous papers on emotion identification (besides neutral). Ekman went on to develop FACS based on this principle, which has since become the gold standard for emotion recognition research. Neutral was then added to most human recognition datasets, resulting in a total of seven basic emotions. Figure 2 shows image samples of various emotions from three datasets.

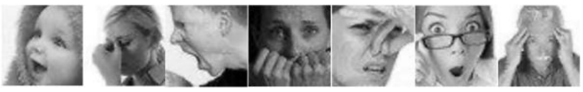


Figure 2: (Left to right) Happiness, sadness, anger, fear, disgust, surprise, and neutral are the six cardinal emotions. The photos are part of the FER dataset.

Earlier work on emotion recognition relied on a standard two-stage machine learning approach, in which some features are extracted from photos in the first step, and then a classifier (such as SVM, or random forest) is used to detect the emotions in the second step. We direct readers to the photos in Figure 2 to gain a clearer feel of some of the possible issues with the images, where the image can have a partial face or the face can be blocked with a hand or eyeglasses.

Following the widespread use of deep learning, particularly convolutional neural networks, for image classification and other vision issues, numerous organizations built deep learning-based face expression recognition models (FER). Aneja et al [2], for example, constructed a deep learning-based model of facial expressions for stylized animation characters by training a network for modeling the expression of human faces, one for that of animated faces, and one to transfer human images into animated ones.

Although the above study improves on previous work on emotion recognition, it appears that a simple piece for attending to the crucial face regions for emotion detection is absent. We attempt to address this challenge in this paper by offering a framework based on an attentional convolutional network that can focus on important facial regions.

III. THE PROPOSED FRAMEWORK

To classify the underlying emotion in face photos, we offer an end-to-end deep learning framework based on attentional convolutional networks. Adding extra layers/neurons, facilitating gradient flow in the network, or better regularizations (e.g. spectral normalization) are frequently used to improve a deep neural network, especially for classification problems with a high number of classes. Due to the small number of classes, we show that a convolutional network with less than 10 layers and attention (trained from scratch) may produce promising results in face expression detection, outperforming state-of-the-art models in various databases.

Given a face image, it is evident that not all portions of the face are relevant in recognizing a given emotion, and in many circumstances, we only need to focus on a few regions to obtain a feel of the underlying mood. Based on this fact, we incorporate an attention mechanism into our framework via a spatial transformer network to focus on relevant face regions.

The proposed model architecture is depicted in Figure 3. The feature extraction section is made up of four convolutional layers, two of which are followed by a max-pooling layer and a ReLU activation function. After that, a dropout layer and two fully-connected layers are added.

To warp the input to the output, several transformations can be utilized; in this case, we used an affine transformation, which is typical in many applications. [7] is a good place to start for more information about the spatial transformer network.

After that, the model is trained by utilizing a stochastic gradient descent approach to minimize a loss function (more specifically Adam optimizer). The classification loss (cross-entropy) plus the regularization term (which is the '2 norm of the weights in the final two fully-connected layers) make up the loss function in this work.

$$\mathcal{L}_{overall} = \mathcal{L}_{classifier} + \lambda \|w_{(fc)}\|_2^2$$

On the validation set, the regularization weight, lambda, is adjusted. We can train our models from scratch by combining dropout and '2 regularization. We also explored a network architecture with over 50 layers, but the accuracy did not increase much. As a result, the following model was chosen in the end.

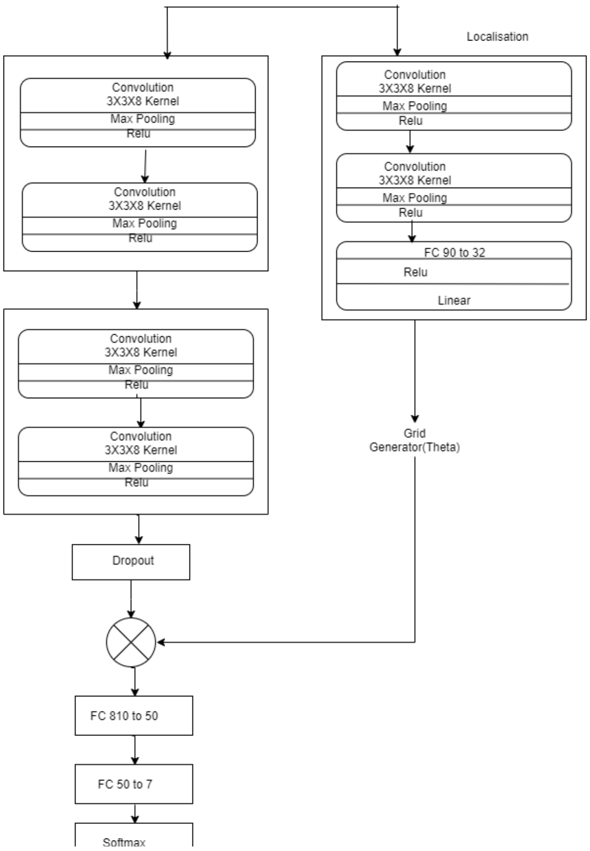


Figure 3: The proposed model architecture

IV. EXPERIMENTAL RESULTS

We present a detailed experimental examination of our model on a face expression recognition database in this part. We begin by providing a quick explanation of the database used in this study, followed by the performance of our model on the database, and finally, the results of our model.

A. Database

In this paper, we present an experimental evaluation of the proposed model using the FER2013 dataset, which is a popular facial expression recognition dataset. Before we go into the results, let's have a look at the database in general.

In the ICML 2013 Challenges in Representation Learning, the Facial Expression Recognition 2013 (FER2013) database was first introduced. This collection contains 35,887 photos with a resolution of 48x48 pixels, the majority of which were captured in natural situations. The training set originally had 28,709 photos, and the validation and test sets each had 3,589 images. Faces are automatically recorded in this database, which was developed using the Google image search API. Any of the six cardinal expressions, as well as neutral, are assigned to faces. FER has greater image variety than the other datasets, including facial occlusion (primarily with the hand), partial faces, low-contrast photos, and eyeglasses. Figure 4 shows four examples of photographs from the FER dataset.

Figure 4: Four images from the FER database

B. *Image Preprocessing*

The first is Grayscale Conversion, which is used to minimize the amount of information in photographs by converting them to grayscale. Each RGB image is made up of three channels that display the red, green, and blue components in RGB space. The sample below illustrates the general ideal of an RGB color image.

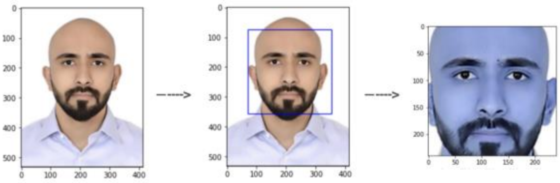


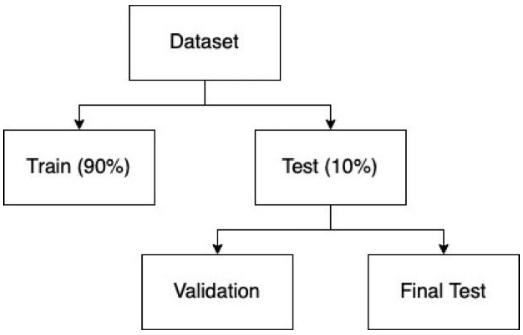
Figure 5: Image Preprocessing

The next step is image resizing. Images are made up of many pixels, which are the smallest units in the image. Furthermore, pictures are a two-dimensional matrix pattern in which each pixel represents a piece of information. In grayscale graphics, for example, '0' represents white and '255' represents black. Because there is so much data to handle, input photos are used in resizing processing to reduce image resolution while maintaining the same amount. The figure below shows how several resolutions can be used to depict the same image.

C. *Experiment Analysis and Comparison*

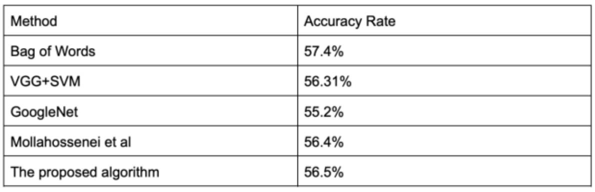
On the above dataset, we will now present the performance of the suggested model. We train the model on a subset of the dataset, validate it on the validation set, then report the accuracy over the test set in this scenario.

We quickly outline our training approach before going into the details of the model's performance on several datasets. In our trials, we only trained one model, but we tried to keep the architecture and hyper-parameters the same throughout all of them. With an Apple M1-8 core GPU, the model is trained for 100 epochs from scratch. We start with random Gaussian variables having a mean of zero and a standard deviation of 0.05. We utilized the Adam optimizer with a learning rate of 0.005 and weight decay for optimization. Our models are trained on the FER dataset in about 1-2 hours. To train the model on a greater number of photos and make the trained model invariant on tiny alterations, data augmentation is utilized for the images in the training sets.



As previously stated, the FER-2013 dataset is more difficult to work with. Aside from the intra-class variance of FER, the unbalanced nature of distinct emotion classes is another major problem in this dataset. Some classes have a lot more examples than others, such as happiness and neutral. We trained the model using the whole 3,230 images in the training set, validated it using 180 validation images, and reported model accuracy using the 360 images in the test set. On the test set, we were able to reach an accuracy rate of roughly 56.42 percent.

Below table shows a comparison of our model's results with some of the earlier works on FER 2013.



Method	Accuracy Rate
Bag of Words	57.4%
VGG+SVM	56.31%
GoogleNet	55.2%
Mollahossenei et al	56.4%
The proposed algorithm	56.5%

V. Analysis

The CNN model developed for image classification is a robust approach. We believe it is stable since it provides a constant accuracy of 55% for the training and validation parts of the dataset with 3500 images. The most prevalent issue is overfitting, which is resolved by removing certain features, bringing the validation and train accuracies closer together. We also employed batch normalization, which is a standard practice for converting numerical data to scalars for use in models. Typically, this CNN model is trained on tangible datasets such as pictures. However, we merged this model with real-time face detection, so it always returns a result for a face.

Theoretically the accuracy of this model is 70 percent, we can try to improve the accuracy. Right now the application only classifies emotion for one face. But in the future, we are planning to update the software which gives results for multiple faces. We are also planning to save data for keeping track and augmenting the dataset.

VI. Conclusion

Using an attentional convolutional network, this study provides a new framework for face expression recognition. We believe that attention is a key component in recognizing facial expressions, and that neural networks with fewer than ten layers can compete with (and even beat) much deeper networks in this regard. We also conducted a thorough experimental investigation of our work using a popular facial expression recognition database, with encouraging findings. In addition, we used a visualization method to highlight the salient portions of face photos, which are the most important parts for recognizing various facial expressions.

ACKNOWLEDGMENT

We'd like to thank Manas Sambre for granting us access to the FERG database via Kaggle. We'd want to express our gratitude to our professor, teaching assistant, and our colleagues at the University of Texas in Arlington for reviewing our work and providing really helpful feedback and recommendations.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] Aneja, Deepali, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. ”Modeling stylized character expressions via deep learning.” In Asian Conference on Computer Vision, pp. 136-153. Springer, Cham, 2016.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. ”Spatial transformer networks.” Advances in neural information processing systems, 2015.

[8] Han, Kun, Dong Yu, and Ivan Tashev. ”Speech emotion recognition using deep neural networks and extreme learning machines.” Fifteenth annual conference of the international speech commu- nication association, 2014.

[9] Petrantonakis, Panagiotis C., and Leontios J. Hadjileontiadis. ”Emotion recognition from EEG using higher order crossings.” IEEE Transactions on Information Technology in Biomedicine 14.2: 186-197, 2010.

[10] Wu, Chung-Hsien, Ze-Jing Chuang, and Yu-Chung Lin. ”Emotion recognition from text using semantic labels and separable mixture models.” ACM transactions on Asian language informa- tion processing (TALIP) 5, no. 2: 165-183, 2006.

[11] Lyons, Michael J. Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. ”The Japanese female facial ex- pression (JAFFE) database.” third international conference on automatic face and gesture recognition, pp. 14-16, 1998.