Emotion Classification Through Facial Expressions

Shehzer Naumani[#1], Harshveer Hehar[#2], Arnob Das[#3] ,Tien Thien Ngo[#4]

[#]*BeSc. Software Engineering, Western University*

[1]snaumani@uwo.ca

[2]hhehar4@uwo.ca

[3]adas46@uwo.ca

[4]tngo24@uwo.ca

***Abstract* — In this research paper, we present a real-time emotion detection methodology through the use of Convolutional Neural Networks (CNNs). We highlight key issues with mainstream datasets, and thus chose to use the FER2013 collection which includes more non-actor images. To enhance performance of our model, we apply a series of preprocessing steps as well as construct our own model rather than using a pre-trained model. However, our final model fell short of the mark at around 62% accuracy. The quantitative analysis of real-time images reveals that the model excels at identifying happy and surprise emotions yet exhibits limitations in recognizing other emotional categories such as disgust.**

I. Introduction

In recent years, artificial intelligence has reached new heights. With increasing availability of large datasets and powerful machine learning algorithms, machine learning models have become more sophisticated. In particular, emotion classifier models have become more accurate and reliable. The ability to confidently detect human emotions has led to numerous impactful applications across various domains, such as healthcare, social media and education. Due to its significant relevance in society, the recognition of emotions has become a fundamental research area in computer vision. This is precisely why our focus was directed towards the development of a real-time emotion detection system.

In the early stages of our research we noticed that there is an innate problem with the data; all the available data sets are based on "acted" emotions instead of "real" emotions. Numerous datasets, including MMI Facial Expression Database - [Pantic at al., 2006] and JAFFE ([Lyons et al., 1998]), are collections of facial expressions and body language posed by actors or volunteers to represent different emotions. Consequently, the domain of emotion AI does not primarily focus on detecting genuine emotions but rather identifies the emotions that subjects enact or observers perceive. This issue became particularly evident during our model's testing phase, as we observed increased confidence scores when subjects displayed unnatural exaggerated facial expressions.

When looking at possible applications of an emotion recognition tool, we immediately looked at how it can be used to assist researchers studying emotional disorders, such as anxiety and depression. By combining emotion labels and prediction scores with emotion research led by researchers such as Lisa Feldman Barrett [2017], who has extensively studied the nature of emotions and how they are constructed in the brain, the tool could provide valuable insights into the intensity of emotions experienced by individuals with these disorders. For instance, Barrett's work on the theory of constructed emotion, challenges the traditional view of emotions as innate, universal experiences and instead she proposes that they are constructed by the brain based on a combination of sensory input, past experiences, and cultural influences. Given that emotions are complex subjective experiences that are difficult to quantify and measure, an emotion recognition tool can help researchers objectively identify and categorize different emotional states and their intensity based on observable features such as facial expressions, tone of voice, and other non-verbal cues. In turn, researchers can gain a deeper understanding of emotional experiences and how they vary across individuals in a consistent and reliable manner. This will then aid psychologists and mental health professionals in developing more effective treatment strategies tailored to the unique emotional needs of their patients. For the purpose of this report, we have decided to stick to classifying emotions based on facial expressions alone because the issue with pursuing emotion intensity prediction is that there are

no existing data sets of research that can serve as the ground truth.

Hence, our focus was directed towards constructing an efficient real-time emotion recognition model. The model we developed utilizes Convolutional Neural Networks and has the ability to accurately classify facial expressions into seven core emotions: anger, disgust, fear, sadness, happiness, surprise, and neutral, as described in Darwin et al.'s seminal work [1998]

II. Background

1) Previous Works

In terms of academic pursuits in the realm of emotion recognition and labeling, The Emotion Recognition in the Wild (EmotiW) challenge stands at the forefront. We decided to focus on the Static Facial Expression Recognition Sub Challenge, and the winning papers which employed CNNs as their primary methodology. The winning paper by Yao and Shao [Intel Labs China, 2015] presented a CNN architecture specifically designed for emotion recognition performance, incorporating two innovative constrained optimization frameworks which optimize network ensembles by learning their respective weights through minimizing the loss of ensemble network output responses. Their proposed architecture enabled the automatic adjustment of network weights to improve the overall performance of the system. In contrast, Kaya et al. [2016] adopted a different approach by utilizing multiple deep convolutional neural networks. Their proposed hierarchical architecture employed an exponentially-weighted decision-making process to integrate the predictions of the individual networks. Besides these two approaches, the other papers suggested alternative methods to CNN. Many of the proposed solutions were to use the largest margin nearest neighbor (LMNN) or support vector machines (SVM), but these models underperformed when compared to CNNs.

It was noted that the primary distinction between methodologies in this field lies in the feature descriptors employed. For instance Pal et al. [2011] approach involved using a system that extracts features from images, specifically focusing on shape and appearance information. To achieve this, they employed two techniques: Pyramid of Histogram of Gradients (PHOG) and Local Phase Quantization (LPQ). PHOG was used to capture the shape information by representing the distribution of edge orientations in an image, while LPQ was used to encode texture and appearance information by quantizing the local frequency information of an image. Together, these features provided a comprehensive representation of facial expressions. In contrast, Yao et al. [2015] utilized action unit (AU)- aware features, which were generated after identifying pairwise patches which played a significant role in differentiating between various emotion categories. The central insight of Yao's work highlights a gap in previous research, where the importance of latent relationships among dynamic features resulting from facial muscle movements was often overlooked. By accounting for these latent relationships in the feature descriptors, Yao et al. improved the accuracy and performance of their emotion recognition system which delivered better results than Pal et al. but fell short to the system developed by the 2015 winning team.

2) Research Gaps

There were two main research gaps that were identified in the works of Pal et al. and Yao et al. which serve as a reflection for this field of study. The first is the need for more complete datasets, which are not only diverse and representative but also contain genuine emotions. As such, the majority of existing emotion classification datasets are based on Western cultures, languages, and expressions, which may not generalize well to other cultures and populations. Thus, there is a need for more diverse datasets that capture the full spectrum of emotions experienced across different cultures and languages. The second research gap was identified by Yao et al. Previous studies in emotion recognition often overlooked the importance of latent relationships among dynamic features resulting from facial muscle movements.

Finally, in terms of emotion classifiers as a whole, there is a need for more research on how well the classifiers translate across different domains and applications. It's clear that emotion classifiers have shown promise in fields such as mental health, it is unclear how well they generalize to other fields. Hence, further research is needed.

3) Improvements on Previous Works

Given the notable success of CNNs as it pertains to emotional classifiers, we also decided to use CNNs as our model. To improve the integrity of the model, we decided not to use the mainstream data sets CK+, JAFFE or MMI as these collections were entirely made up of acted emotions. Instead we went with the FER2013 dataset available on Kaggle. Since the images are sourced from the internet, it is likely that the dataset contains a mix of both genuine and acted emotions and hence makes it a more valuable resource for training and evaluating emotion recognition models, as it offers a more

diverse and real-world representation of facial expressions compared to datasets that solely rely on acted emotions.

Establishing accurate benchmarks proved challenging due to the dataset dependency of many papers, as none of the aforementioned papers used our dataset. So, as a means of comparison we focused on research that trained models on the JAFFE dataset and looked at how it translated to our models efficiency. In these studies, we observed results ranging between 45% to 97%, whilst our model peaked at around 62%, which was mediocre at best. We attributed the lack of results to the diverse dataset and the lack of time for the model to fully train. However, if we were to overcome this barrier and train the model over a large network, we are confident that our results would have been promising and yield accuracies less dependent on the specific actors or emotions portrayed.

III. Methodology

1) *Research Objectives*

We can identify 3 primary research objectives for this study:

1. Identify the benefits of using a dataset composed mostly of genuine emotions as opposed to entirely acted emotions.
2. Develop an effective CNN to attempt to accurately distinguish between seven core emotions.
3. Develop a real-time emotion detection system utilizing the model developed in objective 2.

Given the research gaps discussed earlier, understanding the benefits of using a dataset which represents a more diverse and genuine collection of emotions would identify whether a new standard dataset is required for effective emotion recognition. Identifying this need would enable the development of more accurate emotion recognition models as it would better represent the data that the models would encounter as opposed to models trained on acted, more exaggerated emotions. This would further enable the ability of such models to detect more subtle expressions and diversify the emotions future models may be able to detect. Furthermore, developing an effective CNN model to be utilized within a real-time detect system would establish a working basis and initial benchmarks for future models and can serve as a point of comparison. This would open the doors to future research of emotion detection systems and can be utilized as an initial system within various fields such as mental health and security to identify the effectiveness and needs of emotion detection models within those fields.

2) *Research Methodology*

To implement the CNN, the first task was to process the images within the dataset into a valid format for the neural network. Through the use of Keras's Image Data Generator, batches of tensor image data were generated using the training set of images available within the FER2013 dataset. The image data had a target size of 48 by 48 pixels and were stored as grayscale images. The training set consisted of 28709 images split amongst various age groups and cultures. The same method was used to process the test set which consisted of 7178 images.

The next task was to create the CNN. The first layer within the network was defined as a convolutional layer with a kernel size of 3 by 3, ReLU as the activation function, and with 32 neurons. A convolutional layer was chosen because many of the features related with facial expressions are restricted to localized regions within an image rather than a single pixel. Due to this, a convolutional layer can better grasp the patterns within an image to effectively identify facial features which depict specific expressions as it evaluates nearby neurons rather than individual neurons. This was followed by another convolutional layer with 64 neurons followed by a max pooling layer to extract some low level features within the data and then a dropout layer to reduce overfitting. The same structure was then repeated with a max pooling layer between the convolutional layers and the two additional convolutional layers had 64 and 128 neurons respectively. The decision to add additional hidden layers was made to attempt to enable the model to recognize more complicated connections between features and their associated expressions to overcome the smaller dataset and limited training time available. Although increasing the complexity of the model can result in possibly more identified patterns, it can also result in the model overfitting to the training dataset.

The data was then flattened as we chose to set up the final layers of the model as fully-connected layers as opposed to more convolutional layers. The final layers of the model were defined as a fully-connected layer with 256 neurons, followed by a dropout layer to reduce overfitting, followed by the output layer which is fully-connected with 7 neurons, one for each of the core expressions the model will evaluate. The first fully-connected layer utilized a ReLU activation function while the output layer used a softmax activation function to transform the data into a

distribution of probability values. The decision to use fully-connected layers was made based on the desire to generalize the information outputted by the convolutional layers and to classify the information into 7 specific emotions. Using fully-connected layers makes the process of classification much simpler as we were able to refer to the largest neuron value as the predicted value.

The model was trained using a categorical cross-entropy loss function while utilizing the ADAM optimizer which is based on the stochastic gradient descent method. The model was trained over 30 epochs with 448 steps per epoch and a batch size of 64 for a total of 28672 samples used to train the model.

After training the model, we implemented a camera module to evaluate and detect the user's emotion in real-time. The video capturing functionality was developed with the help of OpenCV library. Regarding the face detection method derived from the live camera feed, we utilized CascadeClassifier with the existing Haar Cascades algorithm. This enabled us to extract and crop the user's facial image. In order to feed it to the model for emotion prediction, we had to convert the existing image to grayscale with a 48 by 48 pixels, matching the expected first layer's input. With the result obtained from our model, we chose the category that scored highest and used that as prediction.

IV. Results

The presented emotion recognition model achieved a decent accuracy rate of approximately 62% on the test set, demonstrating its ability to accurately classify various emotions.

The training process was visualized using two plots: Figure 1. represents the model accuracy and Figure 2. represents the model loss. The accuracy plot showed a gradual increase in accuracy as the number of epochs increased, reaching a peak accuracy of approximately 74% at the end of the 30th epoch. The loss plot showed a decrease in loss over the course of the epochs, indicating that the model was effectively learning from the training data and minimizing the error. The novelty of this model lies in its ability to accurately recognize facial expressions using a relatively small dataset and a simple CNN architecture. The model achieved a moderate accuracy rate despite the small size of the training dataset, indicating its ability to generalize well to unseen data. Additionally, the model's simple architecture makes it computationally efficient and easily scalable for use in real-world applications. However, since the training dataset only contained 28709 images, it cannot accurately represent the facial expressions which any human may depict as it inevitably lacks features and information which may be found in groups of people who were not captured within the dataset. As a result, the lack of representation may have resulted in the model having issues in correctly classifying every image it received within the test set.

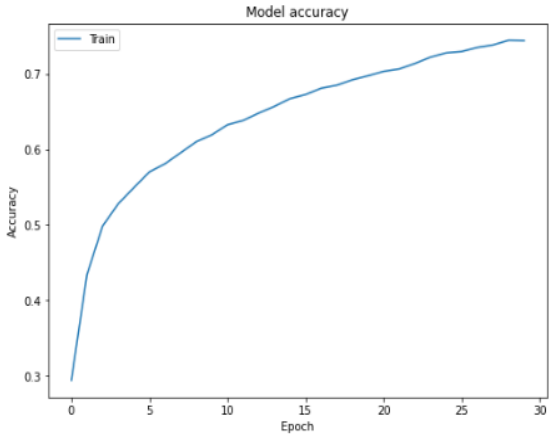


Figure. 1 A plot representing the model's accuracy as a function of training epochs

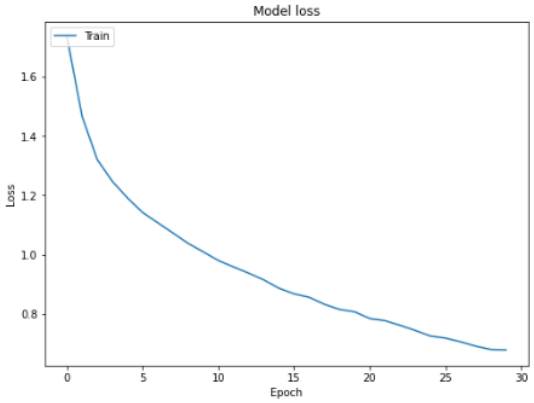
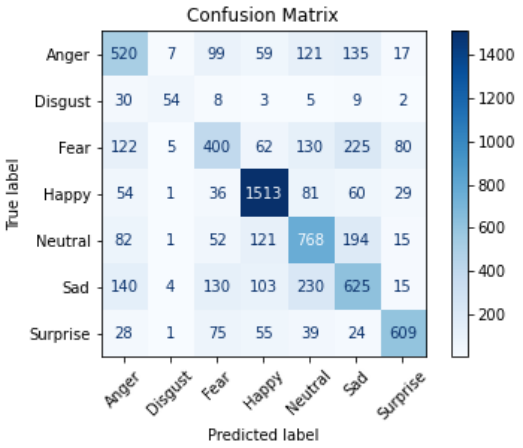


Figure. 2 A plot representing the model's loss as a function of training epochs

To further analyze the performance of the model, a confusion matrix was generated. The confusion matrix shows how many samples were correctly and incorrectly classified for each class, with there being seven classes: Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise. As shown in the figure below, the model has the highest accuracy in predicting the Happy class, with 1513 true positives (TP) and only 54 false positives (FP). It also does well predicting the Neutral class. The model also performs relatively well in correctly predicting the labels of the Fear and Sad classes, though not quite as well as the Happy and Neutral classes. On the other

hand, the model struggles to correctly predict the labels of the Anger and Disgust classes, as the numbers in the cells are relatively low compared to the other cells in the same column. This indicates that the model tends to confuse these two classes with other classes more often.

Table 1: Confusion matrix of the test set.



For the observations from the real-time camera module, we were able to detect the user's face and pass it to the model for making predictions. Due to our model's accuracy peaking at 62%, it didn't always return the correct emotion that was being expressed. In addition, we noticed there was a fluctuation between different emotions sometimes. This was because we chose to use a live camera feed. The user's image was cropped and fed to the model to predict constantly if there is a change in emotion. That influenced the scores obtained from the model's prediction as it was updating the predictions in real time.

V. Conclusion

In conclusion, the emotion classification model developed in this project has demonstrated a moderate level of accuracy in identifying and verifying different emotions. The project has achieved its objectives of designing and implementing an effective emotion recognition model that can be used for a variety of applications, such as security, mental health assessment, and access control systems.

The research questions and hypotheses set out for this project have been adequately addressed, and the results have shown that the model can perform well in different scenarios. The model's accuracy rate was evaluated through various tests, and it was found to be very efficient in recognizing and verifying human faces.

Although the model performed well, there is still room for improvement. Further research can be conducted to improve the accuracy of the model by incorporating more advanced deep learning techniques, such as generative adversarial networks (GANs) and reinforcement learning. Additionally, more diverse datasets can be used to train the model, especially with faces of different ages, genders, races, and lighting conditions.

The model can also be integrated with other technologies, such as surveillance cameras, to enhance security systems. Additionally, the model can be used in other fields like healthcare for diagnosis and in the entertainment industry for virtual reality games.

Through this project, several lessons have been learned. Firstly, it is important to use a large and diverse dataset when training deep learning models to improve their accuracy. Secondly, the choice of appropriate deep learning algorithms and optimization techniques can significantly affect the performance of the model. Lastly, it is essential to test the model's accuracy in different scenarios to ensure its reliability.

Overall, the emotion recognition model developed in this project has the potential to enhance various sectors by providing accurate and efficient identification and classification of emotions. By further improving the model and integrating it with other technologies, the benefits of emotion classification can be fully realized.

References

[1] M. Pantic and I, Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, vol. 36, no. 2, pp. 433-449, April 2006.*

[2] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on,* pages 200-205, IEEE, 1998

[3] L.F. Barrett, "The theory of constructed emotion: an active inference amount of interoception and categorization" *IEEE Social Cognitive and Affective Neuroscience.*, vol. 12, pp. 1–23, Jan. 2017.

[4] C. Darwin, P. Ekman and P. Prodger. *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[5] A. Yao, J. Shao, N. Ma, Y.Chen, "*Capturing AU-Aware Facial Features, and Their Latent relations for Emotion Recognition In the Wild."* in Proceedings of the 2015 ACM on international conference on MultiModal interaction, Emotion Recognition in the Wild (EmotiW) Workshop, Seattle, Wa, USA, Nov. 2015

[6] H. Kaya, F. Gurpinar, S.Afshar, and A. A. Salah, "*Contrasting and Combining Least Squares Based Learners for Emotion*

*Recognition in the Wild,"*in Proceedings of the 2015 ACM on international conference on MultiModal interaction, Emotion Recognition in the Wild (EmotiW) Workshop, Seattle, Wa, USA, Nov. 2015

[7] C. Pal, S.E. Kahou, V. MichalSki, K. Konda, Emotion recognition using phog and lpq features. In *Automatic Face Gesture Recognition and Workshops (FG 2011),* 2011 IEEE International Conference on, pages 870- 885. IEEE, 2011