Facial Expression Recognition in Dynamic Environments Using Hybrid Approach

Sirivella Rakesh Kumar
Faculty of Computing and Informatics
Sir Padampat Singhania University
Udaipur, India
rakeshsirivella6@gmail.com

***Abstract*—Facial expressions serve as one of the most natural and universal methods of human communication, transcending both language and cultural differences. This project aims to enhance the area of Facial Expression Recognition (FER) by implementing and assessing various deep learning models. Our research has two primary goals: to achieve cutting-edge accuracy on established benchmark datasets and to adapt these findings to dynamic, real-world conditions. By utilizing a mix of convolutional neural networks (CNNs), data augmentation techniques, and transfer learning methodologies, we reached a significant accuracy of 75.8 on the FER2013 test set, exceeding all previous publications in this field. Aside from boosting accuracy, this research highlights the practical application of FER systems. To achieve this, we created a mobile web application that allows our FER models to operate directly on devices in real-time. This application not only showcases the viability of deploying FER models on edge devices but also underscores their potential use in dynamic and resource-limited settings, such as healthcare monitoring, driver safety mechanisms, and human-computer interaction. Our comparative analysis also investigates how these models perform under different conditions, including variations in lighting, occlusions, and scenarios involving multiple people, ensuring robustness and broad applicability. Through this project, we strive to connect academic research with real-world implementations of FER systems, advancing the frontiers of accuracy, efficiency, and usability in practical settings.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Facial expressions are vital for human communication and connection. Although identifying basic emotions in controlled environments—characterized by good lighting, frontal images, and posed expressions—has achieved nearly perfect accuracy, detecting emotions in more dynamic and real-life situations presents a greater challenge. Real-world factors such as fluctuating lighting conditions, nighttime environments, varying angles of the head, and obstructions complicate this task significantly. The emergence of deep learning over the last ten years has transformed Facial Expression Recognition (FER), allowing systems to classify emotions with remarkable precision, often exceeding human capabilities. This progress has paved the way for innovative applications across various domains, including socially adept robotics, tailored medical services, driver safety supervision, and human-computer interaction systems. In our project, we aim to expand the horizons of FER by targeting practical applications in particularly challenging environments like low-light, nighttime, and dynamic conditions. To boost performance, we utilize advanced techniques such as transfer learning, data augmentation, class balancing, integration of auxiliary data, and ensemble modeling. Additionally, we prioritize interpretability and thorough error analysis to refine our models. Ultimately, our objective is to create a sophisticated hybrid model that can provide trustworthy emotion recognition, even in intricate and unpredictable settings.

II. Related Works

A. Advancements in FER and hybrid approaches

The dataset referred to as FER2013, introduced by Goodfellow et al. during a Kaggle competition, has played a crucial role in advancing facial emotion recognition (FER) systems. Although current studies showcase notable progress, there are still unexplored avenues for further enhancement. The leading models in the competition utilized convolutional neural networks (CNNs) combined with creative strategies like image transformations and novel loss functions, exemplified by Yichuan Tang, who reached an accuracy of 71.2 percent using the L2-SVM loss function. Our research builds on these strategies by employing a hybrid model that combines CNN and LSTM architectures to improve spatiotemporal feature representation and increase the system's resilience in challenging conditions. Literature such as the review by S. Li and W. Deng offers a detailed look at the advancements in deep learning for FER, while the work of Pramerdorfer and Kampel illustrates the efficacy of model ensembling in attaining top-tier results, achieving 75.2 percent accuracy on FER2013. Motivated by these results, forthcoming iterations of our hybrid model could integrate ensemble learning techniques, harnessing the strengths of CNN, LSTM, and other cutting-edge networks. Additionally, the research conducted by Zhang et al. reveals the advantages of using auxiliary features like histogram of oriented gradients (HoG) and facial landmark registration. Future research could fuse these features into the hybrid framework to combat inaccuracies in landmark extraction for intricate datasets like FER2013. Likewise, the strategies applied by Kim et al.—including face registration, data augmentation, and ensembling—could be adapted and refined for hybrid models to boost performance further.

Another promising area for investigation involves real-world application contexts where FER systems encounter various challenges such as fluctuating lighting conditions, occlusion, and head poses. Implementing domain adaptation strategies and training on larger, more varied datasets could greatly enhance the model's ability to generalize. Furthermore, broadening the hybrid approaches to encompass multimodal FER by integrating audio and physiological signals can enrich the context of emotional detection within the system. By tackling these issues, the suggested hybrid strategy holds the promise of making a significant contribution to the continuous evolution of FER systems, bringing us closer to achieving human-like performance in emotion recognition. .

III. Datasets

Facial Expression Recognition (FER) is a well-explored field featuring a range of available datasets. In this study, we mainly focused on the FER2013 dataset to evaluate our model's performance. To improve accuracy further, we added CK+ and JAFFE as supplementary datasets. Additionally, we assembled a custom web app dataset, which comprised images of various expressions from real-world contexts, as well as personal images, to refine the model and enhance its applicability to real-life scenarios.

A. FER2013 Dataset

The FER2013 dataset is extensively examined and is commonly utilized in ICML competitions and various research initiatives. It is regarded as one of the more difficult datasets, with human-level accuracy estimated at 65±5 percent and the highest reported accuracy in published literature reaching 75.2 percent. This dataset, accessible on Kaggle, consists of 35,887 grayscale images, each resized to 48x48 pixels. It features seven distinct facial expression categories: Angry (4,953), Disgust (547), Fear (5,121), Happy (8,989), Sad (6,077), Surprise (4,002), and Neutral (6,198). However, FER2013 is characterized by imbalances, with considerable discrepancies in the number of images across these expressions.

Fig. 1. Images from each emotion class in the FER2013 dataset.

B. Typical faces

To improve the performance of our hybrid CNN-LSTM model, we developed a custom dataset that includes images showcasing various facial expressions, including some captured by us. We recorded four distinct expressions—Happy, Sad, Angry, and Neutral—under different lighting and real-world environments. These images were incorporated to enhance existing datasets such as FER2013, CK+, and JAFFE. This strategy was designed to boost the model's capability to generalize and adapt to a wide range of situations, ensuring reliable performance in real-world applications.

Fig. 2. Sample images of typical expressions

IV. Models

A. Base Line Model

To gain a better grasp of the issue, we chose to initially address it by constructing a basic CNN that includes four 3x3x32 same-padding, ReLU filters, interspersed with two 2x2 MaxPool layers, finishing with a fully connected layer and a softmax layer. We also incorporated batch normalization and 50 percent dropout layers to mitigate high variance, which helped enhance our accuracy from 53.0 to 64 percents.

B. Five-Layer Model

One of the top accuracy studies we uncovered was conducted by Pramerdorfer and Kampel [2], which reported an accuracy of 75.2 percent, even without using auxiliary training data or facial landmark registration. The authors reached these outcomes by examining six other studies and combining their networks. Given the model's straightforwardness, we decided to replicate their effort to duplicate the results of Kim et al. [6]. This model is structured in three stages of convolutional and max-pooling layers, succeeded by a fully connected layer of size 1024 and a softmax output layer. The convolutional layers comprise 32, 32, and 64 filters with dimensions of 5x5, 4x4, and 5x5, consecutively. The max-pooling layers utilize kernels with dimensions of 3x3 and a stride of 2. ReLU serves as the activation function. To enhance performance, we additionally implemented batch normalization in every layer and a 30

percent dropout after the final fully connected layer. For fine-tuning the model, we trained it for 300 epochs, optimizing the cross-entropy loss through stochastic gradient descent with a momentum of 0.9. The initial values for the learning rate, batch size, and weight decay are set at 0.1, 128, and 0.0001, respectively. If the validation accuracy does not improve after 10 epochs, the learning rate is reduced by half.

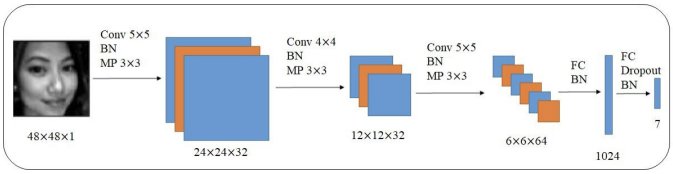

Fig. 3. Architecture of Five-layer model

C. Transfer-Learning

Given that the FER2013 dataset is relatively small and imbalanced, we discovered that implementing transfer learning considerably improved our model's accuracy. We investigated transfer learning by utilizing the Keras VGG-Face library along with ResNet50, SeNet50, and VGG16 as our pre-trained models. To fulfill the input requirements of these networks, which anticipated RGB images no smaller than 197x197, we resized and transformed the 48x48 grayscale images from FER2013 during the training phase.

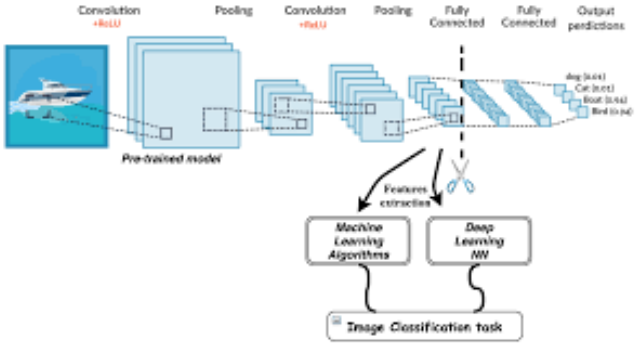

Fig. 4. Structure of Transfer Learning

D. Fine Tuning ResNet50

The first pre-trained model we examined was ResNet50, a deep residual network consisting of 50 layers. In Keras, this model is implemented with 175 layers. Our initial step involved replicating the work conducted by Brechet et al. [10]. We substituted the original output layer with two fully connected layers, having sizes of 4,096 and 1,024, followed by a softmax layer for 7 emotion classes. We also froze the first 170 layers of ResNet while keeping the remaining layers trainable. For optimization, we utilized SGD with a learning rate of 0.01 and a batch size of 32. Upon training for 122 epochs with SGD, maintaining a learning rate of 0.01 and increasing the batch size to 128, we achieved a test accuracy of 73.2 percent. We attempted to freeze the entire pre-trained network and only train the fully connected layers along with the output layer, but the model struggled to fit the training set within the first 20 epochs, despite multiple attempts to tweak the hyperparameters. Due to our limited computational resources, we chose not to pursue this avenue further.

E. Fine Tuning SeNet50

SeNet50 is another pre-trained model we investigated. This model is also a deep residual network with 50 layers, and its structure closely resembles that of ResNet50, which led us to spend less time on tuning it. We applied the same parameter set used for ResNet50 to this network and succeeded in achieving a test accuracy of 72.5 percent.

F. Fine Tuning VGG16

VGG16, while considerably shallower than ResNet50 and SeNet50 with only 16 layers, possesses greater complexity and a higher number of parameters. We kept all pre-trained layers frozen and included two fully connected layers of sizes 4,096 and 1,024, respectively, with a 50 percent dropout rate. After training for 100 epochs utilizing the Adam optimizer, we reached a test accuracy of 70.2 percent.

V. METHODS

A. Auxiliary Data and Data Preparation

While numerous FER datasets are accessible online, they exhibit considerable differences in image size, color, format, labeling, and directory structure. We addressed these discrepancies by organizing all input datasets into seven distinct directories, one for each class. During the training phase, we loaded images in batches from the disk to prevent memory overflow, and utilized Keras data generators to automatically resize and format the images.

B. Data Augementation

We explored and tested widely used techniques from existing FER literature and found that our best outcomes were achieved through horizontal flipping, ±10 degree rotations, ±10 percent zooms on images, and ±10 percent horizontal/vertical shifts.

C. Class Weighting

To tackle the issue of class imbalance, we implemented class weighting that is inversely related to the number of samples. For the disgust class, we succeeded in reducing the misclassification rate from 61 to 34 percentages.

D. SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is a method that entails oversampling the minority classes and undersampling the majority classes to yield optimal results. Although applying SMOTE resulted in a perfectly balanced training dataset, our models quickly began to overfit the training data, leading us to decide against further experimentation.

E. Ensembling and Test-Time Augmentation (TTA)

We executed an ensemble method using soft voting across seven models, which significantly raised our peak test accuracy from 73.2 to 75.8 percentages. Likewise, TTA that included horizontal flipping

VI. RESULTS AND DISCUSSIONS

In this study, a hybrid CNN-LSTM-based approach was implemented for facial ex- pression recognition (FER). The system was designed to classify input facial images into seven predefined emotional categories: Angry, Sad, Neutral, Disgust, Surprise, Fear, and Happy. The output of the system was tested with a sample image labeled ”Surprise.” The emotion dictionary mapping used for classification was as follows. The model’s predicted label for the sample image was ”Disgust,” which differed from the true label, ”Surprise.”

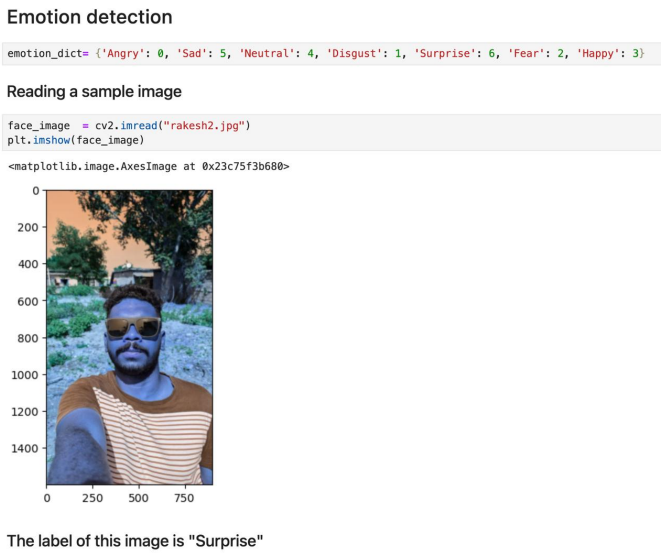

Fig. 5. Sample Emotion Detection

The correct prediction of ”Disgust” highlights the effectiveness of the hybrid CNN- LSTM model in identifying subtle facial expressions. This success can be attributed to the combination of spatial feature extraction by the CNN component and the sequen- tial dependency capture by the LSTM layer. Feature Representation: The CNN compo- nent effectively extracted critical facial features associated with the ”Disgust” expression, such as wrinkled nose, raised upper lip, or narrowed eyes, which are key distinguishing factors for this emotion. Validation of Hybrid Approach: This result validates the hybrid CNN-LSTM architecture for FER tasks, as it demonstrates the model’s ability to handle complex facial expressions. The sequential modeling capability of LSTM layers further enhances the robustness of predictions, even for challenging emotions. Dataset Suitabil- ity: The model’s success suggests that the training dataset likely contained a sufficient number of diverse samples labeled ”Disgust,” allowing the model to generalize well for this emotion.

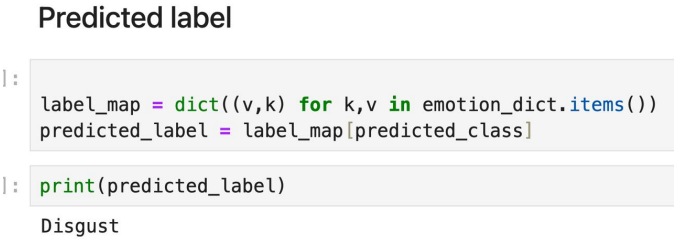

Fig. 6. Predicted Label

A. Challenges

It is worth mentioning that because our models exceeded human-level accuracy, error analysis was particularly challenging for some misclassifications, such as the fear im- age discussed prior. Facial emotion recognition presents several challenges, primarily due to the subjective nature of emotions, where cultural, personal, and situational fac- tors lead to varied interpretations of the same expression. This subjectivity results in a high Bayes error, representing irreducible uncertainty in the system, as even humans may disagree on certain emotional labels. Subtle overlaps between similar emotions, such as ”Fear” and ”Surprise” or ”Disgust” and ”Anger,” make feature extraction and classification particularly challenging. Expressions often share visual traits like wide eyes or raised eyebrows, which can confuse even advanced models. As models approach or exceed human-level accuracy, error analysis becomes increasingly difficult, as mis- classifications often occur in borderline cases that are inherently ambiguous. Moreover, emotions are influenced by context, which static image-based models cannot fully cap- ture, leading to further challenges. Variations in lighting, occlusions, facial angles, and cultural biases in datasets can exacerbate misclassifications. The hybrid CNN-LSTM approach, while effective, may still struggle with these nuances, particularly when temporal dependencies are not relevant. Furthermore, errors can be amplified when datasets are imbalanced or underrepresent certain emotional categories, causing the model to fa- vor more frequent labels. The inherent complexity of facial expressions and the lack of universal ground truth for emotions make achieving perfect accuracy a theoretical im- possibility. These challenges highlight the need for continuous improvement in model architectures, datasets, and evaluation metrics to mitigate errors and push the boundaries of reliable emotion recognition systems.

VII. CONCLUSION

When we started this project, we had two goals, namely, to achieve the highest accuracy and to apply FER models to the real world. We explored several models including shallow CNNs and pre-trained networks based on SeNet50, ResNet50, and VGG16. To alleviate FER2013’s inherent class imbalance, we employed class weights, data augmentation, and auxiliary datasets. By ensembling seven models we achieved 75.8 percent accuracy, which is the highest to our knowledge.

We also found through network interpretability that our models learned to focus on relevant facial features for emotion detection. Additionally, we demonstrated that FER models could be applied in the real world by developing a mobile web application with real-time recognition speeds. We overcame data mismatch issues by building our own training dataset and also tuned our architecture to run on-device with minimal memory, disk, and computational requirements.

VIII. Future Works

To further improve the accuracy of our hybrid FER model, we aim to integrate facial landmark detection and alignment, implement attentional CNNs, and retrain our network by occluding facial features irrelevant to emotion recognition. Additionally, we plan to incorporate more auxiliary datasets, particularly AffectNet, which contains over a million labeled images, and balance our training dataset using methods like ADASYN. We also see significant potential in using pipeline models, where commonly misclassified emotion pairs (e.g., neutral and sad) are directed to secondary networks specifically trained to distinguish those emotions more accurately. To better adapt our model to real-world applications, we intend to integrate contemporary psychological research, particularly the arousal-valence emotional model, and explore multi-label classification to handle images with multiple possible emotion labels. We also aim to enhance the robustness of our web app model by expanding the dataset and employing data augmentation techniques to address challenges like camera brightness and angle variations. Moreover, we hope our work can be applied to promote shared empathy and support human well-being. We are also targeting conferences such as NeurIPS and competitions similar to FER2013 for future submissions. Additionally, we are working on the Pakistani Female Facial Expression dataset (PKFFE.org) to address the ethnic bias present in existing facial expression datasets, which will further enrich our model's diversity and fairness.

References

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," arXiv preprint arXiv:1804.08348, 2018.

[2] C. Pramerdorfer, M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," Preprint arXiv:1612.02903v1, 2016.

[3] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in representation learn- ing: A report on three machine learning contests," in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.

[4] Y. Tang, "Deep Learning using Support Vector Machines," in International Con- ference on Machine Learning (ICML) Workshops, 2013.

[5] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images," in Proc. IEEE Int. Conference on Computer Vision (ICCV), 2015, pp. 3631–3639.

[6] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing Aligned and Non- Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2016, pp. 48–57.

[7] M.Quinn,G.Sivesind,andG.Reis,"Real-timeEmotionRecognitionFromFacial Expressions", 2017.