Facial Emotion Detection using Convolutional Neural Networt

BY, V Harsha Vardhan, G Venkata Pavan Kumar
GUIDE: DR ARUN KUMAR SIVAPURAM
DEPARTMENT OF COMPUTER SCIENE ENGINEERING
SRM UNIVERSITY ,ANDHRA PRADESH

INTRODUCTION

Facial Emotion Recognition aims to automatically identify human emotions from facial images. It is challenging due to variations in pose, lighting, and image quality. Using deep learning, especially CNNs, helps the system learn important facial features automatically. In this project, a Spatial Transformer Network (STN) is combined with a CNN to improve alignment and achieve accurate emotion classification on the FER2013 dataset.

PROBLEM STATEMENT

Human facial expressions play a crucial role in emotional communication, yet automatically recognizing emotions from facial images remains a challenging task. Real-world facial images often suffer from variations in pose, lighting, occlusion, resolution, and expression intensity. Traditional machine learning methods and handcrafted features fail to generalize under such variations. Therefore, the problem is to develop an accurate, robust, and efficient deep learning-based system that can automatically detect and classify human emotions from facial images across seven categories—Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral—using the FER2013 dataset.. The goal is to create a system capable of handling real-world conditions while maintaining high classification accuracy.

DATASET

- *Contains 35,887 grayscale images, each of size 48×48.*
- *Collected from Kaggle*
- *Includes 7 emotion classes:*
  Angry, Happy, Sad, Fear, Disgust, Surprise, Neutral
- *Challenges in the dataset:*
- *Poor lighting*
- *Partial faces*
- *Occlusions (hands, hair, glasses)*
- *Large variation in expressions*

METHODOLOGY

Pre-processing Steps:
To ensure the model learns correctly, several pre-processing techniques are applied:

1.Grayscale Normalization
Pixel values scaled for stable training.

2.Resize Images to 48×48
Maintains uniformity across dataset.

3.Data Normalization
Helps CNN converge faster and avoid exploding gradients.

4.STN-based Alignment
Automatically rotates/transforms images to focus on important regions.

These steps make the input clean and consistent.

The proposed methodology for this project is to build a reliable system that can recognize human emotions from face images. After alignment, the image is passed into a simple Convolutional Neural Network (CNN) that contains several small 3×3 filters. These filters help the model learn important facial features such as the shape of the eyes, eyebrows, and mouth. Max-pooling and dropout help the model reduce noise and prevent overfitting. The final features are flattened and passed into fully connected layers, which classify the image into one of seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, or Neutral.

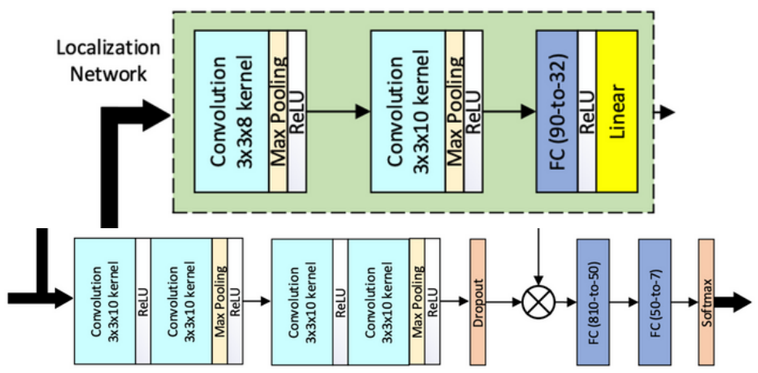

Figure 1: Illustrates the proposed model architecture.

CNN Architecture (Layer-by-Layer)
1.Convolution Layer 1
- **10 filters of size 3×3**
- **Detects basic features (edges, lines, light/dark patterns)**

2. ReLU Activation
- **Introduces non-linearity**
- **Helps model learn complex patterns**

3. Max Pooling
- **Reduces image size**
- **Keeps the strongest features**

4. Convolution Layer 2 (Conv2)
- **10 filters of size 3×3**
- **Learns deeper patterns (eye shape, mouth curves)**

5. ReLU + Max Pooling
- **Same behaviour as above**
- **Extracts robust features**

6. Flattening
- **Converts extracted features into a vector of size 810**

7. Fully Connected Layer (FC1)
- **Reduces features from 810 → 50**
- **Learns emotion-specific combinations of features**

8. Fully Connected Layer (FC2)
- **Final layer with 7 outputs**
- **Each output corresponds to an emotion class**

9. SoftMax
- **Converts outputs to probabilities**
- **Highest probability = predicted emotion**

REFERENCES

[1] Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor. "Emotion recognition in human-computer interaction." IEEE Signal processing magazine 18, no. 1: 32-80, 2001.
[2] Edwards, Jane, Henry J. Jackson, and Philippa E. Pattison. "Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review." Clinical psychology review 22.6: 789-832, 2002.
[3] Chu, Hui-Chuan, William Wei-Jen Tsai, Min-Ju Liao, and YuhMin Chen. "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning." Soft Computing: 1-27, 2017.
[4] Khorrami, Pooya, Thomas Paine, and Thomas Huang. "Do deep neural networks learn facial action units when doing expression recognition?." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015.
[5] Giannopoulos, Panagiotis, Isidoros Perikos, and Ioannis Hatzilygeroudis. "Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013." Advances in Hybridization of Intelligent Methods. Springer, Cham, 1-16, 2018.
[6] Zhang, T., Zheng, W., Cui, Z., Zong, Y., & Li, Y. Spatialtemporal recurrent neural network for emotion recognition. IEEE transactions on cybernetics, (99), 1-9, 2018.

RESULTS

Fig 2: Some of the sample outputs

FER-2013 dataset is more challenging than other facial expression recognition datasets we used. Besides the intra-class variation of FER, another main challenge in this dataset is the imbalance nature of different emotion classes. Some of the classes such as happiness and neutral have a lot more examples than others. We used the entire 28,709 images in the training set to train the model, validated on 3.5k validation images, and report the model accuracy on the 3,589 images in the test set. We were able to achieve an accuracy rate of around 70.02% on the test set. The confusion matrix on the test set of FER dataset is shown in Figure 8. As we can see, the model is making more mistakes for classes with less samples such as disgust and fear. The comparison of the result of our model with some of the previous works on FER 2013 are provided in Table I: Classification Accuracies on FER 2013 dataset

Method	Accuracy Rate
Bag of Words [52]	67.4%
VGG+SVM [53]	66.31%
GoogleNet [54]	65.2%
Mollahosseini et al [19]	66.4%
The proposed algorithm	70.02%

Table I: Classification Accuracies on FER 2013 dataset

CONCLUSION

In this project, we built a robust facial emotion recognition system by integrating a Spatial Transformer Network with a Convolutional Neural Network. The STN improved the quality of the input images by correcting pose and alignment issues, while the CNN extracted meaningful features to classify emotions into seven categories. Our model achieved good accuracy on the FER2013 dataset and performed better than several existing approaches. This shows that combining alignment and feature learning can significantly improve emotion recognition tasks.
For future work, the system can be expanded to handle real-time video emotion detection, multi-face emotion recognition, and higher-resolution datasets. More advanced models like ResNet, EfficientNet, or Vision Transformers can also be tested to further boost accuracy. Additionally, adding audio signals or body gestures could turn the system into a complete multi-modal emotion recognition solution. The model can also be integrated into applications such as smart classrooms, mental health monitoring, or interactive AI systems.