Facial Emotion Recognition Using Convolutional Neural Network

V. Harsha Vardhan
Department of Computer Science and Engineering
SRM University - AP
Andhra Pradesh, India
harshavadlamudi697@gmail.com

G Venkata Pavan Kumar
Department of Computer Science and Engineering
SRM University - AP
Andhra Pradesh, India
venkatapavankumar@srmap.edu.in

***Abstract*—Facial Emotion Recognition (FER) is an important research area in computer vision and human–computer interaction. Traditional approaches rely on handcrafted features and classical machine learning models but fail under variations such as lighting, occlusion, and pose changes. In this paper, we propose a deep learning-based convolutional neural network (CNN) integrated with a Spatial Transformer Network (STN) for automatic alignment and robust feature extraction. The model is trained on the FER2013 dataset containing over 35,000 facial images across seven emotion classes. The proposed framework achieves an accuracy of 70.02% on the test dataset, outperforming several existing baseline methods. The STN module improves recognition performance by normalizing variations in face position and orientation. Experimental results demonstrate the effectiveness of the proposed architecture for real-world facial emotion recognition.**

***Index Terms*—Facial Emotion Recognition, CNN, Deep Learning, FER2013, Spatial Transformer Networks.**

I. Introduction

Facial expressions are one of the most powerful non-verbal communication methods. Automatic detection of emotions from facial images has applications in healthcare, surveillance, virtual assistants, robotics, and human–computer interaction. With advancements in deep learning, Convolutional Neural Networks (CNNs) have become the preferred choice for FER due to their ability to automatically learn hierarchical features.

However, CNNs still face issues when the input faces suffer from rotation, translation, poor lighting, or occlusion. To address this, we integrate a Spatial Transformer Network (STN) to allow the model to automatically align and normalize faces before feature extraction. Our project focuses on building a custom CNN from scratch using PyTorch, trained and evaluated on the FER2013 dataset.

II. Related Work

Earlier studies relied on handcrafted features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), or Gabor filters combined with classifiers like SVM. Deep learning-based works replaced manual feature extraction with CNN architectures such as VGGNet, GoogleNet, and ResNet. Recent studies explore attention mechanisms and spatial transformer modules for improved recognition. Our work contributes a simpler CNN + STN hybrid architecture trained from scratch that achieves competitive accuracy while being computationally efficient.

III. Proposed Methodology

A. Dataset

The FER2013 dataset consists of 35,887 grayscale facial images of size 48×48 pixels. These images are categorized into seven classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset is divided into 28,709 training images, 3,589 validation images, and 3,589 test images.

B. Spatial Transformer Network (STN)

The STN module performs:

- **Localization network** – Predicts affine transformation parameters.
- **Grid generator** – Creates sampling grid.
- **Sampler** – Produces a spatially normalized image.

Mathematically, for target coordinate (x_t, y_t),

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \Theta \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix},$$

where Θ is a 2×3 affine matrix learned during training.

C. CNN Architecture

The proposed CNN contains:

- Four convolutional layers with 3×3 kernels
- Batch Normalization for stable learning
- MaxPooling layers for spatial reduction
- Dropout for regularization
- Two fully connected layers (810→50→7)
- Softmax for final classification

The network learns low-level edges, mid-level shapes, and high-level emotion-specific features such as eye shape, mouth curvature, and eyebrow angle.

IV. Results

The model was trained for 100 epochs with batch size 128 and learning rate 0.005 using the Adam optimizer. Data normalization and augmentation improved generalization.

A. FER2013 Results

The proposed model achieves:

- **Training accuracy:** Increasing steadily across epochs
- **Test accuracy:** 70.02%

B. Comparison with Existing Methods

TABLE I
CLASSIFICATION ACCURACY COMPARISON

Method	Accuracy
Bag of Words	67.4%
VGG + SVM	66.31%
GoogleNet	65.2%
Mollahosseini et al.	66.4%
Proposed Model	**70.02%**

V. CONCLUSION

This project successfully demonstrates a deep learning-based facial emotion recognition system using a custom CNN combined with a Spatial Transformer Network. The STN enhances robustness by aligning faces before convolution, resulting in improved accuracy. The architecture proves efficient, achieving better performance than several previous models. Future work may incorporate attention mechanisms, ensemble models, or transformer-based vision networks to further improve performance.

ACKNOWLEDGMENT

We thank Dr. Arun Kumar, Assistant Professor, Department of Computer Science and Engineering, SRM University–AP, for guidance during the development of this project.

REFERENCES

[1] P. Ekman, “Facial expression and emotion,” American Psychologist.
[2] Pooya Khorrami et al., “Do deep neural networks learn facial action units?”
[3] Mollahosseini et al., “AffectNet: A Database for Facial Expression.”
[4] Barsoum et al., “Training deep networks for ambiguous labels.”
[5] Zhang et al., “Spatial-Temporal Emotion Recognition.”