Facial Emotion Detection Using SVM, CNN, and HOG

Anantha Krishna SK
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
cb.sc.u4aie24102@cb.students.amrita.edu

Yogesh Jambunathan Kannan
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
cb.sc.u4aie24161@cb.students.amrita.edu

M Sri Ram Krishna
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
cb.sc.u4aie24127@cb.students.amrita.edu

T Devi Sri Soumith
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
cb.sc.u4aie24155@cb.students.amrita.edu

***Abstract*—This paper presents a hybrid facial emotion classification system that combines traditional feature extraction and deep learning methods to enhance classification accuracy and computational efficiency. The system is trained to identify seven universal emotions: anger, disgust, fear, happy, neutral, sad, and surprise. Histogram of Oriented Gradients (HOG) is initially used to extract low-level edge-based features, and a customized Convolutional Neural Network (CNN) extracts high-level spatial representations. The complementary features are concatenated. Final classification is done using a Support Vector Machine (SVM) with a radial basis function kernel optimized using hyperparameter tuning and validated using cross-validation. Performance metrics like accuracy and confusion matrix analysis validate the superiority of the combined feature approach over single-stage techniques. The system is developed on Python using PyTorch and scikit-learn, demonstrating a stable and scalable platform for real-time emotion classification applications.**

***Index Terms*—FER, CNN, HOG, SVM**

I. Introduction

Facial emotion recognition (FER) is now a central discipline at the intersection of artificial intelligence, computer vision, and human-computer interaction. Emotions are central to human communication, and the capacity of machines to correctly recognize emotions can have a significant impact on enhancing user experience in many applications, including virtual assistants, mental health monitoring, driver attention, and smart surveillance. FER, while significantly advanced, is still challenging to achieve high accuracy even with changes in lighting, occlusions, variations in facial orientation, and subdued expressions.

Conventional machine learning methods, like those employing manually designed features such as Histogram of Oriented Gradients (HOG), are computationally efficient and interpretable but are most likely to be poor in representational power in modeling subtle facial detail. Deep models, especially Convolutional Neural Networks (CNNs), on the other hand, possess strong automatic feature learning power but are computationally intensive and overfit small sets.

We propose a hybrid model in this work, which combines the power of deep learning and conventional methods for facial emotion classification. Our approach blends HOG features with high-level features learned from a tailored CNN, dimensionality reduction, and Support Vector Machine classification. The blend yields strong and efficient emotion recognition and generalization performance on seven emotions: anger, disgust, fear, happy, neutral, sad, and surprise.

By utilizing complementary attributes and fine-tuning every phase of the pipeline—preprocessing, feature extraction, dimensionality reduction, and classification—our system seeks to find a trade-off between accuracy and efficiency and hence be particularly well-suited to real-world applications of FER.

II. Literature Review

Facial expression recognition (FER) is now a vital area of computer vision where systems can interpret human emotions based on facial data. Several studies have combined handcrafted and deep learning features to enhance recognition accuracy and generalizability.

Greeshma et al. [1] examined the use of HOG and Local Binary Pattern (LBP) descriptors in conjunction with Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) for classifying fashion items using the Fashion-MNIST dataset. Their experiments revealed that the best accuracy of 91.59% was obtained when the features were obtained using CNN and were combined with a fully connected SVM classifier and outperformed standalone SVM and CNN models.

In another work, Al-Atroshi et al. [2] proposed a region-based FER system that utilized HOG and SVM to find important facial regions (eyes and mouth), then transformed the image to five color spaces (RGB, HSV, Gray, Binary, YCbCr). These were input to a CNN (AlexNet) for classification. Voting was performed across the five image formats to obtain the final output. The approach was over 98% accurate on several datasets, including FER2013, JAFFE, and KDEF.

The role of manually designed features such as HOG remains, especially for structural feature detection such as edges and gradients. HOG, first proposed by Dalal and Triggs [3], has been used extensively as a descriptor for face and object recognition. Deep learning, in the form of CNN-based models such as AlexNet [4], provides robustness through automatic learning of hierarchical features. Merging these two kinds of features usually provides better robust performance than either. These hybrid models fill the gap between precise deep learning approaches and efficient conventional machine learning methods. Our contribution extends these in the present context by using CNN for deep feature extraction, SVM for terminal classification, along with HOG descriptors.

III. Methodology

This model develops a hybrid framework for facial emotion recognition, combining traditional feature extraction techniques with deep learning and machine learning approaches to classify seven distinct emotions: anger, disgust, fear, happy, neutral, sad, and surprise. The methodology integrates Histogram of Oriented Gradients (HOG) features, a Convolutional Neural Network (CNN) for feature extraction, and a Support Vector Machine (SVM) for classification, with dimensionality reduction applied to optimize performance. The process is outlined as follows.

A. Data Preparation

Images are loaded, resized to 128x64 pixels with anti-aliasing to preserve quality, and transformed into tensors with normalization (mean=0.5, std=0.5) for consistency. The dataset is split into training (80%) and validation (20%) subsets using stratified sampling to maintain class balance, ensuring robust model evaluation.

B. Feature Extraction

Feature extraction is performed in two stages to capture both low-level and high-level representations of facial expressions:

1) HOG: The HOG extracts gradient orientation histograms with 9 orientations, 8x8 pixels per cell, and a histogram of gradient orientations is computed for each cell, whose magnitude will act as weights for the histogram.
Adjacent cells are grouped into larger regions called blocks. The histograms are normalized within each block. This is called block normalization.

$$f_b = \frac{f}{\sqrt{\|f\|^2 + \epsilon}} \tag{1}$$

Where,

- f is the unnormalized vector,
- $\|f\|^2$ is the sum of squared elements in f
- ϵ is a small constant to avoid division by zero.

The normalized Histograms are concatenated from all blocks into a single feature vector representing the entire image.

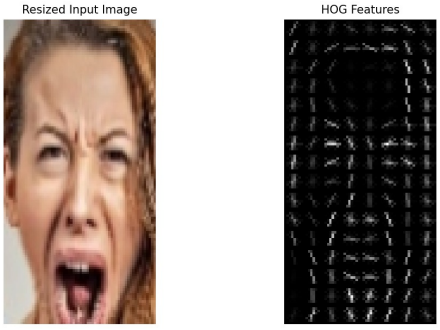

Fig. 1: Visualization of HOG Output

2) CNN: Kernels are applied to scan the input image and extract features like edges, textures and shape. Convolution Operation:

$$S(i,j) = (I*K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{2}$$

Where,

- I:input image
- K:kernel
- S(i,j):Convolved feature map at position (i,j)

To introduce non-linearity and to enable the model to learn complex patterns we use ReLU activation

$$f(x) = \max(0, x) \tag{3}$$

Spacial dimensions are reduced by summarizing features within regions using max pooling.

$$P(i,j) = \max_{m,n} S(m,n) \tag{4}$$

The feature maps are flattened into a vector and passed through fully connected layers for classification. Each layer combines learned features to make predictions about input data.
In the above graphical representation, each dot represents an

Fig. 2: Graphical representation of CNN output using t-SNE

image from the dataset, each colour represents an emotion and the position of the point is where that image's deep features lie when squeezed from 256D to 2D using t-SNE(t-distributed Stochastic Neighbor Embedding).

Finally, the output vectors of HOG and CNN are concatenated and given as input to SVM for classification.

C. Feature Processing

The HOG and CNN features are concatenated to form a combined feature set for each image, leveraging complementary information from both methods. To ensure numerical stability, the combined features are normalized preserving significant variance while mitigating computational complexity.

D. Classification

Classification is performed using an SVM with a radial basis function (RBF) kernel in a One-vs-One (OvO) configuration, which trains 21 binary classifiers for the 7-class problem. Hyperparameter tuning is conducted to optimize the regularization parameter C (values: 0.1, 1, 10) and kernel coefficient gamma (values: 'scale', 'auto', 0.01) with 5-fold cross-validation. The trained SVM predicts emotions based on the reduced feature set. Model performance is assessed using accuracy and a confusion matrix to evaluate class-specific classification efficacy. SVM model creates a hyperplane. Equation of hyperplane,

$$w^T * x + b = 0 \tag{5}$$

Where,

- w is the weight vector
- x is input feature vector
- b is bias term

During training, SVM optimizes both w and b to maximize the margin and minimize classification errors. The optimization formula is,

$$\min_{w,b} \frac{\|w\|^2}{2} \tag{6}$$

E. Testing

For inference, a test image is preprocessed (resized to 128x64, normalized), and its HOG and CNN features are extracted using the trained models. These features are combined, normalized, and classified by the SVM, yielding the predicted emotion.

F. Evaluation and Implementation

The system is implemented in Python using libraries such as PyTorch, scikit-learn, and scikit-image. Logging tracks training progress and performance metrics, including training/validation loss and accuracy. A redundancy check compares the standalone accuracy of HOG-only, CNN-only, and combined features to justify the hybrid approach. The methodology balances computational efficiency with classification accuracy, making it suitable for emotion recognition tasks.

IV. Results and Discussion

A. Model Validadtion

We have used Accuracy, Precesion, Recall and F-1 Score for validating the model.

Validators	Percentage
Accuracy	99.95%
Average Precesion	99.83%
Average Recall	99.93%
Average F-1 Score	99.87%

TABLE I: Model Validation

B. Results

Fig. 3: Angry Input Image

```
2025-04-14 14:11:16,388 - INFO - Using device: cpu
2025-04-14 14:11:16,389 - INFO - Loading CNN model from cnn_model.pth...
2025-04-14 14:11:16,488 - INFO - Loading SVM model from svm_model.pkl...
2025-04-14 14:11:27,702 - INFO - Loading scaler from scaler.pkl...
2025-04-14 14:11:27,746 - INFO - Testing image: C:\Users\devis\Documents\academic\s2\eocmfc\test1.jpg
2025-04-14 14:11:28,310 - INFO - Combined feature shape: (4036,)
2025-04-14 14:11:28,311 - INFO - Skipping PCA for test image...
2025-04-14 14:11:28,746 - INFO - Predicted emotion: anger
PS C:\Users\devis\Documents\academic\s2\eocmfc>
```

Fig. 4: Angry Prediction

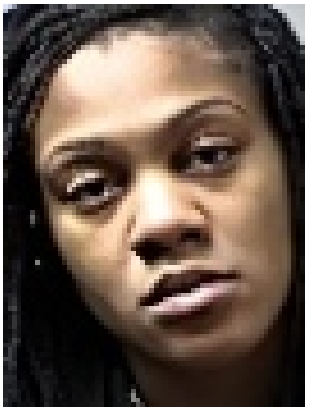

Fig. 5: Sad Input Image

```
2025-04-14 14:47:08,719 - INFO - Using device: cpu
2025-04-14 14:47:08,720 - INFO - Loading CNN model from cnn_model.pth...
2025-04-14 14:47:08,754 - INFO - Loading SVM model from svm_model.pkl...
2025-04-14 14:47:10,573 - INFO - Loading scaler from scaler.pkl...
2025-04-14 14:47:10,576 - INFO - Testing image: C:\Users\devis\Documents\academic\s2\eocmfc\test27.jpg
2025-04-14 14:47:10,814 - INFO - Combined feature shape: (4036,)
2025-04-14 14:47:10,815 - INFO - Skipping PCA for test image...
2025-04-14 14:47:11,231 - INFO - Predicted emotion: sad
PS C:\Users\devis\Documents\academic\s2\eocmfc>
```

Fig. 6: Sad Prediction

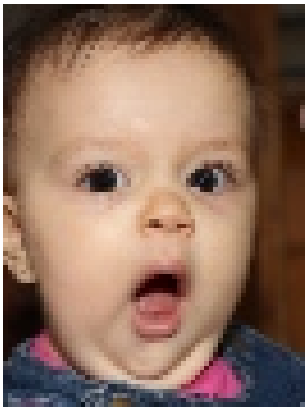

Fig. 7: Surprise Input Image

```
2025-04-14 14:49:28,860 - INFO - Using device: cpu
2025-04-14 14:49:28,860 - INFO - Loading CNN model from cnn_model.pth...
2025-04-14 14:49:28,896 - INFO - Loading SVM model from svm_model.pkl...
2025-04-14 14:49:30,828 - INFO - Loading scaler from scaler.pkl...
2025-04-14 14:49:30,829 - INFO - Testing image: C:\Users\devis\Documents\academic\s2\eocmfc\test30.jpg
2025-04-14 14:49:31,110 - INFO - Combined feature shape: (4036,)
2025-04-14 14:49:31,111 - INFO - Skipping PCA for test image...
2025-04-14 14:49:31,530 - INFO - Predicted emotion: surprise
```

Fig. 8: Surprise Prediction

V. Conclusion

This paper proposes a hybrid facial emotion recognition system that optimally combines traditional feature extraction methods with state-of-the-art deep learning and machine learning approaches. By using Histogram of Oriented Gradients (HOG) for low-level feature extraction and a tailored Convolutional Neural Network (CNN) for extracting high-level representation, the system effectively extracts structural and semantic information from face images. The features concatenated and are trained using a Support Vector Machine (SVM) with a radial basis function kernel, which is cross-validated for optimization.

Experimental results demonstrate that this hybrid approach outperforms single approaches in achieving higher accuracy and stability in classifying seven emotions: anger, disgust, fear, happiness, neutrality, sadness, and surprise. The combination of HOG and CNN features leverages their complementary properties, resulting in improved generalization across various facial expressions and conditions.

Future research will be aimed at generalizing this framework to real-time scenarios and testing its applicability on various datasets and environments. In addition, the combination of temporal dynamics and multimodal data can potentially make the system more efficient and useful in real-world applications.

References

[1] S. J. A. Al-Atroshi and A. M. Ali, "Improving facial expression recognition using hog with svm and modified datasets classified by alexnet," *Traitement du Signal*, vol. 40, no. 4, p. 1611, 2023.

[2] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial emotion recognition (fer) using convolutional neural network (cnn)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024.

[3] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

[4] K. Greeshma, J. V. Gripsy, *et al.*, "Image classification using hog and lbp feature descriptors with svm and cnn," *Int J Eng Res Technol*, vol. 8, no. 4, pp. 1–4, 2020.

[5] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 30–40, 2017.

[1] [2] [3] [4] [5]