Int. J. Nonlinear Anal. Appl. In Press, 1–6
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2025.37902.5478

Analyzing human emotions from facial images using machine learning

Mina Koleini*, Mina Zaminkar

Department of Computer Engineering, Faculty of Engineering and Technology, Shahid Ashrafi Esfahani University, Isfahan, Iran

(Communicated by Asadollah Aghajani)

Abstract

In recent years, human emotion analysis based on facial images has become one of the most important research areas in the fields of machine vision and human-computer interaction. In this paper, a machine learning-based framework is presented to identify the main emotions (happiness, sadness, anger, surprise, fear, disgust, and indifference) from human facial images. Feature extraction algorithms such as Histogram of Oriented Gradients (HOG) and classification algorithms such as SVM, KNN, and deep neural networks were used. Experimental results on the FER2013 database show that the proposed method has high accuracy in emotion classification. In the field of image processing and machine vision, face analysis, especially in the field of emotion recognition, is a technical and research challenge. Several factors, such as variation in individual facial features, lighting changes, camera position, and image resolution, can affect the accuracy of emotion analysis systems. Also, some emotions naturally have a similar appearance and their separation requires algorithms with high learning power and resolution. A novel hybrid framework is proposed for emotion recognition from facial images, integrating traditional handcrafted features (HOG) with deep learning-based features (CNN). The key contributions of this study are as follows: (1) the introduction of a hybrid feature fusion strategy that leverages the complementary strengths of classical and deep methods, (2) a comparative evaluation of classification algorithms (SVM, KNN, DNN) using the FER2013 dataset, and (3) demonstrable improvements in classification accuracy over baseline models, achieving up to 78% accuracy. This comprehensive approach highlights the effectiveness of combining traditional and modern techniques for robust and real-time emotion recognition. This study shows that combining classical and modern methods can significantly increase the accuracy of emotion recognition.

Keywords: emotion analysis, machine learning, face recognition, emotion classification, FER2013, machine vision
2020 MSC: Primary 68T10; Secondary 68T07, 68U10, 62H30

1 Introduction

In today's world, effective human-computer interaction has become one of the most important axes of technological development. One of the crucial aspects of this interaction is the ability of systems to understand and interpret human emotions. Emotions not only play a key role in human communication but can also be used to optimize the performance of intelligent systems in fields such as education, mental health, marketing, and security. Therefore, emotion analysis through the face, which is one of the most natural channels for expressing emotions, has become increasingly important.

*Corresponding author
Email addresses: m.koleini@ashrafi.ac.ir (Mina Koleini), zaminkar@ashrafi.ac.ir (Mina Zaminkar)

Received: April 2025 *Accepted:* June 2025

Traditional emotion analysis methods were mainly based on extracting visual features from images (such as facial lines, wrinkles, or the angle of facial features) and using classifiers such as SVM or KNN. With the advent of deep learning, especially convolutional neural networks (CNN), it has become possible to automatically and more deeply extract emotion-related features. These algorithms are able to identify and analyze complex patterns in images without the need for explicit definition of features.

In this paper, a comprehensive framework for detecting and classifying human emotions based on facial images is presented. This framework includes preprocessing, hybrid feature extraction (traditional and deep), and classification with machine learning algorithms. In order to evaluate the performance of the proposed model, the FER2013 dataset is used, and the results are compared with conventional methods. This study shows that the combination of classical and modern methods can significantly increase the accuracy of emotion recognition.

One of the most widely used methods in facial emotion analysis is the use of appearance features such as lip lines, wrinkles around the eyes, and eyebrow positions. These features are manually extracted and fed into a machine learning classifier to predict the corresponding emotion [7]. Although these methods have been somewhat successful, they have limited performance in the face of high variability in facial expressions, lighting changes, and individual differences.

With the advent of deep neural networks, especially convolutional networks (CNN), sentiment analysis has entered a new phase. These models are capable of extracting high-level and complex features from images without the need for manual feature design [6]. The main advantage of deep learning is its adaptability to diverse data and high generalization power in recognizing complex facial patterns.

CNN-based models have been evaluated in various papers for human emotion recognition. For example, modified models of ResNet and MobileNet have been able to provide excellent performance in real-time environments with lighter parameters [2]. Among the advantages of these models is their compatibility with limited hardware resources, such as mobile phones.

Databases such as FER2013 and AffectNet are common for training and evaluating models. FER2013 contains over 35,000 natural facial images labelled for 7 main emotions, while AffectNet provides a larger dataset with higher racial and optical diversity [7]. Using these databases along with data augmentation techniques can significantly improve the performance of the models.

In addition to CNN, some studies have used recurrent networks (RNN) or CNN-RNN hybrid models to analyze emotions in video. This approach is especially effective when the temporal sequence of mood changes is important [5]. However, in this paper, we focus on still image analysis, which has its own complexities.

Data preprocessing also plays an important role in improving the accuracy of the models. Background removal, illumination normalization, facial feature alignment, and environmental noise reduction can significantly improve the accuracy of neural networks [4]. It is also necessary to use face detectors such as MTCNN in the pre-processing stage.

In the feature extraction process, although deep features are well represented by CNN, some classical methods such as Local Binary Patterns (LBP) and HOG are still used as complementary features in some hybrid models [1]. These features can be helpful in situations where the data has limited diversity.

For the final classification, models such as Fully Connected Networks, LSTM, or even SVM are used in the final layers. Models trained using optimizers such as Adam and Cross-Entropy or focal loss functions perform well in the face of unbalanced data [9]. In this paper, a hybrid model of HOG and CNN features is proposed for emotion analysis. Our model is trained on FER2013 data and is also able to distinguish between similar emotions such as fear and surprise. The final model's accuracy in emotion classification is about 78%, which is a significant increase over the baseline models.

In summary, this study proposes a hybrid framework that integrates handcrafted features—such as Histogram of Oriented Gradients (HOG)—with deep learning-based features derived from Convolutional Neural Networks (CNN). The framework is evaluated using multiple machine learning classifiers, including Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and deep neural networks, on a widely-used emotion recognition dataset. This approach demonstrates enhanced capability in distinguishing between subtle facial expressions by leveraging the complementary strengths of traditional and deep learning techniques. Moreover, the design emphasizes computational efficiency, making it suitable for real-time applications on devices with limited processing power.

According to the results obtained, it can be concluded that the combination of classical and modern methods in emotion analysis from facial images can provide more accurate performance. In the future, multimodal information such as voice and body language can also be used to improve the accuracy of emotion analysis. Also, the use of transformer and self-attention architectures can be a new step in optimizing this area.

2 Research methodology

1. CNN model input: facial image
   Suppose the input face image is a two-dimensional matrix $I \in RH \times WI \in RH \times W$ for a grayscale image or $I \in RH \times W \times 3I \in RH \times W \times 3$ for a color image with $c = 3$.
2. Feature Extraction with CNN
   In each convolution layer, image features are extracted with convolutional filters:
   $$F^{(l)} = W(\sigma^{(l)} * F^{(l-1)} + b^l) \tag{2.1}$$
   where
   $F^{(l)}$: Output of layer l, also called feature maps
   $W^{(l)}$: Convolution kernels in layer l
   $*$: Convolution operation
   b^l: Bias
   σ: Activation function (such as ReLU)
3. Time sequence analysis with RNN or LSTM (if using video or image sequences)
   When analyzing video or image sequences, the output of the CNN at each time instant t is fed into the LSTM:
   $$h_t, c_t = LSTM(f_t, h_{t-1}, c_{t-1}) \tag{2.2}$$
   $$f_t = CNN(x_t) \tag{2.3}$$
   where:
   x_t = input image/frame at time step t, $f_t = CNN(x_t)$ = feature vector extracted from CNN at time t, h_t = hidden state of the LSTM at time t, c_t = cell state of the LSTM at time t, h_t = output (e.g., predicted emotion) at time t.
4. Emotion Prediction
   Finally, the final output (whether from CNN or RNN) is fed into a Fully Connected layer with a Softmax function to give the probability of belonging to each emotion class (e.g. anger, happiness, surprise, sadness, etc.):
   $$y_t = Softmax(Wh_t + b) \tag{2.4}$$

2.1 Performance evaluation criteria

To evaluate the performance of the model on the FER2013 dataset, the Accuracy criterion is used:

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} 1(y_i = \hat{y}_i) \tag{2.5}$$

where:

N: Total number of samples

$\hat{y}_i$:Model prediction for sample i

y_i: True label

1: Indicator function

2.2 Mathematical conclusion

This modeling allows the emotional analysis of images to be described with mathematical and deep learning-based methods and the results to be quantitatively evaluated.

3 Theoretical foundations

Human emotion analysis has always been one of the main goals in the field of human-computer interaction (HCI). Correct understanding of the user's emotional states can lead to improved user experience in intelligent systems. The face, as one of the most obvious sources of emotional expression in humans, plays a key role in this process [6].

From a psychological perspective, emotions refer to psychological and physiological reactions to external or internal stimuli, that in many cases are expressed through facial expressions. Studies have shown that some emotional states such as happiness, sadness, anger, and surprise appear universally and similarly on the face in all cultures [7].

Automatic processing of facial expressions requires methods that can detect subtle changes in facial features, such as lip movement, cheek muscle contraction, or eyebrow raising. In the past, this analysis was performed using traditional algorithms such as geometric or texture feature extraction [4].

Features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) have been widely used to identify facial texture and edge contours. These methods allow the model to detect subtle structural differences in the face [1].

However, these traditional methods have poor performance when faced with lighting changes, different facial angles, and people with high racial diversity. As a result, the use of deep learning as an alternative approach has rapidly expanded in this field [9].

Convolutional neural networks (CNN) are capable of automatically extracting features from raw image data. This advantage has made CNNs a mainstream tool in face and emotion analysis, as they no longer require manual feature design [6].

One of the most successful databases used in this field is the FER2013 database, which contains thousands of facial images with emotion labels. This database has been used as a reference in many prominent studies [7].

Image preprocessing is of great importance in deep learning models. Techniques such as brightness normalization, face alignment, keypoint detection, and face cropping can have a positive impact on the final accuracy [4].

To extract temporal features in video situations, it has been proposed to combine CNNs with recurrent networks such as LSTM or GRU. These hybrid architectures can understand emotional changes over time and analyze dynamic facial behavior [5].

In addition to the architecture, the choice of the cost function (Loss Function) also plays a key role in training the model. In models with unbalanced data, using focal loss instead of cross-entropy yields better results in discriminating between rare emotions [9].

Network optimization is performed with algorithms such as Adam and SGD. The Adam algorithm is widely used in training sentiment analysis models due to its ability to automatically adjust the learning rate [1].

In the field of optimizing model performance, techniques such as dropout, batch normalization, and data augmentation are very effective. These methods prevent overfitting of the model to the data and increase its generalizability [2].

In recent years, the use of attention mechanism has also entered the field of sentiment analysis. This mechanism improves the accuracy of feature extraction by focusing on specific areas of the image [9].

In some studies, combining features extracted from deep networks with traditional features such as HOG and LBP has provided more accurate performance than using each one individually. This shows that combining old knowledge with new technology is still useful [1].

In order to evaluate the models, metrics such as Accuracy, Confusion Matrix, and F1-score are used. In the field of emotion analysis, especially for rare emotions such as "fear" or "disgust," the use of F1-score is more important [7].

In addition to image data, some studies have also used multimodal data fusion including voice, body language, and even written texts to create emotion analysis systems with a deeper understanding of human behavior[5].

With the expansion of applications of this technology in industry, there is a need for systems with high speed and desired accuracy. Models such as MobileNet [3] and EfficientNet [8] have been proposed as lightweight architectures for real-time implementation.

One of the remaining challenges in this field is the issue of data bias. Many databases have racial or gender imbalances, which may reduce the accuracy of models for certain groups [6].

Addressing racial and gender bias is a critical consideration in the development of emotion recognition systems. While the proposed framework utilizes the FER2013 dataset due to its accessibility and widespread use, it is acknowledged that this dataset lacks demographic diversity, potentially leading to biased performance across different ethnicities or gender groups. Since the current model does not explicitly incorporate bias mitigation techniques, there remains a risk of unequal accuracy rates across subpopulations.

To address this, future research can integrate fairness-aware training methods, such as re-weighting underrepresented groups or applying adversarial debiasing strategies. Additionally, evaluating model performance on demographically disaggregated subsets—if available—can help identify disparity gaps. The development of culturally and demographically diverse datasets, including localized databases, is also essential for reducing algorithmic bias and enhancing the inclusiveness of emotion recognition systems.

Finally, incorporating ethical auditing tools and engaging with interdisciplinary perspectives—particularly from the social sciences and psychology—can support the responsible deployment of such systems, especially in sensitive domains such as education, healthcare, or mental health monitoring.

Facial image emotion analysis has made significant progress in recent years, but further research and development is still needed in the areas of data diversity, model generalization, and multifaceted feature integration.

4 Comparative Evaluation with state-of-the-art methods

To evaluate the effectiveness of the proposed hybrid approach, a comparative analysis was conducted with several well-known models that have been tested on the FER2013 dataset. Table 1 summarizes the reported performance of selected state-of-the-art methods from the literature alongside the results obtained in this study.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4.1}$$

Table 1: Comparison of emotion recognition accuracy on FER2013 dataset

Model	**Accuracy (%)**	**Source/Notes**
VGGNet	71.2	[7]
ResNet-18	73.6	[6]
MobileNet-V2	74.5	[2]
Vision Transformer (ViT)	79.0	[9]
Proposed Hybrid (HOG + CNN)	78.0	This study – hybrid feature fusion with SVM

The results indicate that the proposed hybrid framework achieves competitive accuracy compared to more complex deep learning models such as ResNet and ViT. Notably, while ViT shows slightly higher accuracy, it demands significantly more computational resources due to its transformer-based architecture. In contrast, the proposed approach offers a favorable trade-off by combining interpretable handcrafted features (HOG) with deep CNN-based representations, enabling robust performance with lower computational overhead. This makes it suitable for real-time or resource-constrained environments without sacrificing classification quality.

5 Conclusion

This study examined the theoretical foundations and new methods for analyzing human emotions through facial images. Given the increasing importance of human-machine interaction in fields such as education, mental health, security, and marketing, the ability of systems to correctly understand and interpret emotions plays an important role in optimizing the user experience. As one of the main channels for transmitting emotions, the face is a reliable and relatively universal source for emotional analysis.

Advances in the field of deep learning, especially the use of convolutional neural networks, have significantly improved the accuracy and efficiency of automatic facial expression analysis. Lightweight and optimized models have also enabled this analysis to be performed in real time. Also, the use of a combination of classical and modern methods, the use of diverse databases, and attention to the temporal or multimodal dimensions of emotions have paved the way for more accurate and intelligent development.

Despite significant progress, challenges such as data balance, model generalizability to diverse faces, and accurate understanding of complex or mixed emotions still remain. Therefore, continuing research to improve accuracy, reduce model bias, and utilize multichannel information will be an important step in realizing intelligent interaction systems.

In summary, emotion analysis from facial images is not only a technical issue but also an interdisciplinary challenge that, by combining computer science, psychology, and artificial intelligence, can pave the way for fundamental changes in human-technology interactions.

5.1 Suggestions

While this study presents a promising framework for emotion recognition using hybrid features, several directions can be pursued to enhance its performance and applicability. Among the many potential areas of development, two stand out as both highly feasible and impactful:

1. Multimodal Emotion Recognition
   Facial expressions alone may not fully capture the nuances of human emotions. Integrating additional modalities such as voice, text, or physiological signals (e.g., heart rate or EEG) can lead to more robust and context-aware emotion analysis. Multimodal fusion strategies—such as late fusion or attention-based cross-modal models—can improve recognition accuracy, especially for subtle or mixed emotions. This direction is especially promising for real-world applications like mental health monitoring or affective computing in education.
2. Incorporation of Transformer-Based Architectures
   Recent advances in vision transformers (ViT) have demonstrated strong performance in visual recognition tasks by modelling long-range dependencies. Applying transformer-based models—or hybrid CNN-transformer architectures—could further enhance feature representation in facial emotion recognition. Additionally, self-attention mechanisms may allow the model to focus on emotion-relevant facial regions (e.g., eyes or mouth), improving interpretability and precision. Future studies could explore lightweight or distilled transformer variants suitable for real-time deployment.

Other promising directions include:

- Expanding datasets with better demographic balance to reduce racial and gender bias.
- Developing localized datasets tailored to cultural facial expression patterns.
- Applying reinforcement learning for real-time adaptive emotion modelling.
- Deploying models in actual software systems (e.g., mental health apps) to validate practical effectiveness.
- Prioritizing multimodal fusion and transformer architectures offers both high technical value and alignment with current research trends in deep learning and human-computer interaction.

References

[1] J. Chen, Y. Song, and J. Liu, *Facial expression recognition with visual context attention and local-global feature fusion*, Pattern Recogn. Lett. **138** (2020), 276–282.

[2] B. Hasani and M.H. Mahoor, *Facial expression recognition using enhanced deep 3D convolutional neural networks*, Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshops, 2017, pp. 30–40.

[3] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, arXiv preprint arXiv:1704.04861, **2017** (2017).

[4] H. Jiang, H. Zhang, and J. Wu, *Robust facial expression recognition using deep fusion of local and global features*, Neurocomputing **330** (2019), 186–194.

[5] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, *Analysing affective behavior in the first ABAW competition*, IEEE Int. Conf. Comput. Vision Workshops, 2019, pp. 1–10.

[6] S. Li and W. Deng, *Deep facial expression recognition: A survey*, IEEE Trans. Affect. Comput. **13** (2020), no. 3, 1195–1215.

[7] A. Mollahosseini, B. Hasani and M.H. Mahoor, *AffectNet: A database for facial expression, valence, and arousal computing in the wild*, IEEE Trans. Affect. Comput. **10** (2017), no. 1, 18–31.

[8] M. Tan and Q. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, Int. Conf. Machine Learn., PMLR, 2019, pp. 6105–6114.

[9] K. Zhang, Z. Zhang, and Z. Li, *Facial expression recognition via deep learning with attention mechanism*, Neurocomputing **333** (2019), 505–518.