Electronics and Electrical Engineering

AIMS Electronics and Electrical Engineering 8(2): 217–236.
DOI: 10.3934/electreng.2024010
Received: 13 February 2024
Revised: 04 April 2024
Accepted: 22 April 2024
Published: 07 May 2024

https://www.aimspress.com/journal/ElectrEng

*Research article*

# Robust CNN for facial emotion recognition and real-time GUI

**Imad Ali[1,] * and Faisal Ghaffar[2]**

[1] Department of Computer Science, University of Swat, Swat, KP, Pakistan
[2] System Design Engineering Department, University of Waterloo, Waterloo, Canada

*** Correspondence:** Email: imadali@uswat.edu.pk; Tel: +92996850470.

**Abstract:** Computer vision is witnessing a surge of interest in machines accurately recognizing and interpreting human emotions through facial expression analysis. However, variations in image properties such as brightness, contrast, and resolution make it harder for models to predict the underlying emotion accurately. Utilizing a robust architecture of a convolutional neural network (CNN), we designed an efficacious framework for facial emotion recognition that predicts emotions and assigns corresponding probabilities to each fundamental human emotion. Each image is processed with various pre-processing steps before inputting it to the CNN to enhance the visibility and clarity of facial features, enabling the CNN to learn more effectively from the data. As CNNs entail a large amount of data for training, we used a data augmentation technique that helps to enhance the model's generalization capabilities, enabling it to effectively handle previously unseen data. To train the model, we joined the datasets, namely JAFFE and KDEF. We allocated 90% of the data for training, reserving the remaining 10% for testing purposes. The results of the CCN framework demonstrated a peak accuracy of 78.1%, which was achieved with the joint dataset. This accuracy indicated the model's capability to recognize facial emotions with a promising level of performance. Additionally, we developed an application with a graphical user interface for real-time facial emotion classification. This application allows users to classify emotions from still images and live video feeds, making it practical and user-friendly. The real-time application further demonstrates the system's practicality and potential for various real-world applications involving facial emotion analysis.

**Keywords:** facial emotions; prediction; recognition; preprocessing; CNN; GUI

## 1. Introduction

Human facial expressions play a pivotal role in interpersonal communication, serving as a critical indicator of emotions, engagement, and involvement. With the rapid progress in artificial intelligence and machine learning, there is a growing interest in developing models capable of accurately recognizing and interpreting human emotions. Such advancements hold great potential for creating more natural and seamless interactions between humans and machines [1,2]. In this context, facial emotion detection systems have emerged as a promising area of research, aiming to endow machines with the ability to perceive and comprehend human emotions from facial cues. Facial emotion detection systems are comprised of two fundamental components: face detection and emotion recognition. The face detection module identifies and localizes human faces in images or videos, while the subsequent emotion recognition module analyzes the detected faces to determine the underlying emotions [3]. The integration of these components paves the way for machines to perceive and respond to human emotions, leading to a multitude of real-world applications. Recent advancements in computer vision have sparked extensive research in the area of face detection, utilizing large datasets and intricate models. Vaillant et al. [4] made groundbreaking strides in face detection by pioneering the use of neural networks, training a convolutional neural network, and employing a sliding window technique to locate faces in images. Likewise, in [5], a connected neural network approach was developed for face detection in images, making a notable contribution to the field. The progress in face detection has been further enhanced by the availability of publicly accessible benchmarks such as the Wilder Face-Face Detection Benchmark [6], PASCAL FACE [7], and Face Detection Database and Benchmark [8]. These benchmarks have played a crucial role in facilitating the development and evaluation of various face detection algorithms, further propelling advancements in this domain.

In the current landscape, face detection algorithms fall into four major categories, each with its unique approach and characteristics: Cascade-Based Algorithms: Notable works like Viola and Jones [9] and Lienhart and Maydt [10] exemplify these algorithms, which operate through a cascading series of classifiers. Each stage refines the face detection process further, achieving enhanced accuracy and efficiency by combining multiple classifiers. Preprocessing includes variance normalization and integral image computation for efficient feature evaluation, while performance is assessed using detection rate, false positive rate, and ROC curves for accuracy evaluation. Part-Based Algorithms: Belhumeur et al. [11] and Yang et al. [12] are examples of part-based algorithms that break down the face detection task into individual facial components, such as eyes, nose, and mouth. By analyzing these parts separately and combining their detections, these algorithms effectively locate the entire face region. Preprocessing involves building a 3D basis from images captured under diverse lighting. Performance is evaluated by comparing recognition error rates across various lighting and facial expression conditions, using datasets from Harvard Robotics Lab and Yale Center for Computational Vision. Channel Feature-Based Algorithms: Wu et al. [13] and Yan and Kassim [14] represent channel feature-based algorithms that analyze specific color channels or feature maps to identify facial patterns and distinguish them from the background. Leveraging these distinctive features aids in accurate face detection. The preprocessing steps involve synthetic blurring of images with varied blur sizes and noise levels, as well as real photo acquisition with registration. Performance evaluation includes metrics such as Normalized Root Mean Square Error, convergence analysis, comparison with other methods, and visual inspection of reconstructed images and estimated Point Spread Functions. Neural Network-Based Algorithms: The use of neural networks for face detection, demonstrated in

Krizhevsky et al. [15] and Sermanet et al. [16], has gained great popularity due to their ability to learn complex representations from data. Preprocessing involves data augmentation, including image translations, reflections, and intensity alterations, alongside normalization and max-pooling. Performance is evaluated using error rates, compared with previous approaches on the ImageNet dataset. The diverse approaches in these four categories underscore the continuous advancements in face detection research, aiming to provide robust and efficient solutions for a wide range of applications. Each category has its strengths and limitations, and researchers continue to explore innovative approaches to improve the accuracy and efficiency of face detection systems.

While the focus has mainly been on face investigation, considerable efforts have also been directed toward facial expression recognition. Researchers have devised systems based on Facial Action Coding [17–19], and some have explored the use of Hidden Markov Model neural network-based models for facial emotion detection [20]. Facial emotion prediction, often considered the second stage following face detection, plays a pivotal role in understanding human emotions from visual cues. Recent advancements in this domain, as evidenced by papers such as Emotion Recognition Using a Transformer-based Architecture [21] and Learning Dynamic Affective Contexts for Facial Emotion Recognition [22], highlight the significance of this area. Additionally, Efficient Facial Emotion Recognition using Siamese Neural Networks [23] and Facial Expression Recognition using Spatiotemporal Attention Mechanism [24] underscore the ongoing research in improving the accuracy and efficiency of facial emotion prediction. However, a persistent research gap remains. Many approaches in these articles often incorporate unnecessary and misleading features, leading to confusion during training and reduced accuracy. These complexities pose practical challenges, especially in real-world applications, as they can result in substantial delays in transitioning face detection algorithms. Therefore, developing frameworks that effectively filter out irrelevant background information and focus solely on crucial facial features is crucial to achieving accurate face detection and practical deployment in diverse domains. Despite great progress in the field, driven by the availability of massive datasets, sophisticated models, and continuous benchmarking, there are hurdles to overcome. For example, these approaches often incorporate unnecessary and misleading features, leading to confusion during the training and reduced accuracy. Developing frameworks that effectively filter out irrelevant background information and focus solely on facial features, is crucial to achieve accurate face detection. Furthermore, a practical concern arises from the complexity of these models, resulting in substantial delays when transitioning face detection algorithms into real-world applications. Thus, it is vital to reduce the model complexity for practical deployment in diverse domains.

To address the aforementioned challenges in face detection, we abstract our considered problem as designing a robust deep learning framework for predicting facial emotions, with a critical focus on accurately classifying seven fundamental facial expression classes: Happy, surprise, disgust, neutral, fear, sad, and anger. Our objective is to filter out unnecessary and misleading features to reduce the model's complexity and enable it for real-world application. We utilize a specialized CNN designed to prevent irrelevant features during training. The process involves face detection, estimating the facial area, and applying diverse pre-processing steps to enhance feature recognition. This optimizes CNN's understanding of facial expressions for accurate emotion prediction. The system assigns probability scores to emotion categories, classifying the image based on the highest score. The accuracy our proposed CNN model achieves through the rigorous training process on the challenging dataset is, 78.1%, surpassing the performance of baseline models. This remarkable level of accuracy underscores

the effectiveness of the proposed system in accurately predicting facial emotions, thus validating its potential as a robust and reliable tool for emotion recognition tasks. Building upon the success of the model, we develop an application capable of detecting facial emotions in real-time video and still images. This application employs the trained model to provide accurate and reliable emotion recognition for various practical purposes. The article introduces several significant contributions related to face emotion prediction, which are summarized as follows:

1. We propose an efficient system that leverages a CNN to predict facial emotions, which assigns probabilities to each emotion class so that the system can accurately identify each emotion in facial expressions.

2. Understanding the importance of data in deep learning models, the proposed system adopts diverse pre-processing steps for each image. These steps aim to enhance prediction accuracy by allowing the neural network to recognize relevant features effectively.

3. To put the proposed system into practical use, we design a graphical user interface (GUI) for real-time emotion classification. This application allows users to quickly acquire emotion predictions from broadcasted videos and static images.

To ensure that readers grasp the comprehensive understanding, this article is organized as follows: Section 2 provides a detailed explanation of the proposed system for emotion recognition, highlighting the CNN architecture and how probabilities are assigned to different facial expression classes. Section 3 evaluates the performance of the proposed system. It presents the datasets used for training and testing and provides the results of the experiments, showcasing the accuracy achieved by the model. Section 4 presents the designed GUI for a real-time emotion classification application enabling users to interact with the system and obtain emotion predictions effortlessly. Section 5 concludes the article by summarizing the contributions.

## 2. Proposed framework

Figure 1 illustrates the schematic illustration of the proposed framework for facial emotion prediction. The process is explained step by step as follows: Once the face is successfully detected, it is extracted from the rest of the image. This step is crucial as it isolates the region of interest, i.e., the face, for further analysis. After the face is extracted, it undergoes an extensive pre-processing phase. Pre-processing is a critical step in deep learning models as it prepares the input data for the neural network. Various pre-processing techniques are applied to the face image to enhance the neural network's ability to recognize essential features associated with emotions accurately. The pre-processed face image is then fed into the CNN. The CNN is designed to analyze facial features and learn patterns from the input data. It processes the face image through multiple layers of convolutions, pooling, and activations, effectively extracting relevant features to understand the emotional expression. The output of the CNN is a set of probabilities, each corresponding to one of the emotion classes, namely anger, disgust, fear, happy, sad, surprise, and neutral. These probabilities represent the model's confidence in predicting each emotion class based on the given input face image. Finally, the emotion class with the highest probability is determined, indicating the model's prediction for the emotion expressed by the face. The image is classified into the emotion class that the model believes it most likely represents. Further details on each stage are likely provided in subsequent sections of the article to give a comprehensive understanding of the system's working and performance.

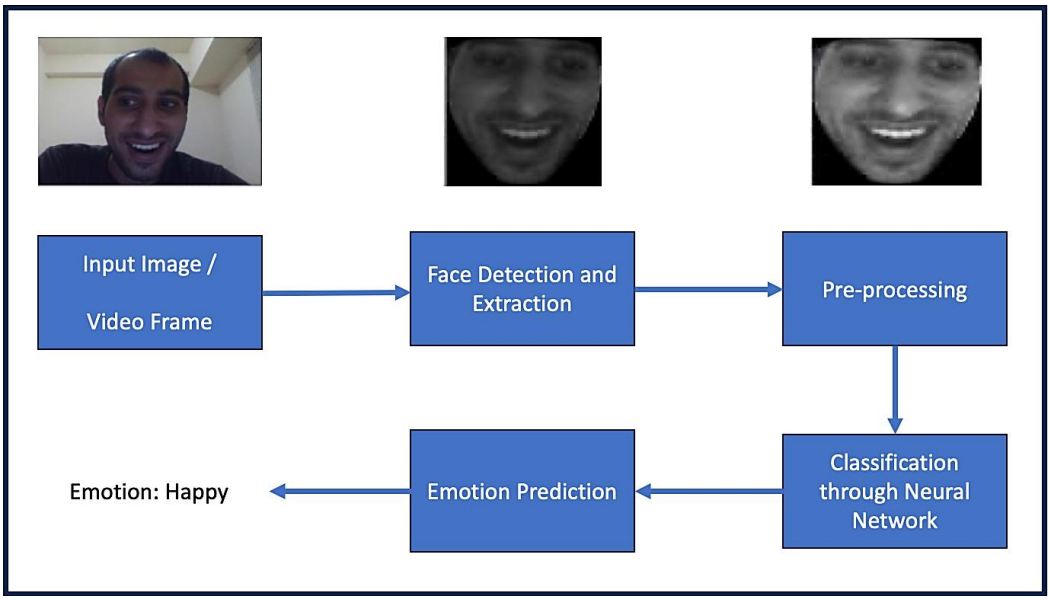


**Figure 1**. Illustration of the proposed framework: a comprehensive schematic representation detailing the architecture and components.

### *2.1. Face detection*

In our proposed framework for facial emotion prediction, face detection plays a pivotal role as the initial step. Various face detection algorithms have been developed, including Haar cascade, HOG +, SVM, and deep learning models. For our system, we have opted to utilize the algorithm proposed in [25] due to its capabilities of speed and real-time usage. This algorithm adeptly detects the locations of 68 (x, y) coordinates on the detected face. These coordinates represent key regions of interest on the face, as illustrated in Figure 2. The 68 points are strategically placed to indicate specific facial features. For instance, points 1-17 define the jawline, points 18-22 correspond to the left eyebrow, points 23-27 to the right eyebrow, points 37-42 to the left eye region, points 43-48 to the right eye region, points 28-36 to the nose region, points 49-60 to the outer lip area, and points 61-68 specify the inner lip structure. To train this face detection algorithm, a labeled training set of facial points on images is used. The (x, y) coordinates of these points are manually labeled to represent various regions of the face. The algorithm utilizes a regression tree model to predict facial landmark points based on the pixel intensities in the image. One notable aspect of this method is the absence of feature extraction, resulting in exceptional speed and real-time performance without compromising accuracy and quality. As facial emotions are primarily expressed through the eyes, nose, eyebrows, and select facial regions, regions above the midpoint of the forehead and ear regions are not required for our emotion prediction task. Consequently, the chosen face detection algorithm accurately identifies facial areas that are relevant to our specific interests. Figure 3 presents some examples of faces detected by our system, with the corresponding regions marked using the 68-point model. The choice of the 68-point model for facial region marking stems from its superior accuracy and detail, enhancing both detection precision and subsequent emotion prediction. This model's effectiveness lies in its ability to capture essential facial landmarks, resulting in improved recognition of subtle emotional cues. Additionally, its detailed approach increases efficiency by reducing false positives and computational demands. This

accuracy-efficiency balance makes our method well-suited for real-time applications like facial emotion recognition. These results demonstrate the effectiveness and accuracy of the face detection algorithm in identifying the crucial facial regions essential for our subsequent facial emotion prediction process. In contrast, some eyebrows may have been marked inconsistently in Figure 3, where the accuracy of facial landmark detection is attributed to various factors such as image quality, pose variations, and the specific facial landmark detection algorithm employed.


**Figure 2.** Facial landmark coordinates visualization.

**Figure 3.** Detecting faces and estimating face areas.

### *2.2. Face extraction*

Following successful face detection in the input images, the subsequent crucial step involves extracting the detected face region from the original image while effectively eliminating the

surrounding background. This process involves isolating the facial area of interest, which is essential for accurate facial analysis and emotion recognition. By precisely extracting the detected face region, the system can focus solely on the facial components, such as the eyes, nose, mouth, and eyebrows, which are pivotal for facial emotion prediction. Eliminating the surrounding background ensures that only the relevant facial features are considered for further processing. This extraction step serves as a vital pre-processing technique, enhancing the efficiency and accuracy of facial emotion recognition algorithms. By isolating the facial region, the subsequent analysis can concentrate on the essential elements that convey emotional expressions, leading to more reliable and robust predictions of human emotions. This face extraction process is essential as it isolates the facial region, which is crucial for analyzing and predicting facial expressions accurately. To achieve face extraction, we begin by using the 68-point facial landmark detection results and selecting the first 27 points (points 1-27). These 27 points cover the overall facial area that primarily deals with facial expressions. Subsequently, we employ a mathematical method called the convex hull to determine the smallest encompassing structure for the set of twenty-seven points. The convex hull algorithm efficiently determines the smallest possible boundary that encompasses all the given points. In our case, applying the convex hull to the 27 facial landmark points creates a boundary that tightly encloses the face region. This convex hull boundary effectively defines the outer limits of the face area, providing a precise and compact representation of the facial structure. This is depicted in Figure 4 (Left), where the convex hull is illustrated around points 0-27.

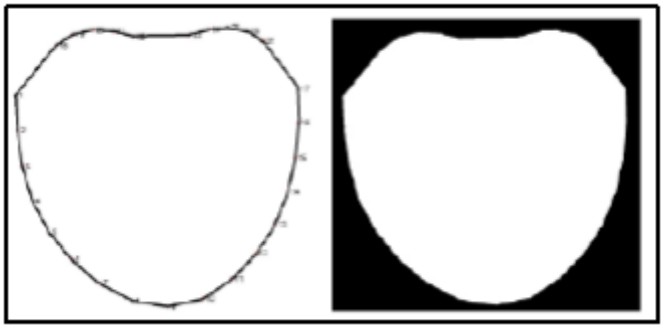

**Figure 4.** Landmarks of 0-27 of the convex hull (left), and its mast (right).

To create a mask facilitating the extraction of the face region, we fill the convex hull using a convex polygon, resulting in a mask that maintains the original image's dimensions. In this mask, the facial area is depicted by white pixels, while the background pixels remain black. This mask effectively aids in isolating the face area for further processing and analysis. Indeed, the mask serves as a crucial guide, delineating the regions of the original image that belong to the face area and distinguishing them from the irrelevant parts. Figure 4 (Right) provides a visual representation of the obtained mask after the convex filling process. With this mask, the face region can be easily isolated from the original image, streamlining the subsequent facial emotion recognition process. This extraction process involves considering only the white-pixel regions defined by the mask, which precisely correspond to the facial area, and discarding the black-pixel regions that represent the surrounding background. As a result, we obtain a clear and isolated facial region that is ideal for subsequent facial emotion recognition and analysis. The extracted face region obtained through this process serves as the input for the subsequent stages of our facial emotion prediction system, facilitating accurate and focused emotion analysis.

### *2.3. Preprocessing*

After locating the mask for face extraction, it is applied to the original images to isolate the face area, as shown in Figure 5. This process involves using the mask as a filter, where the white-pixel regions of the mask correspond to the facial area, and the black-pixel regions represent the background. By applying the mask to the original images, we can effectively extract the face region while eliminating the surrounding background. The resulting images will contain only the facial features, making them ideal for further analysis and facial emotion recognition tasks. This step ensures that the subsequent processing focuses exclusively on the essential facial components, enabling accurate and robust predictions of human emotions. These extracted face images serve as the input data for the CNN in our facial emotion prediction system. To ensure the CNN model receives appropriate and well-prepared data, a series of pre-processing steps are applied to the extracted face images. These pre-processing steps play a crucial role in enhancing the model's ability to learn relevant features and make accurate predictions. The first pre-processing step involves histogram equalization on the cropped face images. Histogram equalization is used for intensity normalization and contrast enhancement, which helps in improving the visibility and clarity of facial features. Histogram equalization is applied as the first pre-processing step on the extracted face images. This technique serves to normalize intensity levels and enhance image contrast. By improving the visibility and clarity of facial features, histogram equalization prepares images for subsequent processing steps.

**Figure 5.** Process of extracting facial regions.

Next, the bilateral filter is utilized to further enhance the face images. The bilateral filter is adept at noise reduction while retaining the integrity of facial feature edges. This filter effectively reduces noise while preserving the edges of facial features. It achieves denoising without compromising crucial details by combining domain filtering and range filtering. By incorporating domain filtering

and range filtering, this filter achieves denoising without compromising essential details. Figure 6 visually demonstrates how the bilateral filter effectively removes noise from the images while retaining high-frequency edges, ensuring that the important facial features are preserved. After applying the bilateral filter, a convolutional 2D filter with a specific kernel is used. The kernel ([-1, -1, -1], [-1, 9, -1], [-1, -1, -1]) is employed to enhance image details and sharpen the features. This filter kernel enhances image details and sharpens features. By further improving the image quality, this step contributes to more accurate feature recognition by the subsequent model. Following the pre-processing steps, all images are resized to a consistent size of 80 x 100 pixels. Standardizing the image size ensures uniformity in data input to the CNN model. Finally, the pre-processed images are organized into arrays for both training and testing purposes. These arrays are then fed into the CNN model, which utilizes this well-prepared data to learn and make predictions about facial emotions.

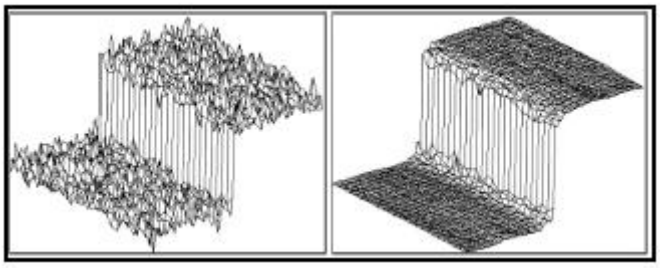

**Figure 6.** Noise reduction and edge preservation using bilateral filtering.

Through the application of these pre-processing steps, the input data presented to the CNN model undergoes optimization, ensuring a conducive environment for accurate and efficient learning. By enhancing the data's quality and reducing noise, the model can effectively discern meaningful patterns and features, leading to improved facial emotion recognition performance. The combination of face extraction, pre-processing, and CNN-based emotion prediction forms a powerful system capable of recognizing and classifying facial emotions with high accuracy. These augmentation, normalization, and feature extraction techniques collectively optimize the input data for the CNN model. By enhancing data quality, reducing noise, and preserving relevant features, the model becomes better equipped to recognize meaningful patterns and achieve improved performance in facial emotion recognition. The systematic application of these techniques results in a robust and accurate system for recognizing and classifying facial emotions.

### *2.4. CNN architecture*

The architecture we propose for facial emotion prediction involves a series of layers tailored to effectively learn and categorize emotions from input facial images. This CNN structure comprises specific components: Five convolutional layers, one max pooling layer, two average pooling layers, and three dense layers. This thoughtful arrangement optimizes feature extraction and classification, resulting in robust facial emotion recognition capabilities. For regularization, we incorporate a 20% dropout rate in the dense layers to counter overfitting and enhance generalization. The input layer accommodates images of dimensions 80 x 100, matching the size of the utilized face images. These images then undergo processing through the initial convolutional layer, composed of 64 filters with a (5, 5) kernel size and employing the ReLU activation function. Post this convolutional layer, the output size becomes (76, 96, 64). A subsequent max pooling layer, with a (5, 5) pooling size and (2, 2) strides,

further reduces dimensions to (36, 46, 64). The following sequence involves two consecutive convolutional layers, each equipped with 64 filters and a (3, 3) kernel size utilizing the ReLU activation function.

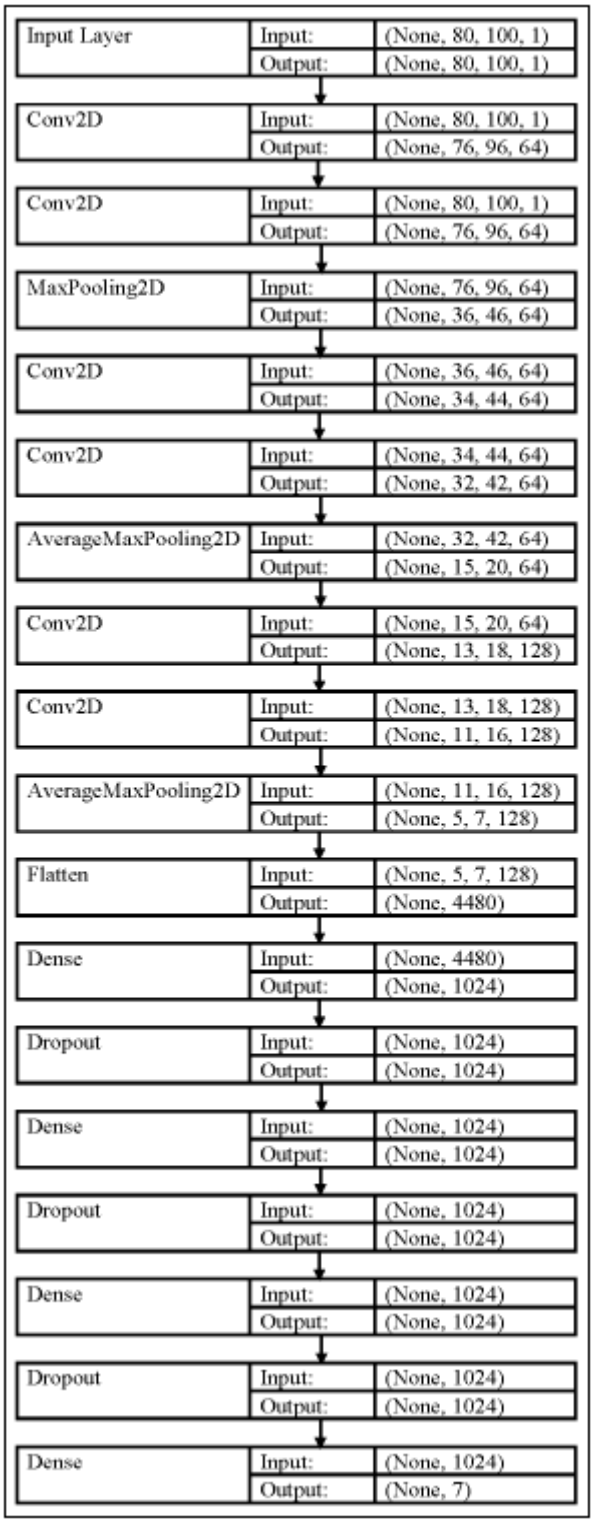


**Figure 7.** The proposed CNN architecture.

The second convolutional layer yields an output size of (32, 42, 64), contributing significantly to capturing relevant features from the input data and enhancing pattern recognition in facial expressions. Once through the average pooling layer with a (3, 3) pool size and (2, 2) strides, the output size

becomes (15, 20, 64). This output progresses through two more convolutional layers, each with 128 filters and a (3, 3) kernel size using ReLU activation. The resultant output is (11, 16, 128). Another average pooling layer, mirroring the prior specifications, produces an output size of (5, 7, 128). This output is then flattened into a one-dimensional vector with a size of 4480, which is fed into three fully connected dense layers, each comprising 1024 filters. For added robustness and to counter overfitting, dropout with a 0.2 rate is applied to each dense layer. The final dense layer employs the SoftMax activation function, culminating in an output size of 7. These seven values correspond to probabilities of each emotion class (anger, disgust, fear, happy, sad, surprise, and neutral). The class with the highest probability signifies the model's prediction for the given facial expression, ensuring precise and dependable emotion recognition. The choice of the ReLU activation function in the convolutional layers over the sigmoid is based on its ability to mitigate the vanishing gradient problem, ensuring more effective learning. Additionally, ReLU promotes sparse representation, which has proven to be beneficial for deep learning tasks. This architecture, as described and depicted in Figure 7, constitutes the backbone of the proposed CNN model, enabling accurate and efficient facial emotion prediction.

## 3. Evaluation

Within the evaluation section, comprehensive insights are furnished regarding the dataset employed in the study, followed by a meticulous analysis of the performance exhibited by their proposed systems.

### *3.1. Dataset*

Within the scope of this article, a fusion of two separate datasets was executed [26,27]. The Japanese Female Facial Expression dataset [26] encompasses a compilation of 213 static grayscale female images featuring 10 distinct models, all presented at a resolution of 256 x 256 pixels. On the other hand, the Karolinska Directed Emotional Faces dataset comprises [27] 4900 grayscale images from 70 different models. The images in both datasets have a resolution of 572 x 762 pixels. All images were captured under consistent lighting conditions, and the models did not wear makeup, glasses, or earrings. Each model represented in both datasets underwent a comprehensive imaging process, capturing a diverse array of images from five distinct perspectives: From full right to half left, half right, entire left, and frontal. The positions of the mouth and eyes were consistently fixed on a grid for all images, ensuring standardized facial features' alignment. Subsequently, the images were cropped to a specified resolution, providing a consistent and uniform dataset. The age range of the models included in both datasets ranged from 20 to 30 years, ensuring a consistent age group for emotion analysis and prediction. The combined datasets encompass images representing seven distinct facial expressions: Anger, disgust, fear, happy, sad, surprise, and neutral. For the purpose of analysis and training in their facial emotion recognition system, we assigned numeric labels (0, 1, 2, 3, 4, 5, 6) to each of these expressions. This labeling scheme facilitates streamlined processing and classification of emotions within the proposed system. The JAFFE and KDEF datasets, while widely used, have limitations in terms of diversity in terms of age, ethnicity, gender, and cultural background. This lack of diversity may lead to biased or limited model performance when applied to broader, real-world scenarios involving a diverse range of individuals. Moreover, the limited size of these datasets can pose challenges in terms of generalization. Models trained on smaller datasets may struggle to generalize well to previously unseen variations in facial expressions, lighting conditions, and other environmental factors. By merging JAFFE and KDEF datasets, we aimed to overcome limitations and

create a diverse training dataset, enhancing the model's ability to generalize to unseen variations in facial expressions, lighting, and environments. This approach addressed dataset size challenges, enriched model training, and improved performance in real-world scenarios

To create a comprehensive dataset, we merged both datasets, organizing the images based on their corresponding emotion labels. Following data pre-processing and augmentation techniques, the dataset underwent division into separate training and testing sets. We employed diverse data augmentation techniques, including rotation, brightness adjustment, geometric transformations, and noise injection, to enhance the model's robustness and generalization, improving its performance in facial emotion recognition tasks. Specifically, the testing data constitutes around 11% of the training data. With a total of 14,200 images, the training dataset serves as the foundation for the system's learning process, while the test dataset, containing 1,580 images, evaluates the model's performance and generalization capabilities. To ensure the dataset's quality and to prioritize facial expression recognition over face detection, images with undetected faces during the pre-processing stage were excluded. This approach focused on refining the model's ability to accurately recognize emotions from the detected facial regions. Additionally, we aimed to maintain a balanced dataset by ensuring that each emotion class had a similar number of samples. The distribution statistics of each emotion class in the final datasets are depicted in Figure 8, demonstrating the dataset's evenness and suitability for emotion recognition training. By combining and processing these datasets, we have created a robust and diverse dataset for training and evaluating their proposed systems. The balanced nature of the dataset ensures that the model can learn effectively from various facial expressions and generalize well to unseen data during testing.

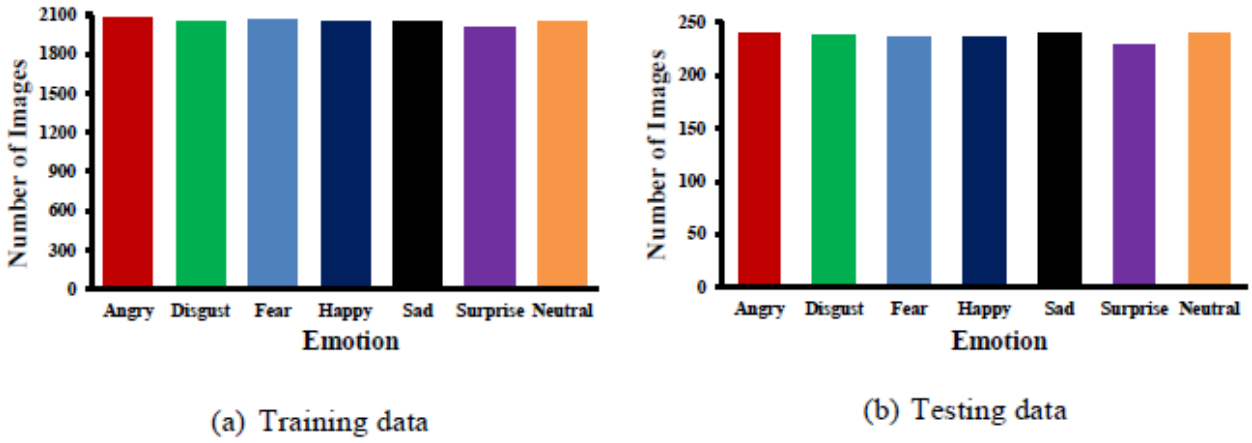


**Figure 8.** (a) Training and (b) Testing data split of the emotion dataset.

### *3.2. Results*

During the evaluation of the proposed system, we conducted a series of experiments to determine the optimal architecture and parameters for the CNN model. Following the initial hyperparameter tuning phase, the selected configuration involved a learning rate of 0.01 coupled with a batch size of 100. Stochastic Gradient Descent was employed as the optimizer. The convergence criterion was defined as a scenario in which the model's accuracy remained stagnant for a continuous span of 20 to 30 epochs. The initial phase of training involved utilizing a batch size of 100 and experimenting with different configurations of CNN layers, ranging from 1 to 8. The objective was to determine the optimal number of layers that would yield the highest accuracy for the facial emotion recognition

system. After thorough experimentation, we discovered that the highest accuracy was achieved with five CNN layers. Increasing the number of layers beyond this point resulted in longer execution times without significant accuracy improvements. The trend of accuracy with varying CNN layers is depicted in Figure 9. It's interesting to observe that the accuracy peaks at 70.2% with five layers in the model. This optimal configuration was extensively explored in subsequent experiments to enhance the system's performance.

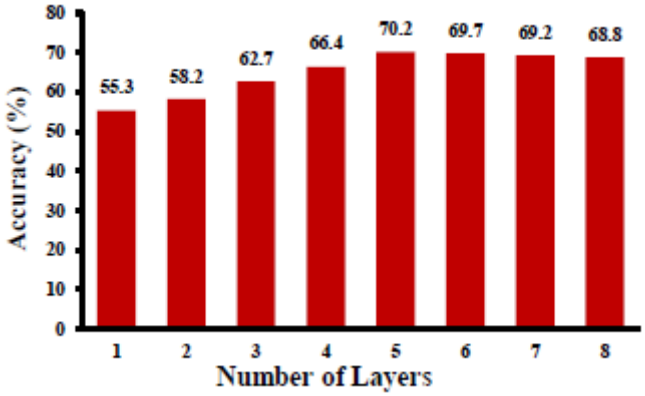


**Figure 9.** Accuracy vs. number of layers.

After determining the optimal number of layers, we proceeded to adjust the number of training epochs within a range of 25 to 500 for fine-tuning the model. After determining the optimal number of layers, we proceeded to adjust the number of training epochs within a range of 25 to 500 for fine-tuning the model. Figure 10 displays the relationship between the number of epochs and the corresponding accuracy. As expected, the accuracy improved with increasing epochs. However, the rate of improvement slowed down considerably after 300 epochs. Considering the balance between accuracy and computational resources, we selected 300 epochs as the optimal choice. At this point, the model achieved an impressive accuracy of 78.1%, making it a competitive contender for comparison with other models in further evaluations.

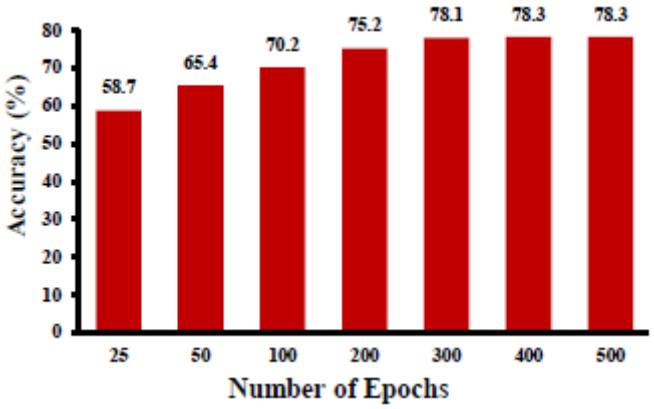


**Figure 10.** Accuracy vs. number of epochs.

Table 1 displays the performance metrics for each emotion class, as well as the average values across all classes. These metrics provide a comprehensive evaluation of the proposed system's performance in recognizing various facial emotions. The model achieves an average precision of 0.78. This indicates that, on average, when the model predicts any emotion, it is correct approximately 78%

of the time. This suggests a relatively high level of accuracy in the model's overall predictions. The average recall score is 0.77. This means that, on average, the model successfully identifies and captures 77% of all instances of facial expressions across all emotion classes. The recall score reflects the model's ability to detect actual instances of emotions. The average F1-Score is 0.77, and the F1-Score is a balanced metric that considers both precision and recall. It measures overall accuracy by harmonizing the trade-off between false positives and false negatives. The average F1-Score of 0.77 indicates a consistent and balanced performance across all emotion classes. These average metrics indicate that the proposed model performs well in recognizing and classifying facial emotions. The balanced values of precision, recall, and F1-Score reflect a robust and reliable performance in accurately predicting emotions while maintaining a reasonable trade-off between precision and recall. Overall, the model's average performance underscores its effectiveness in capturing and interpreting a diverse range of human emotions.

**Table 1.** Performance metrics for all emotions.

| Emotion | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.84 | 0.56 | 0.67 | 230 |
| 1 | 0.82 | 0.86 | 0.84 | 225 |
| 2 | 0.60 | 0.60 | 0.60 | 225 |
| 3 | 0.90 | 0.88 | 0.89 | 225 |
| 4 | 0.79 | 0.83 | 0.81 | 230 |
| 5 | 0.87 | 0.84 | 0.85 | 220 |
| 6 | 0.67 | 0.85 | 0.75 | 225 |
| Avg. | 0.78 | 0.77 | 0.77 | 1580 |

It is worth mentioning that there are several challenges affecting models' performance, including: (i) Changes in lighting conditions which drastically affect the appearance of facial features, making it difficult for a model to identify consistent patterns across different lighting scenarios; (ii) facial expressions can vary based on the pose of the face, such as head orientation; and (iii) different individuals express the same emotion in unique ways, and emotions themselves are complex, often blending multiple expressions. This variability poses a significant challenge to accurate recognition. The proposed framework of CNN has the ability to automatically learn hierarchical features from raw data. This includes the ability to capture invariant features across different lighting conditions. Also, the CNN is inherently equipped to capture spatial hierarchies in images, enabling the model to recognize facial expressions across various poses. This is facilitated by the convolutional and pooling layers, which create feature maps that can capture patterns in different spatial resolutions. Moreover, with the challenge of limited data and enhanced model generalization, data augmentation techniques are employed. By introducing variations in lighting, pose, and other factors during training, the model becomes more adaptable to real-world conditions.

### 3.2.1. Comparative analysis

Figure 11 presents a comparative analysis in terms of the accuracy (%) achieved by the proposed facial emotion recognition model with that of several standard networks, namely AlexNet, GoogleNet, ResNet, VGG, and a custom CNN**.** All models were trained in terms of accuracy, where VGG achieves 68.7%, but it has limitations in facial emotion recognition compared to the proposed CNN model. AlexNet's 69.6% accuracy is lower, possibly due to its suitability for this complex task. ResNet

achieves 71.2% accuracy, outperformed by the CNN model, possibly due to distinct feature learning. GoogleNet's 74.1% accuracy is higher than AlexNet's but lower than the CNN models. The CNN model achieves a higher accuracy of 78.1% due to its tailored architecture, effectively recognizing and classifying facial emotion features, surpassing other established models like AlexNet, GoogleNet, ResNet, and VGG. This success is attributed to its robust CNN techniques and comprehensive training on diverse datasets, enabling nuanced pattern recognition across various facial expressions. The higher accuracy of the proposed model signifies its ability to recognize and classify facial emotions more accurately, making it a promising approach for real-world applications requiring facial emotion analysis. Despite the limitations of a small dataset, we successfully designed a system with good accuracy for facial emotion recognition. The proposed CNN model's performance highlights its potential for real-world applications, especially with more extensive datasets for training. The experiments and evaluation results validate the effectiveness and superiority of the proposed system in recognizing facial emotions compared to state-of-the-art models.

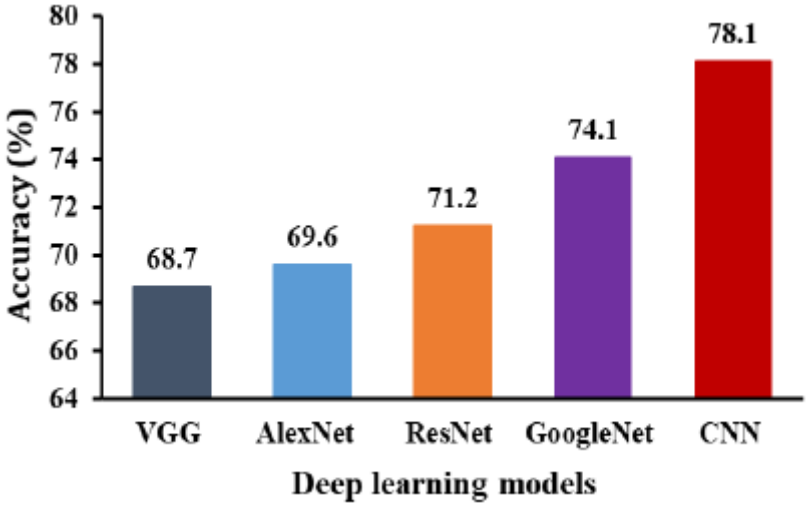


**Figure 11.** Accuracy of different deep learning models.

## 4. Designed GUI

The trained CNN model impeccably integrates into a dynamic application, providing users with the flexibility to analyze both static images and live video streams for facial emotion recognition. Within the intuitive user interface, individuals have the freedom to select their preferred mode of operation, whether it be processing single images or analyzing real-time video footage. Upon receiving input, the application swiftly initiates pre-processing steps to enhance image quality and subsequently predicts facial emotions with remarkable accuracy. Leveraging cutting-edge techniques, the pre-processing and labeling stages are completed within a mere 20-30 milliseconds, ensuring seamless real-time performance across both image and video processing tasks. This streamlined process empowers users with efficient and accurate facial emotion analysis capabilities, making it an invaluable tool for various applications and scenarios.


**Figure 12.** GUI results for happy, angry, and disgusted facial expressions (Top, Middle, and Bottom, respectively).

We face several challenges in real-time application development encompassed achieving real-time performance, designing a user-friendly interface, ensuring compatibility across diverse environments, maintaining accuracy and robustness, and optimizing for resource-constrained devices. These were addressed through techniques such as optimizing the computational efficiency of the CNN model, conducting UI usability testing, compatibility testing, rigorous training, and validation

procedures, and employing optimization techniques like model compression and quantization for resource-constrained devices. The designed GUI showcases both the unaltered image or frame extracted from the video and the cropped facial image employed for the recognition of facial emotions. Additionally, two labels are included in the interface: 1. Probabilities of All Labels: This label shows the probabilities assigned to each emotion label by the CNN model. For example, for a given facial expression, the model may assign probabilities for each emotion class, such as anger, disgust, fear, happy, sad, surprise, and neutral. 2. Predicted Emotion: This label shows the emotion predicted by the model based on the highest probability among all emotion classes. As an example, when the maximum probability is associated with the "happy" emotion class, the label will exhibit "happy" as the forecasted emotion. To enhance the accuracy of emotion recognition, the application employs a strategy of calculating the average probability of related emotions across all processed images. By aggregating the probabilities of different emotions over multiple images, the system aims to obtain a more robust and reliable prediction for each facial expression, leading to improved accuracy in emotion recognition. This approach helps select the proper label, thereby increasing the overall accuracy of the emotion prediction. In Figure 12 (Top), an example of a happy face is showcased, and the corresponding emotion label "happy" is prominently displayed on the top-right label. The misalignment in the user's eyes in Figure 12 (Top) is attributed to the real-time processing lag of 20-30 milliseconds." In Figure 12 (Middle), the image depicts an expression of anger, and the label "angry" is distinctly presented. In the lower section of Figure 12, the system identifies a facial expression conveying disgust, and the associated emotion label "disgust" is showcased on the left. Additionally, the application visually displays the probabilities for all three images at the bottom, providing valuable insights into the model's degree of confidence in its predictions for each distinct emotion class. The user-friendly interface allows users to easily interpret the results and gain a comprehensive understanding of the emotion recognition process. This user-friendly and real-time application is efficient in displaying the predicted emotions and associated probabilities, making it a valuable tool for facial emotion recognition in both still images and live video feeds. The clear and intuitive user interface enhances the overall user experience, allowing for effective and accurate analysis of human emotions in real-world scenarios.

## 5. Conclusions

In conclusion, we introduce a highly efficient and accurate system for facial emotion recognition, leveraging the power of CNN. The proposed system utilizes the power of deep learning to predict and assign probabilities to each facial emotion, enabling precise emotion classification. To ensure the best performance, the system applies diverse pre-processing steps to each image before feeding it into the CNN. These pre-processing steps play a crucial role in optimizing the neural network's ability to recognize essential features from the input data. The experimental results demonstrate the superiority of the proposed system over other models, indicating its effectiveness in facial emotion recognition tasks. The model's performance is further validated through the real-time application with the GUI, showcasing promising results and accuracy. We recognize the challenges posed using complex networks, as they can result in heavy models that perform poorly during live video processing. In response to this concern, the proposed system adopts a strategic approach that strikes a balance between accuracy and efficiency. The system ensures smooth and effective facial emotion prediction without overburdening the processing capabilities by prioritizing real-time performance while preserving robust emotion recognition capabilities. This equilibrium between accuracy and efficiency

enhances the practical applicability of the proposed system, making it a valuable tool for real-world emotion recognition tasks, particularly in dynamic environments where real-time processing is critical. Our research greatly advances facial emotion analysis by achieving a balance between efficiency and accuracy, enhancing robustness and generalization, providing benchmarking and comparative analysis, and suggesting future directions for multi-modal analysis, contextual understanding, novel architectures, and ethical considerations. These efforts promise valuable insights into understanding and analyzing human emotions, enriching human-computer interactions across domains.

## Author contributions

Imad Ali: Conceptualization, Software, Validation, Investigation, Resources, Writing – review & editing, Methodology, Validation, Supervision; Faisal Ghaffar: Writing-original draft, Software, Validation, Formal analysis, Investigation, Project administration, Data curation, Visualization. All authors have read and approved the final version of the manuscript for publication.

## Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Conflict of interest

The authors declare no conflicts of interest in this paper.

## References

1. Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. *Comput Speech Lang* 25: 556–570. https://doi.org/10.1016/j.csl.2010.10.001
2. Erol BA, Majumdar A, Benavidez P, Rad P, Choo KKR, Jamshidi M (2019) Toward artificial emotional intelligence for cooperative social human-machine interaction. *IEEE Transactions on Computational Social Systems* 7: 234–246. https://doi.org/10.1109/tcss.2019.2922593
3. Cohn JF, Ambadar Z, Ekman P (2007) Observer-based measurement of facial expression with the Facial Action Coding System. *The Handbook of Emotion Elicitation and Assessment* 1: 203–221. https://doi.org/10.1093/oso/9780195169157.003.0014
4. Vaillant R, Monrocq C, Le Cun Y (1994) Original approach for the localization of objects in images. *IEE Proceedings-Vision, Image and Signal Processing* 141: 245–250. https://doi.org/10.1049/ip-vis:19941301
5. Rowley HA, Baluja S, Kanade T (1998) Neural network-based face detection. *IEEE T Pattern Anal* 20: 23–38. https://doi.org/10.1109/34.655647
6. Jain V, Learned-Miller E (2010) FDDB: A benchmark for face detection in unconstrained settings. Technical Report UMCS-2010-009, University of Massachusetts, Amherst. Available from: https://people.cs.umass.edu/~elm/papers/fddb.pdf
7. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. *IEEE Conference on Computer Vision and Pattern Recognition*, 2879–2886. https://doi.org/10.1109/cvpr.2012.6248014

8. Yan J, Zhang X, Lei Z, Li SZ (2014) Face detection by structural models. *Image Vision Comput* 32: 790–799. https://doi.org/10.1109/fg.2013.6553703
9. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1–9. https://doi.org/10.1109/cvpr.2001.990517
10. Lienhart R, Maydt J (2002) An extended set of Haar-like features for rapid object detection. *Proceedings of the IEEE International Conference on Image Processing*, 1–4. https://doi.org/10.1109/icip.2002.1038171
11. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: Recognition using class-specific linear projection. *IEEE T Pattern Anal* 19: 711–720. https://doi.org/10.1109/34.598228
12. Yang MH, Kriegman DJ, Ahuja N (2002) Detecting faces in images: A survey. *IEEE T Pattern Anal* 24: 34–58. https://doi.org/10.1109/34.982883
13. Wu J, Zhang C, Xue T, Freeman B, Tenenbaum J (2016) Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. *Advances in Neural Information Processing Systems*, 82–90. https://doi.org/10.1609/aaai.v32i1.12223
14. Sroubek F, Milanfar P (2011) Robust multichannel blind deconvolution via fast alternating minimization. *IEEE T Image Process* 21: 1687–1700. https://doi.org/10.1109/tip.2011.2175740
15. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105. https://doi.org/10.1145/3065386
16. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: Integrated recognition, localization, and detection using convolutional networks. *arXiv preprint arXiv*:1312.6229. https://doi.org/10.48550/arXiv.1312.6229
17. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2006) Fully automatic facial action recognition in spontaneous behavior. *7th International IEEE Conference on Automatic Face and Gesture Recognition*, 223–230. https://doi.org/10.1109/fgr.2006.55
18. Pantic M, Rothkrantz LJ (2004) Facial action recognition for facial expression analysis from static face images. *IEEE T Syst Man Cy B* 34: 1449–1461. https://doi.org/10.1109/tsmcb.2004.825931
19. Tian YI, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. *IEEE T Pattern Anal* 23: 97–115. https://doi.org/10.1109/cvpr.2000.855832
20. Ng HW, Nguyen VD, Vonikakis V, Winkler S (2015) Deep learning for emotion recognition on small datasets using transfer learning. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 443–449. https://doi.org/10.1145/2818346.2830593
21. Chaudhari A, Bhatt C, Krishna A, Travieso-Gonz ález CM (2023) Facial emotion recognition with inter-modality-attention-transformer-based self-supervised learning. *Electronics* 12: 1–15. https://doi.org/10.3390/electronics12020288
22. Yang D, Huang S, Wang S, Liu Y, Zhai P, Su L, et al. (2022) Emotion recognition for multiple context awareness. *Proceedings of the European Conference on Computer Vision*, 144–162. https://doi.org/10.1007/978-3-031-19836-6_9
23. Song C, Ji S (2022) Face Recognition Method Based on Siamese Networks Under Non-Restricted Conditions. *IEEE Access* 10: 40432–40444. https://doi.org/10.1109/access.2022.3167143
24. Qu X, Zou Z, Su X, Zhou P, Wei W, Wen S, et al. (2021) Attend to where and when: Cascaded attention network for facial expression recognition. *IEEE Transactions on Emerging Topics in*

*Computational Intelligence* 6: 580–592. https://doi.org/10.1109/tetci.2021.3070713
25. King DE (2009) Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10: 1755–1758. https://doi.org/10.1145/1577069.1755843
26. Lyons MJ, Kamachi M, Gyoba J (1997) Japanese female facial expressions (JAFFE). Database of Digital Images. https://doi.org/10.5281/zenodo.3451524
27. Goeleven E, De Raedt R, Leyman L, Verschuere B (2008) The Karolinska directed emotional faces: a validation study. *Cognition and Emotion* 22: 1094–1118. https://doi.org/10.1080/02699930701626582


AIMS Press