applied sciences

MDPI

Article

Expression Recognition of Multiple Faces Using a Convolution Neural Network Combining the Haar Cascade Classifier

Jui-Feng Yeh *, Kuei-Mei Lin, Chia-Chen Chang and Ting-Hao Wang

Department of Computer Science and Information Engineering, National Chiayi University, Chiayi City 60004, Taiwan; s1120311@mail.ncyu.edu.tw (K.-M.L.); s1062997@mail.ncyu.edu.tw (C.-C.C.); s1062999@mail.ncyu.edu.tw (T.-H.W.)

* Correspondence: ralph@mail.ncyu.edu.tw

Abstract: Facial expression serves as the primary means for humans to convey emotions and communicate social signals. In recent years, facial expression recognition has become a viable application within medical systems because of the rapid development of artificial intelligence and computer vision. However, traditional facial expression recognition faces several challenges. The approach is designed to investigate the processing of facial expressions in real-time systems involving multiple individuals. These factors impact the accuracy and robustness of the model. In this paper, we adopted the Haar cascade classifier to extract facial features and utilized convolutional neural networks (CNNs) as the backbone model to achieve an efficient system. The proposed approach achieved an accuracy of approximately 70% on the FER-2013 dataset in the experiment. This result represents an improvement of 7.83% compared to that of the baseline system. This significant enhancement improves the accuracy of facial expression recognition. Herein, the proposed approach also extended to multiple face expression recognition; the module was further experimented with and obtained promising results. The outcomes of this research will establish a solid foundation for real-time monitoring and prevention of conditions such as depression through an emotion alert system.

Keywords: convolutional neural network; Haar cascade classifier; facial expression recognition

Citation: Yeh, J.-F.; Lin, K.-M.; Chang, C.-C.; Wang, T.-H. Expression Recognition of Multiple Faces Using a Convolution Neural Network Combining the Haar Cascade Classifier. *Appl. Sci.* **2023**, *13*, 12737. https://doi.org/10.3390/app132312737

Academic Editors: Teen-Hang Meen, Chun-Yen Chang, Po-Lei Lee, Charles Tijus and Kuei-Shu Hsu

Received: 21 October 2023
Revised: 23 November 2023
Accepted: 24 November 2023
Published: 28 November 2023

1. Introduction

In the wake of the COVID-19 pandemic, online video communication has become widespread, and we have come to realize that facial expressions on the screen play a significant role in mutual interactions. In recent years, facial expression recognition has emerged as a critical determinant of danger because of the rapid advancements in artificial intelligence and computer vision. For instance, it is now used for detecting driver fatigue in automotive safety systems [1]. In the system, driver behavior is monitored through an in-car camera. The camera analyzes information such as head posture, eye movements, and facial expressions to detect the level of driver fatigue. In the medical field, emotion recognition has become a preventive measure for conditions such as depression and other mental health disorders. The system analyzes individuals' emotional states by detecting emotions. If it detects prolonged negative emotional states, the system will encourage or connect users to professional mental health resources.

In the past, people primarily relied on experience or a combination of their cognitive understanding of others to judge emotions when receiving emotional signals. However, these methods often carried biases. Today, people are turning to objective machines for expression recognition. These models automatically extract facial features and classify emotional categories. Yang et al. [2] combined VGG 16 and LBP to extract facial features and classify six emotion categories with *softmax*. Agrawal et al. [3] aimed to extract useful facial features and designed two novel convolutional neural network architectures. Then, they also achieved a 69% accuracy result.

Traditional facial expression recognition faces several challenges, including imbalanced data, privacy concerns, and limitations in handling only single-person facial images. For example, datasets containing human faces are often incomplete and limited due to privacy concerns. Additionally, the dataset also suffers from data sparsity due to the limited number of samples for specific emotion categories. Furthermore, existing facial emotion recognition systems are constrained by their ability to process only single-face images. Therefore, we adopted data augmentation to enhance the facial image and improve the data sparsity problem. Subsequently, a CNN with the Haar cascade classifier ensemble was employed as the backbone network for the method. While the CNN with the Haar cascade classifier approach may not necessarily yield the highest accuracy, it possesses the capability to perform real-time detection on multiple individuals. This makes it valuable in applications such as automatic attendance systems in classroom settings, negative emotion alerts for mental health patients, and surgical assessments for facial deformities. We propose a method for facial expression recognition with low-cost hardware. In this method, facial expression recognition is regarded as an emotion classification task. We augment the facial image data, utilize the Haar cascade classifier for face localization, and employ a convolutional neural network (CNN) for emotion classification.

The main contributions of this article are as follows:

1. Integrating the Haar cascade classifier and CNN model to enhance feature extraction and ensure real-time capabilities of the model.
2. Designing a facial expression recognition system in both single-person and multi-person images.

The rest of this article is organized as follows: Section 2 reviews the previous research on facial expression recognition. Section 3 describes the proposed approach. Section 4 presents the experimental results and a discussion of these results. Finally, Section 5 concludes this article.

2. Related Works

Face expression is an important element in human emotional expression. Therefore, facial expression recognition has consistently been a critical research theme in the fields of affective computing and computer vision. Yang et al. [2] proposed a weighted mixture deep neural network. The network combined VGG16 and LBP with weighted fusion to extract facial features and used *softmax* to classify six kinds of emotion categories. They experimented on the JAFFE dataset, CK+ dataset, and Oulu-CASIA dataset, achieving excellent results in each case. Agrawal et al. [3] researched the impact of different combinations of filter sizes and quantities on classification accuracy. They also designed two architectures best suited for facial expression recognition. The FER-2013 dataset was employed to evaluate the performance of these architectures. Sarvakar et al. [4] proposed facial emotion recognition using convolutional neural networks (FERC). FERC is divided into two parts: background removal and facial feature extraction. The CNN model was used to segment the facial region in background removal. In the facial feature extraction stage, FERC utilized the expression vectors as a representation of their facial features. Finally, the expression vectors were further fed into the CNN model and classified the emotions. Khattak et al. [5] enhanced the accuracy of facial expression recognition by augmenting the layers of the CNN model. Additionally, this model demonstrated effective age and gender classification. Zeng et al. [6] proposed facial expression recognition to automatically distinguish the expressions with high accuracy. Facial geometric features and appearance features were introduced into the model. In addition, the deep sparse autoencoder (DSAE) was adopted to learn a robust and discriminative facial expression recognition system. Zeng et al. [7] proposed a real-time facial expression recognition and learning feedback system. The system analyzed students' facial expressions instantly and assessed their learning emotions. Additionally, many facial expression recognition systems utilized CNN models as their backbone [8–10]. Hence, it is evident that CNN models are highly beneficial for facial expression recognition tasks. Gao et al. [11] presented cross-domain facial expression recognition (CD-FER). CD-FER

integrates features across domains and incorporates dynamic label weighting. This enables the model to leverage the potential benefits of transferable cross-domain local features, resulting in improved performance in facial emotion recognition tasks. Tang et al. [12] introduced BGA-Net, a novel approach to facial expression recognition. BGA-Net integrates bidirectional gated recurrent units (BiGRUs), a convolutional neural network (CNN), and an attention mechanism. The primary objective of BGA-Net is to capture dependencies between sub-regions in the image, focusing on the most discriminative areas through an attention mechanism.

In deep learning, the scarcity of data due to challenges in data collection or limited sample sizes often leads to the problem of insufficient data. Such a dataset can potentially have a detrimental impact on the training and learning performance of models. Therefore, data augmentation is employed as a solution to overcome the limitations of limited training data. Lai [13] utilized additional image samples and data augmentation techniques to improve errors in expression recognition. Liu [14] used GAN-based image enhancement techniques, and data augmentation has been employed to enhance image quality and improve the accuracy of the model. Porcu et al. [15] evaluated the impact of various data augmentation techniques on facial expression recognition. Umer et al. [16] proposed novel data augmentation. The novel data augmentation combines a bilateral filter, a sharpening filter, image rotation, and more. Then, this data augmentation can enhance the fine-tuning capabilities of a CNN model and improve the overfitting problem. Psaroudakis et al. [17] proposed a MixAugment data augmentation method based on Mixup. Mixup demonstrated an improvement in the recognition rate for wild data in the context of facial expression recognition tasks. Bobojanov et al. [18] introduced a novel facial emotion recognition model vision transformer (ViT) model. Recognizing the common issue of data imbalance for widely used emotion datasets, the researchers opted for the evaluation of their model on RAF-DB and FER-2013. Additionally, to address the imbalance challenge, they employed data augmentation techniques to construct a balanced dataset.

A wavelet is a mathematical tool used in signal processing and image processing, frequently applied in applications such as facial recognition. Wavelets exhibit advantages in multi-scale analysis and feature extraction. The particular significance of real-time facial recognition is due to its feature compression properties. The Symlet wavelet, Coiflet wavelet, and Haar wavelet are commonly employed wavelet techniques in facial recognition. The Symlet wavelet is characterized by its symmetry and flexibility. With adjustable parameters, it provides improved accuracy in facial identification. The Coiflet wavelet is a family of compactly supported wavelets. Its finite support property effectively reduces computational complexity while preserving essential features. Therefore, it exhibits superior performance in both time and frequency domains. The Haar wavelet is the simplest form of a wavelet. Its straightforward computation gives it a significant advantage in real-time signal processing. While the Symlet wavelet and Coiflet wavelet can depict more detailed facial contour information, their computational demands make them less feasible in resource-constrained scenarios. Conversely, the Haar wavelet is unable to capture intricate facial contour details. However, its minimal resource requirement and precise edge detection capabilities contribute to its excellent performance in capturing facial positions. In our system, the prioritization of facial position detection outweighs the demand for capturing detailed facial contours. Therefore, we opted for the Haar wavelet to be incorporated as a part of our model.

Goyani et al. [19] discussed various applications of Haar in different fields and categorized the proposed methods. Shen et al. [20] employed an enhanced active shape model (ASM) for facial expression recognition, utilizing Haar features to automate face detection in their applications and research. Liu et al. [21] proposed an Adaboost-based face detection algorithm using Haar-like features. The model aimed to improve issues related to extended training times and low detection efficiency. Mishra et al. [22] proposed a facial recognition system. The system combines Haar-like features, face detection, and OpenCV to compare certain facial features in the image with a facial dataset. Guo et al. [23] proposed

a CNN-enhanced multi-level Haar wavelet features fusion network (CNN-MHWF2N). The system combines the spatial features of the 2-D-CNN and the Haar wavelet decomposition features to obtain spectral and spatial information. Wu et al. [24] proposed a face image recognition method. The model is based on Haar-like and Euclidean distance and improved accuracy of facial recognition. Zhang et al. [25] proposed a weak classifier that improved the Haar-like features and the AdaBoost algorithm. The weight update method in the AdaBoost algorithm was changed and achieved an effective accuracy improvement. Farkhod et al. [26] presented a graph-based approach to emotion recognition. In their methodology, a two-step process was employed, combining the Haar cascade classifier and a media-pipe face mesh model. Initially, the Haar-cascade classifier was utilized for face detection, followed by the application of the media-pipe face mesh model for precise landmark localization. This synergistic combination allowed for the extraction of facial features critical for emotion recognition. Notably, the proposed framework by Farkhod et al. demonstrated effective emotion prediction even in scenarios where individuals were wearing masks. Chinimilli et al. [27] introduced an attendance system that combines the Haar cascade and the Local Binary Pattern Histogram (LBPH) algorithm. In the paper, the Haar cascade is employed for face detection, while LBPH is utilized for face recognition. Shafique et al. [28] employed the Haar cascade classifier to implement the detection of social anxiety disorder (SAD). They utilized the detection of gaze interaction/avoidance to determine whether individuals were afflicted by SAD. Takiddin et al. [29] proposed a facial deformity measurement method based on the Haar cascade. Haar was trained on normal facial data and deformity data to learn the differences and features among different faces, providing a score indicating the degree of facial deformity.

3. Proposed Method

In this section, we introduce the details of the proposed method. Our system architecture is illustrated in Figure 1. The system was divided into training and test stages. In the training stage, we trained a Haar cascade classifier to segment facial areas in an image and trained a CNN model to detect human emotion. In the test stage, the Haar cascade classifier detected the presence of a human face in an image and segmented the facial region for the CNN model. Subsequently, the CNN model yielded emotional results based on facial features. Furthermore, we provided a user interface (UI) for users to upload images they want to recognize.

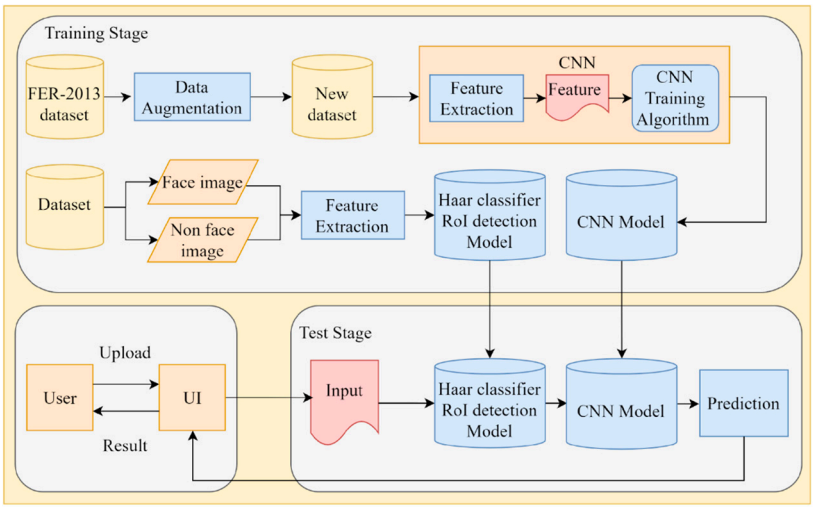

Figure 1. System framework of the proposed approach.

3.1. Data Augmentation

Data bias and data sparsity are challenges in deep learning training. Data bias and data sparsity can lead to inaccuracy, poor generalization, and model bias. Reducing the

impact of these issues has become a crucial concern in the field of deep learning. The causes of these issues include data imbalance, sampling selection bias, and data sparsity. In this study, we used data augmentation to alleviate the impact of data sparsity on deep learning models.

The analysis of the FER-2013 dataset revealed significant scarcity of the "disgust" emotion data compared to other emotion data. This attribute of disgust data led to difficulties in accurately predicting the "disgust" emotion by the model and reduced the overall model performance. The confusion matrix is illustrated in Figure 2. To address this issue, we employed two data augmentation techniques: random flipping and brightness adjustment. These approaches aimed to augment the FER-2013 dataset by increasing the availability of "disgust" emotion samples, thereby reducing the adverse effects of data sparsity.

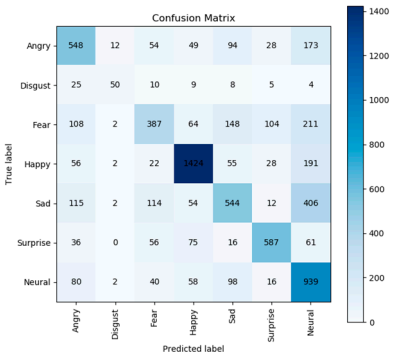

Figure 2. The recognition result using a confusion matrix with the original FER-2013 dataset.

3.1.1. Image Flipping

We adopted random flipping in our data augmentation. Horizontal flipping was implemented by the function library provided by TensorFlow and is illustrated in Figure 3.

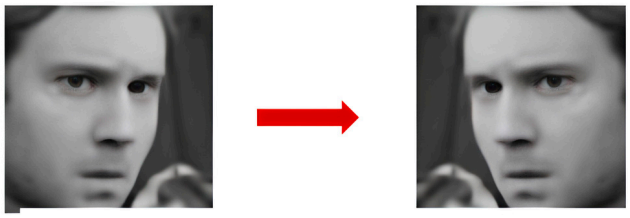

Figure 3. A horizontal flipping example.

3.1.2. Color Adjustment

We adopted the TensorFlow function library to implement color adjustment in our data augmentation. We added a random value to the image that was a basis for increasing or decreasing brightness. The example is illustrated in Figure 4.

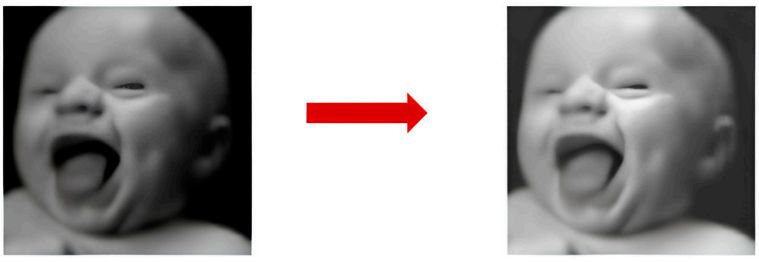

Figure 4. A color adjustment example.

3.2. Haar Cascade Classification

The Haar cascade classifier is a machine learning approach for object detection, especially for face detection. The Haar cascade classifier has the advantage of speed and high accuracy and is often used in biometric identification and security systems. In our system, the Haar cascade classifier was employed to detect the presence of faces and extract facial regions in images. Then, these facial region extractions were fed into the CNN model and trained. The Haar cascade classifier is a Haar wavelet with a robust classifier based on the AdaBoost algorithm. In our implementation, we utilized the Haar cascade classifier from OpenCV (Open Source Computer Vision Library) as the foundational model for our Haar cascade classifier. This model encompasses the Haar-like feature, Integral Image, and AdaBoost classifier modules. The Haar cascade classifier architecture is illustrated in Figure 5.

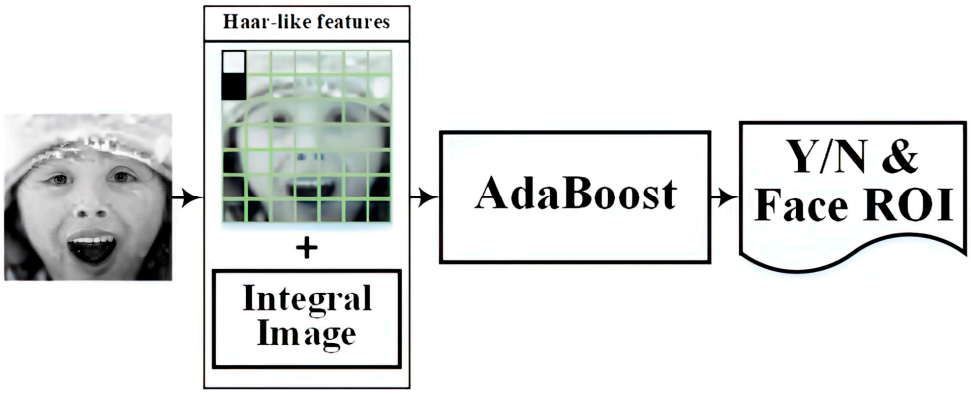

Figure 5. Haar cascade classification dataflow.

3.2.1. Haar-like Features and Integral Image

Haar-like features consist of multiple sets of black and white rectangular masks that slide over the image and calculate the pixel sum between different pixels to extract the grained visual feature. The mask of Haar-like features consists of the white rectangular and black rectangular. The white rectangular represents areas with bright or highly reflective areas of an object. Conversely, the black rectangular represents areas with dark or shadowed areas of an object. An example of Haar-like features is illustrated in Figures 6 and 7. A feature vector can combine with multiple Haar-like features and be used to train a classifier.

Figure 6. Example of Haar-like features proposed in [30].

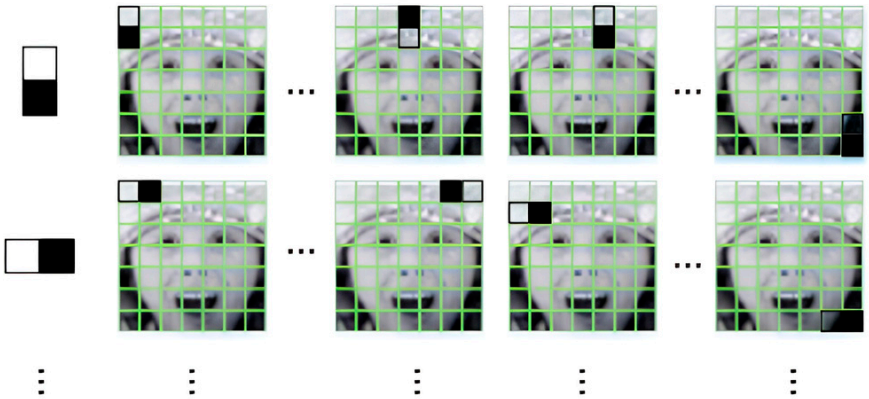

Figure 7. Haar-like feature extraction diagram used in the proposed approach.

However, Haar-like spends too much time calculating the pixel value sum. Therefore, an integral image was designed to accelerate the calculation of pixel value sums within rectangular regions. Equation (1) defines the feature value. An integral image is defined as a matrix of the same size as the image, where each pixel value represents the cumulative sum of pixel values within a rectangular region from the top-left corner to that pixel. Thus, Equations (2) and (3) can be employed to compute the sum of pixel values within a rectangular area, eliminating redundant computations and achieving time savings.

$$featureValue(I) = w_{white} \times \sum_{Pixel \in white} Pixel - w_{black} \times \sum_{Pixel \in black} Pixel \quad (1)$$

$$Integral(I_x, I_y) = \sum_{x=0}^{x} \sum_{y=0}^{y} Pixel(x, y) \quad (2)$$

$$rect = Integral(x + w, y + h) + Integral(x, y) - Integral(x + w, y) - Integral(x, y + h) \quad (3)$$

where *Pixel* represents the pixel value in position (*x*,*y*) of the image. While Haar-like features are a simple feature extraction method, they are widely recognized for their performance and computational efficiency in applications such as facial detection.

3.2.2. AdaBoost (Adaptive Boosting)

AdaBoost was further used to choose the useful Haar-like features and train the classifier. AdaBoost is a machine learning technique that creates a strong classifier by combining multiple weak classifiers. In our system, we employed various masks of Haar-like features paired with thresholds as individual weak classifiers. Subsequently, AdaBoost selected the useful weak classifiers and combined them to form a strong classifier. Finally, we adopted the Haar cascade classifier approach to integrate the strong classifiers into a multi-stage classifier. An advantage of this approach was that each stage underwent threshold-based detection, ensuring recall and reducing false-positive rates. The approach is illustrated in Figure 8.

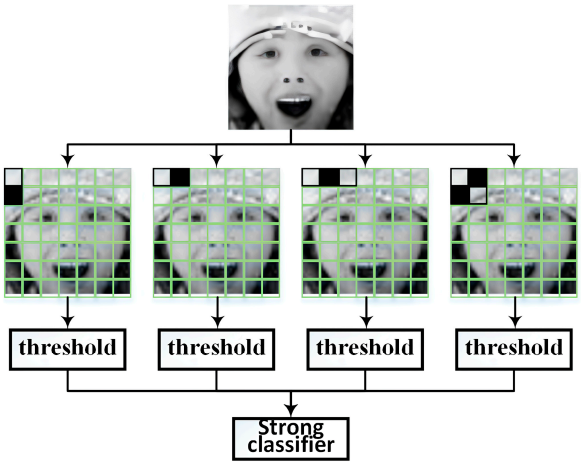

Figure 8. AdaBoost diagram for Haar-like feature selection.

3.3. Convolutional Neural Network (CNN)

A convolutional neural network (CNN) was used as the backbone model in this study. The advantages of CNN include feature extraction ability, spatial hierarchy, and scale invariance. The convolutional layer extracts facial features automatically and trains the model based on facial features. The facial feature is a multiple spatial hierarchy including local features (e.g., ear, eyes) and global features. A CNN model can extract a spatial hierarchy by multiple convolution layers and pooling layers and allow the model to understand images better. Due to the different sizes in input images, our system must

possess the attribute of scale invariance. Multi-layers of convolutional kernels in the CNN with varying field sizes enable the processing of features at different scales, thereby achieving scale invariance. Because of the three attributes of a CNN, we ultimately selected the CNN model as the primary network for achieving facial expression recognition.

Our CNN model was composed of four convolution layers, four pooling layers, a flattened layer, and two fully connected layers. The overall architecture is shown in Figure 9. The filters slide the input image and emphasize object edges to realize the extraction effect of facial features in the convolution layer. The output example of the convolution layer is illustrated in Figure 10. We employed max pooling as the model's pooling method because it can preserve high-frequency features more effectively than average pooling can. Additionally, it reduces the sensitivity of the convolutional layers to edges, thus achieving the goals of feature preservation, dimension reduction, and a decrease in the number of learnable parameters. In the flattened layer, facial features are mapped into a one-dimensional vector and fed into the fully connected layer as input. The flattened layer diagram is illustrated in Figure 11. In the fully connected layer, features are integrated, and calculated probabilities correspond to seven classes of emotions. Lastly, expression recognition was conducted through *softmax*. Cross-entropy loss was computed to facilitate subsequent model training and evaluation. The *softmax* and cross-entropy functions were defined by Equations (4) and (5), respectively.

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \tag{4}$$

$$L = -\sum_i^C t_i log(f(s)_i) \tag{5}$$

Here, C represents the number of neurons in the fully connected layer, s_j represents the values of each neuron, and s_i denotes the value obtained for the i-th class.

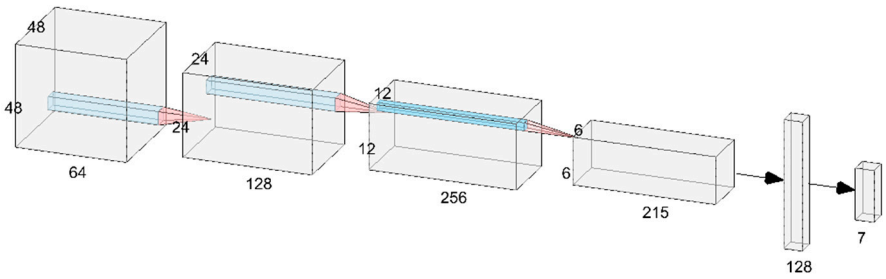

Figure 9. Architecture of the proposed CNN.

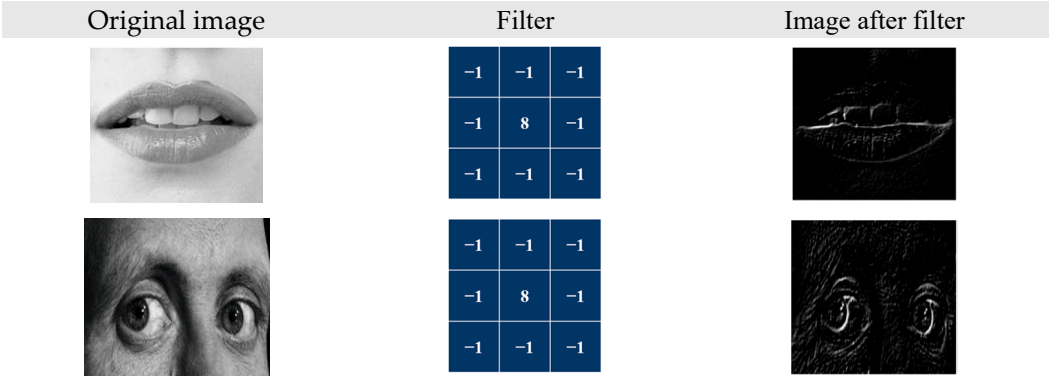

Figure 10. Convolution layer output.

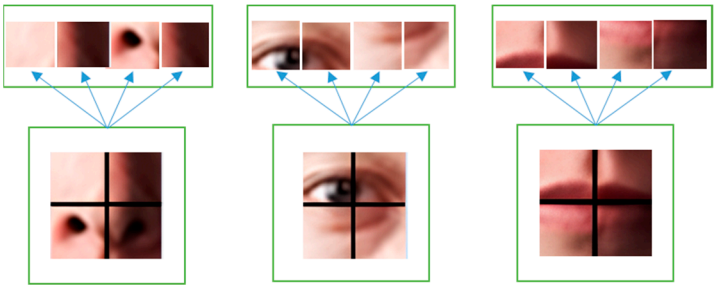

Figure 11. Flatten layer diagram.

4. Experimental Results

4.1. Experimental Setup

To evaluate the performance of the developed facial expression recognition system, we equiped a computer with an Intel Core i5-9500 processor and 16 GB of RAM as the hardware platform for conducting our experiments. The FER-2013 dataset was adopted to evaluate the proposed system. The FER-2013 dataset is a dataset designed for expression recognition. It comprises seven emotion classes, namely Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The FER-2013 dataset is divided into 28,709 images for training and 7178 images for testing (see Table 1). The size of the images in the FER-2013 dataset is 48 × 48 pixels. However, the FER-2013 dataset still faces the challenge of data sparsity. To address this issue, data augmentation techniques were employed to augment the FER-2013 dataset, mitigating the impact of data sparsity.

Table 1. Dataset distribution.

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Training	3995	436	4097	7215	4830	3171	4965	28,709
Testing	958	111	1024	1774	1247	831	1233	7178

The metrics in facial expression recognition tasks are accuracy and time complexity. The recognition result is divided into four parts: true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*). Accuracy is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{6}$$

We adopted the methodology proposed by Shah et al. [28] for calculating the time complexity of CNNs. Time complexity is estimated based on the input channel number, kernel size, number of filters, and length of output feature map for each layer of the CNN model, as shown in Equation (7). For the Haar cascade classifier, time complexity estimation was determined by the input image's size and the number of channels in the features, as shown in Equation (8).

$$complexity(CNN) \equiv \mathrm{O}(\sum_{n=1}^{d} k_{n-1} \cdot s_n^2 \cdot f_n \cdot l_n^2) \tag{7}$$

$$complexity(Haar) \equiv O(I_w \cdot I_h \cdot n) \tag{8}$$

where d is the depth of the convolutional layer, l_n is the length of the output feature map, f_n is the number of filters in the layer, S_n is the length of the filter, and k_{n-1} defines the number of input channels in the l layer. I_w is the width of the input image, I_h is the height of the input image, and n is the number of the feature dimension.

4.2. Filter and Haar Evaluation

Facial regions occupy a significant portion of the dataset's images. Therefore, to enhance the feature extraction capability of the CNN for large objects, we added a larger filter to our CNN model. Furthermore, we aimed to direct the CNN model's focus more towards the facial regions. To achieve this, we employed Haar cascade classifiers to segment the facial areas. The result is illustrated in Table 2.

Table 2. Different filter performance.

Model	Filter Composition	Accuracy
Baseline (CNN)	(64, 128, 256)	60.60%
CNN	(64, 128, 256, 512)	62.31%
CNN + Haar	(64, 128, 256, 512)	**63.67**%

The optimal results are highlighted in bold.

In Table 2, we first show the result of the adjustment composition of filters in the CNN. Filter composition was reconfigured to (64, 128, 256, 512), and accuracy improved by 1.71%. We posit that the addition of an extra convolutional layer and filter led to an improvement in accuracy. Subsequently, we evaluated the impact on the system by focusing on facial regions using the Haar cascade classifier. In terms of results, the Haar cascade classifier contributed to a 1.36% improvement in accuracy. We attribute this enhancement to the reduction of background noise, enabling the model to concentrate more on extracting facial features.

4.3. Data Augmentation Evaluation

Due to the limited amount of data provided by the FER-2013 dataset, we employed data augmentation techniques to expand the dataset. In data augmentation, flipping and color adjustment are among the most common methods; thus, we utilized these two methods as the foundation for data augmentation. The flipping method includes both horizontal and vertical flipping modes, although in practical applications, it is uncommon for human faces to be upside down. Obviously, vertical flipping was not an appropriate method for our system. Therefore, we only employed horizontal flipping for data augmentation in our experiments.

Table 3 shows the effectiveness of data augmentation on the FES-2013 dataset. In practice, the model with data augmentation exhibited improvements of 7.76% and 3.85% compared to the baseline model. We attribute this improvement to two factors. The first factor is the diversification of image extraction. Through data augmentation, the dataset incorporates a wider variety of facial expression images, providing the model with more opportunities to learn features. The second factor is the outcome of color adjustments. Because of saturation adjustment, the facial features within the images were highlighted, making it easier for the CNN model to extract useful features. However, excessive increase in saturation leads to blurring of facial features. In our data augmentation, we observed that some images initially had a lighter color tone, and color adjustments resulted in the blurring of facial features. This explains the phenomenon where using both methods of horizontal flipping and color adjustment led to a decrease in accuracy.

Table 3. Augmentation performance.

Model	Augmentation Method	Accuracy
Baseline (CNN)	-	60.60%
CNN + Haar	-	63.67%
CNN + Haar	Horizontal flipping	**68.36%**
CNN + Haar	Horizontal flipping + color adjustment	64.21%

"-" denotes the model without augmentation. The optimal results are highlighted in bold.

4.4. Optimization Evaluation

To determine the model's optimization, we trained the model multiple times, and the results are presented in Table 4.

Table 4. Epoch optimization result.

Epoch	Accuracy
30	68.36%
40	69.08%
50	69.13%
60	**69.47%**
75	69.14%

The optimal results are highlighted in bold.

From Table 4, with an increasing number of epochs, the model's accuracy gradually improved and reached its peak at epoch 60. However, we observed a decrease when the number of epochs reached 75, indicating a possible issue of overfitting. Consequently, we decided to choose the model at epoch 60 as our model. In Table 5, the results of our model in terms of other evaluation metrics are presented. The achieved metrics were as follows: accuracy of 69.47%, precision of 76.28%, recall of 69.45%, F1-score of 72.70%, and an approximate time complexity of 2.1×10^8. These findings contribute to a comprehensive assessment of our model's performance across various evaluation criteria.

Table 5. Other evaluation metrics with an optimization model.

Model	Accuracy	Precision	Recall	F1-Score	O(·)
CNN + Haar + Augmentation	69.47%	76.28%	69.45%	72.70%	$O(2.1 \times 10^8) + O(1.5 \times 10^5)$

4.5. Comparison with Other Models

In this section, we compare with the baseline (CNN) the models proposed by Agrawal [3] and Zhang [25]. Accuracy and time complexity were adopted as the evaluation metrics, and FER-2013 was used as the training and test data. First, we compared the baseline with our model. As shown in Table 6, while the addition of an extra convolutional layer and the Haar cascade classifier increased the time complexity of our model, it also resulted in a significant improvement of 8.87% in accuracy. Therefore, we considered this trade-off acceptable. This improvement was evident as the baseline extracted images that included the background, resulting in the inclusion of a considerable amount of noise in the features. Moreover, the limited size of the dataset did not provide sufficient support for the model to converge completely, which led to the model's inferior performance. The two models proposed by Agrawal achieved higher accuracy compared to that of the original CNN by adjusting the composition of layers and hyperparameters. However, blindly deepening the model sometimes failed to extract useful facial features. Additionally, these models tended to capture noise features from the background during feature extraction, resulting in a 3.7% lower accuracy compared to that of our model. Additionally, while Agrawal achieved a lightweight model in Model 2 by modifying the number of layers and hyperparameters, this resulted in a reduction in accuracy. Furthermore, in terms of time complexity, their approach slightly lags behind our system's overall efficiency. We also compared the model proposed by Minaee et al. [3] with our model. Minaee et al. proposed a two-branch convolution network system. A branch was defined as an attention branch. The attention branch paid more attention to emotionally relevant facial regions of the images than our Haar cascade classifier did. Therefore, they achieved higher accuracy than we did, and their two branches significantly reduced the time cost incurred by the

deep model. Although our model fell behind Minaee's model in terms of both accuracy and time complexity, our model successfully achieved real-time multi-person expression recognition. Real-time multi-person emotion recognition remains essential for applications such as monitoring classroom attentiveness. Despite the trade-offs, our model addresses the necessity of real-time multi-person emotion recognition in specific applications.

Table 6. Experimental results in comparison with those of other models.

Model	Accuracy	Real-Time Multiple Faces	O(·)
Baseline (CNN)	60.60%	-	$O(2.1 \times 10^8)$
Agrawal et al. [3] model 1	65.77%	-	$O(3.8 \times 10^9)$
Agrawal et al. [3] model 2	65.23%	-	$O(3.7 \times 10^8)$
Minaee et al. [9]	70.02%	-	$O(2.7 \times 10^5)$
Our (CNN + Haar)	69.47%	+	$O(2.1 \times 10^8) + O(1.5 \times 10^5)$

"-" and "+" denote the model without and with real-time multiple faces, respectively.

4.6. Multi-Person Expression Recognition

Since there is no formal dataset available for multi-person expression recognition, we utilized randomly collected multi-person images from the internet as test data. The results of the actual experiments are presented in Figure 12. The system detects facial regions and recognizes the emotions displayed on their faces, which are displayed above the bounding boxes. In this experiment, 10 images were selected as the test dataset, containing a total of 102 people. Among these, our system detected 75 people with an accuracy of 73.53%. It correctly identified the emotions of 57 people with an accuracy of 55.88%.

Figure 12. An example of the proposed model for multi-person expression recognition.

5. Conclusions

In this study, we proposed a multi-person facial expression recognition system. Our approach aimed to improve the limitation of sparse data and the restriction to single-person facial images. Leveraging the attribute of Haar cascade classifiers, we identified facial regions while simultaneously focusing on the face during subsequent model training. This attention to facial regions rather than the background aided in more precise feature extraction. Additionally, the multi-face detection capability of the Haar cascade classifiers helped us overcome the constraint of single-person facial images. Subsequently, we employed a convolutional neural network (CNN) as the backbone network to implement facial feature extraction and create an efficient classification system. Furthermore, we implemented experiments to optimize the model and confirmed that the model achieved its 69.47% accuracy at epoch 60 when using data augmentation with horizontal flipping. When compared to other methods, our approach outperformed most of them, which demonstrated the effectiveness of our method. In a multi-person expression recognition experiment, the proposed approach achieved an accuracy of 55.88%.

However, our system has its limitations. Firstly, to achieve higher accuracy, we employed methods with lower time complexity, such as the Haar cascade classifier, but we did not reduce the time complexity by altering the architecture of the CNN model. As a result, compared to the approach of modifying the CNN model by Minaee et al. [9], we slightly lagged in both accuracy and time complexity. In future work, we want to incorporate the concept of attention, allowing the model to focus more on learning emotion-related features. We plan to expand the training dataset for multi-person expression recognition and develop more algorithms tailored specifically for this task to achieve higher accuracy.

Author Contributions: All authors contributed to the design and implementation of the research, to the analysis of the state of the art, and the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: National Science and Technology Council: 111-2221-E-415-012-MY3.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://paperswithcode.com/dataset/fer2013; accessed on 26 November 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ashlin Deepa, R.; Sai Rakesh Reddy, D.; Milind, K.; Vijayalata, Y.; Rahul, K. Drowsiness Detection Using IoT and Facial Expression. In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Hyderabad, India, 11–13 December 2021; Springer: Berlin/Heidelberg, Germany, 2023; Volume 2, pp. 679–692.
2. Yang, B.; Cao, J.; Ni, R.; Zhang, Y. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access* **2017**, *6*, 4630–4640. [CrossRef]
3. Agrawal, A.; Mittal, N. Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy. *Vis. Comput.* **2020**, *36*, 405–412. [CrossRef]
4. Mehendale, N. Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci.* **2020**, *2*, 446. [CrossRef]
5. Khattak, A.; Asghar, M.Z.; Ali, M.; Batool, U. An efficient deep learning technique for facial emotion recognition. *Multimed. Tools Appl.* **2022**, *81*, 1649–1683. [CrossRef]
6. Zeng, N.; Zhang, H.; Song, B.; Liu, W.; Li, Y.; Dobaie, A.M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* **2018**, *273*, 643–649. [CrossRef]
7. Cheng, Y.P.; Wang, Y.; Tseng, F.H. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. In Proceedings of the Conference on Engineering, Technological and STEM Education ETS 2021, Porto, Portugal, 7–8 October 2021; pp. 36–51.
8. Xie, S.; Hu, H.; Wu, Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognit.* **2019**, *92*, 177–191. [CrossRef]
9. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef]
10. González-Lozoya, S.M.; de la Calleja, J.; Pellegrin, L.; Escalante, H.J.; Medina, M.A.; Benitez-Ruiz, A. Recognition of facial expressions based on CNN features. *Multimed. Tools Appl.* **2020**, *79*, 13987–14007. [CrossRef]
11. Gao, Y.; Cai, Y.; Bi, X.; Li, B.; Li, S.; Zheng, W. Cross-Domain Facial Expression Recognition through Reliable Global–Local Representation Learning and Dynamic Label Weighting. *Electronics* **2023**, *12*, 4553. [CrossRef]
12. Tang, C.; Zhang, D.; Tian, Q. Convolutional Neural Network–Bidirectional Gated Recurrent Unit Facial Expression Recognition Method Fused with Attention Mechanism. *Appl. Sci.* **2023**, *13*, 12418. [CrossRef]
13. Lin, H.Y. Applying Convolutional Neural Networks and Transfer Learning to Classify PCB Defects. Available online: https://hdl.handle.net/11296/695vqe (accessed on 19 October 2023).
14. Liu, Y.J. The Study on Recognizing Learning Emotion Based on Image Enhancement Combined with Convolutional Neural Network. Available online: https://hdl.handle.net/11296/vq434s (accessed on 19 October 2023).
15. Porcu, S.; Floris, A.; Atzori, L. Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics* **2020**, *9*, 1892. [CrossRef]
16. Umer, S.; Rout, R.; Pero, C. Facial expression recognition with trade-offs between data augmentation and deep learning features. *J Ambient. Intell. Humaniz.Comput.* **2021**, *13*, 721–735. [CrossRef]
17. Psaroudakis, A.; Kollias, D. Mixaugment & mixup: Augmentation methods for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 2367–2375.

18. Bobojanov, S.; Kim, B.M.; Arabboev, M.; Begmatov, S. Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets. *Appl. Sci.* **2023**, *13*, 12271. [CrossRef]
19. Goyani, M.M.; Patel, N.M. Evaluation of various classifiers for expression recognition using multi level Haar features. *Int. J. Next Gener. Comput.* **2018**, *9*, 131–150.
20. Shen, J.F.; Shi, S.W.; Zuo, X.; Xu, D. Multi View Face Detection Based on Multi-channel Discriminative Projection HAAR Features. *Data Acquis. Process.* **2018**, *33*, 270–279.
21. Liu, Y.X.; Zhu, Y.; Sun, J.B.; Wang, Y.B. Improved Adaboost face detection algorithm based on Haar-like feature statistics. *Image Graph.* **2020**, *25*, 1618–1626.
22. Mishra, A. An Artificial Neural Network-based Security Model for Face Recognition Utilizing HAAR Classifier Technique. *Int. J. Adv. Res. Comput. Sci.* **2023**, *14*, 8–16. [CrossRef]
23. Guo, W.; Xu, G.; Liu, B.; Wang, Y. Hyperspectral image classification using CNN-enhanced multi-level haar wavelet features fusion network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6008805. [CrossRef]
24. Wu, H.; Cao, Y.; Wei, H.; Tian, Z. Face recognition based on Haar like and Euclidean distance. *J. Phys. Conf. Ser.* **2021**, *1813*, 012036. [CrossRef]
25. Zhang, C.; Liu, G.; Zhu, X.; Cai, H. Face detection algorithm based on improved AdaBoost and new haar features. In Proceedings of the 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou, China, 19–21 October 2019; IEEE: New York, NY, USA, 2019; pp. 1–5.
26. Farkhod, A.; Abdusalomov, A.B.; Mukhiddinov, M.; Cho, Y.I. Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces. *Sensors* **2022**, *22*, 8704. [CrossRef]
27. Chinimilli, B.T.; Anjali, T.; Kotturi, A.; Kaipu, V.R.; Mandapati, J.V. Face recognition based attendance system using Haar cascade and local binary pattern histogram algorithm. In Proceedings of the 2020 4th international conference on trends in electronics and informatics (ICOEI), Tirunelveli, India, 16–18 April 2020; pp. 701–704.
28. Shafique, S.; Khan, I.A.; Shah, S.; Jadoon, W.; Jadoon, R.N.; ElAffendi, M. Towards Automatic Detection of Social Anxiety Disorder via Gaze Interaction. *Appl. Sci.* **2022**, *12*, 12298. [CrossRef]
29. Takiddin, A.; Shaqfeh, M.; Boyaci, O.; Serpedin, E.; Stotland, M. Gauging facial abnormality using HAAR-cascade object detector. In Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, 11–15 July 2022; pp. 1448–1451.
30. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; IEEE: Washington, DC, USA, 2001; Volume 1, p. 1.