IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 June 2023, accepted 30 June 2023, date of publication 10 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294099

APPLIED RESEARCH

CNN Learning Strategy for Recognizing Facial Expressions

DONG-HWAN LEE AND JANG-HEE YOO, (Senior Member, IEEE)
Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea
Artificial Intelligence Major, University of Science and Technology (UST), Daejeon 34129, South Korea

Corresponding author: Jang-Hee Yoo (jhy@etri.re.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government (MSIT) through the Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders Based on Cognition of the Psychological Behavior and Response under Grant 2019-0-00330.

ABSTRACT The ability to recognize facial expressions using computer vision is a crucial task that has numerous potential applications. Although deep neural networks have achieved high performance, their use in the recognition of facial expressions is still challenging. This is because different facial expressions have varying degrees of similarities among themselves, and numerous variations cause diversity in the same facial images. In this study, we propose a novel divide-and-conquer-based learning strategy to improve the performance of facial expression recognition (FER). The face area in an image was detected using MobileNet, and a ResNet-18 model was employed as a backbone deep neural network for recognizing facial expressions. Subsequently, groups containing similar facial expressions were categorized by analyzing the confusion matrix, which represents the inference results of the trained ResNet-18 model, and these similar facial expression groups were then utilized to re-train the deep learning model. In the experiments, the proposed method was evaluated using two thermal (Tufts and RWTH) and two RGB (RAF and FER2013) datasets. The results demonstrate improved FER performance, with an accuracy of 97.75% for Tufts, 86.11% for RWTH, 90.81% for RAF, and 77.83% for FER2013. As such, the proposed method can accurately classify large amounts of facial expression data.

INDEX TERMS Convolutional neural network, divide-and-conquer, facial expression recognition, learning strategy.

I. INTRODUCTION

Facial expression is one of the key elements for conveying human emotions and is used in various fields such as human-computer interaction [1], [2], [3], [4], advertising [5], [6], education [7], [8], and healthcare [9], [10]. Moreover, facial expression recognition (FER) using RGB images collected in a laboratory environment achieves high performance. Several recent studies have used RGB datasets, including large-scale and in-the-wild images. However, to improve FER performance, variations that are unrelated to facial expressions must be addressed [11]. Additionally, thermal infrared sensors and thermal imaging systems have been rapidly developed to prevent the spread of COVID-19. Thermal images are illumination-invariant and contain temperature information from emotional changes in the forehead, cheeks, nose, and maxillary areas [12].

There are two major approaches to FER techniques in computer vision. The first approach performs recognition using handcrafted feature vectors [13], [14], [15], whereas the other is an end-to-end approach that performs recognition using automatically extracted features based on deep neural networks [16], [17], [18], [19]. Shan et al. [13] compared the time and memory costs of the feature extraction process between a local binary pattern (LBP) and a Gabor filter, both being handcrafted feature-based methods. They showed that the LBP-based method could decrease the time and memory costs by approximately 10 times compared to when using the Gabor filter. Liu et al. [14] used a fusion feature including an LBP and a histogram of oriented gradient

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

(HOG) extracted from salient areas in a face. They utilized principal component analysis to reduce the feature dimensions and classified six facial expressions using a support vector machine (SVM). For the extended Cohn–Kanade (CK+) dataset [20], the fusion method yielded an accuracy of 98.3%. Kopaczka et al. [15] introduced an RWTH dataset that included high-resolution thermal-facial images and performed FER using several feature descriptors and classifiers on the RWTH dataset. The highest accuracy of 75.43% for the four facial expressions was obtained using the dense scale-invariant feature transform descriptor and SVM classifier. They also classified eight facial expressions using the HOG and SVM methods and achieved the highest accuracy of 46.7%.

In deep neural networks-based approaches, Tang [16] proposed a structure combining a convolutional neural network (CNN) and SVM. The proposed model achieved the highest recognition rate of 71.2% in the FER2013 competition [21]. Vignesh et al. [17] proposed a new segmentation block that was utilized to extract additional informative feature maps and obtained a classification accuracy of 75.97% on the FER2013 dataset. The residual masking network proposed by Pham et al. [18] consists of two main components: a residual layer and a masking block. The residual layer is used to make feature maps, and the masking block is used to produce activation maps of the same size as the feature maps. Subsequently, the maps that are yielded through the two components are refined. Compared to a previous study [16], they showed that their model could improve the accuracy by 2.98%, and the ensemble of six models yielded the highest accuracy of 76.82% on the FER2013 dataset. Kamath et al. [19] proposed TERNet, which is a custom CNN model for thermal FER. They utilized three image preprocessing techniques, such as adaptive histogram equalization, guided filtering, and intensity adjustment, to increase the amount of data. For the Tufts dataset [22], this model achieved the highest accuracy of 96.2% for the four classes of facial expressions.

On the other hand, several CNN learning strategies have been studied to improve the generalization performance of deep neural network models [23], [24], reduce computational costs [25], [26], and address class-imbalance problems [27], [28]. Wu et al. [23] employed dropout [29] to demonstrate its effect on convolutional and max-pooling layers. They showed that the best result could be obtained by simultaneously applying dropout in the max-pooling and fully connected layers. This model shows a state-of-the-art error rate of 0.39% on the MNIST dataset [30]. Zhang et al. [24] proposed a FER method for improving performance by handling noisy samples in the CNN training. They achieved a high classification accuracy of 90.35% on the RAF dataset [31], which contained seven facial expression classes. Recently, Xue et al. [32] proposed the APViT model that utilizes feature maps extracted by a CNN extractor and two attentive pooling modules. They achieved a state-of-the-art accuracy of 91.98% on the RAF dataset.

In addition, CNN parameter compression is important for maximizing the limited hardware resources in embedded systems. Han et al. [25] presented a pruning method that reduced memory and computation costs by eliminating the unimportant weights of the CNN model. This method achieved parameter compression rates of 9 to 13% in the LeNet [30], AlexNet [33], and VGG [34] models without loss of generalization performance. Han et al. [26] also proposed a deep compression method to remove parameters from a CNN model. The pruned network is compressed by the quantization and weight-sharing stages. Afterward, the frequently appearing quantized weights are assigned fewer bits during the Huffman coding stage. The proposed method was applied to three models (LeNet, AlexNet, and VGG) and achieved compression rates in the range of 35 to 40% with no loss of accuracy.

Also, the class imbalance is an important problem because the CNN model tends to yield a high generalization performance only for the majority of classes in the dataset. Hensman and Masko [27] analyzed the training effect of a CNN model with an imbalanced data. They applied a random distribution to the CIFAR-10 dataset [35] to generate 10 imbalanced sets; oversampling was thereafter performed to create balanced sets. The performance difference between the models trained on balanced and imbalanced datasets was approximately 63%. Khan et al. [28] proposed a method for simultaneously optimizing the class-dependent costs and CNN parameters through training. The performance of this method was compared with those of various sampling methods, such as over-, under-, and hybrid sampling. They achieved the best performance of 98.6% on the MNIST dataset, which comprised unbalanced class distributions.

Despite numerous studies aimed at improving the CNN performance, training the CNN model with data collected under various conditions remains a challenge in several domains, such as object detection [36], activity recognition [37], and visual speech recognition [38]. Similarly, FER has been disturbed by inter-class similarity and intra-class differences [39]. For instance, fear and surprise confuse FER because they present similar changes in the areas of the eyes and mouth. Furthermore, images of the same facial expression may be difficult to classify because the extraction of general features is hindered by illumination changes, head-pose variance, and occlusion. Therefore, these problems should be solved to achieve successful FER results. Accordingly, this study aims to solve the problem of performance degradation due to facial expression similarity.

In this study, we propose a new divide-and-conquer CNN learning strategy to increase the FER accuracy by minimizing the intra-class distance and maximizing the inter-class distance. To achieve this, the facial area in the original input image was first detected and adjusted to the same resolution. After optimizing the ResNet-18 [40] model, similar facial expressions were grouped into the same class to reduce classification difficulty. Finally, the final facial expressions of the groups were re-trained by the optimized ResNet-18 model and tested. The experimental results demonstrate that, compared to previous CNN learning, the proposed learning strategy can improve the accuracy of both thermal and RGB datasets.

II. METHODOLOGY

An overall flow diagram of the FER process is shown in Fig. 1. The proposed FER scheme consists of four stages: i) preprocessing for face detection and normalization; ii) CNN model optimization; iii) grouping for similar facial expressions using confusion matrix analysis, and iv) classification of the categorized similar facial expression groups into final facial expressions. Details of the techniques used in the FER process are discussed below.

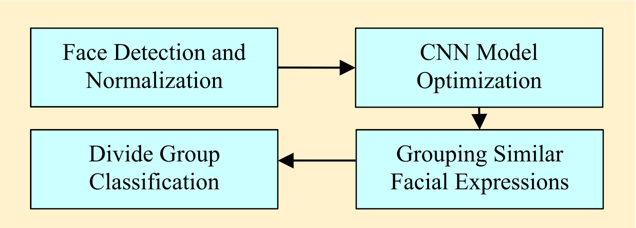

Face Detection and Normalization
CNN Model Optimization
Grouping Similar Facial Expressions
Divide Group Classification

FIGURE 1. Flow diagram of the FER process.

A. FACE DETECTION AND NORMALIZATION

Most facial expression datasets include facial and non-facial areas, and the non-facial areas probably not be an essential region [41]. Therefore, face detection is necessary to remove the non-essential areas from the input image. In this study, RetinaFace [42] was employed as the face detector that demonstrates promising accuracy and high-speed performance on RGB datasets, including various head poses, blurs, and illumination variations. RetinaFace is also expected to have excellent performance on grayscale thermal images. In cases where the facial area was not automatically detected, the facial images were manually cropped. After detecting the facial area, all images were normalized to the same resolution because detected facial regions differ from person to person. Moreover, data augmentation was adapted to prevent a decrease in the generalization performance of untrained data [43]. Augmented data can be generated by applying methods such as rotation, flipping, and color-space transformation to the images.

B. CNN MODEL OPTIMIZATION

Training the CNN model from scratch using a small amount of data is challenging [44]. The initial weights of the ResNet-18 model trained on ImageNet [45], including RGB images with 224 × 224 pixels, were used. To improve the FER performance, the layers of the pre-trained model were modified. First, the output size of the input convolution layer was maintained, and the filter size was decreased. This is because the features were extracted from the facial images with spatial resolutions smaller than that of ImageNet. Second, the second layer was converted from max-pooling to convolution to extract more features through training. Finally, the number of output nodes in the fully connected layer coincided with that of facial expressions to be classified. The details of the optimized CNN model used in this study are presented in Table 1.

TABLE 1. Optimized CNN architecture used in this study.

Type	Filter	Channel	Stride	Output size
Input image	-	1	-	130×160
Conv1	3×3	64	1	130×160
Conv2	3×3	64	2	65×80
Conv3_×	3×3	64	1	65×80
	3×3	64	1	65×80
	3×3	64	1	65×80
	3×3	64	1	65×80
Conv4_×	3×3	128	2	33×40
	3×3	128	1	33×40
	3×3	128	1	33×40
	3×3	128	1	33×40
Conv5_×	3×3	256	2	17×20
	3×3	256	1	17×20
	3×3	256	1	17×20
	3×3	256	1	17×20
Conv6_×	3×3	512	2	9×10
	3×3	512	1	9×10
	3×3	512	1	9×10
	3×3	512	1	9×10
Average pool.	-	512	-	1×1
Fc layer	-	512	-	Expressions

C. DIVIDE-AND-CONQUER LEARNING STRATEGY

A common CNN learning method is to utilize labeled data to train and test the neural networks. Solving the FER problem using all labeled data is difficult because of the similarities and variations in the facial expression images. The divide-and-conquer algorithm separates a given problem into two or more subproblems of the same or related type until these become adequately enough to be solved directly [46], [47], [48]. Thereafter, it solves the subproblems and combines the solutions into the original problem [46], [48]. The overall process of the proposed method is illustrated in Fig. 2. The training of CNN models with reduced-complexity subgroups is expected to improve classification accuracy. The proposed algorithm consists of two main phases.

*Phase 1: Grouping similar facial expressions.*First, a confusion matrix was generated by the inference result of training the optimized CNN model using all training datasets. In the confusion matrix, false positives (FPs) and false negatives (FNs) represent the proportion of incorrectly predicted test images. Performance improvement can be expected through the grouping of facial expressions that show a high error rate in the matrix. The confusion matrix analysis process is illustrated in Fig. 3. Figure 3(a) shows an example confusion matrix of the classification results for a dataset ($= E$) with n facial expressions. As shown in Fig. 3(b), the confusion

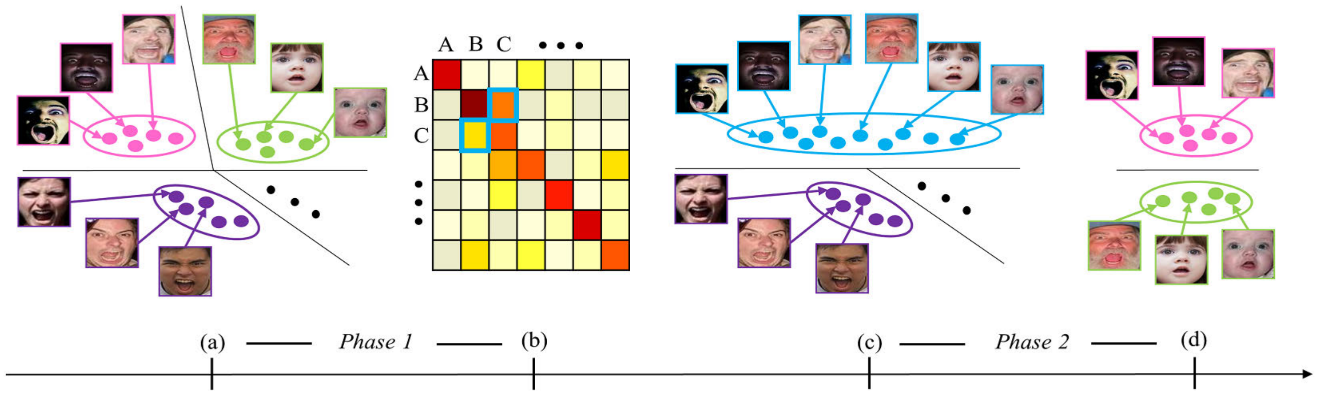


FIGURE 2. Overall process of the CNN learning strategy. Facial expression images have high inter-class similarity and high intra-class difference. In *Phase*1, similar facial expressions are grouped by confusion matrix analysis. In *Phase*2, the final facial expressions of the groups are trained by the CNN model and tested.

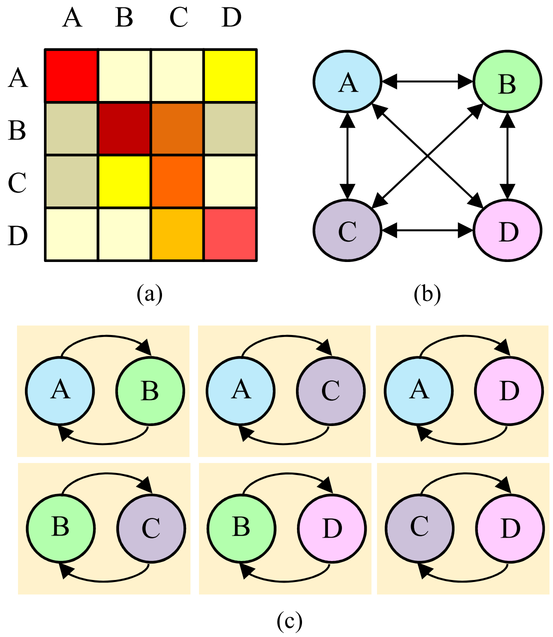


FIGURE 3. Process of the confusion matrix analysis. (a) the confusion matrix can be generated by training four facial expressions, (b) the confusion matrix can be converted to the weighted directed graph, and (c) the sum of weights in each graph is compared to find the most similar expressions.

```
Algorithm 1 Divide and Conquer based CNN Learning Strategy
 Input: a facial expression dataset $E$
          $E \leftarrow [C_1, C_2, \ldots, C_n]$, $n$ = the label of facial expression
 Phase 1 - Grouping similar facial expressions
 Train dataset $E$ using the optimized CNN for creating the confusion matrix
 Analyze the confusion matrix for grouping similar facial expressions
     $S \leftarrow [S_1, S_2, \ldots, S_m]$, $m$ = the label of similar facial expression group
 Phase 2 - Classifying similar facial expression groups
 Train generated set $S$ using the optimized CNN
 $S_m \leftarrow [R_{1,F_1}, R_{2,F_2}, \ldots, R_{m,F_m}]$
 $F_m$ = the number of final facial expressions in $S_m$
 for all attribute $R_{m,F_m} \in S_m$ do
   while $F_m > 1$ do
     if $F_m > 2$ then
       if $F_m$ is even then
         Generate new facial expression group $S'_m$
         $S'_m \leftarrow [R_{1',F_m/2}, R_{2',F_m/2}]$
         $R_{m,F_m} \leftarrow S'_m$
         $F_m \leftarrow F_m/2$
       else
         Train final facial expression in $R_{m,F_m}$ using the optimized CNN
         $F_m \leftarrow F_m/2$
     else
       Train final facial expression in $R_{m,F_m}$ using the optimized CNN
       $F_m \leftarrow F_m/2$
   end
 end
```

FIGURE 4. Algorithm of the proposed CNN learning strategy. The optimized CNN model trains from scratch the entire layer whenever the number of facial expressions changes. This implies that the CNNs in phases 1 and 2 have a different number of nodes in the fully connected layer, and the parameters in all layers are updated during training.

matrix can be converted into a weighted directed graph [49]. The total number of cases required to be analyzed for grouping k facial expressions was ${}_nC_k$. Figure 3(c) shows all the groups to be analyzed, and the number of graphs is six when $n = 4$ and $k = 2$. Each graph G comprises k vertices and ${}_kP_2$ edges, and the sum of the weights was calculated. The facial expressions in G with the largest sum of weights, which is the error rate, were newly mapped into the same group. The set S in Fig. 4 consists of similar facial expression groups corresponding to the vertices in graph G and non-grouped facial expressions.

Phase 2: Classifying grouped similar facial expressions. This phase focuses on addressing the subproblems. The created set S was used to train another optimized CNN model. The total number of images between set S and the original set E is the same; however, the target labels that the CNN needs to classify are decreased to m. Also, each set S containing two or more final facial expressions F_m must be classified. For instance, the set can be separated into smaller subgroups of half the size when $F_m > 2$ and even. Except under this condition, all the final facial expressions in the set were classified. Finally, the classification results obtained in *Phase 2* were combined and compared with the results in *Phase 1*.

Figure 4 shows the proposed algorithm used in this study. The values of k can be manually set to $\lceil n/2 \rceil$. The goal of this method is to improve the accuracy by classifying a similar facial expression set with reduced complexity and the final facial expression of the set.

III. EXPERIMENTAL RESULTS

In the experiments, details of the thermal and RGB datasets used for evaluating the proposed method were described. Four datasets, Tufts, RWTH, RAF, and FER2013, containing images with all facial expressions labeled in advance, were used in the experiments. The experiments were conducted to establish that the proposed method can improve performance in terms of accuracy. A description of the grouped facial expressions for each dataset and comparison results with previous methods are also provided. In addition, we used PyTorch and OpenCV libraries to implement the optimized CNN model and apply data augmentation.

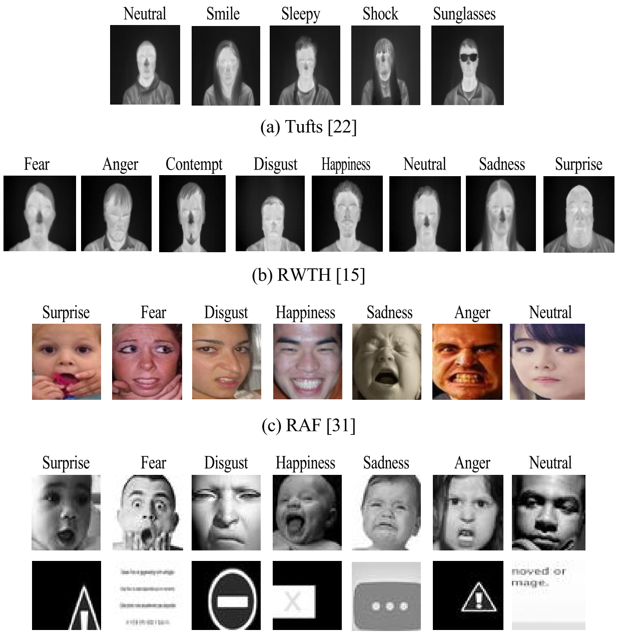
A. DATASETS

Two most recent public thermal datasets were selected for comparing the thermal FER results. Tufts and RWTH datasets containing frontal face images were collected using a long-wave infrared camera in a controlled environment. The two popular public RGB datasets, RAF and FER2013, contained images under various in-the-wild conditions collected by crawling the web. The Tufts face dataset [22] comprises multimodality face images from 113 subjects. Each thermal image for FER has a resolution of 336 × 256 and is labeled with one of five facial expressions: a neutral expression, a smile, eyes closed, an exaggerated shocked expression, and sunglasses. The RWTH dataset [15] contained only thermal images from 90 subjects; however, not all subjects participated in the process of filming facial expressions (fear: 70, anger: 77, contempt: 63, disgust: 75, happiness: 78, neutral: 77, sadness: 77, and surprise: 77 subjects). A total of 1,782 grayscale images had a high resolution of 1024 × 768 pixels.

The RAF dataset [31] provided 15,339 aligned facial images labeled with seven facial expressions: surprise, fear, disgust, happiness, sadness, anger, and neutral. All images were resized to 100 × 100 pixels and separated into 12,271 training and 3,068 test images. The FER2013 dataset [21] has 35,887 segmented face images with the same facial expression labels as the RAF dataset. Each grayscale image was normalized to 48 × 48 pixels and separated into 28,709 images for training and 3,589 images from the public and private test sets. A few original samples from each dataset are shown in Fig. 5, and a summary of the facial expression datasets is shown in Table 2.

TABLE 2. Summary of facial expression datasets.

Datasets	Res.	Subject	Expr.	Samples
Tufts [22]	336×256	113	5	565
RWTH [15]	1024×768	90	8	1,782
RAF [31]	100×100	-	7	15,339
FER2013 [21]	48×48	-	7	35,887



(a) Tufts [22]

(b) RWTH [15]

(c) RAF [31]

(d) FER2013 [21]

FIGURE 5. Sample images of facial expression datasets.

In the Tufts dataset, a few incorrectly labeled images were updated, and sunglass images were removed owing to a loss of temperature information. A total of 446 images were collected from 112 subjects. Among these, 22 subjects were selected as the test set. In addition, the dataset was separated into two sets: 1,074 for training and 88 for testing. The 892 images, with data augmentation by histogram equalization, were separated into two sets: 2,142 for training and 178 for testing. In the RWTH dataset, all images were split based on the largest number of 78 subjects in which facial expressions appeared. A total of 15 subjects who participated in filming all scenarios were selected to make a test set. The dataset, consisting of eight facial expressions, was separated into two sets: 4,266 for training and 360 for testing. In addition, four facial expressions with neutral, happy, sad, and surprise images were separated into two sets: 2,241 for training and 180 for testing. In the RAF dataset, 36,813 images were used for training and 3,068 images for inference.

In the FER2013 dataset, the images that did not represent facial expressions were identified. A total of 62 images were removed from the training set, and 7 images were removed from the private test sets. Consequently, a total of 85,941 images were used for training and 3,582 images for testing. A summary of the data used in the experiments is shown in Table 3.

B. EXPERIMENT AND ANALYSIS

All experiments were conducted in the order of the dataset listed in Table 3. The thermal datasets were divided into training and test sets at a ratio of 8:2. For all experiments, the detected facial images were normalized to 130 × 160 pixels and automatically augmented using horizontal flipping and random rotation techniques. In the experiment for the

TABLE 3. Summary of data used in experiments.

Datasets	Res.	Train	Test	Expr.
Tufts	130×160	1,074	88	4
		2,142	178	4
RWTH	130×160	4,266	360	8
		2,241	180	4
RAF	130×160	36,813	3,068	7
FER2013	130×160	85,941	3,582	7

Tufts dataset, histogram equalization was used for comparison with the previous result [19]. To train the CNN model, the Adam optimizer was used with a learning rate of 0.001, with the epoch set to 10^4, and the loss value was calculated using the L1 function. The comparative results for the accuracy of the proposed method and previous FER methods are summarized in Table 4. Additionally, Table 4 includes the learning results of the similar expression group in *Phase 1* and the final results achieved up to *Phase 2*.

TABLE 4. Comparative analysis on four FER datasets.

Datasets	Expr.	Methods	Accuracy (%).	
			Phase 1	Final
Tufts	4	Optimized model	-	92.05
		Proposed method	97.73 ⓐⓑ, ⓒ, ⓓ	**93.18**
	4	Kamath *et al.* [19]	-	96.20
		Optimized model	-	96.07
		Proposed method	98.31 ⓐⓑ, ⓒ, ⓓ	**97.75**
RWTH	8	Kopaczka *et al.* [15]	-	46.70
		Optimized model	-	54.72
		Proposed method	77.22 ⓐⓔⓕⓖ, ⓗ, ⓘ, ⓙ, ⓚ	**62.50**
	4	Kopaczka *et al.* [15]	-	75.43
		Optimized model	-	80.56
		Proposed method	96.11 ⓐⓖ, ⓘ, ⓙ	**86.11**
RAF	7	Zhang *et al.* [24]	-	90.35
		Optimized model	-	85.10
		Proposed method	94.36 ⓐⓖⓗⓙ, ⓔ, ⓘ, ⓚ	**90.81**
FER2013	7	Pham *et al.* [18]	-	76.82
		Optimized model	-	69.04
		Proposed method	90.01 ⓐⓔⓖⓚ, ⓗ, ⓘ, ⓙ	**77.83**

ⓐ: Neutral, ⓑ: Smile, ⓒ: Sleepy, ⓓ: Shock, ⓔ: Fear, ⓕ: Contempt, ⓖ: Sadness, ⓗ: Disgust, ⓘ: Surprise, ⓙ: Happiness, ⓚ: Anger

In the first experiment using the Tufts dataset, the optimized CNN model was evaluated using a relatively small number of samples. Compared to the commonly used cross-entropy loss, the accuracy was improved by 2.28%. This may be because the L1 loss was calculated by reflecting the prediction value corresponding to the classes that are not the correct answer. To confirm the accuracy of the proposed learning method, facial expressions similar to neutral and smiles were grouped. The proposed method achieved 97.73% and 90.91% accuracy in phases 1 and 2, respectively. Consequently, we achieved an accuracy of 93.18%, which is 1.13% better than training four facial expressions simultaneously as a single CNN model. In the second experiment, the performance of the proposed method was compared with the results presented by Kamath et al. [19]. Determining the optimal window size and threshold for image processing techniques used as the resolution changes are difficult. Therefore, only the histogram equalization technique, considering each pixel intensity of the entire image area, was applied to increase the data. The proposed method achieved the highest accuracy of 97.75%. This is because facial expression data of the same subject could exist in the training and test sets when separating the dataset based on the total number of images. Therefore, this method can reduce the accuracy when predicting the facial expressions of subjects not included in the dataset.

Kopaczka et al. [15] reported the classification accuracy for eight and four expressions with neutral, happiness, sadness, and surprise on the RWTH dataset. In our first experiment, among the eight facial expressions, the four facial expressions of fear, contempt, neutrality, and sadness were mapped to the same group. Then, the group was divided into two small subgroups: (fear and contempt) and (neutral and sadness). The classification accuracy of the facial expression group was 77.22% in Phase 1 and 63.16% and 72.22% in Phase 2. Second, among the four facial expressions, neutral and sadness were grouped to evaluate the method, and the accuracy of phases 1 and 2 was 96.11% and 80%, respectively. Consequently, the highest accuracies of 62.50% and 86.11% were achieved for eight and four facial expressions, respectively.

To show that the proposed method also improves the performance on large FER datasets, the experiments were conducted on two public RGB datasets (RAF and FER 2013). Through the analysis of the confusion matrix on the RAF dataset, the highest error rate of 74% were disgust, sadness, anger, and neutral, and these were classified as the first similar group. Two subgroups (disgust and sadness) and (anger and neutral) were then categorized, and 89.24% accuracy was obtained. The error rate of the second similar group that included four facial expressions (disgust, happiness, sadness, and neutral) was 71%, which was almost the same as that of the first group. Therefore, the second group was also used to evaluate the proposed method.

Figure 6 shows the recall confusion matrix on the RAF dataset. Similar facial expressions can be found in the confusion matrix by analyzing elements other than the diagonal components that signify the correct answer rate. By applying the proposed method, the participants were divided into two subgroups: (disgust and sadness) and (happiness and neutral). Consequently, the final learning result using the second facial expression group achieved 90.81% accuracy, which improved by 1.57% over the first group. This implies that not only the most similar facial expression groups can help improve accuracy, but other groups can also improve accuracy.

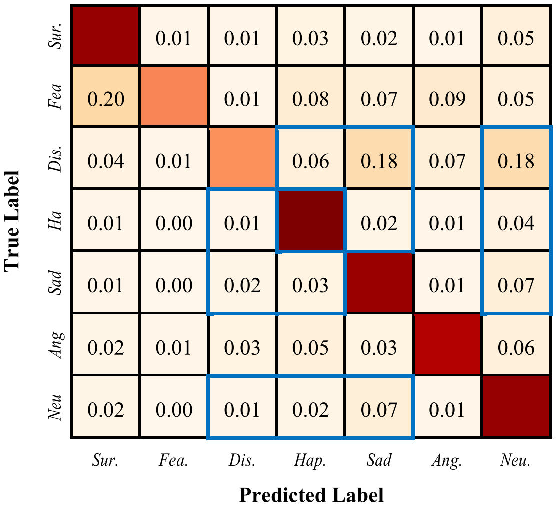


FIGURE 6. **Confusion matrix of the optimized CNN model on the RAF dataset. Blue squares: error rates of selected facial expressions by confusion matrix analysis result. These similar facial expressions belong to the same group.**

Finally, the proposed method was tested using the FER2013 dataset. The optimized CNN model achieved a low accuracy of 69.04% when classifying all facial expressions. The four most similar facial expressions (sadness, neutral, fear, and anger) were grouped, and two subgroups (sadness and neutral) and (fear and anger) were created. The facial expression groups achieved an accuracy of 90.01% in Phase 1 and 78.54% and 81.25% in Phase 2. Consequently, the proposed method achieved the highest accuracy of 77.83%. As such, the experimental results on the four FER datasets confirmed that owing to similar facial expressions and variations that are not related to facial expressions, performance improvement with the previous CNN learning method is difficult.

The proposed method can generalize facial expressions that are difficult to recognize because the CNN model utilizes a subgroup that is easier to train than the original data. The FER results with our approach achieved performance improvements of 1.55%, 7.78%, 0.46%, and 1.01% on the Tufts, RWTH, RAF, and FER2013 datasets, respectively. Hence, the proposed approach will efficiently solve problems arising from other types of data, as well as facial expressions. Although the proposed approach showed 1.17% lower accuracy than the Vision Transformer method [32], which is a state-of-the-art model on the RAF dataset, our approach showed that solving the problems by dividing similar classes into subgroups could contribute to improving recognition performance. Therefore, applying the proposed method to the Vision Transformer model in further studies will be necessary.

IV. CONCLUSION

We have described a divide-and-conquer-based CNN learning strategy that utilizes the grouping of similar data by analyzing the classification results for training and testing the CNN model. Facial regions have been detected and normalized to a consistent size to facilitate data assignment to the CNN model. The model was efficiently initialized using pre-trained weights and updated for FER. The proposed learning strategy, implemented with the newly designed model, was evaluated using the Tufts, RWTH, RAF, and FER2013 datasets. The experimental results demonstrate that the proposed method can improve classification accuracy in comparison to previous methods that discriminate all facial expressions simultaneously. Further, it is necessary to reduce the computational costs associated with solving subproblems; thus, developing an integrated CNN architecture that can solve subproblems simultaneously is an important task. In addition, more studies are required to identify the most similar facial expression group for the highest performance improvement. In the future, we expect the applications of the proposed method not only to Vision Transformer-based models, but also to utilization in other recognition tasks such as voice, action, object, and text.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human–computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] L. Yin, X. Wei, P. Longo, and A. Bhuvanesh, "Analyzing facial expressions using intensity-variant 3D data for human computer interaction," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Hong Kong, 2006, pp. 1248–1251.

[3] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, "Weight-adapted convolution neural network for facial expression recognition in human–robot interaction," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 3, pp. 1473–1484, Mar. 2021.

[4] M. D. Putro, D. Nguyen, and K. Jo, "A fast CPU real-time facial expression detector using sequential attention network for human–robot interaction," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7665–7674, Nov. 2022.

[5] P. Lewinski, M. L. Fransen, and E. S. H. Tan, "Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli," *J. Neurosci., Psychol., Econ.*, vol. 7, no. 1, pp. 1–14, Mar. 2014.

[6] N. Hamelin, O. E. Moujahid, and P. Thaichon, "Emotion and advertising effectiveness: A novel facial expression analysis approach," *J. Retailing Consum. Services*, vol. 36, pp. 103–111, May 2017.

[7] Y. Li and Y. Hung, "Feature fusion of face and body for engagement intensity detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 3312–3316.

[8] M. Dindar, S. Järvelä, S. Ahola, X. Huang, and G. Zhao, "Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1390–1400, Jul. 2022.

[9] Z. Jiang, S. Harati, A. Crowell, H. S. Mayberg, S. Nemati, and G. D. Clifford, "Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 2, pp. 664–672, Feb. 2021.

[10] C. Bisogni, A. Castiglione, S. Hossain, F. Narducci, and S. Umer, "Impact of deep learning approaches on facial expression recognition in healthcare industries," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5619–5627, Aug. 2022.

[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul./Sep. 2022.

[12] I. A. Cruz-Albarran, J. P. Benitez-Rangel, R. A. Osornio-Rios, and L. A. Morales-Hernandez, "Human emotions detection based on a smart-thermal system of thermographic images," *Infr. Phys. Technol.*, vol. 81, pp. 250–261, Mar. 2017.

[13] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[14] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors*, vol. 17, no. 4, p. 712, Mar. 2017.

[15] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof, "A thermal infrared face database with facial landmarks and emotion labels," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 5, pp. 1389–1401, May 2019.

[16] Y. Tang, "Deep learning using linear support vector machines," 2013, *arXiv:1306.0239*.

[17] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation VGG-19 architecture," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 1777–1787, Mar. 2023, doi: 10.1007/s41870-023-01184-z.

[18] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 4513–4519.

[19] S. K. M. Kamath, R. Rajendran, Q. Wan, K. Panetta, and S. Agaian, "TER-Net: A deep learning approach for thermal face emotion recognition," *Proc. SPIE*, vol. 10993, May 2019, Art. no. 1099309.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[21] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015.

[22] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.

[23] H. Wu and X. Gu, "Towards dropout training for convolutional neural networks," *Neural Netw.*, vol. 71, no. 1, pp. 1–10, Nov. 2015.

[24] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 418–434.

[25] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, Dec. 2015, pp. 1135–1143.

[26] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.

[27] D. Masko and P. Hensman, "The impact of imbalanced training data for convolutional neural networks," Bachelor thesis, School Comput. Sci. Commun., KTH, Chennai, India, 2015.

[28] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[30] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[31] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[32] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Trans. Affect. Comput.*, early access, Dec. 5, 2022, doi: 10.1109/TAFFC.2022.3226473.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[35] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., May 2009.

[36] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[37] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: A comprehensive survey," *IEEE Access*, vol. 8, pp. 83791–83820, 2020.

[38] C. Sheng, G. Kuang, L. Bai, C. Hou, Y. Guo, X. Xu, M. Pietikäinen, and L. Liu, "Deep learning for visual speech analysis: A survey," 2022, *arXiv:2205.10839*.

[39] D. Ruan, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Deep disturbance-disentangled learning for facial expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 2833–2841.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[41] L. Yan, Y. Shi, M. Wei, and Y. Wu, "Multi-feature fusing local directional ternary pattern for facial expressions signal recognition based on video communication system," *Alexandria Eng. J.*, vol. 63, pp. 307–320, Jan. 2023.

[42] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.

[43] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, Świnoušcie, Poland, May 2018, pp. 117–122.

[44] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[46] (2023). *Wikipedia, Divide-and-Conquer Algorithm*. [Online]. Available: https://en.wikipedia.org/wiki/Divide-and-conquer_algorithm

[47] D. R. Smith, "The design of divide and conquer algorithms, *Sci. Comput. Program.*, vol. 5, no. 1, pp. 37–58, Dec. 1985.

[48] D. Lee and J.-H. Yoo, "Divide and conquer strategy for CNN model in facial emotion recognition based on thermal images," *J. Softw. Assessment Valuation*, vol. 17, no. 2, pp. 1–10, Dec. 2021.

[49] W. E. Gilbraith, C. P. Celani, and K. S. Booksh, "Visualization of confusion matrices with network graphs," *J. Chemometrics*, vol. 37, no. 3, p. e3435, Mar. 2023.

DONG-HWAN LEE received the B.Sc. degree in computer engineering from Chungnam National University, Daejeon, South Korea, in 2020, and the M.Sc. degree from the University of Science and Technology (UST), Daejeon, in 2022, where he is currently pursuing the Ph.D. degree in artificial intelligence. His research interests include computer vision, deep learning, pattern recognition, and human–robot interaction.

JANG-HEE YOO (Senior Member, IEEE) received the B.Sc. degree in physics and the M.Sc. degree in computer science from the Hankuk University of Foreign Studies, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in electronics and computer science from the University of Southampton, U.K., in 2004. Since November 1989, he has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, as a Principal Researcher. He has also been a Professor with the Department of Artificial Intelligence, University of Science and Technology, South Korea. He was a Visiting Scientist with the University of Washington, Seattle, USA, from August 2014 to July 2015. His research interests include computer vision, human motion analysis, biometric systems, HCI, and intelligent robot. He is a member of the IEIE.