symmetry

MDPI

Article

HOG-ESRs Face Emotion Recognition Algorithm Based on HOG Feature and ESRs Method

Yuanchang Zhong [1,*], Lili Sun [1], Chenhao Ge [1] and Huilian Fan [2]

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400030, China; sll@cqu.edu.cn (L.S.); gch@cqu.edu.cn (C.G.)
[2] School of Big Data and Intelligent Engineering, Yangtze Normal University, Chongqing 408100, China; fanhl@yznu.edu.cn
* Correspondence: zyc@cqu.edu.cn

Abstract: As we all know, there are many ways to express emotions. Among them, facial emotion recognition, which is widely used in human–computer interaction, psychoanalysis of mental patients, multimedia retrieval, and other fields, is still a challenging task. At present, although convolutional neural network has achieved great success in face emotion recognition algorithms, it has a rising space in effective feature extraction and recognition accuracy. According to a large number of literature studies, histogram of oriented gradient (HOG) can effectively extract face features, and ensemble methods can effectively improve the accuracy and robustness of the algorithm. Therefore, this paper proposes a new algorithm, HOG-ESRs, which improves the traditional ensemble methods to the ensembles with shared representations (ESRs) method, effectively reducing the residual generalization error, and then combining HOG features with ESRs. The experimental results on the FER2013 dataset show that the new algorithm can not only effectively extract features and reduce the residual generalization error, but also improve the accuracy and robustness of the algorithm, the purpose of the study being achieved. The application of HOG-ESRs in facial emotion recognition is helpful to solve the symmetry of edge detection and the deficiency of related methods in an outdoor lighting environment.

Keywords: face emotion recognition; HOG; original pixel; ESRs

Citation: Zhong, Y.; Sun, L.; Ge, C.; Fan, H. HOG-ESRs Face Emotion Recognition Algorithm Based on HOG Feature and ESRs Method. *Symmetry* **2021**, *13*, 228. https://doi.org/10.3390/sym13020228

Academic Editor: Theodore E. Simos
Received: 10 January 2021
Accepted: 26 January 2021
Published: 30 January 2021

1. Introduction

As we all know, human beings communicate mainly through speech, and use body language to emphasize some parts of speech and express their emotions [1]. Facial expression is one of the most natural, powerful, and direct ways to express emotions and intentions in human communication. Of course, emotion can also be expressed through voice and text, among others, but face is the most popular [2]. In 1974, Mehrabadu [3] showed that about 50% of people in daily communication convey information through facial expressions, only about 40% of people communicate through voice and assist in face, and the remaining 10% express through words. The main reason is that the face contains many effective emotional features, and it has more advantages in data collection [3]. Around the 20th century, emotions were defined as seven states, namely, fear, happy, anger, disgust, surprise, sad, and normal. Studies have found that different emotional states are closely related to actions, such as gnashing teeth when angry, dancing when happy, and full of tears when sad. Therefore, most of the facial emotions directly express the emotional state at that time, and these seven states are widely used in face emotion recognition research at this stage.

In recent years, with the development of machine learning, computer vision, behavior science, and face emotion recognition is an interesting and challenging application, so it has become an important research field. Facial emotion recognition can be widely used in driver safety, medicine, human–computer interaction, and so on. In medicine, patients who have their own defects or psychological problems may not be able to express

their emotions normally in some cases. Therefore, facial emotion recognition technology can solve this problem and achieve effective communication [4]. In human–computer interaction, because Siri, Cortana, Alexia, and other IPAs (Intelligent Personal Assistant) can only use natural language to communicate with human beings, in order to improve effective communication, emotion recognition can be added. In terms of safety, facial emotion recognition can be used to identify the driver's emotion. Through non-invasive monitoring of the driver's emotional state, it can timely and effectively judge whether the driver should make dangerous behavior, so as to prevent the occurrence of dangerous events, or to monitor and predict the fatigue state and attention, so as to prevent the occurrence of accidents [5]. The application of facial emotion recognition technology in medical, human–computer interaction, monitoring driver's emotional state, and other real environment is of great significance for the treatment of special patients and the maintenance of traffic safety.

In addition, according to the in-depth study of facial emotion recognition, deep learning is an important part of facial emotion recognition, especially the use of convolutional neural network; through training massive data, many effective features can be extracted and learned, so as to improve the accuracy of face emotion recognition. It is found that most of the features of facial emotion come from the muscle movement of the face driven by eyes and mouth, while hair, ears, and other parts have little influence on facial emotion [4]. Therefore, in order to obtain ideal output results, the machine learning framework of face emotion recognition is not sensitive to other parts of the face, and only focuses on the important parts of the face. In recent years, with the deepening of human research and the rapid development of related disciplines, facial emotion recognition technology is still a research hotspot. However, there are still two problems to be solved in face emotion recognition algorithm. First of all, most feature extraction methods are still similar to the traditional manual feature extraction methods, which cannot effectively extract features. Secondly, because the emotion recognition algorithm cannot effectively reduce the residual generalization error, it seriously affects the accuracy and robustness of the algorithm.

In conclusion, according to the above two problems, although great progress and breakthrough have been made in recent years, the improvement and perfection of face emotion recognition algorithm is still a hot spot for many scholars in the future. For example, there is still a rising space in effective feature extraction and recognition accuracy [6]. It can be seen that histogram of oriented gradient (HOG) features can effectively extract face features, and ensemble methods can effectively improve the accuracy and robustness of the algorithm [7]. Therefore, this paper improves the traditional ensemble methods to the ensembles with shared representations (ESRs) method, effectively reducing the residual generalization error, and proposes an face emotion recognition algorithm based on HOG features and ensembles with shared representations (ESRs), namely HOG-ESRs. The experimental results on FER2013 dataset show that the new algorithm model can not only effectively extract features and reduce residual generalization error, but also improve the accuracy of the algorithm. Specifically, this paper makes the following three contributions:

1. Based on [8], an ensemble with shared representations method is proposed, four network branches are used, and each branch is based on the original pixel data features and HOG features;
2. The new algorithm model is not only based on the original pixel data features of the data set, HOG features are added in the last layer of each branch of the convolution layer. Finally, the extracted mixed feature set is sent to the FC (Fully Connected) layer for calculation;
3. According to the results of five and six convolution layers explored by CNN (Convolutional Neural Networks) in [9], the recognition accuracy is not improved. It is found that the model with four convolution layers and two FC layers is the optimal network model for the FER2013 dataset. Therefore, the CNN model with four convolution layers and two FC layers is used in the network branch of the HOG-ESRs model.

The organizational structure of this paper is as follows. The first part introduces the development of facial emotion recognition and related knowledge, and introduces the main contributions of this paper. In the next section, the research status is mainly introduced. In the third part, HOG features and ensembles with shared representations method are introduced in detail, and the model method proposed in this paper is introduced. In the fourth part, we first introduce several classical datasets and explain the selected datasets. Then, we describe the experiment and result analysis in detail. In the fifth part, the experimental results are discussed. The final section summarizes the paper and briefly introduces the idea of perfecting the model.

2. Related Works

In fact, the research history of facial emotion recognition is related to the history of emotion research. It is found that the research on emotion began in the 1970s. Therefore, the research on facial emotion recognition is fairly recent [8]. It is mainly limited by the development of new generation information technology in the 21st century. The details are as follows.

According to the study, the research on facial emotion recognition can be traced back to the 1970s [9]. At first, Paul Ekman, a famous international psychologist, studied the main emotions of human beings, and proposed six basic emotions: surprise, happiness, fear, anger, disgust, and sad [10]. A few years later, Paul Ekman et al. created FACS (facial action coding system), based on different facial expressions corresponding to different facial muscle movements [11]. The determination of FACS not only contributes to the researchers of facial expression muscle movement, but also lays the foundation for facial emotion recognition [12]. Subsequently, in 1978, the research on facial emotion video sequence was started, among which Suwa et al. carried out the first research [13]. A. Pentland et al. combined optical flow data with facial emotion recognition to estimate facial muscle movements, and achieved an accuracy rate of 80% in four expressions of happiness, anger, disgust, and surprise [14]. Shan et al. developed a boosted-LBP (local binary pattern) to extract the features of LBP, and achieved a better recognition effect [15]. Wan et al. proposed a method of locating facial feature points with ASM, and used it to identify continuous facial emotions [16]. Praseeda et al., mainly based on eyes, mouth, and other parts, using the PCA (Principal Component Analysis) method for recognition, also achieved better results [17]. In [18], a face recognition method using 68 kinds of markers was proposed based on face marker features. The system detects emotions based on 26 geometric features (13 differential features, 10 centrifugal features and 3 linear features) and 79 new features in [19]. The experimental results show that the average accuracy reaches 70.65%. Similarly, the work of [20], based on 20 marker features and 32 geometric facial features (centrifugal, linear, slope, and polygon), has been applied to automatic facial emotion recognition with great success. At the beginning of the research, most of the recognition methods are based on common methods to solve the face image preprocessing, geometric manual feature extraction, feature classification, and so on [21]. These conventional methods have an obvious effect in an indoor environment, but their performance decreases in a real environment [22].

With the development of information technology in the new era, in recent years, face emotion recognition with high recognition accuracy has been widely used in real-time systems in machine vision, behavior analysis, video games, and other fields. Therefore, human emotion expression is easy to be "understood" by an HMI (Hman Machine Interface) system [23–26]. With the development of computer vision, artificial intelligence, pattern recognition, and image processing technology, the shortcomings of traditional methods have been overcome. In particular, the use of deep neural network [27,28], on the one hand, enables the network to automatically learn image features and avoid the disadvantages of manual feature engineering; on the other hand, the learning of facial features is more extensive, such as brightness change, rotation change, and so on. Khorrami et al. [29] show that the learning features of a CNN network based on face emotion recognition

training are more consistent with the face features found by psychologist Paul Ekman [30] through general facial expression. Therefore, face emotion recognition has not only formed an independent research field, but also made outstanding achievements in the field of face emotion recognition. Hamester [31] and others proposed a framework based on a multi-channel convolution neural network. Two channels, unsupervised and supervised, were used to train the convolutional autoencoder (CAE) and multi convolution layer to extract implicit features of images, and the effect was far greater than that of the manual feature method. Hu [32] and others not only proposed the deep synthesis multi-channel aggregate convolution neural network, but also improved the transformation-invariant pooling (TI pooling) of Laptev [33] to the expression transformation-invariant pooling (ETI pooling). The experimental results show that the model has strong robustness and high accuracy.

However, with the limitations of static images becoming more and more prominent, researchers are interested in non-stationary facial behavior, 3D video, and stationary data recorded from different perspectives, and thus began to study the spatiotemporal motion characteristics of changing emotions in video sequences on static emotion recognition. Danelakis et al. proposed a 3D video recognition method based on a 3D video face emotion data set, and obtained better results [34]. According to the facial emotion dynamics and morphological changes, Zhang [35] and others proposed a part-based hierarchical bidirectional recurrent neural network (PHRNN) and multi signal convolution neural network; the former is used for face features in continuous sequences, while the latter is used to extract the spatial features of static images—its function is to achieve the complementary characteristics of space and time. The two networks greatly improve the performance of model recognition. Generally, the recognition network model based on video has high computational complexity. Li et al. [36] proposed a multi-channel deep neural network using the gray-scale image of expression image to express the spatial characteristics, and the change of neutral and peak emotions to represent the temporal change characteristics. The experimental results are quite excellent. Kuo [37] and others added a long-short term memory (LSTM) network on the basis of CNN to extract the temporal characteristics of changing emotions and achieve a considerable recognition effect. In addition, it is found that CNN is more suitable for 3D convolution network than RNN (Recurrent Neural Network), and C3D (convolutional 3D) [38] is generated. Typically, Kawaai et al. [39] fused the LSTM-RNN model for audio characteristics, image classification model for extracting geometric features of irregular points in video set, and C3D model for processing temporal and spatial characteristics of images, with an accuracy rate of 17% higher than baseline. In addition, the EEGs (Electroencephalogram) and EMGs (electromyogram) of biosensors in the field of biology have also achieved great success in brain activity and facial muscle behavior perception. The work of [40] integrates DTAN and DTGN features using the joint fine tuning method, which proves that the integration method is better than other weighted sum integration methods.

3. Methods

3.1. HOG Features

Histogram of oriented gradient—namely, HOG feature—is calculated and counted through the histogram of gradient direction of local area of the image, and finally constitutes features, which can effectively extract facial emotional features [7]. HOG feature and scale-invariant feature transform (SIFT) [41] are both calculated on a dense image grid with uniform interval, and overlapped local contrast normalization is used to improve performance. At present, HOG is mainly combined with an SVM (Support Vector Machine) classifier, which is mainly used for image recognition, and improves the performance in pedestrian detection. The implementation process of HOG feature is shown in Figure 1. After normalization, better results can be obtained for the change of shadow and illumination.

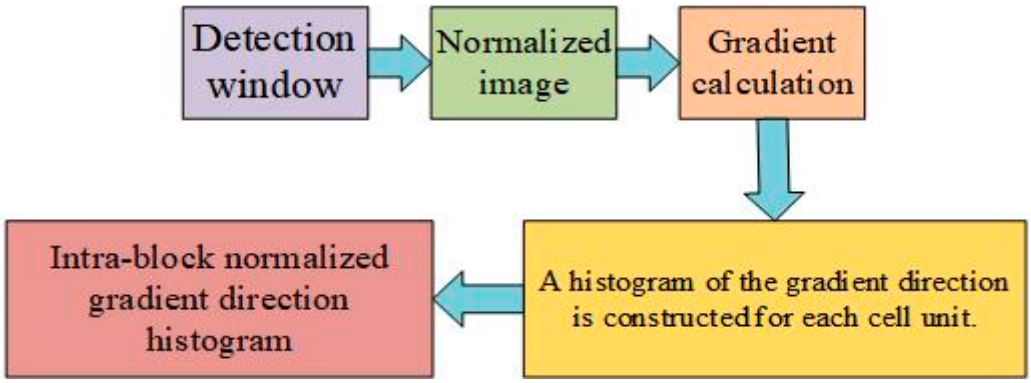


Figure 1. Histogram of oriented gradient (HOG) feature implementation process.

In face emotion recognition, the HOG description operator can be obtained by the following four steps.

3.1.1. Gradient Calculation

Firstly, two Sobel filters and expression images are convoluted to calculate the vertical and horizontal gradient maps. The vertical edge operator is $[-1,0,1]^T$, and the horizontal edge operator is $[-1,0,1]^T$. In particular, gamma and smooth normalization operations can be omitted [10].

3.1.2. Calculation of Amplitude and Direction

The amplitude and direction maps are calculated based on the vertical and horizontal gradient maps in step 1. Assuming *dx* and *dy* represent the gradient values in the horizontal and vertical maps, the amplitude and gradient of the pixel can be obtained according to Equation (1).

$$\begin{aligned} Magnitude &= \sqrt{(dx)^2 + (dy)^2} \\ Orientation &= \tan^{-1}\left(\frac{dy}{dx}\right) \end{aligned} \quad (1)$$

3.1.3. Unit Quantization

The emotional face image is divided into several small units. In Figure 2, the value range of gradient direction is 0~180, which is equally divided into 9 intervals, each of which is 20 degrees. The gradient amplitude is used as the weight of projection (i.e., mapped to a certain direction interval).

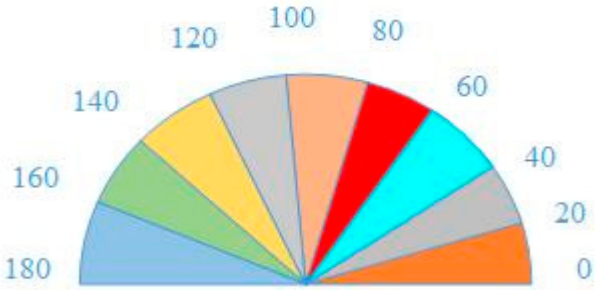


Figure 2. Nine Bins.

3.1.4. Block Normalization

In most cases, uneven illumination will affect the amplitude of the gradient, resulting in different value ranges, and local contrast normalization can improve the robustness

due to illumination changes and improve performance. The normalization process can be obtained by Equation (2).

$$v \to \sqrt{v/(\|v\|_1 + \varepsilon)} \tag{2}$$

In the equation, v represents the eigenvector before normalization and ε represents the constant that makes the denominator non-zero.

To sum up, according to a large number of studies in the literature, HOG features have many advantages in image detection. Firstly, slight body movement can be allowed in sampling and ignored without affecting the results; secondly, HOG features keep good invariance in optical and geometric deformation of image data. Therefore, the HOG feature is introduced into this model.

3.2. ESRs (Ensembles with Shared Representations)

In machine learning, the ensemble method is a method that can effectively reduce residual error and improve the accuracy and robustness of practical application. The integrated method represents a set of models in which collective inference can be made based on a single prediction [42]. Traditionally, the integration method needs to establish a decorrelation model trained independently, which is composed of a single type of neural network method integration [43]. However, in order to enhance its diversity, we can build an integration with different library methods [44,45]. At present, as a kind of resource, neural network integration needs higher computing power, but it is necessary to explore the method to reduce the aggregation redundancy and allow it to be used in practice. Meshgi et al. used the active learning method to reduce training time and redundancy. Meshgi did not use the entire data set, but used the most useful data for training [46,47]. In addition to active learning, the input space is decomposed into multiple regions, and the divide and conquer strategy of training a convolutional neural network in each region can also reduce redundancy. It can be seen that this method is composed of independent models, so it belongs to the "explicit" integration method. Compared with the "implicit" integration method, the "explicit" integration method still has a high redundancy in low-level visual features, while the "implicit" method makes a single network generalization like integration by extracting knowledge [48]. Shen et al. used the output of the CNN set to train the convolutional neural network, which has the advantages of not only maintaining the generalization ability and similar intermediate representation, but also reducing the training redundancy and time [49]. In the context of deep learning, training a network set is often high redundancy, low efficiency, and high cost. However, based on [50], this paper improves the traditional integration method as ensembles with shared representations (ESRs). By changing the branch level of ESR, the ensembles with shared representations (ESRs) based on convolutional neural network can reduce the computational complexity and redundancy without losing the generalization ability and diversity. As shown in Figure 3, the ESRs method is neither a complete "explicit" integration method, nor a complete "implicit" integration method [50]. Its shared layer belongs to the "implicit" part, and the latter belongs to the "explicit" part. Specifically, the implicit part is mainly responsible for reducing redundancy, reasoning and training time, learning some low-level features, and sharing the low-level features with the convolution branch set, while the explicit part is mainly responsible for the overall diversity [51]. The starting level of branch integration in ESRs has an important impact on generalization ability, computational load, redundancy, and diversity [52]. If level 1 is started too early, it may lead to high redundancy of low-level face features, and too late branching may reduce the diversity of integration because the shared layer no longer corresponds to spatial facial features (Level 5).

In addition, ESRs make full use of the two basic characteristics of the translation invariance of cumulus learning mode and the spatial hierarchical structure of multiple cumulus learning modes. In the early days of ESRs, layers learned local and simple visual patterns such as lines, edges, and colors. Then, when each layer was layered, the local patterns of the front layer were integrated into complex concepts such as nose, eyes, mouth, and so on, until the feature map was no longer visually interpretable. Finally, these feature

maps of the last layer were coded as the concept of emotion. As shown in Figure 3, ESRs are composed of gray blocks and purple blocks. Gray blocks represent the basis of the network and are mainly used for the convolution layer of low and intermediate feature learning, while purple blocks are responsible for independent convolution branches. In particular, the information features learned by gray blocks need to be shared with the independent convolution branches represented by purple blocks, which constitute a whole. Each branch not only needs independent learning features, but also competes for the common resources of the shared layer. The sum of the loss functions of each branch in competitive training is as follows, Equation (3):

$$L_{ESRs} = \sum_{b} \sum_{i} L[P(f(x_i) = y_i | x_i, \theta_{shared}, \theta_b), y_i] \quad (3)$$

In the formula, b represents the branch index, (x_i, y_i) represents randomly sampling from the training set, θ_{shared} represents the parameters of the shared layer of the ESRs regulating network, and θ_b represents the parameters constituting the integrated convolution branch.

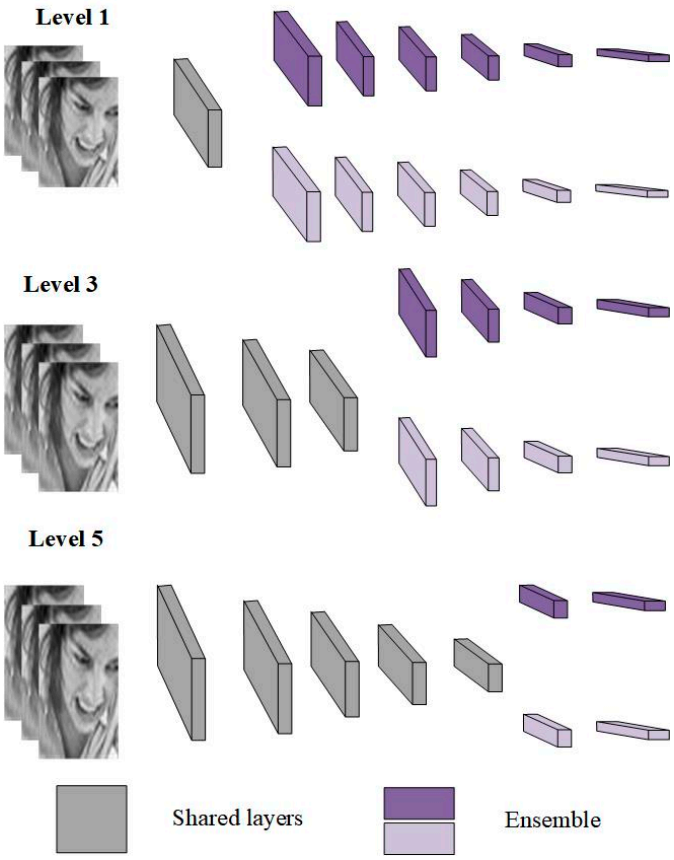


Figure 3. Ensemble with shared representations (ESRs).

The shared layer is an effective transfer learning mechanism, which can accelerate and guide learning when the ensemble grows, and add new convolution branches in order during training, as described in Algorithm 1 [50].

Algorithm 1: Training ESRs.

initialize the shared layers with θ_{shared}
for b **to** *maximum ensemble size* **do**
initialize the convolutional branch B_b with θ_b
add the branch B_b to the network *ESRs*
sample a subset D' from a training set D
foreach *mini-batch* $(x_i.y_i)$~D' **do**
perform the forward phase
initialize the combined loss function L_{esr} to 0.0
foreach *existing branch* $B_{b'}$ *in ESR* **do**
compute the loss $L_{b'}$ with respect to $B_{b'}$
add $L_{b'}$ to L_{esr}
end
perform the backward phase
optimize *ESRs*
end
end

3.3. HOG-ESRs

The proposed method is based on HOG features and ESRs, namely the HOG-ESRs method. Firstly, for the ESRs method, according to [9], we construct a set of four networks, namely, four convolution branches. According to the model exploration in [1], each network branch uses a model with four convolutional branches (ESR-4 LVL. 4) with four convolutional branches, as shown in Figure 4. The CNN model of each branch of the network is based on the original pixel data. HOG features are added to the last convolution layer, and then the mixed feature set enters the full connection layer, as shown in Figure 5. To sum up, in the new model of HOG-ESRs, the convolution layer uses the original pixel data as the main feature of the classification task, connects the features generated by the convolution layer with the HOG features, and sends the composite features into the FC layer network, which is regarded as the single branch network of ESRs.

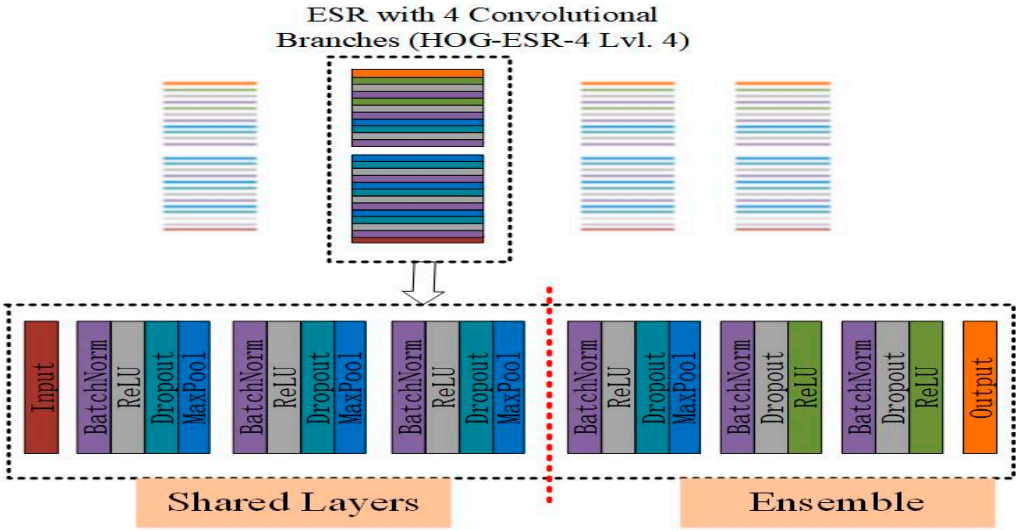


Figure 4. Schematic diagram of histogram of oriented gradient (HOG)-ESR-4 Lvl. 4.

Specifically, for the HOG-ESRs method proposed in this paper, because it is based on HOG features and ESRs, we first analyze the single branch convolutional neural network in the integrated network (Figure 4). According to [1], the single branch network adopts the network architecture of [Conv-(BN)–ReLU-(Dropout)-(Max-pool)] M-[Affine-(BN)-(Dropout)-ReLU] N–Affine-Softmax, where M is 4 and N is 2. The single branch network is based on the original pixel data and added HOG features in the last convolution layer, and

then enters the fully connected layer. In addition to the convolution layer, the single branch network architecture also includes batch normalization (BN), ReLU, dropout, and maxpool, passing through M layers, i.e., 4 layers. For the remaining two fully connected layers, BN, Dropout, and ReLU are always included. In addition, L2 regularization is added to the single branch model architecture. Finally, according to [9], we construct a set of four single branch network models mentioned above (Figure 5).

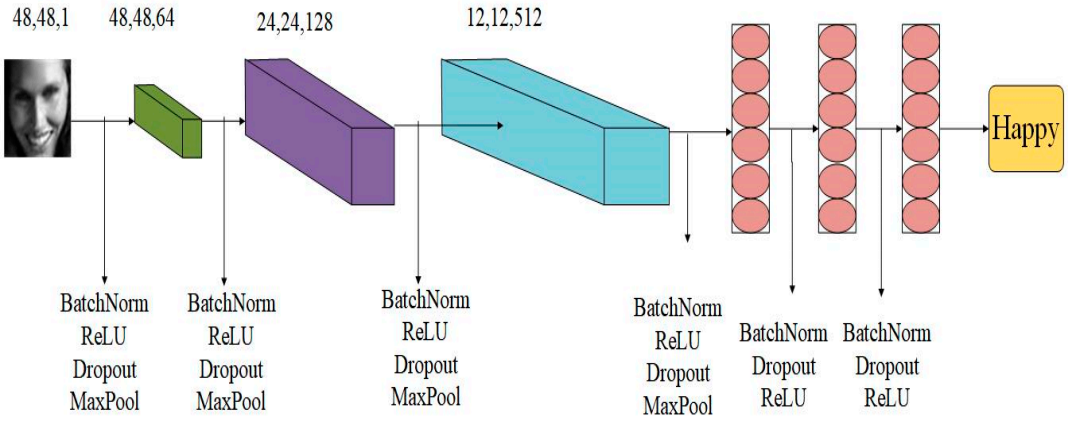


Figure 5. The architecture of CNN (Convolutional Neural Networks) is as follows: four convolution layers and two full connection layers.

The shared layer of HOG-ESRs needs to test bagging before adding a new convolution branch. The shared layer (lr_{sl}) and the trained branch (lr_{tb}) need to be tested before adding new convolution branches. Then, the remaining data are trained. Three different learning rates are used: one is the same initial learning rate (*fixed lr.;* $lr_{sl} = lr_{tb} = 0.1$), the other is a smaller learning rate (*varied lr.;* $lr_{sl} = 0.1$ *and* $lr_{tb} = 0.02$), and the third is not training at all (*frozen layers;* $lr_{sl} = lr_{tb} = 0.0$).

The two main indexes of the evaluation model are loss history and accuracy. The loss history is calculated by Equation (4), and the accuracy is calculated by Equation (5).

$$L_{HOG\text{-}ESRs} = \sum_{b}\sum_{i} L[P(f(x_i) = y_i|x_i, \theta_{shared}, \theta_b), y_i] \quad (4)$$

In the formula, b represents the branch index, (x_i, y_i) represents randomly sampling from the training set, θ_{shared} represents the parameters of the shared layer of the ESRs regulating network, and θ_b represents the parameters constituting the integrated convolution branch.

$$Accuracy = \sum_{i=1}^{n} \frac{N_{correct}^{i}}{N_{all}^{i}} \quad (5)$$

In formula (5), n represents the cross validation multiple, N_{all}^{i} represents the total quantity in the i folds, and $N_{correct}^{i}$ represents the accurately predicted quantity in the i folds.

4. Experiments and Analysis

4.1. Dataset and Features

According to the accumulation of previous scholars, a large professional database of facial emotion has been established, which provides a rich database for future research on facial emotion recognition, and the typical database also provides the basis for the test of face emotion recognition algorithm. For example, JAFFE (Japan female facial expression) database of Kyushu University [53], MMI (man machine interaction) database of Delft University of technology in the Netherlands [54], CK (Cohn Kanade) [55] of Carnegie Mellon University in the United States, and CAS-PEAL database of Chinese Academy of

Sciences [56]. Among them, Jaffe, CK+, fer2013, and affectnet are the most classic databases in the study of facial emotion. The details are as follows.

CK + (Extended Cohn-Kanade) [57,58] database is from Carnegie Mellon University in the United States. It was published in 2010 and established by the Department of psychology and Robotics Research Institute. It is one of the first databases selected by many scholars. The CK + database is based on Cohn Kanade (CK) published in 2000. The number of sequences increased by 22%, and the number of subjects increased by 27%. Among 210 adult subjects aged 18–50, 31% were male, 69% were female, 13% were black, 81% were European and American, and 6% were other races. This emotional dataset contains 593 sequences, each of which is provided with a complete FACS code of peak frame, including seven emotional tags: happy, surprised, angry, afraid, sad, neutral, disgusting, and contempt. Among them, contempt is generally not a kind of emotion. Many experiments have excluded these emotional data [58], and in addition, 593 sequences are neutral frames start, and peak frame ends. In this dataset, the image sequences of the front view and the 30 degree view are digitized into 640 × 480 or 640 × 480 pixel arrays with 8-bit gray scale or 24 bit color values.

JAFFE [53,59] database is a female face emotion database from Kyushu University in Japan. It was established by psychology department and ATR Human Information Processing Laboratory of Japan and released in 1998. It is a relatively old emotional database, but it also provides a reliable database source for Asian emotion recognition research. Because the database is open and the emotion is calibrated according to strict standards, it is also one of the classic facial emotion databases. Compared with 210 subjects in CK + database, Jaffe database was only from 10 Japanese female students, and the database was relatively small. It also contained seven basic emotional expressions, namely, anger, disgust, happiness, surprise, sadness, neutrality, and fear. The database contains 213 positive 256 × 256 gray images. At present, the recognition rate of this database is very high. Now, it is only used for some basic knowledge of facial emotion recognition, such as feature extraction, classification, and so on.

AffectNet database is a field face emotion database. Before the emergence of AffectNet database, the existing database of field facial emotion annotation was very small, and most of them covered discrete emotions, so it was not applicable in continuous dimension model [53]. Therefore, we created the AffectNet database, which collected and annotated more than 1 million emotional images from the Internet through three major search engines, plus 1250 emotion related keyword queries in six different languages. About half of the facial images (about 440,000) were manually labeled with seven discrete facial expressions (category model) and valence and arousal intensity (dimension model). Therefore, there are three models for facial emotion recognition based on this database: First, the categorical model. The identified expressions are selected from the relevant lists, such as Ekman's six basic expressions. Second, the dimensional model, whose values are in a continuous scale, such as valence, and arousal. Valence refers to the positive or negative degree of the event, and arousal refers to whether the event is excited/excited or calm/soothing. Thirdly, FACS (facial action coding system) model, in which facial movements are represented by Au.

The FER2013 dataset is from the kaggle competition, conducted by Pierre Luc carrier and Aaron Courville, and is part of an ongoing research project [60]. They provided a preliminary version of their dataset to the organizers of the seminar for use in the competition. The FER2013 dataset contains 35,887 facial emotion images. However, the dataset saves the expression, image data, and purpose data as a CSV (Comma-Separated Values) file, rather than directly as the given images. There are seven kinds of emotions in the data set, and the corresponding labels are 0–6. The specific labels and emotions are as follows: 0 = anger; 1 = dislike; 2 = fear; 3 = happy; 4 = sad; 5 = surprise; 6 = neutral. The 35,887 image data are composed of 48 * 48 pixel gray scale facial images. As the faces have been registered automatically, they are basically in the middle, and each image occupies about the same amount of space. In addition, the dataset contains 28,709 training images (training), 3589 public test images (public test), and 3589 private test images (private test).

To sum up, Figure 6 depicts an example of each facial expression category in the above four datasets [50]. Co stands for contempt, but, according to a large number of studies in the literature, it is basically eliminated. Therefore, this label is not used in this paper. Compared with other datasets, the FER2013 dataset is selected as the experimental dataset in this paper, and the experiments are based on the original pixel data. In this paper, after reading the original pixel data of FER2013, the average value of the training image is subtracted from the image for normalization, and the image is flipped horizontally in the training set to generate an image to increase the data. In this paper, we not only associate features generated from original pixel data, but also associate HOG features with ensembles with shared representations as a new learning model.

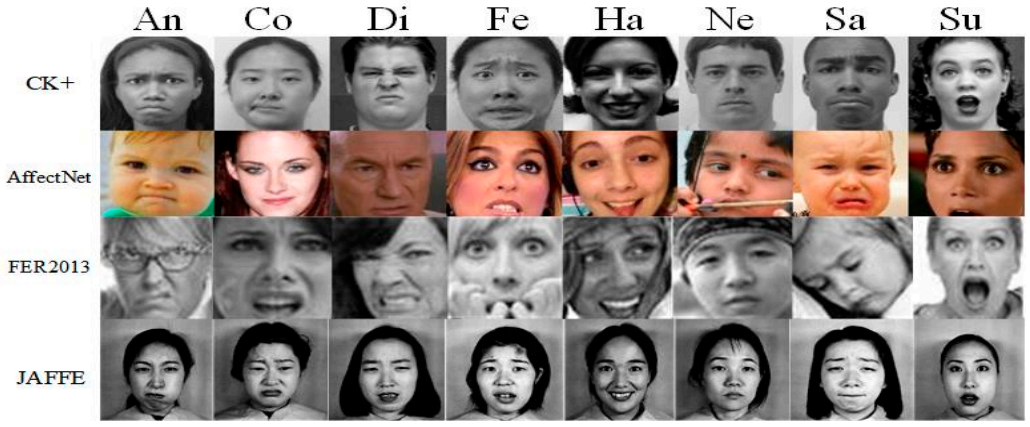


Figure 6. Extended Cohn-Kanade (CK+), AffectNet, FER2013, and Japan female facial expression (JAFFE), from top to bottom.

4.2. Experiments

First of all, each network in each HOG-ESRs is based on the exploratory training results of [1]. The results of [1] show that the network with five and six convolution layers does not improve the classification accuracy. The model with four convolution layers and two FC layers is the best network for the FER2013 dataset. Specifically, the first convolution layer has 64 3 × 3 filters, the second has 128 5 × 5 filters, the third has 512 3 × 3 filters, and the last one has 512 3 × 3 filters. In all convolution layers, the step size is 1 and the activation functions are batch normalization, dropout, max pooling, and ReLU. There are 256 neurons in the hidden layer of the first FC layer and 512 neurons in the second FC layer. In the FC layer, as in the convolution layer, batch normalization, dropout, and ReLU are used as activation functions. Softmax is used as the loss function in this paper. Figure 6 shows the architecture of CNN. Users can specify the number of branch filters, spans, and zero padding for each CNN network, but if not indicated in this article, the default values are used. In this paper, the above model is implemented in Torch, and the GPU accelerated deep learning feature is used to speed up the training process. Because the HOG-ESRs method is based on the ESRs method, it not only inherits the advantage that ESRs can reduce the residual generalization error, but also inherits the advantages of short training time and low computational cost. For the HOG-ESRs method, 40 epochs and 128 batch sizes are used to train the HOG-ESRs network using all the images in the training set, and the super parameters are cross-verified to obtain the most accurate model. Although the training process involves single branch network and integrated network, and involves a lot of preparation work and parameter adjustment process, because of the advantage of short training time, this algorithm can quickly achieve the expected model. Table 1 describes the values with the highest accuracy for each parameter in the model.

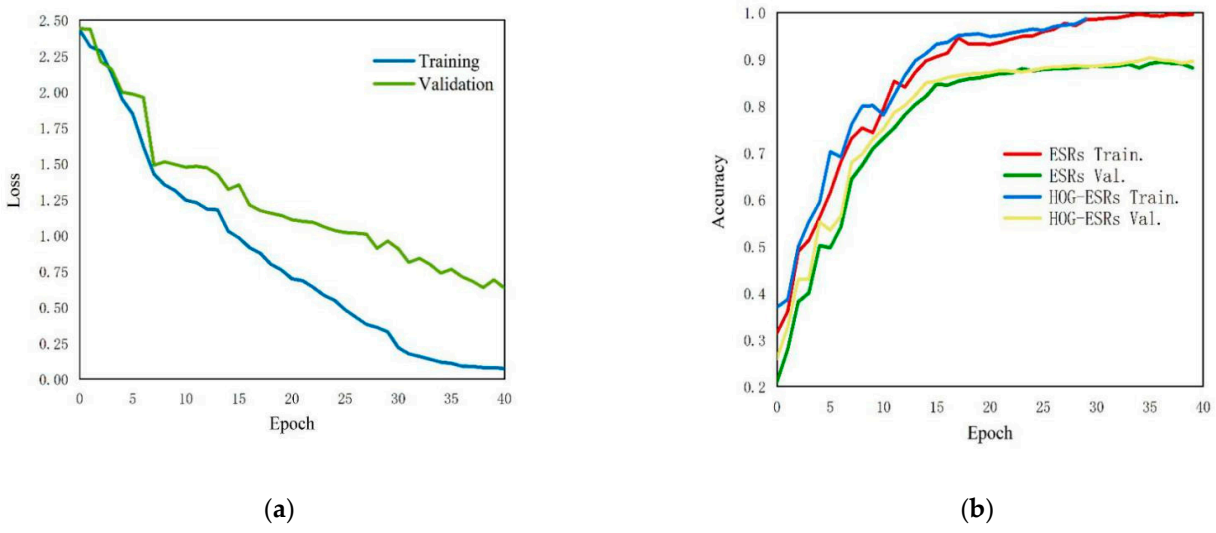
Table 1. The hyper-parameters obtained by cross validation for the deep model.

Parameter	Value
Learning Rate	0.01
Regularization	1×10^{-7}
Hidden Neurons	256,512

4.3. Results

All experiments described in this paper were conducted on a computer with Intel(R) Pentium(R) CPU G4560 @3.5 GHz, 8192 MB RAM, Intel(R) HD Graphics 610, and Windows 10. The algorithms were developed in Python 3.6. The main python libraries used in our framework were the following: Matplotlib 3.0.3, Numpy 1.17.4, OpenCV-python 4.1.2.30, Pillow 5.0.0, Torch 1.0.0, Torchvision 0.2.1, and so on.

In order to evaluate the performance of the new HOG-ESRs model, the loss history and accuracy of the model are drawn in this paper. The results are shown in Figure 7a,b. First of all, the results in Figure 7a show that the training accuracy reaches the highest value quickly, and the convergence speed of the model is very fast. Secondly, Figure 7b shows the accuracy of the model under different iteration times. From the observation of Figure 7b, the model can reduce the over fitting behavior of the model by adding more anti-over fitting technology and non-linear over fitting technology, which is actually the use of dropout and batch normalization. In addition, from the comparison between the model with and without HOG features, it can be seen that the accuracy of the HOG-ESRs model is different from that of the model without HOG features, which can fully reflect that the new model has a strong enough ability to extract enough information using the original pixel data and HOG features.



(**a**) (**b**)

Figure 7. (**a**) The loss history of the HOG-ESRs models; (**b**) the accuracy of model for different numbers of iterations.

In addition, this model is compared with the best precision of the model proposed by other authors, and several representative models are selected according to the year, as shown in Table 2. Two points can be seen from the table. First, the ESRs model is superior to other models in face emotion recognition. Second, the accuracy of the HOG-ESRs model with HOG added to ESRs model is higher than ESRs. Therefore, it can be concluded from these two points that the ESRs model is superior to other models, and adding HOG to ESRs can also improve the accuracy of the model. Therefore, it also reflects the effectiveness and rationality of this work.

Table 2. The best accuracy (%) comparison between different models. HOG, histogram of oriented gradient; ESRs, ensembles with shared representations.

Approach	Year	Accuracy
TFE-JL [44]	2018	84.3%
SHCNN [61]	2019	86.54%
ESRs [50]	2020	87.15 ± 0.1%
HOG-ESRs	2020	89.3 ± 1.1%

After comparing different models, in order to show the rationality of the selected dataset, the accuracy of the model in different datasets is also compared, as shown in Table 3. First of all, it can be seen from the table that the accuracy on the AffectNet dataset is the lowest, probably because of the complexity of this dataset. When introducing this dataset, it is also said that this dataset can be divided into three models in face emotion recognition: categorical model, dimensional model, and FACS model. However, compared with the accuracy of about 59% in other algorithm models [50], the accuracy of this dataset in this paper algorithm model is also high. Secondly, the CK + and JAFFE datasets belong to laboratory datasets, while AffectNet and FER2013 are datasets in natural state, so it can be seen that the algorithm in this paper also has good accuracy in wild datasets, which also shows that the model in this paper has good adaptability.

Table 3. The best accuracy (%) comparison between different dataset based on the HOG-ESRS model. CK+, extended Cohn-Kanade; JAFFE, Japan female facial expression.

Approach	Dataset	Accuracy
HOG-ESRs	CK+	88.3 ± 0.8%
	JAFFE	87.9 ± 2.1%
	AffectNet	62.13 ± 0.5%
	FER2013	89.3 ± 1.1%

Figure 8 shows the average accuracy of adding branch level on the FER2013 data test set, as well as the baseline (dotted line). From the figure, the integration method has higher accuracy than the single network. The accuracy of the HOG-ESRs method at level 4 is as high as that of the traditional integration method, but there are great differences in the number of trainable parameters between the two methods, as shown in Table 4. Table 4 shows that, compared with the traditional set, HOG-ESRs require much less trainable parameters, and the fourth and fifth levels are significantly reduced. Compared with a single network, the recognition performance of HOG-ESRs is significantly improved. At the same time, it also shows that HOG-ESRs have strong generalization ability, while significantly reducing redundancy and computing load. In addition, it can be found from Figure 8 that the performance of the interleaved approach needs to be improved. It is preliminarily considered that the reason may be the low diversity, because the diversity in cross training is only related to the mixing of different data and different starting points.

Finally, the confusion matrix of the model is calculated. Figure 9 shows the visualization of the confusion matrix. Looking at the displayed data, it is easier to learn the features of happy faces than to express other facial emotions, because the model has good accuracy in predicting happiness tags. In addition, the confusion matrix also reflects that the trained network may confuse some tags, such as anger tags and sad tags. By observing their correlation, it is found that the classifier classifies the "anger" tag as "fear" or "sad" in many cases. In fact, even human beings may have difficulty distinguishing anger from sadness because they express their emotions in different ways. In addition, in the process of the experiment, it was found that, if the HOG-ESRs model increased the sample size of each emotion category too much, the diversity of the model would be reduced. On the other hand, the accuracy of different emotion categories would also affect the overall accuracy rate. Therefore, it is best not to blindly improve the sample size and the accuracy

of different emotion categories. In short, the experimental results show that the HOG-ESRs model has good results in accuracy and robustness, and reduces the deviation problem of machine learning.

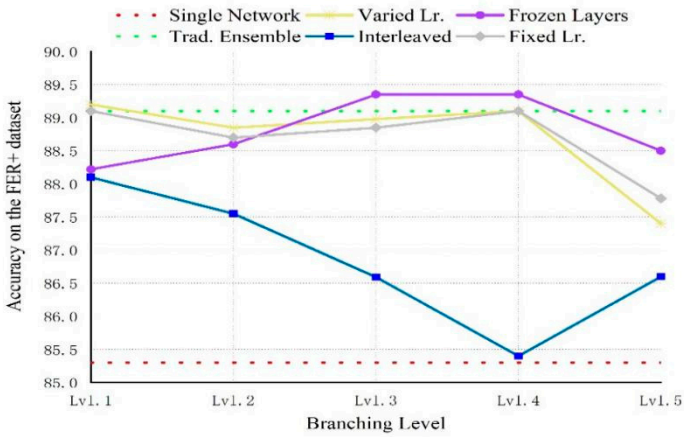


Figure 8. Accuracy on the FER+ increasing dataset branching level for different training strategies.

Table 4. Test accuracy (%) of most accurate networks and baselines on FER+ and their number of parameters.

Approach	#	Accuracy
Single Network	140,802	85.3 ± 2.1%
Traditional Ensemble	560,328	89.1 ± 1.5%
HOG-ESR-4 Lvl.4	370,411	89.3 ± 1.1%
HOG-ESR-4 Lvl.5	250,005	88.4 ± 3.5%

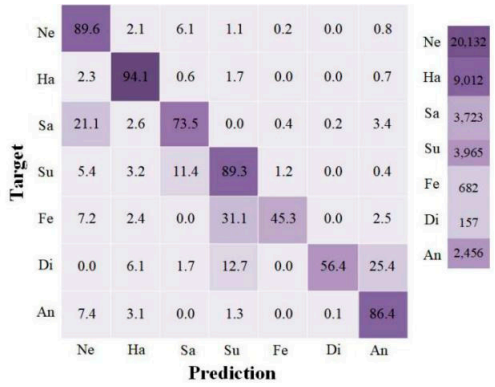


Figure 9. Normalized confusion matrix of the ensemble predictions on FER+ and the emotion label distribution.

To sum up, this paper also tests the recognition time of the algorithm on the experimental platform. In order to show the emotion recognition time of the algorithm, that is, the time from the image entering the model to the model providing the results according to the situation (as shown in Table 5), it is found that the time is not fixed, about 0.83 s to 1.2 s, so the reasoning time of the model is about 1000 milliseconds. Therefore, the model also shows the advantage of time, which confirms the rationality of this model.

Table 5. Model inference time (ms) on the test-bed.

	Fastest Time	Slowest Time
Time	about 830 ms	about 1240 ms

5. Discussion

Facial emotion recognition is a hot research field based on machine vision, artificial intelligence, image processing, and so on. Predecessors have achieved some success. However, with the development of modern science and technology and higher requirements for facial emotion recognition algorithms, there is still room for improvement in the extraction of effective features and recognition accuracy of previous algorithms. According to a large number of studies in the literature, HOG features can effectively extract face features, and the ensembles method can effectively improve the accuracy and robustness of the algorithm. Therefore, this paper proposes a new algorithm, HOG-ESRs, which improves the traditional ensembles method to the ensembles with shared representations method, effectively reducing the residual generalization error, and then combines the HOG with the ESRs. The experimental results show that the accuracy of the new algorithm is not only better than the best accuracy of the model proposed by other authors, but it also shows good results on different data sets. Further, based on a large number of experimental results, the algorithm can not only effectively extract features and reduce residual generalization error, but also improve the accuracy and robustness of the algorithm.

In addition, it is worth noting that the facial emotion recognition algorithm proposed in this paper is widely used, including in driver emotion detection, mental patients' facial emotion detection, intelligent education, and so on. However, the HOG-ESRs model has some limitations. The model network may confuse some tags, such as anger tags and sadness tags. By observing their correlation, it is found that the classifier mistakenly classifies the "anger" tag as "fear" or "sad" in many cases. In fact, even human beings may have difficulty in distinguishing anger and sadness because they express their emotions in different ways. Even with the same facial expression, people may recognize different emotions, which is also the direction of future research. Despite the above limitations, the HOG-ESRs method contributes to improving the accuracy and robustness of the algorithm, which is an attempt to develop a facial emotion recognition algorithm based on deep learning and convolutional neural network in the future.

6. Conclusions

This paper proposes a hybrid strategy based on HOG features and ESRs, which is called the HOG-ESRs method. Histogram of oriented gradient can effectively extract face features, and improving the ensembles with shared representations method can effectively reduce the residual generalization error and improve the accuracy and robustness of the algorithm. All the images in the training set are used to train the HOG-ESRs network, and the model parameters with the highest accuracy are obtained by cross validation of the super parameters. The experimental results on the FER2013 facial expression database show that the proposed method achieves good performance. The model can effectively extract face features, effectively reduce the residual generalization error, and improve the accuracy and robustness of the algorithm. In the future, image data should be added under different illumination to verify and improve the algorithm in order to further improve the model of facial expression recognition.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, Writing—Original draft preparation, L.S. and C.G.; resources, Writing—Review and editing, supervision, project administration, funding acquisition, Y.Z. and H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities, grant number No. 2020CDCGTX055, No. 2019CDCGTX302 and No. 2018CDPTCG000141.

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

References

1. Bettadapura, V. *Face Expression Recognition and Analysis: The State of the Art*; CoRR: New York, NY, USA, 2012.
2. Ding, J.; Chen, K.; Liu, H.; Huang, L.; Chen, Y.; Lv, Y.; Yang, Q.; Guo, Q.; Han, Z.; Ralph, M.A.L. A unified neurocognitive model of semantics language social behaviour and face recognition in semantic dementia. *Nat. Commun.* **2020**, *11*, 2595. [CrossRef] [PubMed]
3. Anagnostopoulos, C.N.; Iliou, T.; Giannoukos, I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif. Intell. Rev.* **2015**, *43*, 155–177. [CrossRef]
4. Dobs, K.; Isik, L.; Pantazis, D.; Kanwisher, N. How face perception unfolds over time. *Nat. Commun.* **2019**, *10*, 1258. [CrossRef] [PubMed]
5. Kumar, M.P.; Rajagopal, M.K. Detecting facial emotions using normalized minimal feature vectors and semi-supervised twin support vector machines classifier. *Appl. Intell.* **2019**, *49*, 4150–4174. [CrossRef]
6. Siddiqi, M.H. Accurate and robust facial expression recognition system using real-time youtube-based datasets. *Appl. Intell.* **2018**, *48*, 2912–2929. [CrossRef]
7. Navneet, D.; Triggs, B. His-tograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 27 June–2 July 2005; Volume 1.
8. Zhao, Y.; Zhang, Y.; Cheng, R.; Wei, D.; Li, G. An enhanced histogram of oriented gradients for pedestrian detection. *Intell. Transp. Sys. Mag. IEEE* **2015**, *7*, 29–38. [CrossRef]
9. Alizadeh, S.; Fazel, A. Convolutional neural networks for facial expression recognition. *arXiv* **2017**, arXiv:1704.06756v1.
10. Donia, M.M.F.; Youssif, A.A.A.; Hashad, A. Spontaneous facial expression recognition based on histogram of oriented gradients descriptor. *Comp. Inform. Ence.* **2014**, *7*, 31–37. [CrossRef]
11. Wang, S.; Yu, R.; Tyszka, J.M.; Zhen, S.; Kovach, C.; Sun, S.; Huang, Y.; Hurlemann, R.; Ross, I.B.; Chung, J.M.; et al. The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nat. Commun.* **2017**, *8*, 14821. [CrossRef]
12. Fernández-Dols, J.M.; Crivelli, C. Recognition of facial expressions: Past, present, and future challenges. In *Understanding Facial Expressions in Communication*; Springer: New Delhi, India, 2015; pp. 19–40.
13. Prylipko, D.; Roesner, D.; Siegert, I.; Guenther, S.; Friesen, R.; Haase, M.; Vlasenko, B.; Wendemuth, A. Analysis of significant dialog events in realistic human–Computer interaction. *J. Multimodal User Interf.* **2014**, *8*, 75–86. [CrossRef]
14. Niese, R. Facial expression recognition based on geometric and optical flow features in colour image sequences. *IET Comp. Vis.* **2012**, *6*, 79–89. [CrossRef]
15. Shan, C.; Gong, S.; Mcowan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comp.* **2009**, *27*, 803–816. [CrossRef]
16. Wan, C. Facial expression recognition in video sequences. *Intell. Control Automat. IEEE* **2012**, *6*, 4766–4770.
17. Praseeda Lekshmi, V.; SasiKumar Vidyadharan, D.S.; Naveen, S. In Proceedings of the 2008 International Conference on Audio, Language and Image Processing–Analysis of Facial Expressions using pca on Half and Full Faces, Shanghai, China, 7–9 July 2008; pp. 1379–1383.
18. Nguyen, B.T.; Trinh, M.H.; Phan, T.V.; Nguyen, H.D. An efficient real-time emotion detection using camera and facial landmarks. In Proceedings of the Seventh International Conference on Information Science & Technology, IEEE, Da Nang, Vietnam, 16–19 April 2017.
19. Loconsole, C.; Miranda, C.R.; Augusto, G.; Frisoli, A.; Orvalho, V. Real-time emotion recognition novel method for geometrical facial features extraction. In Proceedings of the 2014 International Conference on Computer Vision Theory (VISAPP), IEEE, Lisbon, Portugal, 5–8 January 2014; Volume 1, pp. 378–385.
20. Palestra, G.; Pettinicchio, A.; Coco, M.D.; Pierluigi Carcagn, I.; Distante, C. Improved Performance in Facial Expression Recognition Using 32 Geometric Features. In *Image Analysis and Processing–ICIAP 2015, Proceedings of the International Conference on Image Analysis & Processing, Genoa, Italy, 7–11 September 2015*; Springer International Publishing: New York, NY, USA, 2015; pp. 518–528.
21. Tian, Y.L.; Kanade, T.; Cohn, J.F.; Li, S.Z.; Jain, A.K. Facial expression analysis. In *Handbook of Face Recognition*; Springer: London, UK, 2005; pp. 247–275.
22. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comp.* **2017**, *10*, 18–31. [CrossRef]
23. Keltiner, D.; Ekrman, P. Darwin and facial expression: A century of research in review. In *Unemotion, Handbook of Emotions*; Lewis, M., Haviland Jones, J.M., Ekman, P., Eds.; Gilford Press: New York, NY, USA, 2000; pp. 236–249.
24. Ekman, P. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
25. Ekman, P. *Darwin and Facial Expression: A Century of Research in Review*; Academic Press Ishk: Cambridge, MA, USA, 2006; p. 1973.
26. Ekman, P.; Freisen, W.V.; Ancoli, S. Facial signs of emotional experience. *J. Personal. Soc. Psyc.* **1980**, *39*, 1123–1134. [CrossRef]
27. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. *Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution, Proceedings of the ICMI 18th ACM International Conference, Tokio, Japan,12-16 November 2016*; ACM: New York, NY, USA, 2016; pp. 279–283.
28. Hewitt, C.; Gunes, H. Cnn-based facial affect analysis on mobile devices. *arXiv* **2018**, arXiv:1807.08775.

29. Khorrami, P.; Paine, T.L.; Huang, T.S. Do deepneural networks learn facial action units when doing expres-sion recognition? In Proceedings of the IEEE on CVPR–Workshops, Boston, Massachusetts, USA, 11–12 June 2015; Volume 1, pp. 19–27.
30. Ekman, P. The argument and evidence about universals in facial expressions. In *Handbook of Social Psychophysiology*; John Weily & Sons Ltd: Hoboken, NJ, USA, 1989; pp. 143–164.
31. Hamester, D.; Barros, P.; Wermter, S. Face expression recognition with a 2-channel convolutional neural network. international joint conference on neural networks. In Proceedings of the IEEE 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–8.
32. Xie, S.; Haifeng, H. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Trans. Multimed.* **2018**, *21*, 211–220. [CrossRef]
33. Laptev, D.; Savinov, N.; Buhmann, J.M.; Pollefeys, M. TI-POOLING: Transformation-invariant pooling for feature learning in convolutional neural networks. CVPR 2016. *IEEE Comp. Soc.* **2016**, 289–297.
34. Li, D.; Wen, G.; Li, X.; Cai, X. Graph-based dynamic ensemble pruning for facial expression recognition. *Appl. Intell.* **2019**, *49*, 3188–3206. [CrossRef]
35. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. Image Proces. Publ. IEEE Signal Proces. Soc.* **2017**, *26*, 4193–4203. [CrossRef] [PubMed]
36. Sun, N.; Li, Q.; Huan, R.; Liu, J.; Han, G. Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognit. Lett.* **2019**, *119*, 49–61. [CrossRef]
37. Li, T.H.S.; Kuo, P.H.; Tsai, T.N.; Luan, P.C. Cnn and lstm based facial expression analysis model for a humanoid robot. *IEEE Access PP* **2019**, *99*, 1. [CrossRef]
38. Grossman, S.; Gaziv, G.; Yeagle, E.M.; Harel, M.; Mégevand, P.; Groppe, D.M.; Khuvis, S.; Herrero, J.L.; Irani, M.; Mehta, A.D.; et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **2019**, *10*, 4934. [CrossRef]
39. Ouyang, X.; Kawaai, S.; Goh, E.G.H.; Shen, S.; Ding, W.; Ming, H.; Huang, D.Y. Audiovisual emotion recognition using deep transfer learning and multiple temporal models. In Proceedings of the 19th ACM International Conference, Glasgow, Scotland, 13–17 November 2017; pp. 577–582.
40. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 16–17 December 2015; pp. 2983–2991.
41. Nguyen, T.; Park, E.A.; Han, J.; Park, D.C.; Min, S.Y. Object detection using scale invariant feature transform. In *Genetic and Evolutionary Computing*; Springer International Publishing: New York, NY, USA, 2014; Volume 238, pp. 65–72.
42. Dietterich, T.G. Ensemble methods in machine learning. In *Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
43. Hansen, L.K. Neural network ensemble. *IEEE Trans Pattern Anal. Mach. Intell.* **1990**, *12*, 993–1001. [CrossRef]
44. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Proces. IP* **2018**, *28*, 2439–2450. [CrossRef]
45. Wang, C.Z.; Zhang, L.; Wang, B. Sparse modified marginal fisher analysis for facial expression recognition. *Appl. Intell.* **2019**, *49*, 2659–2671. [CrossRef]
46. Meshgi, K.; Oba, S.; Ishii, S. Efficient diverse ensemble for discriminative co-tracking. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
47. Seung, H.S.; Opper, M.A.; Sompolinsky, H. Query by committee. In Proceedings of the 5th Workshop on Computional Learning Theory, Pittsburgh, PA, USA, 27–29 July1992; Volume 284, pp. 287–294.
48. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *Comp. ence.* **2015**, *14*, 38–39.
49. Shen, Z.; He, Z.; Xue, X. Meal: Multi-modelensemble via adversarial learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2019.
50. Siqueira, H.; Magg, S.; Wermter, S. Efficient facial feature learning with wide ensemble-based convolutional neural networks. *arXiv* **2020**. [CrossRef]
51. Siqueira, H.; Barros, P.; Magg, S.; Wermter, S. An ensemble with shared representations based on convolutional networks for continually learning facial expressions. In Proceedings of the IEEE/RSJ International Conference on IROS, Madrid, Spain, 1–5 October 2018; pp. 1563–1568.
52. Chollet, F. Deep learning with python and keras. In *The Handbook by the Developer of the Keras Library*; MITP-Verlag GmbH & Co. KG.: Wachtendonk/Nordrhein-Westfalen, Germany, 2018.
53. Kamachi, M.; Lyons, M.; Gyoba, J. *Japanese Female Facial Expression Database*; Psychology Department in Kyushu University: Fukuoka, Japan, 1998.
54. Valstar, M.F.; Pantic, M. Induced disgust, happiness and surprise: An addition to the MMI facial expression database. Proc.intern.workshop on Emotion Corpora for Research on Emotion & Affect. 2010, pp. 65–70. Available online: http://lrec.elra.info/proceedings/lrec2010/workshops/W24.pdf#page=73 (accessed on 29 January 2021).
55. Gehrig, T. Action unit intensity estimation using hierarchical partial least squares. In *Feminist Theory and Literary Practice*; Foreign Language Teaching and Research Press: Beijing, China, 2015; pp. 1–6.
56. Gao, W.; Cao, B.; Shan, S.; Chen, X.; Zhou, D.; Zhang, X.; Zhao, D. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2007**, *38*, 149–161.

57. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Matthews, I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Vision & Pattern Recognition–Workshops, San Francisco, CA, USA, 13–18 June 2010.
58. Sauter, D.A.; Eisner, F.; Ekman, P.; Scott, S.K. Cross-cultural recognition of basic emotions through emotional vocalizations. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3086. [CrossRef] [PubMed]
59. Lyons, M.J.; Kamachi, M.; Gyoba, J. The Japanese female facial expression. *(JAFFE) Database J.* **1997**, 1–25. [CrossRef]
60. Ullman, S. Using neuroscience to develop artificial intelligence. *Science* **2019**, *363*, 692–693. [CrossRef] [PubMed]
61. Miao, S.; Xu, H.; Han, Z.; Zhu, Y. Recognizing facial expressions using a shallow convolutional neural network. *IEEE Access* **2019**, *7*, 78000–78011. [CrossRef]