information

MDPI

Article

Facial Expression Recognition Based on Random Forest and Convolutional Neural Network

Yingying Wang [1], Yibin Li [1], Yong Song [2,*] and Xuewen Rong [1]

[1] School of Control Science and Engineering, Shandong University, Jinan 250061, China; yywang89@126.com (Y.W.); liyb@sdu.edu.cn (Y.L.); rongxw@sdu.edu.cn (X.R.)

[2] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China

* Correspondence: songyong@sdu.edu.cn

Received: 27 October 2019; Accepted: 25 November 2019; Published: 28 November 2019

Abstract: As an important part of emotion research, facial expression recognition is a necessary requirement in human–machine interface. Generally, a face expression recognition system includes face detection, feature extraction, and feature classification. Although great success has been made by the traditional machine learning methods, most of them have complex computational problems and lack the ability to extract comprehensive and abstract features. Deep learning-based methods can realize a higher recognition rate for facial expressions, but a large number of training samples and tuning parameters are needed, and the hardware requirement is very high. For the above problems, this paper proposes a method combining features that extracted by the convolutional neural network (CNN) with the C4.5 classifier to recognize facial expressions, which not only can address the incompleteness of handcrafted features but also can avoid the high hardware configuration in the deep learning model. Considering some problems of overfitting and weak generalization ability of the single classifier, random forest is applied in this paper. Meanwhile, this paper makes some improvements for C4.5 classifier and the traditional random forest in the process of experiments. A large number of experiments have proved the effectiveness and feasibility of the proposed method.

Keywords: facial expression recognition; feature extraction; convolutional neural network; random forest

1. Introduction

Facial expressions include rich emotional information and play a very important role in interpersonal communication. Facial expression recognition has become one of the most promising biometric recognition technologies due to its characteristics of nature, intuition, non-contact, safety, and rapidity. According to a famous expression theory [1]: in the 100% emotional expression that is composed of facial expression, voice, and language, expression makes up 50 percent of the total information, 40 percent is composed by voice, and language only includes 8 percent, which shows the importance of facial expressions in interpersonal communication. In order to realize a more intelligent and natural human–machine interaction, facial expression recognition has been widely studied in recent decades [2,3], and it has attracted more and more researchers' attention.

The study of human expression can be extended to many other disciplines, such as behavioral science, psychology, and machine intelligence. The mature expression system is beneficial to human life. For example: (1) The monitoring system with the function of recognizing facial expressions can be used to identify the abnormal expression (hate, irritable, insecurity, etc.) in many large public places, such as supermarkets, train stations, airports, and crowded shopping streets. It is very effective to prevent the crime. (2) If the facial expression recognition system can be combined with the driver's safe driving, it can analyze the driver's facial expression at any time to determine whether the driver is tired, which can avoid the potential danger of fatigue driving. (3) Facial expressions can also promote

the development of the service industry. The pleased feedback of customers can be captured in time by facial expression recognition system. (4) If the facial expression recognition system is placed in the hospital, the patient's expression can be monitored and analyzed in real time, which can ensure timely treatment.

According to Ekman and Friesen [4], there is a universality of six basic emotions: happiness, surprise, sadness, fear, anger, and disgust. These emotions can be found in all cultures. Generally, facial expression recognition includes three steps: preprocessing, feature extraction, and classification [5]. Feature extraction is a key step in the whole recognition work [6] and the feature that we expect should minimize the distance of within-class variations of expression while maximizing the distance of between-class variations [7]. In feature extraction, there are two common approaches: appearance-based methods [8,9] and geometric feature-based methods [10,11]. Geometric features are extracted to form a feature vector that represents the shape and locations of facial components. This method can reduce a lot of computation time and make the system response faster, but the extraction of geometric features requires higher accuracy in feature point selection, and it is difficult to achieve accurate positioning of feature points in the case of low image quality or complex background. Meanwhile, geometric feature-based methods can only outline the change of the whole face body shape, and it ignores other parts of the face information such as skin texture changes, which will cause a low accuracy rate in the recognition of subtle changes in facial expressions. Surface properties of the image can be described with appearance-based methods. Because texture can make full use of image information, it can be used as an important basis for image description and recognition either from theory or common sense. However, the texture cannot completely reflect the essential properties of the object, because it is only a feature of the surface of an object. Therefore, it is impossible to obtain high-level image content by using only texture features. Facial expression features extracted by appearance-based methods have strong anti-interference ability, but most of them have complex computational problems and lack the ability to extract comprehensive and abstract features.

Either the geometric feature-based method or the appearance-based method belongs to single feature-based method. For facial expression feature information, single feature can only reflect a local or single feature of the expression, and it cannot fully reflect the overall features of the expression. For the expression classifier, the comprehensive degree of feature attribute value is very important, which can greatly affect the recognition accuracy. Hence, in order to minimize the final error rate, lots of fusion features-based methods have been proposed in recent years. The fusion features-based method can fully and effectively reflect the overall features for an expression image. Although the fusion features-based method proposed in the later stage has better performance than the single feature-based method, handcrafted features are often not well considered. If features are inadequate, even the best classifier would fail to achieve good performance [12]. Meanwhile, the time and labor cost of finding new ways to extract artificial features is very expensive. Studies have shown that neural networks have strong expressive ability, and a reasonable three-layer neural network can express any complex nonlinear mapping relations. In a deep network, each hidden layer can perform nonlinear transformation on the output of the previous layer. This means that each hidden layer is the synthesis and abstraction of the data of the previous layer, and the features learned through layer by layer can better depict the essence of the description object. Therefore, the deep network has better expression ability and learning ability than the shallow network. More hidden features can be automatically acquired to learn more useful features for improving the accuracy of prediction or classification. Therefore, deep learning is an inevitable trend in many machine recognition algorithms. However, deep learning-based methods require a large number of training samples and tuning parameters, and the hardware requirements are also high.

Expression feature extraction is the most important part of the facial expression recognition (FER) system, and effective extraction of facial features will greatly improve the recognition performance. Considering the disadvantages of handcrafted features and the complexity of deep learning model in the process of training and tuning, a new expression recognition system model that combines

the feature extracted from the deep neural network model with the traditional learning classifier is proposed in this paper, which can address the incompleteness of handcrafted features as well as avoid the long training time of the deep learning model.

As an important part of facial expression recognition, the design of the facial expression classifier greatly affects the accuracy rate of facial expression recognition; therefore, the selection and application of classifier is important to determine the final result. The facial expression recognition classifier should have a high computational efficiency and a powerful ability to handle a large number of data sets.

Decision tree [13], as a popular method of pattern recognition and data mining, has been deeply used in various fields due to its simple operation characteristics. Meanwhile, this algorithm has no requirements for the samples. These advantages are available for facial expression classification. Among many decision tree algorithms, C4.5 classifier [14] has been widely used in the field of image recognition and has the ability to classify and recognize facial expressions. Therefore, C4.5 classifier is selected as the classifier for expression recognition in this paper. Because some problems such as overfitting and weak generalization ability of single classifier are considered, ensemble Learning is applied into the decision tree algorithm in order to improve the classification accuracy. Random forest is the most representative algorithm among ensemble learning methods, and it can solve the bottleneck problem of the decision tree, which has a good scalability and parallelism to high-dimensional data in classification. Therefore, the random forest algorithm is selected as the facial expression classifier in this paper.

The remainder of this paper is organized as follows. Section 2 describes previous works which are related to our work. Section 3 describes the proposed method in detail. Section 4 presents experiment results and analysis. Section 5 presents conclusions and future work.

2. Related Work

Generally, a face expression recognition system has three steps including face detection, feature extraction, and feature classification (Figure 1), where feature extraction is the most crucial step. The expression classification accuracy is largely dependent on the effectiveness of the extracted feature. If the features extracted from expressions make within-class distance as small as possible and between-class distance as large as possible, functions of the classifier will be demanded less.

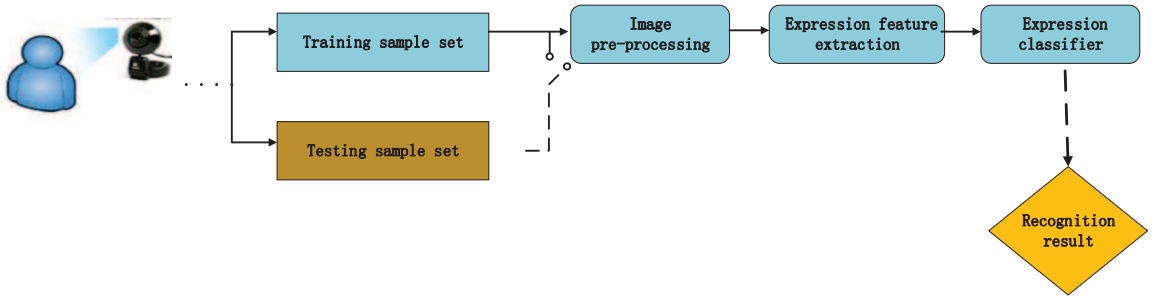


Figure 1. The general structure of facial expression recognition system.

Luo et al. [15] proposed a hybrid method of principal component analysis (PCA) and local binary pattern (LBP). Principal component analysis was used to extract the global grayscale features of the whole image, and LBP was used to extract the local features. The support vector machine (SVM) was used for facial expression recognition. In this paper, the preprocessing methods include geometry normalization, brightness normalization, histogram equalization, image filtering, and facial effective area segmentation. Each image of the training set was normalized into small size (24×24). The recognition result was 93.75%.

Chen et al. [16] applied the HOG to encode these facial components as features. A linear SVM was then trained to perform the facial expression classification. They evaluated the proposed method

on the JAFFE dataset and an extended Cohn–Kanade dataset. In the experiment of JAFFE, the size of the image was 256 × 256. After acquiring the face region from the face image, the size was adjusted to 156 × 156. The leave-one-sample-out strategy was used to test this method and compare with the other methods. The average classification rate on this dataset reached 94.3%. In the experiment of CK+, they divided the images into two sets. One was the training set and the other was the test set. About one-fifth of the images of each group were randomly selected for the test set. The remaining images were chosen as the training set. This method achieved an average of 88.7% with a variance of ±2.3% classification rate at last.

Although the above methods have achieved good recognition results, the data sets used in these experiments are small samples. Moreover, handcrafted feature is not comprehensive. The emergence of deep learning breaks the traditional pattern (feature extraction followed by facial expression classification), which deals with feature extraction and classification simultaneously. Convolutional neural network is the most widely used in image classification, which can map deeper information that can further improve the accuracy rate.

Mollahosseini1 et al. [17] proposed a deep neural network architecture to address the FER problem across multiple standard face datasets, viz. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013. The network included two elements (two traditional CNN and two "Inception" style modules), whereas inception style modules were made up of 1 × 1, 3 × 3, and 5 × 5 convolution layers (Using ReLU) in parallel. All the images were resized to 48×48 pixels for analysis. In order to augment the existed data, this paper extracted five crops of 40 × 40 from the four corners and the center of the image. They evaluated the accuracy of the proposed deep neural network architecture in two different experiments; viz. subject-independent and cross-database evaluation. In the subject-independent experiment, databases are split into training, validation, and test sets. The 5-fold cross validation technique was used to evaluate the results. In FERA and SFEW, the training and test sets were defined in the database release, and the results are evaluated on the database defined test set without performing K-fold cross validation. Accuracy rates are: 94.7%, 77.9%, 55.0%, 76.7%, 47.7%, 93.2%, and 66.4% on each data set respectively.

Wen et al. [18] proposed a method that integrated many convolutional neural networks with probability-based fusion for facial expression recognition. In the all designed CNN models, the softmax classifier was used in the last layer to estimate the probabilities of the testing sample belonging to each class. When many CNNs were generated as base classifiers, their probabilities' outputs were merged using the probability-based fusion method. Because the diversity among the base classifiers is regarded as a key issue in performance for any ensemble learning method, this paper applied the implicit method to generate CNNs with rich diversity. Four databases were used in their experiments, viz. JAFFE, CK+, Emotiw2015, FER2013. In experiments, they obtained 100 CNNs, whose accuracies ranged from 65% to 68% on FER2013-VAL. This paper illustrated that ensemble learning could be applied to further improve performance. It was also seen that both approaches did not obtain very good performance across databases. The main reason for this is that the training database for the approach was not large enough to contain all kinds of samples with rich diversity. Therefore, as many samples as possible should be included in the training database to achieve richer diversity. Furthermore, when the base classifiers were weak, the ensemble method failed to further improve performance.

However, in the above literature, the feature information extracted by machine learning algorithm is not comprehensive. A lot of training time for the CNN model is needed in experiments, and the hardware requirements are very high. Considering the above problems, this paper proposes a method that combines CNN feature with machine learning classifier for facial expression recognition, which can not only increase the coverage of extracted feature information but also can reduce the training time of the model.

3. Proposed Method

3.1. The Acquisition of Features Based on Convolutional Neural Network

Convolutional neural network (CNN) [19] is a supervised learning method that can perform the feature extraction and classification process simultaneously and can automatically discover the multiple levels of representations in data, which has been widely used in the field of computer vision. The general structure of the basic CNN model is shown in Figure 2. The detailed introduction of CNN model can be seen in [20,21].

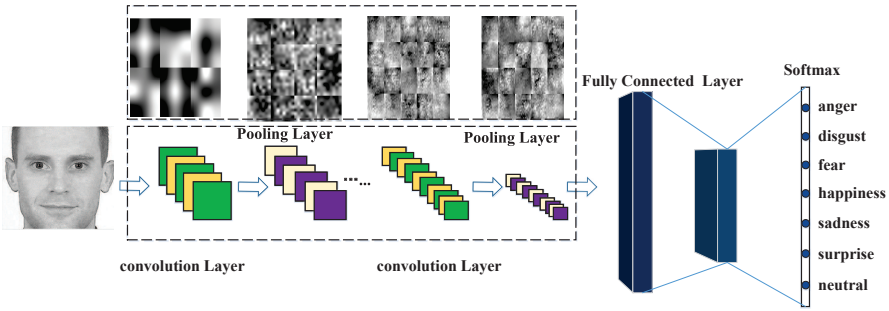


Figure 2. The general structure of convolutional neural network. Convolutional neural network mainly includes convolution layer, pooling layer, and full connection layer. The convolution layer is responsible for feature extraction, the pooling layer is used for feature selection, and the full connection layer is used for classification.

Selecting the CNN model that is used in a task and how to build an appropriate model for feature learning requires complexity analysis based on the current task. The main factors that affect the convolutional neural network include network's depth, the selection of different convolution kernels, the choice of the activation function and so on. Considering the research task of this paper, a simple CNN model framework is built and shown in Table 1.

Table 1. The diagram of convolutional neural network structure.

Layer	Input	Kernel Size	Output
Conv	96 × 96	5 × 5	92 × 92
Conv	92 × 92	5 × 5	88 × 88
Pool	88 × 88	2 × 2	44 × 44
Conv	44 × 44	3 × 3	42 × 42
Pool	42 × 42	2 × 2	21 × 21
Conv	21 × 21	3 × 3	19 × 19
Conv	19 × 19	3 × 3	17 × 17
Conv	17 × 17	5 × 5	13 × 13
Conv	11 × 11	2 × 2	5 × 5
FC			
Softmax			

This paper focuses on expression recognition by combining the random forest with features extracted from the CNN model and finds the most suitable fusion method based on the real experiment environment. The original data will have a certain information loss after going through every layer of the CNN model. When the original data reaches the full connection layer, the nature of the raw data has become distorted, therefore, random forests cannot be directly put on the CNN structure. The most reasonable way is to put the random forest into the last pool layer, which can be seen in Figure 3.

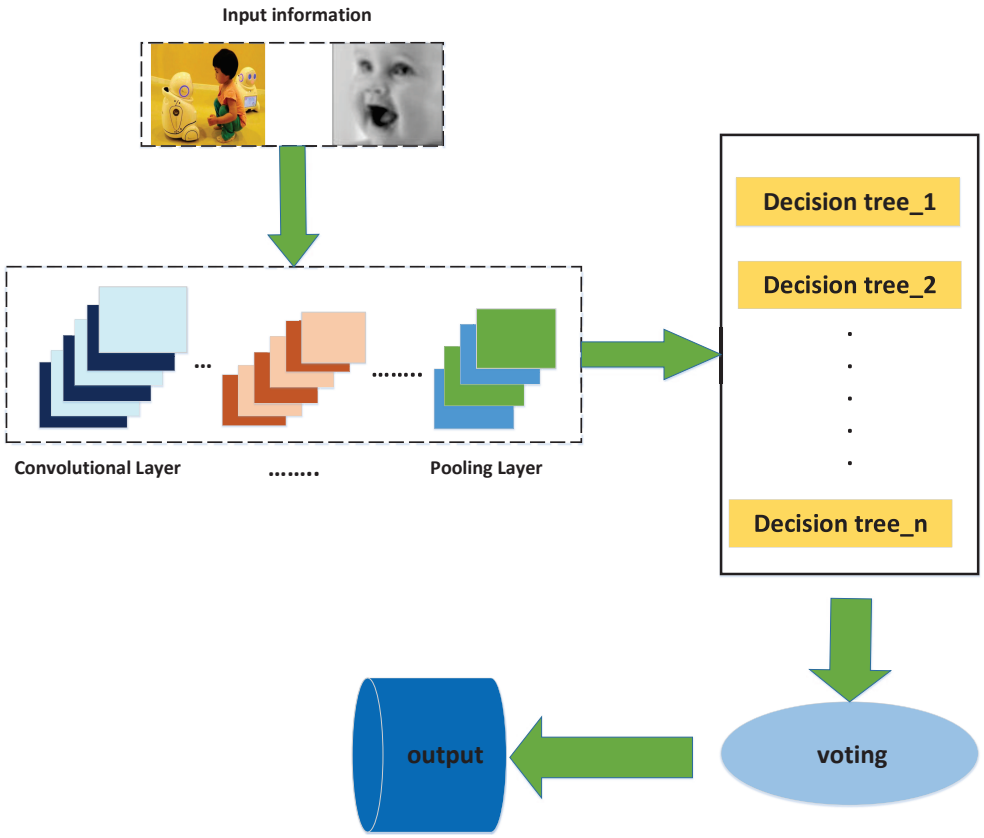


Figure 3. The structure of the new model. The former work of this model is the acquisition of convolutional neural network (CNN) features, while the latter work is the connection between CNN features and the improved random forest for facial expression classification.

3.2. Introduction and Improvement of C4.5 Decision Tree

Among decision tree algorithms, C4.5 classifier has been widely used in the field of image recognition due to its high computational efficiency and the ability to process a large number of data sets as well as its simple and easy characteristics. Therefore, C4.5 is selected as the classifier for facial expression recognition in this paper.

In the process of processing the classification of continuous values, the most core calculation of C4.5 classifier is the acquisition of information gain rate, which can be seen in Equation (1). The information gain rate not only affects the classification efficiency of decision tree but also determines the choice of nodes in the process of decision tree generation.

$$\begin{aligned} Gain(D, a_V) &= \max_{t \in T_a} Gain(D, a_V, t) \\ &= \max_{t \in T_a} \frac{Ent(D)}{IV(a_v)} - \frac{\sum\limits_{\lambda \in \{-1,1\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda)}{IV(a_v)} \end{aligned} \tag{1}$$

where,

$$IV(a_v) = -\sum_{v=1}^{V} \frac{|D^V|}{|D|} \log_2 \frac{|D^V|}{|D|}$$

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = -\frac{m_1}{\sum\limits_{i=1}^{7} m_i} \log_2 \frac{m_1}{\sum\limits_{i=1}^{7} m_i} - \frac{m_2}{\sum\limits_{i=1}^{7} m_i} \log_2 \frac{m_2}{\sum\limits_{i=1}^{7} m_i} - \ldots - \frac{m_7}{\sum\limits_{i=1}^{7} m_i} \log_2 \frac{m_7}{\sum\limits_{i=1}^{7} m_i}$$

$$Ent(D_t^\lambda) = -\frac{l_{t_1}^\lambda}{\sum\limits_{i=1}^{7} l_{t_i}^\lambda} \log_2 \frac{l_{t_1}^\lambda}{\sum\limits_{i=1}^{7} l_{t_i}^\lambda} - \frac{l_{t_2}^\lambda}{\sum\limits_{i=1}^{7} l_{t_i}^\lambda} \log_2 \frac{l_{t_2}^\lambda}{\sum\limits_{i=1}^{7} l_{t_i}^\lambda} - \ldots - \frac{l_{t_7}^\lambda}{\sum\limits_{i=1}^{7} l_{t_i}^\lambda} \log_2 \frac{l_{t_7}^\lambda}{\sum\limits_{i=1}^{7} l_{t_i}^\lambda}$$

Let us introduce some notation. $[a_1, a_2, a_3, ..., a_v, ..., a_V]$ stands for V attribute sets. $|y|$ represents the total number of categories. $IV(a_v)$ stands for the fixed value of the attribute a_v. $|D| = \sum_{i=1}^{|y|} m_i$ stands for the sample number in the whole dataset. $\left|D^V\right|$ stands for the sample number in attribute value a_v. Suppose there are n different values for attribute a_v, then sort these values from smallest to largest (i.e., $\{a_1^v, a_2^v, ..., a_n^v\}$). D can be divided into two different subsets (D_t^- and D_t^+) based on partition point t. Where, $T_a = [\frac{a_i^v + a_{i+1}^v}{2} | 1 \leq i \leq n-1]$, which means that the middle point $\frac{a^i + a^{i+1}}{2}$ of the interval $[a_i^v, a_{i+1}^v)$ is used as the division. Then, we can treat these points as discrete attribute value. Suppose $\bigcup_{i=1(\lambda=-1)}^{7} l_{t_i}^{\lambda}$ and $\bigcup_{i=1(\lambda=1)}^{7} l_{t_i}^{\lambda}$ stands for the number of seven expression labels in these two datasets D_t^- and D_t^+, respectively.

We find that logarithm operation appears very frequently in the process of computing the information gain rate [22] and almost exists in the operation process of each equation. A lot of logarithm operations will affect the computational speed of the system. This paper improves the formula of information gain by introducing Taylor series expansion, which simplifies the operation time and improves the real-time response of the system. The new equation of information gain rate can be seen as follows.

$$\begin{aligned} Ent(D) = & -\frac{m_1}{\sum_{i=1}^{7} m_i} \log_2 (1 - \frac{\sum_{i=2}^{7} m_i}{\sum_{i=1}^{7} m_i}) - -\frac{m_1}{\sum_{i=1}^{7} m_i} \log_2 (1 - \frac{m_1 + \sum_{i=3}^{7} m_i}{\sum_{i=1}^{7} m_i}) - \\ & ... - \frac{m_7}{\sum_{i=1}^{7} m_i} \log_2 (1 - \frac{\sum_{i=1}^{6} m_i}{\sum_{i=1}^{7} m_i}) \\ = & \frac{m_1 \times \frac{\sum_{i=2}^{7} m_i}{2}}{(\sum_{i=1}^{7} m_i) \ \ln 2} + \frac{m_2 \times (m_1 + \frac{\sum_{i=2}^{7} m_i)}{2}}{(\sum_{i=1}^{7} m_i) \ \ln 2} + ... + \frac{m_7 \times \frac{\sum_{i=1}^{6} m_i}{2}}{(\sum_{i=1}^{7} m_i) \ \ln 2} \end{aligned} \tag{2}$$

The new equation of information gain rate can be determined by applying Equation (3) to Equation (2).

$$
\begin{aligned}
Gain(D, a_V) &= \max_{t \in T_a} Gain(D, a_V, t) \\
&= \max_{t \in T_a} \frac{\frac{1}{(\sum_{i=1}^{7} m_i)^2 \ln 2}[m_1 \times \sum_{i=2}^{7} m_i + m_2 \times (m_1 + \sum_{i=2}^{7} m_i) + ... + m_7 \times \sum_{i=1}^{6} m_i]}{\frac{|D^V| \times \sum_{v=1}^{V-1} |D^v|}{|D|^2 \times \ln 2}} \\
&\quad - \frac{\sum_{\lambda \in \{-1,1\}} \frac{|D_t^\lambda|}{|D|} \frac{1}{(\sum_{i=1}^{7} l_{t_i}^\lambda)^2 \ln 2}[l_{t_1}^\lambda \times \sum_{i=2}^{7} l_{t_i}^\lambda + l_{t_2}^\lambda \times (l_{t_1}^\lambda + \sum_{i=3}^{7} l_{t_i}^\lambda) + ... + l_{t_7}^\lambda \times (\sum_{i=1}^{6} l_{t_i}^\lambda)}{\frac{|D^V| \times \sum_{v=1}^{V-1} |D^v|}{|D|^2 \times \ln 2}} \\
&= \max_{t \in T_a} \frac{|D|^2}{(\sum_{i=1}^{7} m_i)^2 \times |D^V| \times \sum_{v=1}^{V-1} |D^v|}[m_1 \times \sum_{i=2}^{7} m_i + m_2 \times (m_1 + \sum_{i=2}^{7} m_i) + ... + m_7 \times \sum_{i=1}^{6} m_i] \\
&\quad - \sum_{\lambda \in \{-1,1\}} \frac{|D_t^\lambda|}{|D|} \frac{|D|^2}{(\sum_{i=1}^{7} l_{t_i}^\lambda)^2 \times |D^V| \times \sum_{v=1}^{V-1} |D^v|}[l_{t_1}^\lambda \times \sum_{i=2}^{7} l_{t_i}^\lambda + l_{t_2}^\lambda \times (l_{t_1}^\lambda + \sum_{i=3}^{7} l_{t_i}^\lambda) + ... + l_{t_7}^\lambda \times (\sum_{i=1}^{6} l_{t_i}^\lambda)
\end{aligned}
\tag{3}
$$

Compared with Equation (1), the complicated log calculation is replaced by the four simple operations in Equation (3), which greatly improves the operation efficiency and the real-time performance of the system.

3.3. Generation of the New Random Forest

Considering that a single classifier is prone to overfit and the generalization ability of one classifier is weak, the ensemble learning method is introduced to improve the classification accuracy.

Random forest (RF) [23], which was proposed in 2001, is composed of multiple decision trees. In the idea of ensemble learning [24], the base learner should be "good but different", that is, individual learners should have a relatively good recognition rate that is different from the others. However, in the process of selecting a single decision tree, the number of the decision tree is set in advance. Decision trees are established randomly, and the final result is determined by voting in an integrated way. In the process of building many single decision trees by the traditional approach, decision trees may not be very different from each other or the recognition rate of the generated individual decision tree is not high, which will affect the final result. This paper proposes a probability selection-based method to determine all the acquired individual decision trees, which not only meets the requirements of good and different but also meets the requirements of diversity. The specific algorithm is shown in Algorithm 1.

Algorithm 1. Generate new random forest

Input: training set D; attribute set A
Output: multiple expression classification decision trees;

Count=0; number=0;
Create the root node node;
If all samples in D belong to the same category C , **then**,
    Mark node as class C leaf node, **return**,
end if
If A=ϕ, **OR** the sample values on A are the same, **then**
    Mark node as a leaf node and its category as the class with the largest number of samples, **return**
end if
For each attribute, information gain rate is calculated by Equation (2).
Select the optimal partition attribute from A, and assume that the test attribute A * has the highest information gain rate during the experiment.
Find the segmentation point of the attribute;
A new leaf node is separated from node a*;
If the sample subset corresponding to this leaf node is empty, then this leaf node is divided to generate a new leaf node, which is marked as the expression with the highest number.
Else,
    continue to split this leaf node;
end if;
One decision tree is created.
make the test sample into the established tree and calculate the recognition rate,
if accuracy<0.6, count=count,
else
    count=count+1;
end if
if count <M,
    repeat step(2)-step(22)
else
    count =M,
break;
end if
Set the threshold value δ
If random <δ
    Select the optimal decision tree from all the currently established decision trees as the alternative decision tree. number=number+1;
else
    The decision tree is randomly selected from all the currently established decision trees as an alternative decision tree. number=number+1;
if number<m,
    repeat step(29)-step(33)
if number=m,
    break
end if
All the selected decision trees are combined to form a random forest
The test samples are put into the random forest, and the classification results of each decision tree are collected. The results with the most votes will be used as the prediction classification of the current sample.

4. Experiments and Results

4.1. Database

Figure 4 shows the four databases that are used in this paper.

JAFFE database [25]: The Jaffe facial expression database was published in 1998. The database was created by 10 Japanese women who were asked to make a variety of facial expressions in a given background and then photographed by a camera. This is a relatively small data sample library, a total of 213 facial expression images that were produced by 10 women. There are seven expressions, such

as: disgust, anger, fear, happy, sad, surprised, and neutral. Each person has three to four images for one expression label, respectively.

CK+ database [26]: CK+ database is an extension of the Cohn–Kanade database, which was released in 2010. The CK+ database has more data than the JAFFE database, which includes 123 subfolders, with a total of 593 expression images. The information in the last image of each sequence contains classification labels, and 327 images have expression classification labels. This database is one of the most widely used in the field of facial expression recognition.

FER-2013 database [27]: The Facial Expression Recognition 2013 database includes 35,887 different images. The training set consists of 28,709 examples. There are two test sets: public test set and private test. The public test is set used for selecting the optimal CNN model, which consists of 3589 examples. The private test set is used for verifying the accuracy rate of the optimal CNN model, which consists of 3589 examples.

RAF-DB database [28]: The Real-world Affective Faces Database is a large-scale facial expression database with about 30,000 facial images. Images downloaded from the Internet in this database are of great variability in terms of subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions, (e.g., glasses, facial hair, or self-occlusion), post-processing operations (e.g., various filters and special effects), etc. RAF-DB database has two folders: basic and compound. The basic folder is used in this paper.

Figure 4. Some examples of four experimental databases. Every database has seven different emotions respectively.

4.2. Data Augmentation

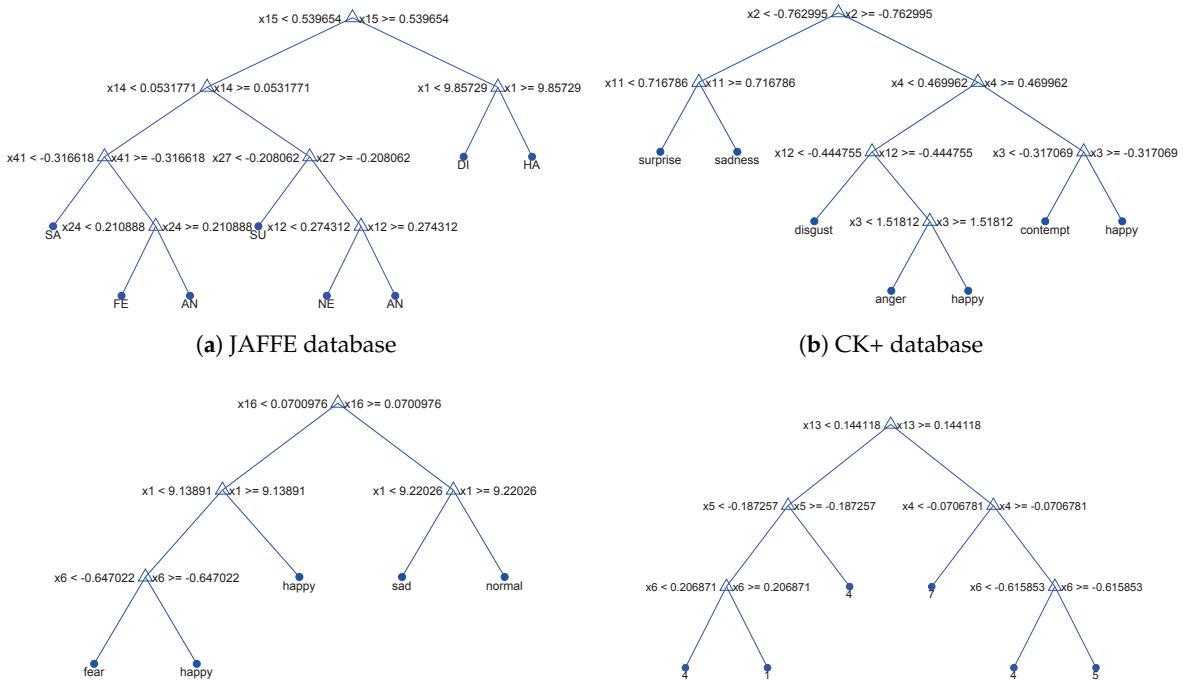
When the CNN model is chosen as the recognition model, the larger the amount of original data there is, the higher the accuracy and the generalization ability of the trained model. Therefore, data augmentation algorithm is very important, especially for some data sets with uneven distribution. A large training database is the advantage of training a good model. There are some common methods for data enhancement, such as: rotating the image, cutting the image, changing the color difference of the image, distorting the image features, changing the size of the image, and enhancing the image noise. The numbers of samples about these four databases are shown in Table 2.

Table 2. The experimental number of four databases: Extended Cohn–Kanade Dataset (CK+), JAFFE, FER2013, and Real-world Affective Faces Database (RAF-DB).

Ck+ Expression Label	Number	JAFFE Expression Label	Number
anger	5941	anger	4840
contempt	2970	disgust	4840
disgust	9735	fear	4842
fear	4125	happy	4842
happy	12,420	neutral	4840
sadness	3696	sad	4841
surprise	14,619	surprise	4840
FER2013 Expression Label	**Number**	**RAF-DB Expression Label**	**Number**
anger	4953	1	1619
normal	6198	2	355
disgust	547	3	877
fear	5121	4	5957
happy	8989	5	2460
sadness	6077	6	867
surprise	4022	7	3204

4.3. Results

For JAFFE database and CK+ database, 70% of the augmented data is taken as the training set and 30% of the data is taken as the test set. For FER2013 database and RAF-DB database, the training set and the testing set are used by the existing samples. Figure 5 shows some decision tree models that were generated in our experiments. Table 3 shows the experimental results obtained by different feature extraction methods and classification methods in this paper. Figure 6 shows the obvious differences between these different methods.



(**a**) JAFFE database (**b**) CK+ database

(**c**) FER2013 database (**d**) RAF-DB database

Figure 5. Some decision tree models that were generated in experiments.

Table 3. The experimental results obtained by different feature extraction methods and different classifiers.

Method (JAFFE)	Accuracy(%)	Running Time (s)
hog+C4.5	55.7	**689.5**
hog+an improved C4.5	74.6	**384.6**
cnn	97.3	**11,940.2**
cnn+one decision tree	95.3	**13,229.4**
cnn+random forest	96.7	**13,158.9**
cnn+new random forest	98.9	**12,715.5**
Method (CK+)	**Accuracy(%)**	**Running Time (s)**
hog+C4.5	61.0	**2595.3**
hog+an improved C4.5	59.7	**1597.7**
cnn	99.9	**19,604.4**
cnn+one decision tree	96.6	**25,398.6**
cnn+random forest	97.6	**22,369.1**
cnn+new random forest	99.9	**20,606.4**
Method (FER2013)	**Accuracy(%)**	**Running Time (s)**
hog+C4.5	46.1	**2061.6**
hog+an improved C4.5	45.9	**1290.8**
cnn	59.2	**14,720.3**
cnn+one decision tree	58.8	**17,880.5**
cnn+random forest	67.1	**17,601.9**
cnn+new random forest	84.3	**15,860.5**
Method (RAF-DB)	**Accuracy(%)**	**Running Time (s)**
hog+C4.5	51.2	**351.6**
hog+an improved C4.5	57.5	**140.8**
cnn	82.6	**6509.2**
cnn+one decision tree	79.8	**7122.5**
cnn+random forest	90.2	**6997.8**
cnn+new random forest	92.3	**6729.1**

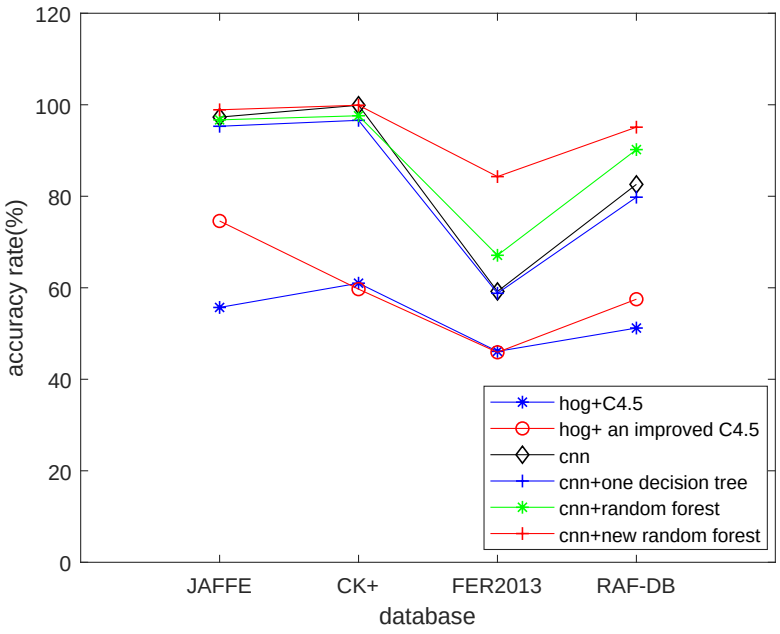
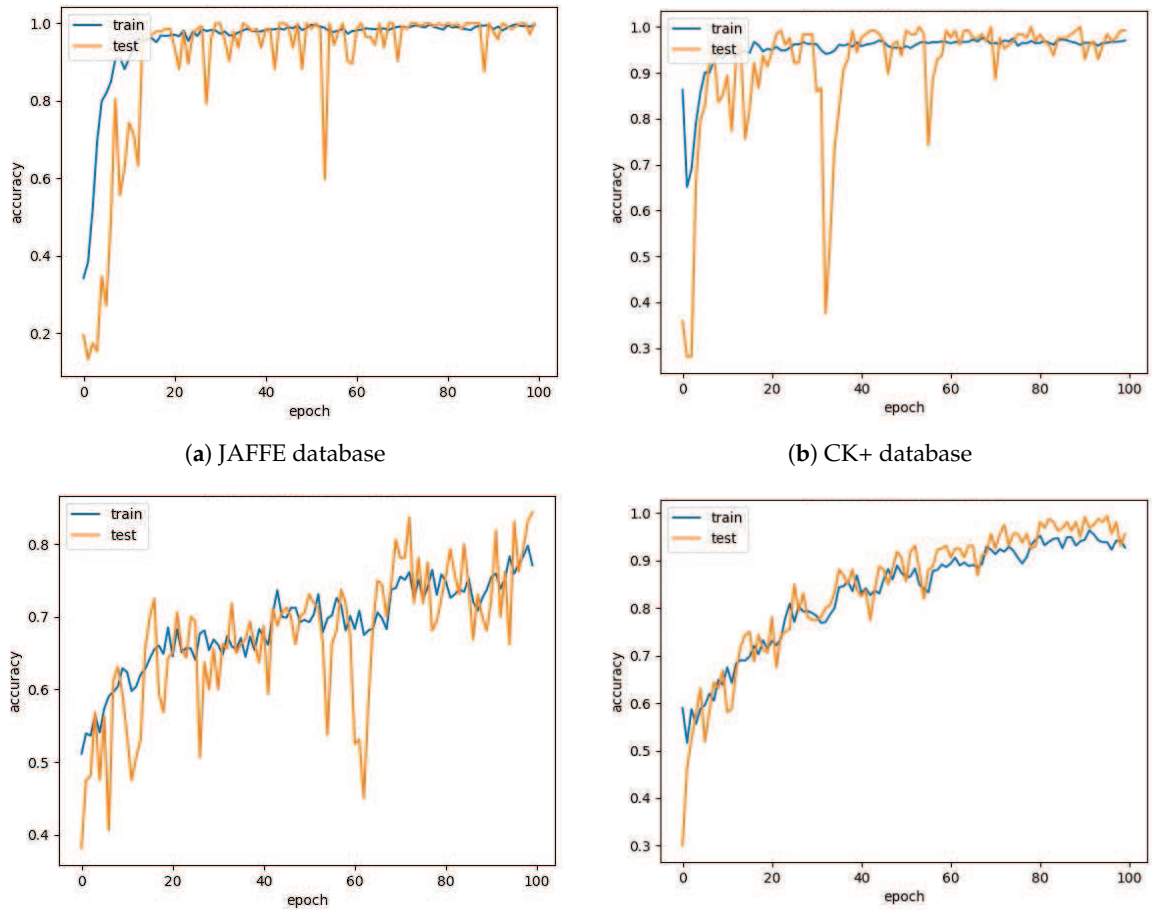


Figure 6. The accuracy rate comparison of the six different methods based on four databases.

As can be seen from the results in Figure 6, the expression classification ability based on "handcrafted" feature is generally lower than that based on CNN feature under the same classifier. The recognition performance of random forest based on probability selection is higher than that of traditional random forest. The experimental curves can be seen in Figure 7. Figure 8 shows the comparison of classification capabilities obtained by the new method that was proposed in this paper and other methods' performance on JAFFE database, CK+ database, FER2013 database, and RAF-DB database, which verifies that this method has certain advantages over other methods.



(**a**) JAFFE database (**b**) CK+ database

(**c**) FER2013 database (**d**) RAF-DB database

Figure 7. Experimental results based on four databases.

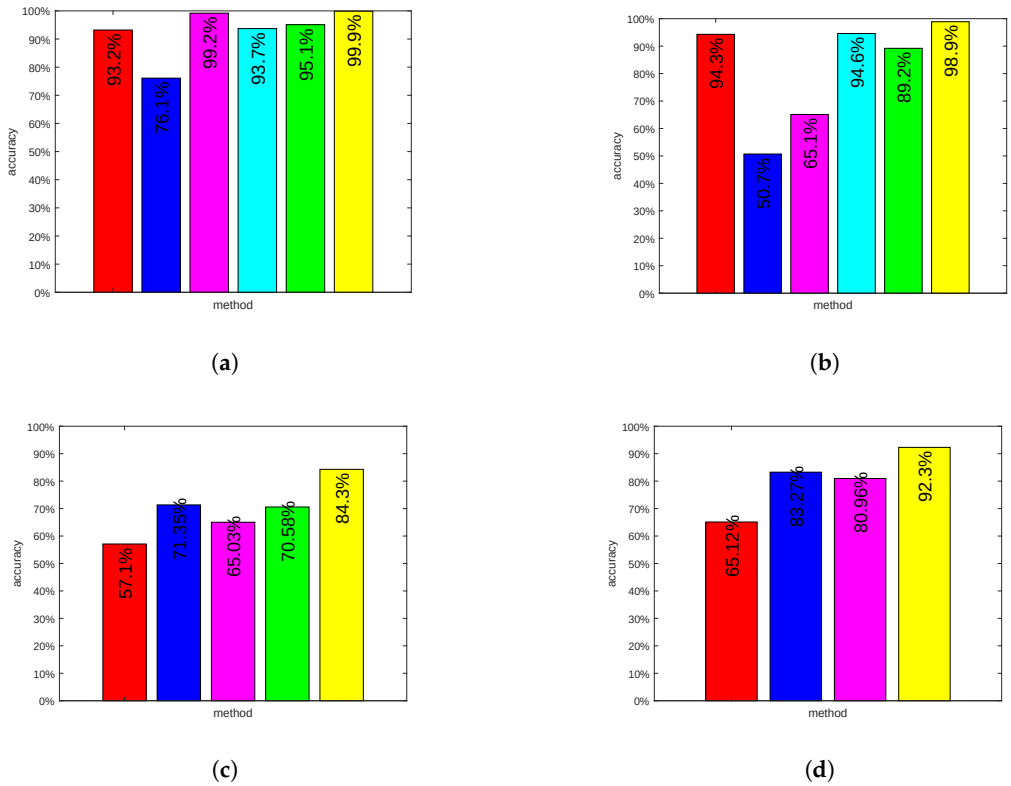


Figure 8. The comparison between the new method and some state-of-the-art methods in papers on four experimental databases. (**a**) JAFFE database. From left to right, results come from [17,18,26,28,29] and the new method proposed in this paper; (**b**) CK+ database. From left to right, results come from [16,18,27,30,31] and the new method proposed in this paper; (**c**) FER2013 database. From left to right, results come from [32–35] and the new method proposed in this paper; (**d**) RAF-DB database. From left to right, results come from [36–38] and the new method proposed in this paper.

5. Conclusions

Human beings, as advanced animals, often communicate emotions through rich and different expressions most of the time. Expression is a very sensitive problem: we can quickly perceive other people's situation and inner activities by observing their different expressions, and it is necessary to observe the changes of other people's expressions in interpersonal communication. Only in this way can we better understand each other and quickly know other people's emotions and thoughts.

In the era of rapid development in science and Internet technology, the demand for human–computer interaction in life has increased quickly. If researchers can make a breakthrough on the question of the recognition of human emotions, this will be a leapfrog development in the intelligent era of human–computer interaction.

Considering the complexity and other issues of the system, this paper does not make experiments based on some special conditions such as make-up and occlusion. Further research is needed on how to recognize facial expressions under these extreme conditions. In addition, for the convolutional neural network, it is necessary to collect as many samples as possible and make the trained network have a good generalization performance.

Author Contributions: Y.L. and Y.W. conceived the research and conducted the simulations; Y.W. designed and implemented the algorithm; Y.S. analyzed the data, results and verified the theory; X.R. collected a large number of references and suggested some good ideas about this paper; all authors participated in the writing of the manuscript.

Funding: This work was supported by the National Nature Science Foundation of China Grant Nos. 61673245 and 61573213.

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

References

1. Mehrabian, A. Communication without words. *Psychol. Today* **1968**, *2*, 4.
2. Darwin, C.; Ekman, P. Expression of the emotions in man and animals. *Portable Darwin* **2003**, *123*, 146.
3. Ying, Z.; Fang, X. Combining LBP and Adaboost for facial expression recognition. In Proceedings of the International Conference on Signal Processing, Beijing, China, 26–29 October 2008.
4. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef] [PubMed]
5. Kalansuriya, T.R.; Dharmaratne, A.T. Facial image classification based on age and gender. In Proceedings of the Fourteenth International Conference on Advances in ICT for Emerging Regions, Colombo, Sri Lanka, 10–14 December 2014.
6. Pang, Y.; Liu, Z.; Yu, N. A new nonlinear feature extraction method for face recognition. *Neurocomputing* **2006**, *69*, 949–953. [CrossRef]
7. Lopes, A.T.; Aguiar, E.D.; Souza, A.F.D.; Oliveira-Santos, T. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. *Pattern Recog.* **2017**, *61*, 610–628. [CrossRef]
8. Liu, W.; Song, C.; Wang, Y. Facial expression recognition based on discriminative dictionary learning. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba Science City, Japan, 11–15 November 2012.
9. Ali, G.; Iqbal, M.A.; Choi, T.S. Boosted NNE collections for multicultural facial expression recognition. *Pattern Recog.* **2016**, *55*, 14–27. [CrossRef]
10. Zhang, Z.; Lyons, M.; Schuster, M.; Akamatsu, S. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998.
11. Bartlett, M.S.; Littlewort, G.; Frank, M.; Lainscsek, C. Recognizing facial expression: machine learning and application to spontaneous behavior. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005.
12. Qi, W.; Jaja, J. From Maxout to Channel-Out: Encoding Information on Sparse Pathways. In Proceedings of the 24th International Conference on Artificial Neural Networks, Hamburg, Germany, 15–19 September 2014.
13. Quinlan, J.R. Induction on decision tree. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
14. Setsirichok, D.; Piroonratana, T.; Wongseree, W.; Usavanarong, T.; Paulkhaolarn, N.; Kanjanakorn, C.; Sirikong, M.; Limwongse, C.; Chaiyaratana, N. Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. *Biomed. Signal Process. Control* **2012**, *7*, 202–212. [CrossRef]
15. Luo, Y.; Wu, C.; Zhang, Y. Facial expression recognition based on fusion feature of PCA and LBP with SVM. *Optik-Int. J. Light Electron Optics* **2013**, *124*, 2767–2770. [CrossRef]
16. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition based on facial components detection and hog features. In Proceedings of the International Workshops on Electrical and Computer Engineering Subfields, Istanbul, Turkey, 22–23 August 2014; pp. 884–888.
17. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going Deeper in Facial Expression Recognition using Deep Neural Networks. In Proceeding of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7–9 March 2016.
18. Wen, G.; Zhi, H.; Li, H.; Li, D.; Jiang, L.; Xun, E. Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition. *Cognit. Comput.* **2017**, *9*, 1–14. [CrossRef]
19. Lawrence, S.; Giles, C.L.; Tsoi, A.C.; Back, A.D. Face recognition: a convolutional neural-network approach. *IEEE Trans. Neural Netw.* **1997**, *8*, 98–113. [CrossRef] [PubMed]
20. Kim, P. Convolutional Neural Network. In *MATLAB Deep Learning*; Apress: Berkeley, CA, USA, 2017.
21. Ming, L.; Hu, X. Recurrent convolutional neural network for object recognition. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.

22. Xu, M.; Wang, J.L.; Chen, T. Improved Decision Tree Algorithm: ID3 +. In *Intelligent Computing in Signal Processing and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 141–149.
23. Butaye, J.; Jacquemyn, H.; Hermy, M. Differential colonization causing non-random forest plant community structure in a fragmented agricultural landscape. *Ecography* **2001**, *24*, 369–380. [CrossRef]
24. Oza, N.C. Online Ensemble Learning. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, Austin, TX, USA, 30 July–3 August 2000.
25. Shih, F.Y.; Chuang, C.F.; Wang, P.S.P. Performance comparisons of facial expression recognition in JAFFE database *Int. J. Pattern Recog. Artif. Intell.* **2011**, *22*, 445–459. [CrossRef]
26. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010.
27. Kankanamge, S.; Fookes, C.; Sridharan, S. Facial analysis in the wild with LSTM networks. In Proceedings of the 25th IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018.
28. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.* **2009**, *27*, 803–816. [CrossRef]
29. Liu, M.; Li, S.; Shan, S.; Chen, X. Au-inspired deep networks for facial expression feature learning. *Neurocomputing* **2015**, *159*, 126–136. [CrossRef]
30. Ahmed, H.; Rashid, T.; Sidiq, A. Face Behavior Recognition through Support Vector Machines. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 101–108.
31. Rashid, T.A. Convolutional Neural Networks based Method for Improving Facial Expression Recognition. In *The International Symposium on Intelligent Systems Technologies and Applications*; Springer: Cham, Switzerland, 2016; pp. 73–84.
32. Tumen, V.; Soylemez, O.F.; Ergen, B. Facial emotion recognition on a dataset using convolutional neural network. In Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 16–17 September 2017.
33. Chang, T.; Wen, G.; Yang, H.; Ma, J.J. Facial Expression Recognition Based on Complexity Perception Classification Algorithm. *arXiv* **2018**, arXiv:1803.00185.
34. Kuang, L.; Zhang, M.; Pan, Z. Facial Expression Recognition with CNN Ensemble. In Proceedings of the International Conference on Cyberworlds, Chongqing, China, 28–30 September 2016.
35. Kim, B.K. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *J. Multimodal User Interfaces* **2016**, *10*, 173–189. [CrossRef]
36. Li, S.; Deng, W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Trans. Image Process.* **2018**, *28*, 356–370. [CrossRef]
37. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef]
38. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.