algorithms

MDPI

Article

Facial Expression Recognition Based on Auxiliary Models

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.); rongxw@sdu.edu.cn (X.R.)
[2] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China; songyong@sdu.edu.cn
* Correspondence: liyb@sdu.edu.cn

Received: 20 September 2019; Accepted: 28 October 2019; Published: 31 October 2019

Abstract: In recent years, with the development of artificial intelligence and human–computer interaction, more attention has been paid to the recognition and analysis of facial expressions. Despite much great success, there are a lot of unsatisfying problems, because facial expressions are subtle and complex. Hence, facial expression recognition is still a challenging problem. In most papers, the entire face image is often chosen as the input information. In our daily life, people can perceive other's current emotions only by several facial components (such as eye, mouth and nose), and other areas of the face (such as hair, skin tone, ears, etc.) play a smaller role in determining one's emotion. If the entire face image is used as the only input information, the system will produce some unnecessary information and miss some important information in the process of feature extraction. To solve the above problem, this paper proposes a method that combines multiple sub-regions and the entire face image by weighting, which can capture more important feature information that is conducive to improving the recognition accuracy. Our proposed method was evaluated based on four well-known publicly available facial expression databases: JAFFE, CK+, FER2013 and SFEW. The new method showed better performance than most state-of-the-art methods.

Keywords: expression recognition; human–computer interaction; sub-regions; ensemble

1. Introduction

Facial expression recognition plays a great role in human–computer interaction. In the course of human communication, 55% of the information is conveyed by different facial expressions, voice constitutes 38% of a communicated message, and language only constitutes 7% [1], therefore facial expression recognition has attracted much attention in recent years [2,3], and has many important applications in, e.g., remote education, safety, medicine, psychology and human–robot interaction systems. Although great progress has been made [4], it is difficult to acquire a facial expression recognition system with a satisfactory accuracy rate due to a variety of complex external conditions such as head pose, image resolution, deformations, and illumination variations. Hence, facial expression analysis is still a challenging work.

Generally, facial expression recognition is composed of three steps: preprocessing, feature extraction and classification [5]. The image preprocessing plays two roles: firstly, the system access to the original image is not generally perfect in practical application, such as the effects of noise, illumination and contrast, hence it is necessary to enhance the image processing with a view to increasing the quality requirements. Secondly, the acquired image information does not meet the specific requirements of subsequent operations, such as size and angle, therefore it is necessary to perform the image processing. The image preprocessing serves as a transition, which needs to be

considered comprehensively. 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. If features are inadequate, even the best classifier would fail to achieve good performance. Among machine learning algorithms, features are extracted by hand, such as local binary patterns (LBP) [7], Gabor [8], local Gabor binary patterns (LGBP) [9], scale invariant feature transforms (SIFT) [10], and histograms of oriented gradient (HOG) [11]. Handcrafted features such as LBP, HOG, and SIFT have been widely used in the traditional approach owing to their proven performances under specific circumstances and their low computational cost in the feature extraction process. After feature extraction, the classification method should be applied to perform facial expression recognition, such as SVM [12], random forest [13], sparse coding [14], neural network [15], etc. Although these methods have achieved great success in specific fields, the handcrafted feature [16] has its inherent drawbacks. When we use handcrafted features, either unintended features that have no effects on classification may get included or important features that have a great influence on the classification may get omitted. This is because the features are "crafted" by human experts, and the experts may not be able to consider all possible cases and include them in the feature. Meanwhile, it is difficult to realize a good recognition result for big datasets with large inter-personal differences in facial expression appearance.

To cope with the above disadvantages, deep learning methods [17,18] are considered, especially the emergence of convolutional neural networks. Convolutional neural network (CNN) [19] is a very effective method to recognize facial emotions. They can perform the feature extraction and classification process simultaneously, and can automatically discover the multiple levels of representations in data. This is why they succeed in breaking the most world records in recognition tasks.

The structure of early convolutional neural network was relatively simple. With the development of the relative research, the structure of convolutional neural network has been continuously optimized and its application field has been extended. In recent years, the research on the structure of convolutional neural network is still very hot, and some network structures with excellent performance have been proposed. The research results of convolutional neural network in various fields make it one of the most concerned research hotspots.

In the 1980s and 1990s, some researchers published relevant research work of CNN, and achieved good recognition results in several pattern recognition fields. Shan et al. proposed a CNN model called LeNet-5, and the success of handwriting character recognition about this model has aroused the attention of academia on convolutional neural network. However, CNN is only suitable for small image recognition at this time. For large-scale data, the recognition effect is not good. At the same time, Convolutional neural network is gradually developing in many fields such as speech recognition, object detection, face recognition etc. In 2012, Krizhevsky et al. used the extended CNN model (AlexNet) to win the champion in ImageNet Large Scale Visual Recognition Challenge (LSVRC) with a huge advantage of accuracy over the second place of 11%, making convolutional neural network become the focus of academia. Since AlexNet, many new convolution models have been proposed: such as the VGG (Visual Geometry Group) proposed by Oxford University, Google's GoogleNet, Microsoft's ResNet, etc., and these models have been constantly breaking AlexNet's ImageNet record.

Since Convolutional Neural Network (CNN) has already proved its excellence in many image recognition tasks, we expect that it can show better results than already existing methods in facial expression prediction problems. Most CNN -based facial recognition tasks use the entire face image as the input information, but what we found in our observations is that the judgment of facial expression is usually completed based on the information of several sensitive components in some areas of the face, such as eye, nose, and mouth. Other areas of the face contribute very little to the main feature of expression. If we use the entire face image to extract the features, the extracted feature vectors might lose some important information because it failed to catch the focus of the facial features. If the above feature information is adopted in the test experiment, it will make the test results very irrational, because the extracted feature information by using the above methods is greatly different from the real

feature information. This is often caused by two factors: (1) There are too few data. Unlike large scale visual object recognition databases such as ImageNet [17], most existing facial expression recognition databases do not have sufficient training data, which leads to the overfitting problem. (2) The learning mechanism of the single-task CNN network itself has some limitations, and problems cannot be solved by the only one CNN model.

Aiming at the above problems, some improvements have been proposed in this paper. To solve the problems caused by too few data, this paper divides some organ images that have important contributions to facial expression recognition from the raw images, which can not only improve the quantity of datasets, but also improve the quality of extracted information. Meanwhile, since it is difficult for the CNN model based on single task to improve the overall accuracy rate in the recognition task, this paper proposes a multi-task learning-based recognition model, which can modify the expression features extracted from the raw images with the help of the auxiliary model, so that the final extracted feature information is more in line with the ideal expression feature information.

The paper is arranged as follows: After this introduction, Related Work (Section 2) is presented which focuses on the various approaches with better performance that scientists have taken recently. Section 3 focuses on the main components of the architecture proposed in this paper. Section 4 presents the experiments and its results. Finally, Section 5 summarizes and concludes this paper.

2. Related Work

A detailed overview for expression recognition was given by Shan [20] and Cohen [21]. This section discusses some recent methods that achieve high accuracy in facial expression recognition using a comparable experimental methodology.

Shan et al. [20] proposed an approach called Boosted-LBP to extract the most discriminant LBP features, and the best recognition performance is obtained by using Support Vector Machine classifiers with Boosted-LBP features. They conducted experiments on the Cohn–Kanade database, MMI database and JAFFE database. They showed that the LBP-based SVMs perform slightly better than the Gabor-wavelet based SVMs by using the 10-fold cross-validation on each dataset.

S L et al. [22] proposed a method which uses Haar classifier for face detection purpose and Local Binary Pattern (LBP) histograms of different block sizes of a face image as feature vectors, and classifies various facial expressions using Principal Component Analysis (PCA). They used grayscale frontal face images of a person to classify six basic emotions, namely happiness, sadness, disgust, fear, surprise and anger.

Zhang et al. [23] proposed a novel facial expression recognition method using local binary pattern (LBP) and local phase quantization (LPQ) based on Gabor face image. Firstly, Gabor wavelets can capture the prominent visual attribute by extracting multi-scale and multi-direction spatial frequency features from the face images, which is separable and robust to illumination changes. Then, the LBP and LPQ feature based on the Gabor wavelet transform are fused for face representation. Considering the dimension of the fused feature is too large, the PCA-LDA algorithm is used to extract complex features. The method is finally tested and verified by multi-class SVM classifiers. This approach was implemented on JAFFE database. Two methods were used to test the effect of the classification. The first validation method was "leave one out". All expression images of one subject were selected as testing samples and the rest images as training samples, and it achieved a recognition accuracy of 81.82%. The other validation method was that two samples of each facial expression for each person were used to form the training set, and the remaining samples were used for testing. The proposed method showed the recognition rate of 98.57%.

Lisai et al. [24] proposed a novel algorithm for Facial Expression Recognition (FER), which is based on fusion of Gabor texture features and Local Phase Quantization (LPQ). Firstly, the LPQ feature and Gabor texture feature are, respectively, extracted from ever expression image. The image is first transformed by LPQ, and then divided into 3×5 blocks. Then, the LPQ histograms are calculated from each block. LPQ histograms of 15 blocks are concatenated into a long series of histogram as a

single vector. Then, Five scales and eight orientations of Gabor wavelet filters are used to extract Gabor texture features and adaboost algorithm is used to select Gabor features. Gabor features are obtained by 40 filters. Then, adaboost algorithm is used to select the 100 most effective features from each Gabor features image. Finally, the final concatenates the 4000 features from the 40 Gabor features images used as facial expression features. They obtained two expression recognition results on both expression features by Sparse Representation based Classification (SRC) method. Finally, the final expression recognition is performed by fusion of residuals of two SRC algorithms. The experiment results on Japanese Female Facial Expression (JAFFE) database demonstrated that the new algorithm was better than the original two algorithms, and this algorithm had a much higher recognition rate of 73.33%.

Minchul et al. [25] used the Convolutional neural network model to realize facial expression recognition. They cropped faces from each dataset and aligned the faces with respect to the landmark position of the eye, and the original 482 $\times$ 48 facial images were cropped into a size of 42 $\times$ 42. The training data were augmented 10 times by flipping them. Five types of data input (raw, histogram equalization, isotropic smoothing, diffusion-based normalization, and difference of Gaussian) were tested on four different network structures, respectively. They selected the one that showed the highest accuracy as a target structure for fine parameter tuning. For the performance evaluation, five different datasets were chosen: FER-2013, SFEW2.0, CK+ (extended CohnKanade), KDEF (Karolinska Directed Emotional Faces), and Jaffe. Finally, Tang's simple network with Hist-eq images was chosen as a baseline CNN model for further research.

Yu et al. [26] proposed a method that contains a face detection module based on the ensemble of three state-of-the-art face detectors, followed by a classification module with the ensemble of multiple deep convolutional neural networks (CNN). Each CNN model is initialized randomly and pre-trained on a larger dataset provided by the Facial Expression Recognition (FER) Challenge 2013. The pre-trained models are then fine-tuned on the training set of SFEW 2.0. To combine multiple CNN models, they presented two schemes for learning the ensemble weights of the network responses: by minimizing the log-likelihood loss, and by minimizing the hinge loss. Their proposed method achieved 55.96% and 61.29%, respectively, on the validation and test set of SFEW 2.0.

Heechul Jung et al. [27] proposed a new CNN method based on two different models. The first deep network model can extract temporal appearance features from image sequences, while the other deep network model can extract temporal geometry features from temporal facial landmark points. The faces in the input image sequences are detected, cropped, and rescaled to 64 $\times$ 64. IntraFace algorithm is used to extract facial landmark points, and accurate facial landmark points are provided consisting of 49 landmark points, including two eyes, a nose, a mouth, and two eyebrows. Finally, these two models are combined using a new integration method. Through several experiments on the CK+ and Oulu-CASIA databases, as well as many data by various data augmentation techniques, this new model showed that the two models cooperate with each other.

Most of the previous methods have processed the entire facial region as the input information, and pay less attention to the sub-regions of human faces, which will lead to a large difference between the extracted features and the expected features. If the extracted information obtained from the entire face image is not ideal, the final recognition result will be affected. Because the judgment of facial expression is usually based on the information of several sensitive components in some areas of the face, such as eye, nose, and mouth, this paper proposes a new method that combines several important sub-regions (i.e., eye, nose, and mouth) and the entire image, which not only can modify the extracted feature information of the entire image, but also can further improve the overall recognition rate of the system.

3. The Proposed Method

3.1. Data Pre-Pocessing

Images of most face databases include not only faces, but also much background information; therefore, removing background is an important step in face expression recognition preprocessing. Although there are many face databases online, the face regions in most of them are not cut out and cannot be directly used in face expression recognition experiments. If the uncropped image is directly used as the input image, it will not only bring a huge amount of calculation, but also affect the final expression recognition results. Therefore, whether the human face can be correctly detected has a great impact on the expression recognition accuracy. To improve the accuracy of face detection, this paper introduces the method of eye positioning to improve the result of face detection. Eyes are the most prominent facial feature; once their position in the face is relatively fixed, the entire useful face will be obtained easily. Because there is a certain distance between the eyes and the size of the face, many face detection algorithms have a strong dependence on the location of the eye, taking the location of the eye as an important step in the recognition process. Other prominent facial features, such as mouth, nose and eyebrows, can be easily obtained from fixed geometric relations after positioning eyes, so the accurate positioning of human eyes is helpful for the face positioning. The algorithm to obtain the cropped face region image is shown in Figure 1, and experimental results are shown in Figure 2.

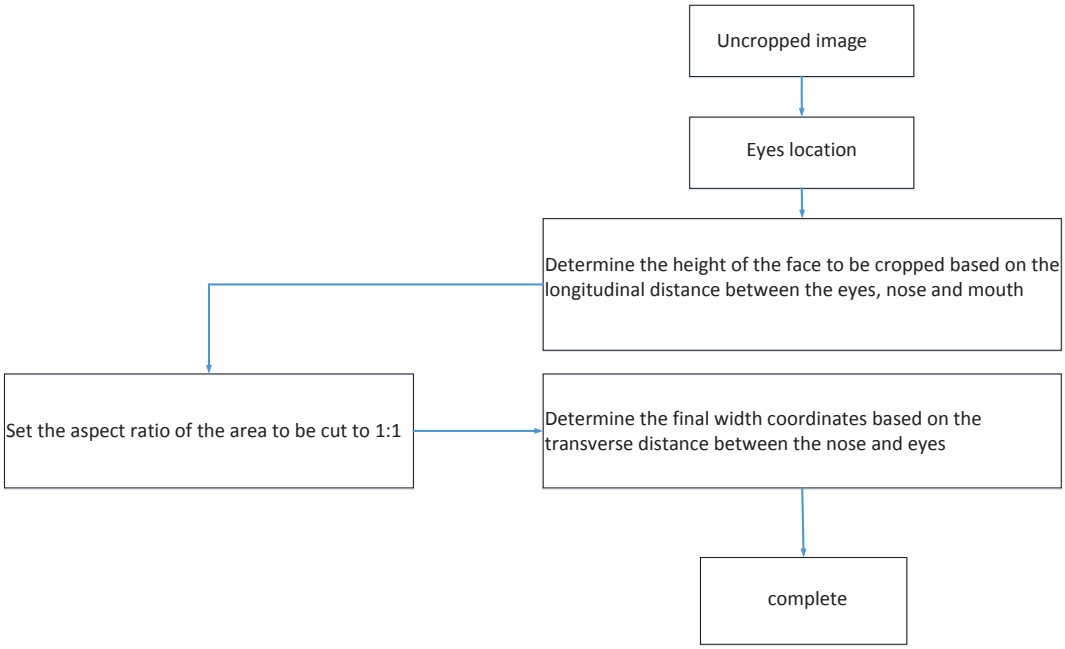


Figure 1. Algorithm to crop face region. An important pre-processing method is used in this paper. Eye orientation is the first step, then other important components are located based on the eye orientation, and finally the picture size scale of face area is set to 1:1.

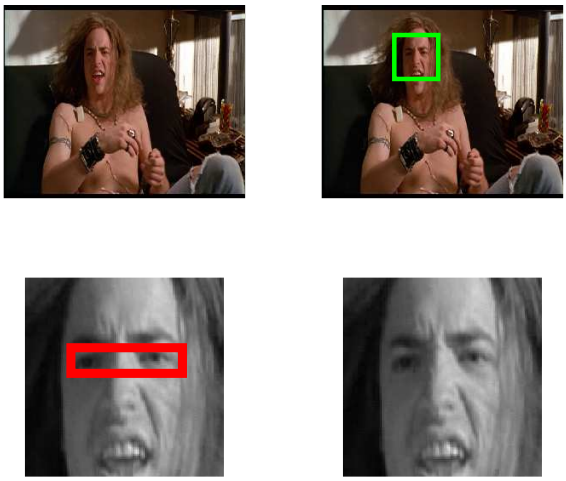


Figure 2. Examples of image pre-processing based on the algorithm of cropping face region: the top-left corner is the original facial expression image; the bottom left is the image of eye orientation; the top-right corner is the image of face positioning based on eye orientation; and the last one is the face region that used in experiments.

3.2. Convolutional Neural Networks

Convolutional neural network is a non-fully connected multi-layer neural network, which is generally composed of convolution layer (Conv), down-sampling layer (or pooling layer) and full-connection layer (FC). Firstly, the raw image is convoluted by several filters on the convolution layer, which can get several feature maps. Then, the feature is blurred by the down-sampling layer. Finally, a set of eigenvectors is obtained through a full connection layer. The architecture of Convolutional Neural Network is represented in Figure 3.

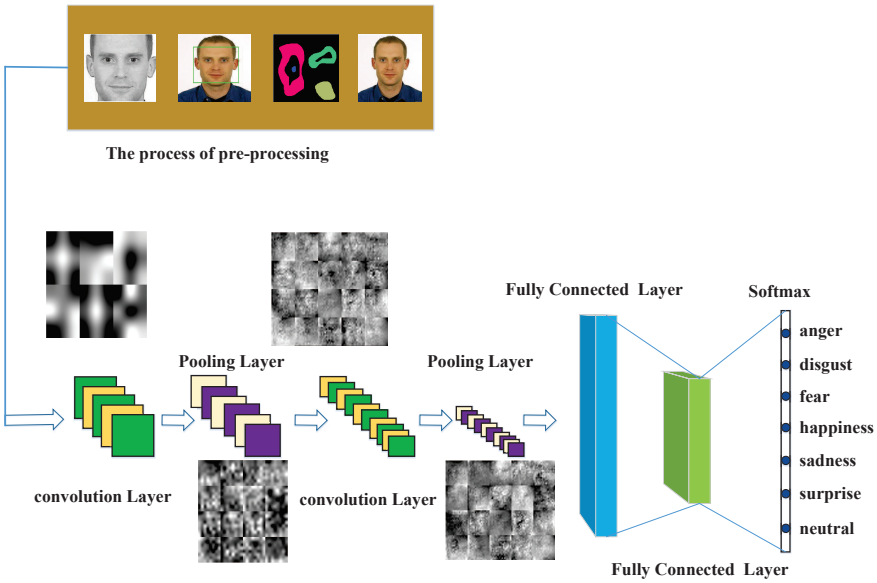


Figure 3. The structure of the common convolutional neural network. The convolutional neural network mainly includes convolution layer, pooling layer and full connection layer, and different layers have different functions. The convolution layer is responsible for feature extraction, the pooling layer is used for feature selection, and the full connection layer is used for classification.

Convolutional Layer: In convolutional layer, multiple convolutional kernels f_k with a kernel size n $\times$ m are applied to the input x to calculate a more rich and diverse representation of the input. It is not sufficient to have only one convolution kernel for feature extraction, hence multiple convolution kernels can be used in this step. If there are 50 convolution kernels, 50 features will be learned correspondingly. No matter how many channels there are in the input image, the total number of channels in the output image is equal to the number of convolution kernels. Figure 4 shows the process of computing the convolution region.

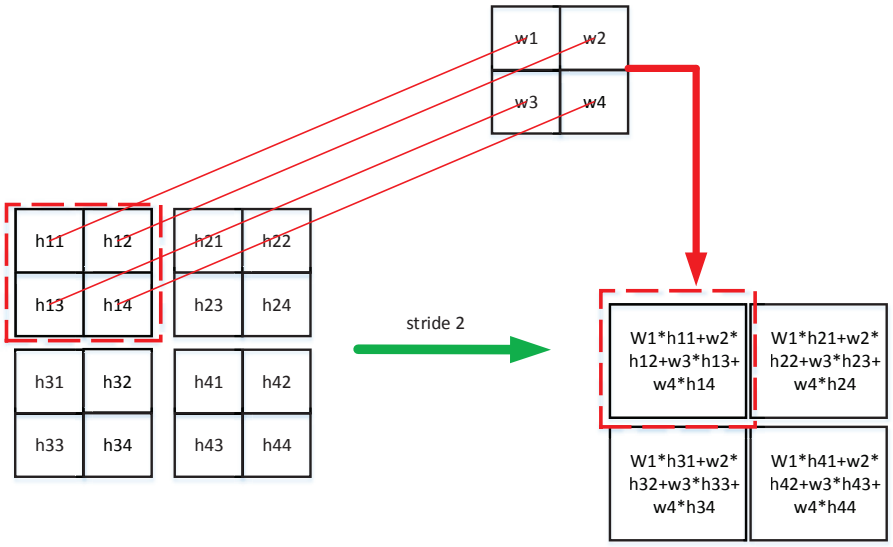


Figure 4. The basic operation of convolution layers. A new feature representation can be obtained by a certain operation, which can be used to obtain a deeper feature expression.The more convolution kernels there are, the more features can be learned. Different convolution kernels will produce different images.

Pooling Layer: The main function of the pooling layer is to lower sampling, and further reduce the number of parameters by removing unimportant samples in feature map. A large image can be downsized by the pooling layer, while retaining much important information. There are many methods for pooling operation: Max Pooling, Mean Pooling, etc. Max Pooling is the most commonly used method. In fact, Max Pooling is to take the maximum value of n $\times$ n samples as the sample value. Figure 5 is the computation process based on 2 $\times$ 2 Max Pooling.

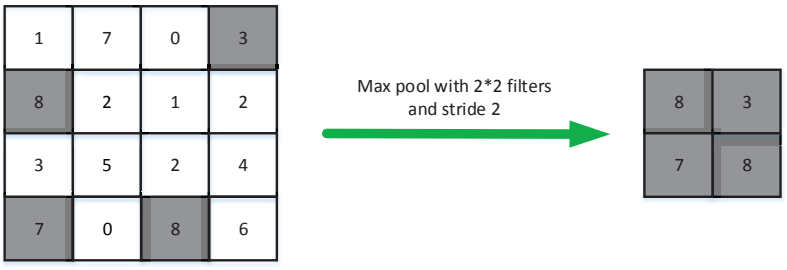


Figure 5. The computation process based on 2 $\times$ 2 Max Pooling.

Activation function: In the process of facial expression classification based on the convolutional neural network, the selection of an activation function plays a great role in the whole system, which is mainly used to introduce nonlinear factors. The sigmoid function, tanh function and rule function are commonly used. Relu function is more efficient than most other activation functions. It has a relatively cheap computation, because no exponential function has to be calculated. This function also can prevent the vanishing gradient error, since the gradients are linear functions or zero but in no case non-linear functions. Figure 6 shows the curve of this activation function.

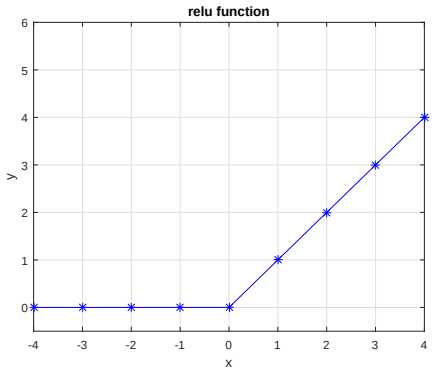


Figure 6. The curve of the relu function.

Fully Connected Layer: The fully connected layer connects all neurons of the prior layer to every neuron of its own layer.

3.3. The Acquisition of Some Important Components of the Face

When all datasets are ready, we align and crop regions of the two eyes, nose, mouth, and whole face. Then, four images are all resized into 96 × 96 pixels. Figure 7 shows a part of images of some sub-regions.

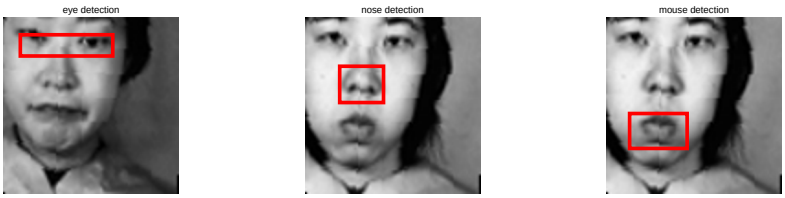


Figure 7. The part of some important components of the face image.

3.4. New Structure

In our daily life, the judgment of one facial expression is mostly based on several sensitive organs in some areas of the face, such as eyes, nose and mouth. Considering the advantages of ensemble learning and the importance of sensitive components' features in facial expression classification, this paper designs a new recognition system based on an auxiliary model. The structure of the model is shown in Figure 8.

There is no specific formula to build a convolutional neural network to ensure that it can work for all scenarios. Different problems require different architectures to produce our desired verification accuracy. Therefore, this paper designs a CNN structure for facial expression recognition according to the requirement of the research task. Figure 9 shows four different architecture of the designed CNN used in this task.

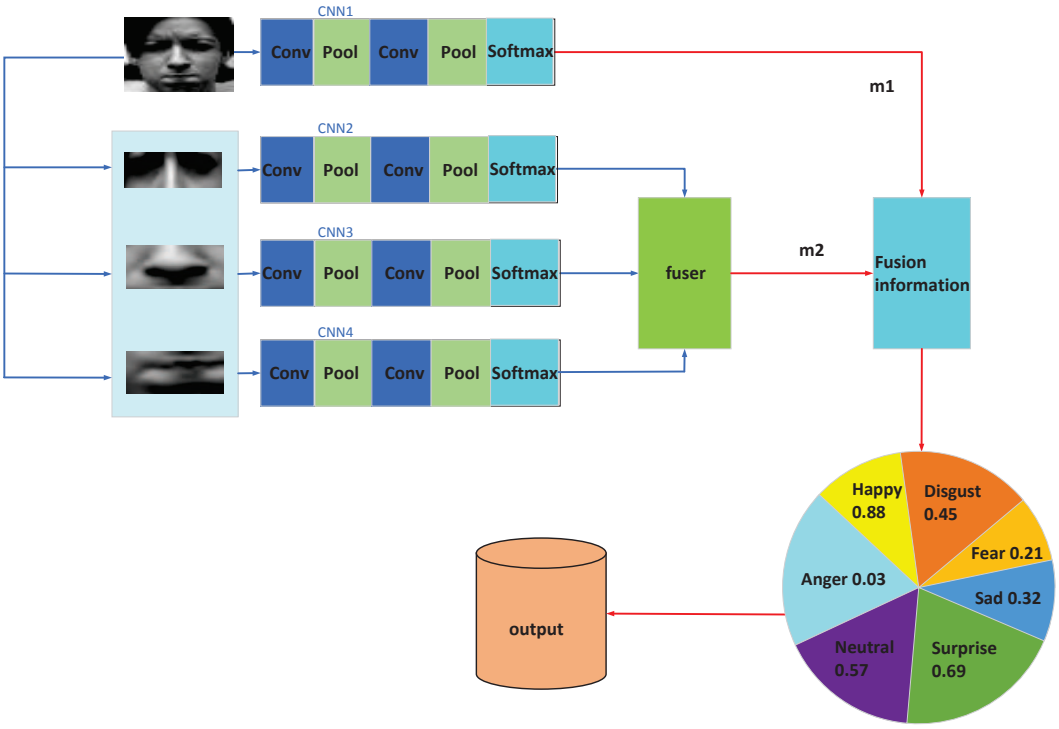


Figure 8. The structure of facial expression recognition system based on auxiliary model.

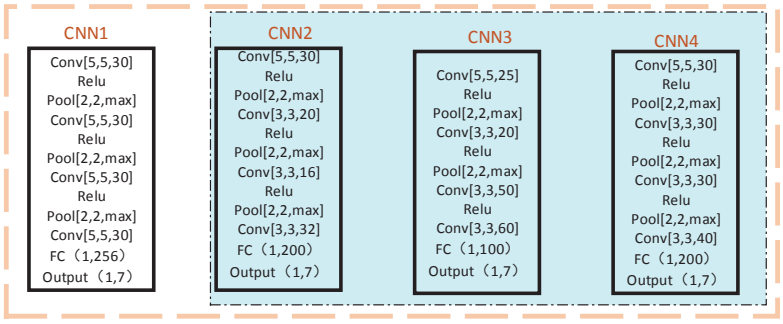


Figure 9. Different architectures of the four designed CNNs. The four models work in parallel. The probability vector expression of an expression is obtained for each model, and the final expression probability value is obtained through the weighted fusion algorithm.

The detailed algorithm of the fuser in Figure 8 is as follows. Let us introduce some notation: CNN_i $(i = 1, 2, 3, 4)$ stands for the name of four CNN model that used in this work. $p^i = [p_1^i, p_2^i, p_3^i, p_4^i, p_5^i, p_6^i, p_7^i]$ is a vector that has seven rows and one column, which stands for the probability of one CNN_i classifier assigned to seven classes. For example, p_2^1 stands for the probabilities that one test sample belongs to the first emotion class in CNN_2 model.

Then, p^i is normalized by the following equation: $p^i = \frac{p^i}{\max(p^i)}$. The final recognition result is determined by the following formula:

$y = \arg\max[m_1 \cdot p_1^i + m_2 \cdot \prod_{i=2}^{4} p_1^i, m_1 \cdot p_2^i + m_2 \cdot \prod_{i=2}^{4} p_2^i, ..., m_1 \cdot p_7^i + m_2 \cdot \prod_{i=2}^{4} p_7^i]$, where $m_1 = 1$ and the initial value of m_2 is 0.01. Seven values can be obtained from the above equation; the final recognition label can be get from the location of these seven values.

As shown in Figure 8, this new structure not only takes into account the strong abstract feature extraction ability for face images, but also takes into account the strong expression ability of important

components for the facial expression. Meanwhile, the application of the probability-based ensemble learning can further improve the performance of the system.

4. Experiments

4.1. Database

JAFFE database was published in 1998 [28], and it is a relatively small database. This database includes 213 images produced by 10 Japanese women, and each person has seven emotional images: disgust, anger, fear, happy, sad, surprise and neutral. Figure 10 shows parts samples of JAFFE.

Figure 10. Examples of images in the JAFFE database. The emotions from left to right are: anger, disgust, fear, happy, sad, surprise, and neutral.

CK+ database is expanded based on Cohn–Kanade database, which was published in 2010. This dataset was introduced by Lucey et al. [29]. There are 593 images and 123 sub-folders, and 327 images have their facial expression labels. This database is one of the most widely used in the field of facial expression recognition. There are 123 university students ranging from 18 to 30 years old, where 65% are female, 15% are African-American and 3% are Asian or South American. The emotions consist of anger, disgust, fear, happiness, sadness, surprise, and contempt. In our experiments, we used the first frame as the neutral category and the last four frames as one of the seven emotional categories for training the network as a frame-based classifier. Some examples of the CK+ database images are shown in Figure 11.

Figure 11. Examples of images in the CK+ database. The emotions from left to right are: anger, contempt, disgust, fear, happy, sadness, and surprise.

The Facial Expression Recognition 2013 (FER-2013) database [30] includes 35,887 different images. The training set consists of 28,709 examples. The public test set used for the leaderboard consists of 3589 examples. The private test set consists of another 3589 examples. The data consist of 48 $\times$ 48 pixel grayscale images of faces. Seven expressions are labeled in this database: normal, happy, sadness, surprise, anger, disgust, and fear. Some examples of the FER-2013 database images are shown in Figure 12.

Figure 12. Examples of images in the FER-2013 database. The emotions from left to right are: anger, disgust, fear, happy, normal, sadness and surprise.

SFEW database [31] is a part of a temporal facial expressions database acted facial expressions in the wild, which we extracted from movies. This database is close to the real world illumination. There are 958 images in the training set and 436 images in the validation set. Some examples of the SFEW database images are shown in Figure 13.

Figure 13. Examples of images in the SFEW database. The emotions from left to right are: anger, disgust, fear, happy, normal, sadness and surprise.

4.2. Data Augmentation

In deep learning, data enhancement is generally carried out on the database in order to enrich the training set and better extract facial expression features. The more original data there are, the higher the accuracy and generalization ability of the trained model will be; therefore, data enhancement is very important, especially for some datasets with uneven distribution. A good training dataset is the prerequisite of training an advanced model. When the training data are done well, it is often twice the result with half the effort in the following model training. However, data annotation is time-consuming, and it is hard to collect enough data. 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. In this study, the above methods were used to enhance the data of the original dataset. Finally, 33,885 images were produced in JAFFE, and about 4840 sample images were contained in each of seven expression folders. CK+ eventually obtained 53,506 images. The number of the experimental databases have been shown in Table 1.

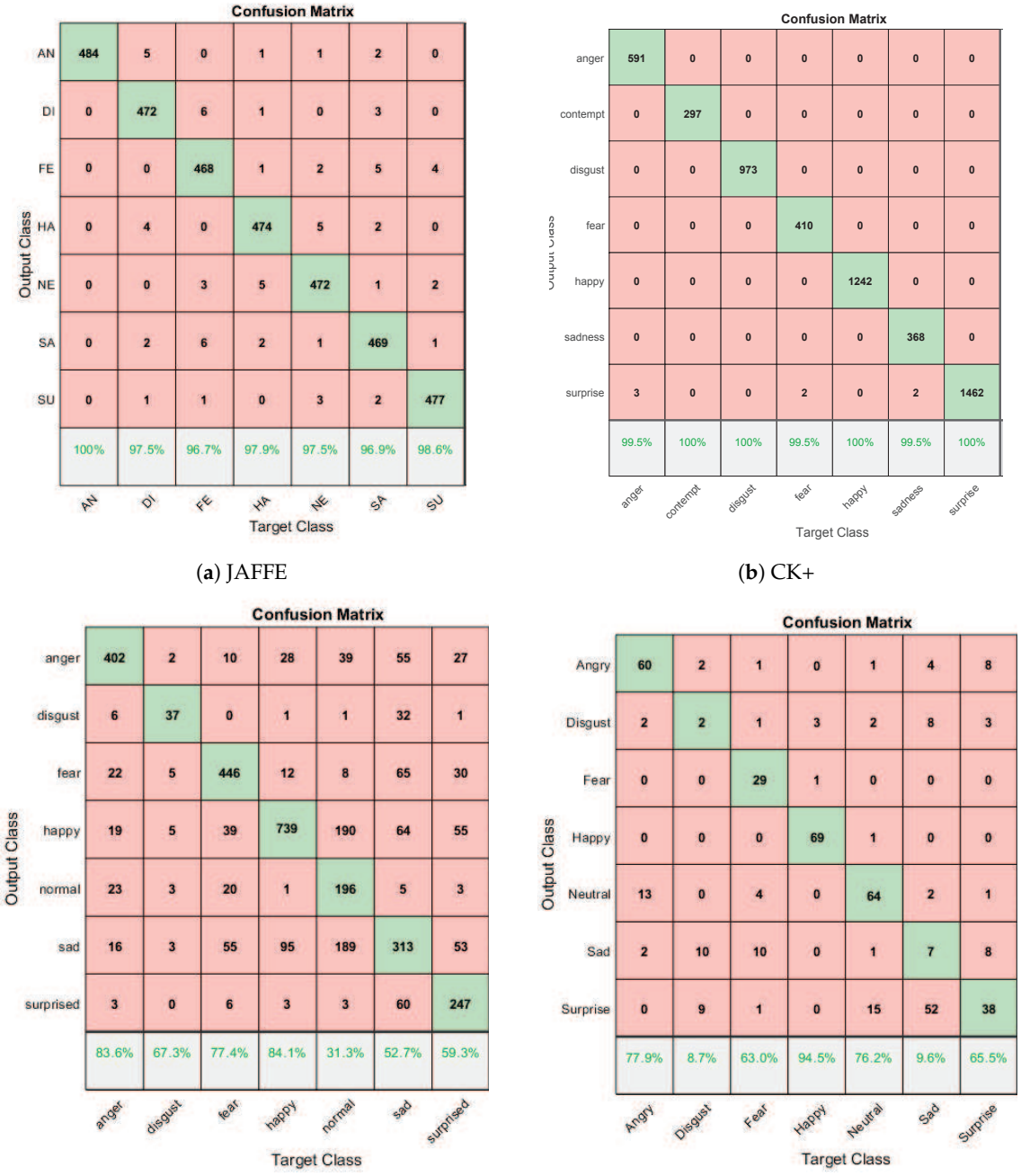
Table 1. The number of data augmentation on CK+, JAFFE, FER2013 and SFEW.

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**	**SFEW Expression Label**	**Number**
anger	4486	anger	244
contempt	491	disgust	73
disgust	4625	fear	123
fear	8094	happy	252
happy	5591	neutral	226
sadness	5424	sad	228
surprise	3587	surprise	152

4.3. Results

Figure 14 shows the confusion matrixes based on the new model. Furthermore, we compared the results for five different inputs (i.e., the whole face input, the combination of face and eyes, the combination of face and nose, the combination of face and mouth, and the whole face region based on sub-regions), as shown in Figure 15. In Figure 15, we can see that it is superior to the same model given in Figure 8 with only the entire face region as the input. Meanwhile, the new method proposed in this paper is still better than the current state of the art in emotion recognition on the JAFFE, CK+,

FER2013 and SFEW datasets, as can be seen in Table 2. In Table 2, we can see that nose has a small contribution to the final accuracy, and mouth has the biggest contribution to the accuracy rate.



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

(**c**) FER2013 (**d**) SFEW

Figure 14. Confusion matrixes based on the new method.

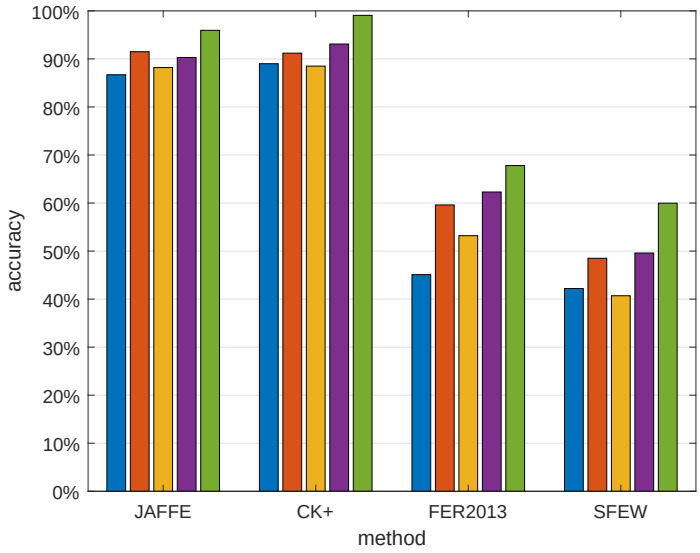


Figure 15. The experimental comparison of different combinations. From left to right, the methods are: the whole face input, the combination of face and eyes, the combination of face and nose, the combination of face and mouth, and the method proposed in this paper.

Table 2. This table summarizes the current state of the art in facial expression recognition on four databases.

Database	**Author**	**Method**	**Accuracy** (%)
CK+	Liu [32]	3DCNN	85.9
CK+	Jung [27]	DTNN	91.44
CK+	This paper	new method	99.07
JAFFE	Chen [33]	ECNN	94.3
JAFFE	Wen [34]	Probability-Based	50.7
JAFFE	This paper	new method	95.95
FER2013	Chen [33]	ECNN	69.96
FER2013	This paper	new method	67.7
SFEW	Li [35]	attention mechanism	53
SFEW	Liu [36]	Adaptive Deep Metric	54
SFEW	This paper	new method	59.97

5. Conclusions

In many cases, human beings communicate their emotions and intensions by their facial expressions, which is one of the most powerful, natural and immediate means. Facial expression analysis is an interesting and challenging task, and it has been applied in many fields such as human–computer interaction and remote education. Although much progress has been made in expression recognition field by researchers, it is not yet easily performed by computers or intelligent robots. In most research tasks, the whole face is used as the input information. In people's daily life, when one person judges the expression of the other person, they usually capture the characteristics of several key parts of the face to judge the final expression. The eyes, nose and mouth are some sensitive parts that play a decisive role in determining one's expression, while others play a small role in the final result. To solve the above problem, we propose a novel CNN framework based on the sub-region auxiliary model in this paper, which takes full advantage of three important regions, and modifies the learning results of the main task by setting different s to improve the final accuracy rate. In the experimental verification on JAFFE dataset, $m_1 = 1$ and $m_2 = 0.68$. In the experimental verification on

CK+dataset, m_1 = 1 and m_2 = 0.59. In the experimental verification on FER2013 dataset, m_1 = 1 and m_2 = 0.72. In the experimental verification on SFEW dataset, m_1 = 1 and m_2 = 0.63.

Future Work: The recognition accuracy of the system is improved through the auxiliary role of the sub-region model. In fact, there are many factors that affect the expression. In addition to the few special regions of the facial image, there are many other key factors that need to be studied continuously.

Author Contributions: Formal analysis, Y.L.; Funding acquisition, Y.S.; Methodology, Y.W.; Resources, X.R.

Funding: This research was funded by the National Nature Science Foundation of China Grant grant number 61673245 and 61573213.

Acknowledgments: 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*, 193–200.
2. Darwin, C.; Ekman, P. expression of the emotions in man and animals. *Portable Darwin* **2003**, *123*, 146.
3. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
4. Alshamsi, H.; Meng, H.; Li, M. Real time facial expression recognition app development on mobile phones. In Proceedings of the International Conference on Natural Computation, Changsha, China, 13–15 August 2016.
5. Jatmiko, W.; Nulad, W.P.; Matul, I.E.; Setiawan, I.M.A.; Mursanto, P. Heart beat classification using wavelet feature based on neural network. *WSEAS Trans. Syst.* **2011**, *10*, 17–26.
6. Kumar, B.V.; Ramakrishnan, A.G. Machine Recognition of Printed Kannada Text. *Lect. Notes Comput. Sci.* **2002**, *2423*, 37–48.
7. 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.
8. Gu, W.; Xiang, C.; Venkatesh, Y.V.; Huang, D.; Lin, H. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognit.* **2012**, *45*, 80–91. [CrossRef]
9. Ming, Y.U.; Quan-Sheng, H.U.; Yan, G.; Xue, C.H.; Yang, Y.U. Facial expression recognition based on LGBP features and sparse representation. *Comput. Eng. Des.* **2013**, *34*, 1787–1771.
10. Berretti, S.; Amor, B.B.; Daoudi, M.; Bimbo, A.D. 3D facial expression recognition using SIFT descriptors of automatically detected keypoints. *Vis. Comput.* **2011**, *27*, 1021. [CrossRef]
11. Wang, X.; Chao, J.; Wei, L.; Min, H.; Xu, L.; Ren, F. Feature fusion of HOG and WLD for facial expression recognition. In Proceedings of the IEEE/SICE International Symposium on System Integration, Kobe, Japan, 15–17 December 2013.
12. Yuan, L.; Wu, C.M.; Yi, Z. Facial expression recognition based on fusion feature of PCA and LBP with SVM. *Opt. Int. J. Light Electron Opt.* **2013**, *124*, 2767–2770.
13. Pu, X.; Ke, F.; Xiong, C.; Ji, L.; Zhou, Z. Facial expression recognition from image sequences using twofold random forest classifier. *Neurocomputing* **2015**, *168*, 1173–1180. [CrossRef]
14. Lin, Y.; Song, M.; Quynh, D.T.P.; He, Y.; Chen, C. Sparse Coding for Flexible, Robust 3D Facial-Expression Synthesis. *IEEE Comput. Graph. Appl.* **2012**, *32*, 76–88. [PubMed]
15. Meng, Z.; Ping, L.; Jie, C.; Han, S.; Yan, T. Identity-Aware Convolutional Neural Network for Facial Expression Recognition. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, Washington, DC, USA, 30 May–3 June 2017.
16. Liu, Y.; Li, Y.; Ma, X.; Song, R. Facial Expression Recognition with Fusion Features Extracted from Salient Facial Areas. *Sensors* **2017**, *17*, 712. [CrossRef] [PubMed]
17. Lv, Y.; Feng, Z.; Chao, X. Facial expression recognition via deep learning. In Proceedings of the International Conference on Smart Computing, Hong Kong, China, 3–5 November 2014; pp. 303–308.
18. Sun, W.; Zhao, H.; Zhong, J. A Complementary Facial Representation Extracting Method based on Deep Learning. *Neurocomputing* **2018**, *306*, 246–259. [CrossRef]

19. Li, H.; Jian, S.; Xu, Z.; Chen, L. Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network. *IEEE Trans. Multimed.* **2017**, *19*, 2816–2831. [CrossRef]
20. Shan, C.; Gong, S.; Mcowan, P.W. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
21. Cohen, I.; Sebe, N.; Garg, A.; Chen, L.S.; Huang, T.S. Facial expression recognition from video sequences: Temporal and static modeling. *Comput. Vis. Image Underst.* **2003**, *91*, 160–187. [CrossRef]
22. Happy, S.L.; George, A.; Routray, A. A real time facial expression classification system using Local Binary Patterns. In Proceedings of the International Conference on Intelligent Human Computer Interaction, Kharagpur, India, 27–29 December 2012.
23. Zhang, B.; Liu, G.; Xie, G. Facial expression recognition using LBP and LPQ based on Gabor wavelet transform. In Proceedings of the IEEE International Conference on Computer & Communications, Chengdu, China, 16–17 October 2016.
24. Li, L.; Ying, Z.; Yang, T. Facial expression recognition by fusion of Gabor texture features and local phase quantization. In Proceedings of the International Conference on Signal Processing, Hangzhou, China, 19–23 October 2014.
25. Shin, M.; Kim, M.; Kwon, D.S. Baseline CNN structure analysis for facial expression recognition. In Proceedings of the IEEE International Symposium on Robot & Human Interactive Communication, New York, NY, USA, 26–31 August 2016.
26. Yu, Z.; Zhang, C. Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In Proceedings of the Acm on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015.
27. 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 IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
28. Abdulrahman, M.; Eleyan, A. Facial expression recognition using Support Vector Machines. In Proceedings of the Signal Processing & Communications Applications Conference, Malatya, Turkey, 16–19 May 2015.
29. 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 & Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010.
30. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7–10 march 2015.
31. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011.
32. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014.
33. Chen, J.; Chen, Z.; Chi, Z.; Hong, F. Facial Expression Recognition in Video with Multiple Feature Fusion. *IEEE Trans. Affect. Comput.* **2018**, *9*, 38–50. [CrossRef]
34. 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]
35. 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] [PubMed]
36. Liu, X.; Kumar, B.V.K.V.; You, J.; Ping, J. Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.