applied sciences

MDPI

Article

Facial Expression Recognition Based on Multi-Features Cooperative Deep Convolutional Network

Haopeng Wu [1], Zhiying Lu [1,*], Jianfeng Zhang [2], Xin Li [1], Mingyue Zhao [1] and Xudong Ding [1]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; wuhpfree@tju.edu.cn (H.W.); xinlitu@tju.edu.cn (X.L.); mingyuezhao@tju.edu.cn (M.Z.); dingxd@tju.edu.cn (X.D.)

[2] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China; zhjf@tju.edu.cn

* Correspondence: luzy@tju.edu.cn

Abstract: This paper addresses the problem of Facial Expression Recognition (FER), focusing on unobvious facial movements. Traditional methods often cause overfitting problems or incomplete information due to insufficient data and manual selection of features. Instead, our proposed network, which is called the Multi-features Cooperative Deep Convolutional Network (MC-DCN), maintains focus on the overall feature of the face and the trend of key parts. The processing of video data is the first stage. The method of ensemble of regression trees (ERT) is used to obtain the overall contour of the face. Then, the attention model is used to pick up the parts of face that are more susceptible to expressions. Under the combined effect of these two methods, the image which can be called a local feature map is obtained. After that, the video data are sent to MC-DCN, containing parallel sub-networks. While the overall spatiotemporal characteristics of facial expressions are obtained through the sequence of images, the selection of keys parts can better learn the changes in facial expressions brought about by subtle facial movements. By combining local features and global features, the proposed method can acquire more information, leading to better performance. The experimental results show that MC-DCN can achieve recognition rates of 95%, 78.6% and 78.3% on the three datasets SAVEE, MMI, and edited GEMEP, respectively.

Keywords: classification; deep learning; facial expression recognition; 3D convolutional neural networks; sequence correlation processing; target location

Citation: Wu, H.; Lu, Z.; Zhang, J.; Li, X.; Zhao, M.; Ding, X. Facial Expression Recognition Based on Multi-Features Cooperative Deep Convolutional Network. *Appl. Sci.* **2021**, *11*, 1428. https://doi.org/10.3390/app11041428

Academic Editor: Shi-Jinn Horng
Received: 29 December 2020
Accepted: 27 January 2021
Published: 4 February 2021

1. Introduction

The face contains a large amount of information such as identity, age, expression and ethnicity. The amount of information contained in facial expressions in the communication process is second only to language [1]. Facial expressions are subtle signals of the communication process. Ekman et al. identified the six facial expressions (happiness, sadness, disgust, fear, angry and surprise) as basic facial expressions that are universal among human beings, while other researchers added neutral, which, together with the previous six emotions, constitutes seven basic emotions [2–5].

1.1. Related Work

Part of the research into this problem has focused on recognizing facial expressions in static images [6–8]. These methods use image features such as texture analysis. Although these approaches are effective methods in extracting spatial information, they fail to capture morphological and contextual variations in the expression process. Recent methods aim to solve this problem by using massive datasets to obtain more efficient features of FER [9–15]. Some researchers use multimodal fusion to recognize emotions, such as voices, expressions, and actions [16].

Recently, with the flourishing of the neural network, a few attempts have tried to use deep neural networks to replace the feature extraction [17]. Inferring a dynamic facial appearance from a single 2D photo is arduous and ill-posed, since the expression formation process blends multiple facial features (mouth, eyes) as well as environment (voice, lighting) into an expression for each moment. To better handle the transformation, one must rely on multi-part independent changes, such as a smile causing the corners of the mouth to rise.

As others show [18], FER can be solved using temporal image sequences and utilizing both spatial and temporal variations.

1.2. Motivations and Contributions

In this paper, we present a new algorithm that performs facial expression recognition with video and achieves satisfactory accuracy on standard datasets. Our work is inspired by an ensemble of regression trees [19] and 3-Dimension Convolutional Neural Networks (3DCNN) [20].

Some of the above algorithms have shown a promising performance in facial expression recognition. Notwithstanding, in terms of video with tiny movement, as well as a database with fewer video or pictures, face recognition is still a challenge. Therefore, exploiting the limited dataset to more effectively improve the recognition accuracy is a problem worth exploring.

This paper proposes using an ensemble of regression trees to annotate corresponding facial features. The network, which is called Multi-features Cooperative Deep Convolutional Network (MC-DCN), captures global and local features by using two small 3DCNNs. These parallel networks provide a better balance between global and local features. The batch normalization was connected behind the convolutional layer to adapt to the characteristics of a small sample dataset.

The algorithm uses a cascaded network to address FER problem. The main contributions of this work can be summarized as follow:

- Firstly, the ensemble of regression trees is applied to achieve facial location. Furthermore, an alternative to the attention mechanism is added. The influence of different facial organs on expressions was analyzed, and the features of facial organs selected. This net can extract the contours of face accurately. The application of facial features allows the network to be fully trained under tiny movements. Meanwhile, the weight analysis of different organs and the entire face can effectively improve the recognition ability of expressions which are not obvious;
- Secondly, a new network was proposed, called Multi-features Cooperative Deep Convolutional Network (MC-DCN), that can dynamically obtain expression features from image sequences. The network combines the face part and the local feature map, and can sense the deformation process and trends of important expression features excellently. Meanwhile, a part called the CNN block is used. The CNN block is improved on the basis of Resnet, which means that the network as a whole has a stronger generalization ability, enhances the compatibility of the algorithm in different scenarios, and improves recognition accuracy.

The rest of this paper is organized as follows: In Section 2, the source of the datasets is given. Section 3 details the entire framework of the algorithm. Qualitative and quantitative experimental results, obtained from three public datasets, are shown in Section 4.

2. Materials

To evaluate the network, this paper conducted extensive experiments on the three popular facial expression recognition datasets, as shown in Table 1: the Surrey Audiovisual Emotions (SAVEE) [21–23], MMI facial expression database [24] and GEneva Multimodal Emotion Portrayals (GEMEP) [25].

2.1. SAVEE Database

The database was captured in The Centre for Vision, Speech and Signal Processing (CVSSP) 3D vision laboratory over several months during different times of the year from four actors. It contains a total of 480 short videos which were recorded by four actors showing seven different emotions. The length of these videos varied from 3 to 5 s, and they include anger, disgust, fear, happiness, neutral, sadness, and surprise. Classification accuracy for visual and audio-visual data for seven emotion classes over four actors by evaluators is given in Table 1. KL, JK, JE, DC in the table are the abbreviation of the actor's names.

Table 1. Average human classification accuracy (%).

Modality	KL	JE	JK	DC	Mean (±CI)
Visual	89.0	89.8	88.6	84.7	88.0 ± 0.6
Audio-visual	92.1	92.1	91.3	91.7	91.8 ± 0.1

2.2. MMI Database

MMI has more than 2900 videos, including a total of 236 videos containing emoticons. Each video contains a complete change process. MMI contains six basic emotions (except neutral) and many other action descriptors which are activated by the Facial Action Coding System (FACS). It begins with Neutral, goes through a series of onset, apex, and offset phases, and returns to a neutral face. Expert assessment was used to expand the emotional video of neutral for the database. Some of the videos in MMI are recorded by dual cameras at the same time, and only one of them was chosen for the video.

2.3. GEMEP Database

As a database for FERA Challenge, GEneva Multimodal Emotion Portrayals (GEMEP) contains more than 7000 audio-video portrayals recorded by 10 different actors, representing 18 emotions with the help of professional theater directors. GEMEP was restructured to unify standards; details will be introduced below.

2.4. Data Augmentation

In order to enable the network to obtain sufficient training and parameter adjustments, the training set database will be horizontally flipped and rotated with tiny angles (specifically including the following angles: $\pm 9^{\circ}$, $\pm 6^{\circ}$, $\pm 3^{\circ}$) and other data enhancement methods.

3. Methodology

A facial expression recognition approach was proposed from a video that employs 3D convolution nets (C3D) framework and ensemble of regression trees (ERT) framework. A detailed description of the algorithm is given in this section. Figure 1 is a flowchart of the algorithm proposed in this paper. In order to obtain features from the dynamic video, we extracted the images in the video to generate an image sequence. This can be done by using two 3DCNNs, which consist of five convolutional layers with an ReLU activation function. Batch normalization was followed by every convolution layer. Further, we combined the two networks by an element-wise average of the output of the fully connected layers, which was then connected to a final softmax layer for classification. The forecast result is represented by (1)

$$y_i = \text{softmax}(p_i) = \frac{\exp(p_i)}{\sum_{j=1}^{N} \exp(p_j)} \quad (1)$$

where p_i is a input sample and y_i is ith type of emotion.

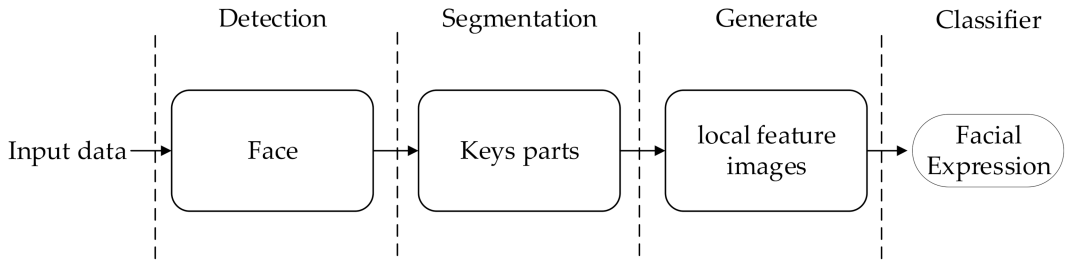


Figure 1. The algorithm flow chart proposed in this paper. The input video data obtain the face image through the face detection module. Then, the region of interest is selected to generate key parts feature image. Finally, the face image and the generated feature image are sent to the network for expression classification.

These local parts were extracted from face in each frame; each of these shallow subnetworks were trained on global-local features. Subnetworks used the CNN Block. Finally, all the fully trained subnetworks were integrated for fine-tuning. The network can comprehensively learn dynamic changes in time and in the global–local features of space.

3.1. Face Alignment with an Ensemble of Regression Trees

Generally, people's complete (or partial) body and surrounding environment were shown in the original video, especially for the database that emphasizes body movements. Therefore, to eliminate body and environment as much as possible, while ensuring the integrity of the face, this paper adopts the ensemble of regression trees to preprocess the database in order to obtain information about critical parts of the face [17]. Each part of the face makes a different contribution to FER, and we hope that the part that contributes more to FER can be used to train the network [26].

By learning and combining these features of each critical part, tree-based local binary features (LBF) use linear regression to detect them. Different from LBF, an ensemble of regression trees (ERT) stored the updated value of the shape directly into the leaf node during their process of learning. The mean shape plus of all passing leaf nodes can obtain the final facial key point position after learning all the trees at the initial position, as shown in (2)

$$S^{(t+1)} = S^{(t)} + r_t\left(I, S^{(t)}\right) \tag{2}$$

where t is the number of cascade layers, and $r_t(\bullet)$ denotes the current regressor. The input parameters of the regressor are the shape of the image I updated by the previous regressor. The features used can be grayscale or other [27].

In order to train all the $r_t(\bullet)$, the gradient tree boosting algorithm is used to reduce the sum of the squared errors of the initial shape and the ground truth. Each regressor consists of many trees. A pixel pair was selected randomly to ensure these parameters of these trees by the coordinate difference between the current shape and ground truth. As the regressor is updated, the initial estimated shape $S^{(0)}$ will eventually be updated to the true shape $S^{(t+n)}$. Algorithm 1 [19] is the update algorithm of the regressor. It is assumed that the input image is I, learning rate $v \in (0, 1)$.

Algorithm 1. The ERT algorithm is used to form the initial position S passes through all the trees and mean shapes plus the ΔS of the passed leaf nodes to obtain the final key point position of the face.

1: Initialize regression function f_0 through the I,S
2: Set up the regression of each layer
3: Fit a regression tree to the targets r_{ik} giving a weak regression function $g_k\left(I, \hat{S}^{(t)}\right)$
4: Update regression function
5: Get the final regression $r_t(\bullet)$

The ERT method pays more attention to the contour of the face, ignoring the distinctive information of texture and wrinkles in facial expressions. In order to improve the performance of extracting facial features, the proposed algorithm combines the detection of texture and wrinkles with ERT. Figure 2, which was generated by style Generative Adversarial Networks (styleGAN) [28], shows that the wrinkles on the eyebrows and forehead that clearly reflect the characteristics of expression vary with age.

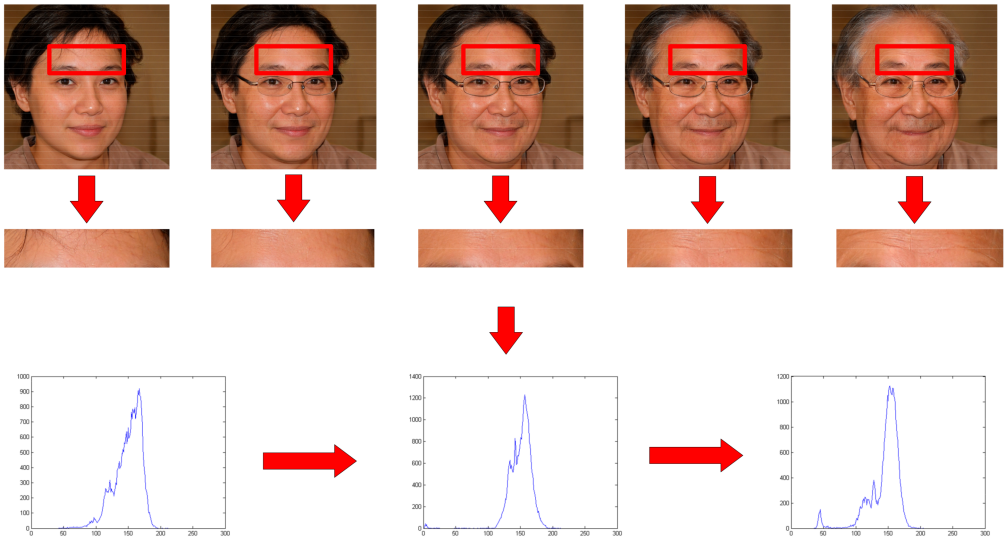
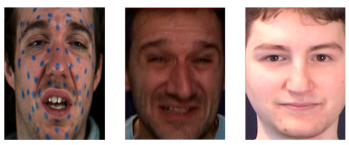
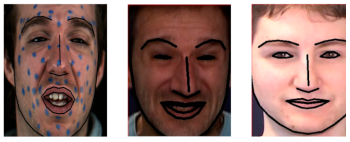
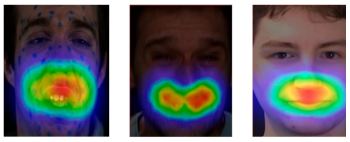


Figure 2. The appearance of the same person at different ages, and the forehead of each period, the bottom is the histogram of the forehead.

Thus, detecting brow and forehead wrinkles deteriorates the generalization ability of the model. To acquire better generalization ability, the proposed scheme detects wrinkles on the corners of mouth and nasolabial folds rather than the eyebrow and forehead. Facial features are shown in Table 2.

Table 2. Facial landmark.

Preprocessing with Image	
Facial	
Features landmark	
Heatmap	
Facial landmark	

3.2. Multi-Features Cooperative Deep Convolutional Network

3.2.1. The Architecture of Net

Recently, the dynamic recognition method has become the backbone of many FERs due to the spatiotemporal characteristics of this task. At present, the methods of the task are mainly concentrated in two aspects. The first is using the optical flow method to find the dynamic trend of the target. Secondly, using the 3D net in the C3D network as an example, a three-dimensional convolution kernel, which is used to convolve the spatiotemporal blocks formed by the video to capture the dynamic features of expressions.

This paper draws on the above two methods. Two parallel 3D networks were used to extract global and local features, respectively. This algorithm first aligns faces in videos and performs size normalization on each face. Secondly, each 16 frames of face images and local feature images were used to generate spatiotemporal blocks with the same size. Thirdly, two 3D networks with the same structure performed feature extraction on different spatiotemporal blocks, respectively. Finally, the average layer was used to merge feature generated by two networks. The framework is shown in Figure 3.

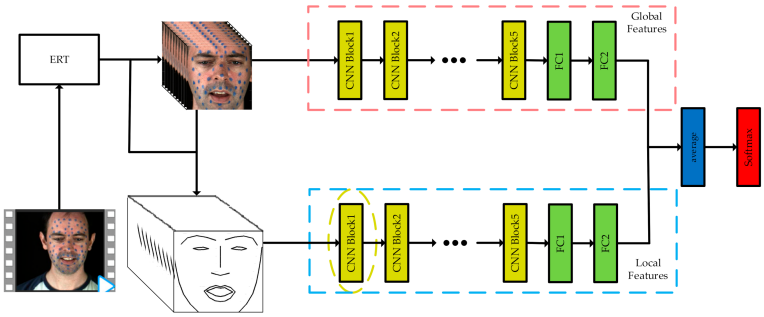


Figure 3. The proposed MC-DCN network architecture. The face data are divided into a face part and a local feature part. They are sent to two identical parallel networks, which are finally connected to an average pooling layer. the CNN block will be introduced in detail in Section 3.2.2.

It can be observed in Figure 3 that input data was preprocessed with a dual-channel as input for the 3D net

$$I = (X, \hat{X}) = (x_i, \hat{x}_i)_{i=1}^{n} \tag{3}$$

where X is the input image sequence, and n is the total number of labeled-image sequence. Considering the 3D convolutional layer, Batch Normalization, Relu, and a pooling layer as a convolutional block, (1) can be represented by (4):

$$y = \text{softmax}\left(h\left(f(X)_{i=1}^{m}, g\left(\hat{X}\right)_{i=1}^{m}\right)\right) \tag{4}$$

where $f(\bullet)$ and $g(\bullet)$ are the convolution block calculation processes, m is the number of convolution blocks, and h is the fusion calculation process of these two networks. Without considering Batch Normalization, $f(\bullet)$ and $g(\bullet)$ give similar results. $f(\bullet)$, as an example, can be expressed as

$$P^{(l)} = f\left(P^{(l-1)}\right) = W^{(l)} \times R\left(P^{(l-1)}\right) + b^{(l)} \tag{5}$$

where P is the output of one of these layers. W, b are the weight coefficients of the layer, R is the calculation of activation function and pooling layer, and l is the number of layers. Further considering Batch Normalization [29], (5) is rewritten as (6).

$$\widetilde{P}^{(l)} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\left(P^{(l)}\right)^{(i)} - \frac{1}{m}\sum_{i=1}^{m}\left(P^{(l)}\right)^{(i)}\right)^2} \times \frac{P^{(l)} - \frac{1}{m}\sum_{i=1}^{m}\left(P^{(l)}\right)^{(i)}}{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\left(P^{(l)}\right)^{(i)} - \frac{1}{m}\sum_{i=1}^{m}\left(P^{(l)}\right)^{(i)}\right)^2 + \varepsilon}} + \frac{1}{m}\sum_{i=1}^{m}\left(P^{(l)}\right)^{(i)} \tag{6}$$

The network is formed by stacking five convolutional blocks. The kernel of the convolution layer is $3 \times 3 \times 3$, the step size is $1 \times 1 \times 1$. The initial pool core size is $1 \times 2 \times 2$, the step size is $1 \times 2 \times 2$, and the subsequent pool core is $2 \times 2 \times 2$, the step size is $2 \times 2 \times 2$. A total of 4096 output units were set in the fully connected layer. The number of output channels is 64, 128, 256 and 512, respectively. Because these two subnetworks have been set before, the number of channels was constant from the fully connected layer to the average layer, as shown in Figure 4.

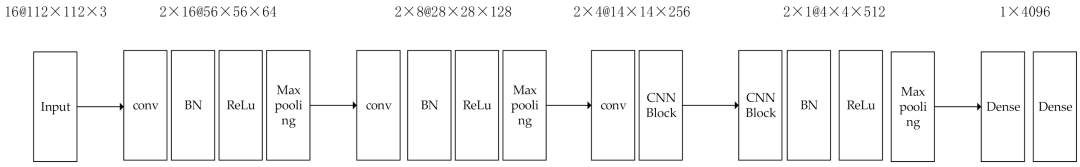


Figure 4. The architecture of the model, the size of the feature map and the number of channels is marked in the figure. Since we are using a parallel network, the number of channels needs to be multiplied by two.

3.2.2. CNN Block

The preprocessed video will produce face images and contour images. In order to adapt to the two types of images and improve the generalization ability of the network, a module was used called the CNN block, which is similar to the Residual Block [30]. It can be observed in Figure 5; the difference with Resnet is the simplest path of information dissemination as the main path. Such a structure has stronger generalization ability and can better avoid the vanishing gradient. Keeping the clear of the shortcut path, the information can be transmitted smoothly in the forward and backward propagation; the BN and ReLU are unified before the weight as pre-activation, which could result in ease of optimization and reduce overfitting on the residual path.

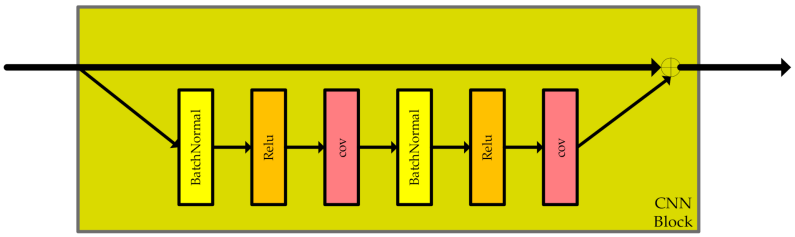


Figure 5. The structure of CNN Block.

3.2.3. The Objective Function

The loss function has a faster convergence rate because its gradient for the last layer of weight has nothing to do with the derivative of the activation function, and is only proportional to the difference between the output label and the true label. Backpropagation is continuous multiplication, so the update of the entire weight matrix will be accelerated. The derivation of multi-class cross entropy loss is simpler, and the loss is only related to the probability of the correct class. The loss is very simple to use to derive the input of the softmax activation layer, as shown in Equation (7)

$$Loss = -\sum_{j=0}^{n-1} y_i \times \ln \left[\frac{e^{\log its_i}}{\sum_{j=0}^{n-1} e^{\log its_i}} \right] \tag{7}$$

where y_i is ith label, and n equals 7 or 6 in this paper due to different databases. This can be obtained in Equation (8), when assigning different weights λ_1 and λ_2 to the losses of the two parallel networks:

$$Loss = \lambda_1 \times Loss_{global} + \lambda_2 \times Loss_{local} \quad (8)$$

4. Experiments

4.1. Implementation Details

In this paper, the training and inference of the proposed algorithm were implemented with Tensorflow backend. The details of the equipment are given as follows: Intel® Core™ i7, 3.00 GHz processors, 32 GB of RAM, and 1 NVIDIA GeForce RTX 2080 SUPER Graphics Processing Unit (GPU).

The initial learning rate is set as 0.01 to 0.0001; for different databases, the network reached a satisfactory Loss after from 8 to 12 h.

4.2. Results on Different Databases

The input image sequence was taken, and included 16 frames. The neighboring frames were used as a supplement when the video was less than 16 frames in length. Using a fixed time of frames instead of a fixed number of frames has the advantage of being more suitable for practical applications [31]. Our ratio of training set to test set is 8 to 2.

The convergence curves using the proposed methods were shown in Figure 6. The blue and the yellow were the result of using original 3DCNN; the red and the black were the result with MC-DCN with SAVEE and MMI. From Table 3, the run-times with different database can be observed.

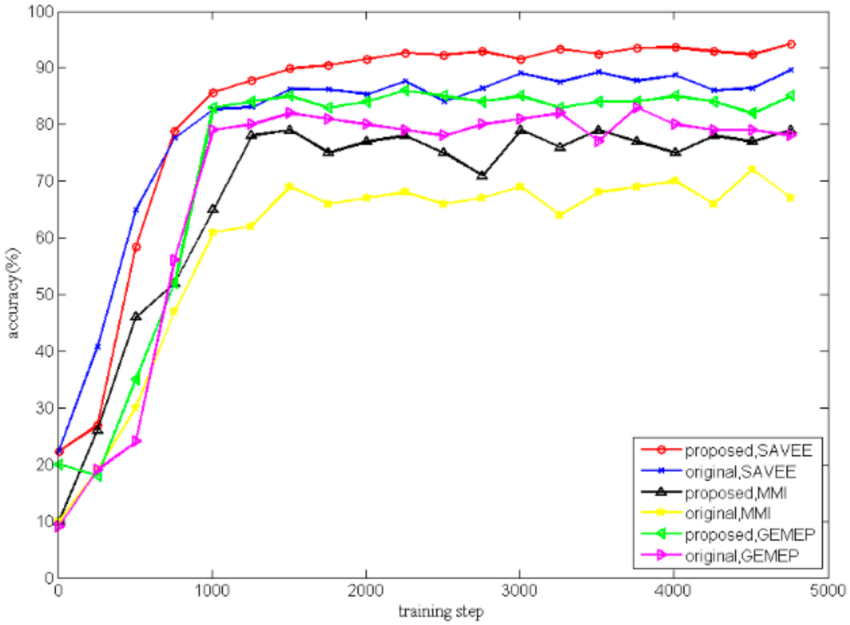


Figure 6. Training convergence curves on different database.

Table 3. The performance of different methods on three databases.

Video/Clip	Method	Performances(ACC:%)		
		SAVEE	**MMI**	**GEMEP**
video	Gaussian (PCA) [22]	91		
video	FAMNN [32]	93.75		
clip	FTL-ExpNet [33]		64.50	
video	F-Bases [34]		73.66	
video	DNN [35]		77.6	
clip	CNN	90.4	54.3	48.7
video	C3D	87	69.4	65.7
video	C3D-DAP [36]	93.8	63.40	50.7
video	Our Method	95.5	78.6	78.3
Preprocessing with database		×	×	✓
Training (per sequence) (in seconds)		1.9 ± 0.2	0.9 ± 0.2	0.9 ± 0.1

4.2.1. Results on SAVEE

To evaluate the performance of the proposed MC-DCN, we compared it with four FER methods, including Gaussian (PCA) [22] and FAMNN [32]. Results are shown in Table 3. Details of the experimental results of our method on the SAVEE database are given in Figure 7.

Figure 7. Test confusion matrix of SAVEE.

4.2.2. Results on MMI

The total accuracy of our model on the MMI dataset is shown in Table 3. Detailed experimental results are given as Figure 8. Obviously, the proposed algorithm achieves better performance in SAVEE than in MMI. It was inferred that there are two main reasons for these results. One reason is that the expression of the MMI dataset is a gradual process; input data are taken from the video for a fixed length time. It is possible to select a sequence where the emotional change has not reached its peak. The other reason was caused by the sample imbalance problem, especially the "neutral". Some training samples are added the manual way to solve the problem of insufficient sample size.

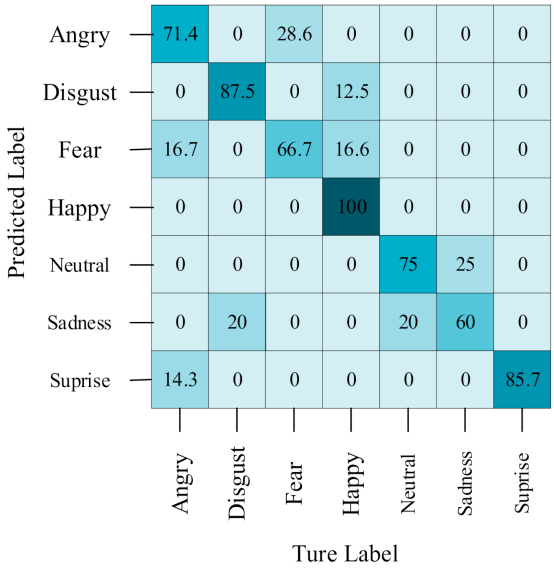


Figure 8. Test confusion matrix of MMI database.

4.2.3. Results on GEMEP

The GEneva Multimodal Emotion Portrayals (GEMEP) are a collection of audio and video recordings featuring 10 actors portraying 18 affective states, with different verbal contents and different modes of expression. A total 105 videos were used with expert estimates. In order to unify the experiment with other databases, these videos were relabeled with six new labels: happiness, sadness, disgust, fear, angry and surprise. The recognition rate was 78.3%.

The database has an uneven sample size because of the reconstruction. A video with different frames was tested to deal with the small samples. The confusion matrix of GEMEP is shown in Figure 9.

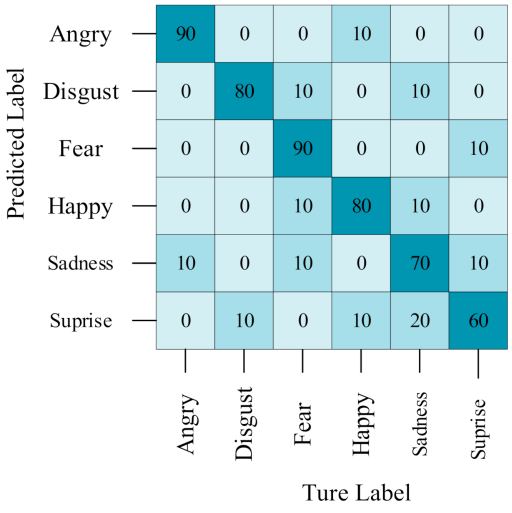


Figure 9. Test confusion matrix of GEMEP.

5. Conclusions

This paper presents a parallel network, MC-DCN, for facial expression recognition. Considering the different contributions of different facial organs in facial expression recognition, firstly, the ensemble of regression trees (ERT) is used to locate facial features. Secondly, the data are further classified by a network containing CNN blocks. Due to facial features simultaneously being affected by expression, age, identity and other aspects, this paper introduces the attention mechanism to the back end of ERT. This method makes the network

pay more attention to the corners of the mouth, nasolabial folds and other parts, which will show very big changes under the influence of expression. At the same time, it ignores crow's feet, which is more susceptible to age. In addition, we have added CNN blocks to the network to improve the generalization ability of the overall network to deal with different scenarios. In addition, we have added CNN blocks to the network to improve the generalization ability of the overall network to deal with different scenarios. The parallel structure allows the network to perceive mutual information while highlighting the learning ability of motions that are not obvious in facial expressions, effectively improving the accuracy of the network recognition. Experimental results show that our proposed algorithm achieves an accuracy of 95.5% on SAVEE (exaggerated expression), and accuracy of 78.6% and 78.3% on MMI (about half of the time in a state of no expression) and GEMEP (no obvious expression, not concentrated), respectively. Compared with other methods, including 3DNN, our method has improved the recognition accuracy. In particular, the results for GEMEP show that the choice of facial features increases the network's ability to learn unobvious movements. In the future, we plan to explore important parts of expressions, such as the state of muscles. In this way, the accuracy of expression recognition can be improved through more precise and accurate features.

Author Contributions: Conceptualization, Z.L. and H.W.; methodology, H.W.; software, H.W.; validation, H.W. and X.D.; formal analysis, H.W. and X.L.; investigation, Z.L. and H.W.; resources, H.W.; data curation, H.W. and M.Z.; writing—original draft preparation, H.W.; writing—review and editing, Z.L. and J.Z.; visualization, H.W.; supervision, Z.L.; project administration, Z.L.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (Grant No. 61972282).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in SAVEE, MMI, GEMEP at http://kahlan.eps.surrey.ac.uk/savee/, https://mmifacedb.eu/, https://www.unige.ch/cisa/gemep.

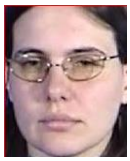
Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. The face part of the GEMEP database.

Database							
	Image Sequence					Label	Relabel
SAVEE						Angry	
						Happiness	
						Sadness	

Table 1. *Cont.*

Database	Image Sequence	Label	Relabel
MMI		Angry	
		Disgust	
		Fear	
GEMEP		Admiration	Surprise
		Angry	Angry
		Anxiety	Sadness
		Concept	Disgust

References

1. Othmani, A.; Taleb, A.R.; Abdelkawy, H.; Hadid, A. Age Estimation from Faces Using Deep Learning: A Comparative Analysis. *Comput. Vis. Image Underst.* **2020**, *196*, 102961. [CrossRef]
2. Ekman, P.; Friesen, W.V. Constants across Cultures in the Face and Emotion. *J. Pers. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
3. Harms, M.B.; Martin, A.; Wallace, G.L. Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies. *Neuropsychol. Rev.* **2010**, *20*, 290–322. [CrossRef]
4. 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]
5. Fasel, B.; Luettin, J. Automatic Facial Expression Analysis: A Survey. *Pattern Recognit.* **2003**, *36*, 259–275. [CrossRef]
6. Zhong, L.; Liu, Q.; Yang, P.; Huang, J.; Metaxas, D.N. Learning Multiscale Active Facial Patches for Expression Analysis. *IEEE Trans. Cybern.* **2015**, *45*, 1499–1510. [CrossRef] [PubMed]

7. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-Piloted Deep Network for Facial Expression Recognition. In *Comput. Vis.—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Germany, 2016; Volume 9906, pp. 425–442. ISBN 978-3-319-46474-9.
8. Sun, X.; Xia, P.; Zhang, L.; Shao, L. A ROI-Guided Deep Architecture for Robust Facial Expressions Recognition. *Inf. Sci.* **2020**, *522*, 35–48. [CrossRef]
9. Ding, Y.; Zhao, Q.; Li, B.; Yuan, X. Facial Expression Recognition From Image Sequence Based on LBP and Taylor Expansion. *IEEE Access* **2017**, *5*, 19409–19419. [CrossRef]
10. Yu, Z.; Liu, G.; Liu, Q.; Deng, J. Spatio-Temporal Convolutional Features with Nested LSTM for Facial Expression Recognition. *Neurocomputing* **2018**, *317*, 50–57. [CrossRef]
11. Kumawat, S.; Verma, M.; Raman, S. LBVCNN: Local Binary Volume Convolutional Neural Network for Facial Expression Recognition From Image Sequences. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019; pp. 207–216.
12. Zhang, H.; Huang, B.; Tian, G. Facial Expression Recognition Based on Deep Convolution Long Short-Term Memory Networks of Double-Channel Weighted Mixture. *Pattern Recognit. Lett.* **2020**, *131*, 128–134. [CrossRef]
13. Yi, J.; Chen, A.; Cai, Z.; Sima, Y.; Zhou, M.; Wu, X. Facial Expression Recognition of Intercepted Video Sequences Based on Feature Point Movement Trend and Feature Block Texture Variation. *Appl. Soft Comput.* **2019**, *82*, 105540. [CrossRef]
14. Yaddaden, Y.; Adda, M.; Bouzouane, A.; Gaboury, S.; Bouchard, B. User Action and Facial Expression Recognition for Error Detection System in an Ambient Assisted Environment. *Expert Syst. Appl.* **2018**, *112*, 173–189. [CrossRef]
15. Xie, S.; Hu, H.; Wu, Y. Deep Multi-Path Convolutional Neural Network Joint with Salient Region Attention for Facial Expression Recognition. *Pattern Recognit.* **2019**, *92*, 177–191. [CrossRef]
16. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; Pantic, M. A Survey of Multimodal Sentiment Analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
17. Sari, M.; Moussaoui, A.; Hadid, A. Simple Yet Effective Convolutional Neural Network Model to Classify Facial Expressions. In *Modelling and Implementation of Complex Systems*; Chikhi, S., Amine, A., Chaoui, A., Saidouni, D.E., Kholladi, M.K., Eds.; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Germany, 2021; Volume 156, pp. 188–202. ISBN 978-3-030-58860-1.
18. Lopes, A.T.; de Aguiar, E.; De Souza, A.F.; Oliveira-Santos, T. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. *Pattern Recognit.* **2017**, *61*, 610–628. [CrossRef]
19. Kazemi, V.; Sullivan, J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, OH, USA, 24–27 June 2014; pp. 1867–1874.
20. Molchanov, P.; Gupta, S.; Kim, K.; Kautz, J. Hand Gesture Recognition with 3D Convolutional Neural Networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 1–7.
21. Haq, S.; Jackson, P.J.B.; Edge, J. Audio-Visual Feature Selection and Reduction for Emotion Classification. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia, 10–11 August 2019.
22. Haq, S.; Jackson, P.J.B. *Speaker-Dependent Audio-Visual Emotion Recognition*; AVSP: Norwich, UK, 2009; pp. 53–58.
23. Wang, W. (Ed.) *Machine Audition: Principles, Algorithms, and Systems*; Information Science Reference: Hershey, PA, USA, 2011; ISBN 978-1-61520-919-4.
24. Valstar, M.F.; Pantic, M. Induced Disgust, Happiness and Surprise: An Addition to the MMI Facial Expression Database. In Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, UK; 2010; p. 65.
25. Bänziger, T.; Mortillaro, M.; Scherer, K.R. Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception. *Emotion* **2012**, *12*, 1161–1179. [CrossRef]
26. Hazourli, A.R.; Djeghri, A.; Salam, H.; Othmani, A. Multi-Facial Patches Aggregation Network for Facial Expression Recognition and Facial Regions Contributions to Emotion Display. *Multimed. Tools Appl.* **2021**, 1–24. [CrossRef]
27. Franklin, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. *Math. Intelligencer.* **2005**, *27*, 83–85. [CrossRef]
28. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 4401–4410. [CrossRef] [PubMed]
29. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, PMLR, San Diego, CA, USA, 9–12 May 2015; pp. 448–456.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*; Springer: Cham, Germany, 2016; pp. 630–645.
31. Zhou, Z.; Zhao, G.; Pietikainen, M. Towards a Practical Lipreading System. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 137–144.
32. Gharavian, D.; Bejani, M.; Sheikhan, M. Audio-Visual Emotion Recognition Using FCBF Feature Selection Method and Particle Swarm Optimization for Fuzzy ARTMAP Neural Networks. *Multimed. Tools Appl.* **2017**, *76*, 2331–2352. [CrossRef]

33. Li, J.; Huang, S.; Zhang, X.; Fu, X.; Chang, C.-C.; Tang, Z.; Luo, Z. Facial Expression Recognition by Transfer Learning for Small Datasets. In *Security with Intelligent Computing and Big-Data Services*; Yang, C.-N., Peng, S.-L., Jain, L.C., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Germany, 2020; Volume 895, pp. 756–770. ISBN 978-3-030-16945-9.
34. Sariyanidi, E.; Gunes, H.; Cavallaro, A. Learning Bases of Activity for Facial Expression Recognition. *IEEE Trans. Image Process.* **2017**, *26*, 1965–1978. [CrossRef] [PubMed]
35. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going Deeper in Facial Expression Recognition Using Deep Neural Networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–9 March 2016; pp. 1–10.
36. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In *Computer Vision—ACCV 2014*; Cremers, D., Reid, I., Saito, H., Yang, M.-H., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Germany, 2015; Volume 9006, pp. 143–157. ISBN 978-3-319-16816-6.