information

Article

Multi-Attention Module for Dynamic Facial Emotion Recognition

Junnan Zhi [1,2], Tingting Song [1], Kang Yu [1], Fengen Yuan [1,2], Huaqiang Wang [1,2], Guangyang Hu [1,2] and Hao Yang [1,*]

[1] Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China; zhijunnan@ime.ac.cn (J.Z.); songtingting@ime.ac.cn (T.S.); yukang@ime.ac.cn (K.Y.); yuanfengen@ime.ac.cn (F.Y.); wanghuaqiang@ime.ac.cn (H.W.); huguangyang@ime.ac.cn (G.H.)
[2] School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: yanghao@ime.ac.cn; Tel.: +86-10-8299-5602

Abstract: Video-based dynamic facial emotion recognition (FER) is a challenging task, as one must capture and distinguish tiny facial movements representing emotional changes while ignoring the facial differences of different objects. Recent state-of-the-art studies have usually adopted more complex methods to solve this task, such as large-scale deep learning models or multimodal analysis with reference to multiple sub-models. According to the characteristics of the FER task and the shortcomings of existing methods, in this paper we propose a lightweight method and design three attention modules that can be flexibly inserted into the backbone network. The key information for the three dimensions of space, channel, and time is extracted by means of convolution layer, pooling layer, multi-layer perception (MLP), and other approaches, and attention weights are generated. By sharing parameters at the same level, the three modules do not add too many network parameters while enhancing the focus on specific areas of the face, effective feature information of static images, and key frames. The experimental results on CK+ and eNTERFACE'05 datasets show that this method can achieve higher accuracy.

Keywords: facial emotion recognition; lightweight; attention module

Citation: Zhi, J.; Song, T.; Yu, K.; Yuan, F.; Wang, H.; Hu, G.; Yang, H. Multi-Attention Module for Dynamic Facial Emotion Recognition. *Information* **2022**, *13*, 207. https://doi.org/10.3390/info13050207

Academic Editors: Danilo Avola, Daniele Pannone and Alessio Fagioli

Received: 3 March 2022
Accepted: 6 April 2022
Published: 19 April 2022

1. Introduction

In recent years, automated dynamic facial emotion recognition (FER) has been widely used in psychology, security, and healthcare and has become a research area of great interest in both academia and industry. In medical research, Saravanan [1] found that patients with Parkinson's disease have reduced spontaneous facial expressions and have difficulties in posing emotional expressions and imitating non-emotional facial movements. Patients with some other neurological disorders such as Alzheimer's disease [2], multiple sclerosis [3], and others have some similar deficits in facial movements. In addition, emotion recognition also affects human daily life in perception, learning, life, and other aspects. For various application scenarios, facial emotion recognition has emerged from many related studies on efficient computing, multimodal analysis, and other aspects. Although researchers have done a lot of work around this task, automated FER that can be applied to medical scenarios remains a challenging problem due to the small magnitude of facial movements, irregular timing of appearances, and scene differences.

Given a video, the currently popular FER pipeline (FER based on multimodal computation is not within this scope) mainly consists of three steps, namely video preprocessing, feature extraction, and sentiment classification [4]. Among them, video preprocessing refers to video frame sampling, data enhancement, face alignment, and the normalization of illumination and poses. Feature extraction is a key part of FER, which quantizes and encodes each frame and image sampled into a dense feature vector. Finally, these dense feature vectors are fed into the classifier for sentiment classification.

Ekman classified human facial emotions into six basic categories, namely anger, disgust, fear, happiness, sadness, and surprise, and proposed FACS to describe facial movements as motor units [5]. The classic facial emotion recognition method needs to combine the corresponding problems, manually perform feature extraction [6,7], and then perform classification. With the application of deep learning in computer vision, in recent years many studies have applied deep learning to facial emotion recognition. Kahou et al. [8] used a fused CNN-RNN structure for facial emotion analysis. A classifier with RNN or LSTM structures can more efficiently model motion information in the temporal dimensions of the video. While this method is much more effective than taking a time-domain average of the CNN output as a classification result or using fully connected network (FC) as a classifier, it also significantly increases the size of the network and the demand on computational resources. Byeon et al. [9] applied C3D, originally developed for video motion analysis, to FER, causing 3D convolution blocks to slide and perform convolution operations on blocks of data consisting of video frames. In the feature extraction phase, this method takes into account both temporal and spatial dimensional information. Fan et al. [10] used CNN-RNN and C3D as two branches of the model for emotion recognition, and finally introduced the idea of multimodal analysis to splice the model output with the speech analysis results and then to perform classification. Noroozi et al. [11] also adopted the multimodal analysis method, which combined the results of three branches of feature point analysis based on prior knowledge, a convolutional neural network, and audio analysis to obtain the final emotion classification result. Fei et al. [12] tried to use GAN to solve the category imbalance problem existing in sentiment data so as to improve the final recognition effect, and also achieved some positive results.

One shortcoming of the above approach is that some of the general video analysis methods are applied to the facial emotion recognition task. Since these methods are often initially proposed for applications in more mainstream tasks such as action recognition and scene classification, they lack optimization for the features of FER, such as the small magnitude of facial actions, random timing of appearances, and complexity of scenes. Additionally, some of the above methods were proposed in the EmotiW challenge. The authors focused more on the final accuracy while ignoring some other factors, such as the computational cost, system complexity, and difficulty of tuning the parameters.

In this article, we hope to propose a FER method that makes up for the shortcomings of the above methods and that is optimized for the characteristics of the FER task without increasing the computational cost too much. Therefore, we propose a FER method that augments a multi-attention module. Based on ResNet, the method uses attention modules in three dimensions of space, channel, and time to select the parts of the video that are important for identifying specific emotions. For the choice of baseline, we show that through experiments ResNet18 does not underperform against other more complex models on the FER task. We conduct experiments on the CK+ [13] and eNTERFACE'05 [14] datasets. Compared with the baseline and other recognition methods, the emotion recognition method with an increased attention module shows significantly improved effects.

2. Related Works

In the Introduction, we mentioned the shortcomings of some FER methods. Recently, some researchers have paid attention to this problem. Meng et al. [15] believed that the importance of different frames in the video should be different, and that ignoring this difference will lead to recognition errors. Their proposed FAN introduces self-attention and relation–attention between frames, giving weights to different frames, achieving good results, and maintaining a lightweight model. Alireza et al. [16] also focused on the frame-to-frame relationship. They added an MLP-based attention learning layer to the output of the Bi-LSTM. Masih et al. [17] added a spatial attention module based on a self-attention mechanism to a CNN network with a Vgg16 backbone and used softmax pooling to filter the frames that may contain possession information in the frame images. Min et al. [18] used the popular residual attention network in image classification as a

feature extraction network and innovatively proposed TS-SATCN, a two-stage TCN with added spatiotemporal attention. The attention mechanism in both the image space and frame dimensions was finally identified. In multimodal analysis, Wang et al. [19] proposed a method of analyzing the attention between fusion mode and mode. On the one hand, the representative emotional features are extracted in the single frame sequences, while on the other hand, based on their importance, the specific pattern features are automatically protruded based on their importance.

3. Materials and Methods

Here, we propose a multi-attention module for the emotion recognition task based on video analysis. The module calculates the spatial attention and channel attention on the feature map of each frame and weights them. At the end of the feature extraction network, the frame attention is calculated based on the information contained in the feature vector for each frame output. We choose ResNet18 [20] as the baseline for the convolutional neural network part. We add spatial and channel attention modules to each basic block of the baseline and frame attention modules between the ResNet18 and classification network. Figure 1 also shows the location and function of the spatial attention module and the channel attention module in the baseline. The two modules act on the feature map in a serial manner.

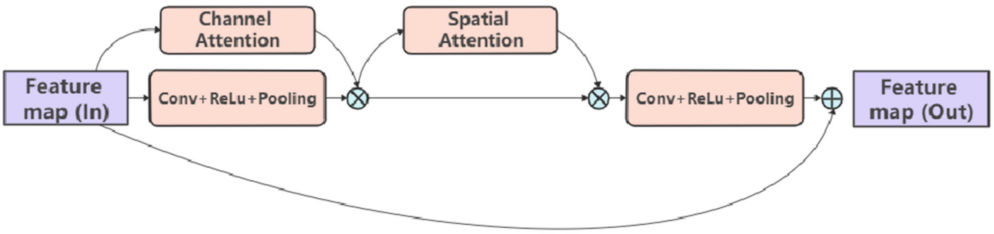


Figure 1. ResNet's basic block that increases the spatial attention and channel attention. The two concatenated convolution modules are the original parts of ResNet's basic block. In the middle of the bipolar convolution module, the channel attention module and the spatial attention module are added to adjust the value of the feature map adaptively.

3.1. Spatial Attention

Here, we design a spatial attention module that uses the spatial correlation of feature maps for computation. Spatial attention focuses on the distribution of useful information in space. In order to obtain a 2D weight map, the feature map needs to be compressed in the channel dimension. First of all, the initial compression would be used to obtain several feature maps containing key information, and then to calculate the importance of weights for each spatial location based on these feature maps. The details of this attention module are listed below.

Suppose we have a feature map as F of N $\times$ N $\times$ C, we use NIN [21], i.e., 1 $\times$ 1 convolution, to extract key information in the channel dimension from the feature map. After experiments, we think it is appropriate to keep the feature maps of 16 channels in this step. After the convolution in the channel domain, we will obtain the key information map of N $\times$ N $\times$ 16, F_{info}, followed by the activation layer ReLU. Then, we perform a convolution on F_{info} to obtain the attention weight matrix of N $\times$ N $\times$ 1. Figure 2 shows the structure of this attention module.

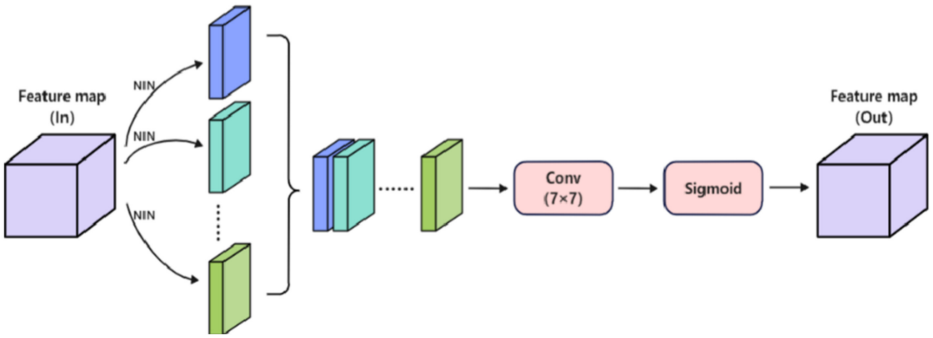


Figure 2. Schematic diagram of the spatial attention module. As mentioned in the text, the spatial attention module uses the "network in network" approach to extract information, while the obtained key information matrix is calculated by convolution and activation, to obtain the spatial attention weight map.

Finally, the attention weight matrix is the inner product on the feature map caused by broadcasting, and the result of applying spatial attention can be obtained. The formula for calculating spatial attention is as follows:

$$\begin{aligned} A_{sp} &= \sigma(Conv(ReLU(NIN(F)))) \\ &= \sigma\Big(Conv\Big(ReLU\Big(F_{info}\Big)\Big)\Big) \end{aligned} \tag{1}$$

where A_{sp} is the attention weight matrix of the spatial dimension, σ is the sigmoid activation function, and F is the feature map of the input attention module.

3.2. Channel Attention

Similar to spatial attention, here we design a module that utilizes feature maps to compute channel attention. We think that each channel of the feature map contains a certain feature. The purpose of the channel attention is to discover which features are useful for our task. We use max pooling and average pooling to extract the key information from each channel to obtain the channel key information vector. In order to improve the computational efficiency, we refer to the methods used by Woo and use a bottleneck multi-layer perceptron (MLP) to operate on the extracted channel key information vector. The details of the channel attention module are as list below.

We still assume that there is a feature map of N × N × C as F. We extract the key information in the spatial dimension from the feature map by performing average pooling and max pooling in the spatial dimension. After the pooling operation, we obtain two sets of 1 × 1 × C key information vectors F_{max} and F_{avg}. The obtained key information vectors F_{max} and F_{avg} are added to the input of a three-layer MLP to obtain the weight vector of 1 × 1 × C channel attention. Figure 3 visualizes this process. Finally, the attention weight vector is the inner product on the feature map caused by broadcasting, and the result from applying the channel attention is obtained. The calculation formula of the channel attention is as follows:

$$\begin{aligned} A_{ch} &= \sigma(MLP(MaxPool(F) + AvgPool(F))) \\ &= \sigma\big(MLP\big(F_{max} + F_{avg}\big)\big) \end{aligned} \tag{2}$$

where A_{ch} is the attention weight matrix of the channel dimension, σ is the sigmoid activation function, and F is the feature map of the input attention module.

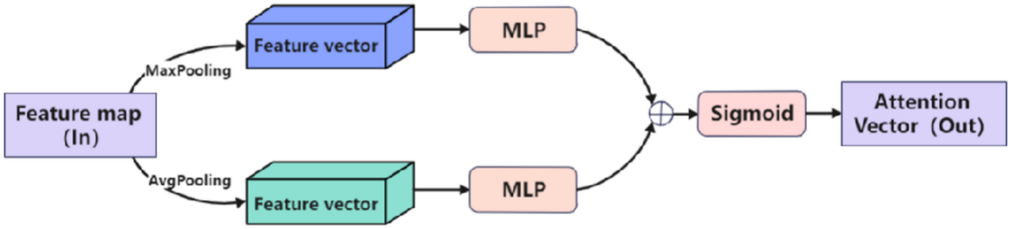


Figure 3. Schematic diagram of the channel attention module. As mentioned in the paper, the channel attention module uses max pooling and average pooling to extract information and uses MLP and other methods to calculate the channel attention weight vector.

3.3. Frame Attention

We also propose a frame attention module for video-based facial emotion recognition. We believe that the biggest difference between video-based facial emotion recognition and static facial emotion recognition lies in the relationship and continuity of the previous and next frames. In most cases, not every frame of a video will have a distinct emotional label. Based on this idea, we hope to be able to give more attention to frames with obvious emotional features. The details of this attention module are as listed below.

The flow of the frame attention module is shown in Figure 4. The convolutional neural network part of the model will output a one-dimensional feature vector of 256×1 for each frame of the video. The feature vector contains a variety of information, and we use an MLP with shared parameters to compute the feature vector, simply extract the emotional features it may contain, and compute the attention weight to the current frame, as follows:

$$a_i' = M_f(f_i) \tag{3}$$

where f_i is the final feature vector of each frame, M_f is the MLP used for frame weights, and a_i' is the unnormalized frame weight. After the attention weights for all frames are calculated, we perform a normalization calculation based on the sum of the attention weights. The normalized weights are superimposed on the feature vector of each frame of the video.

$$a_i = \sum_{j=1}^{n} \frac{M_f(f_i)}{M_f(f_j)} \tag{4}$$

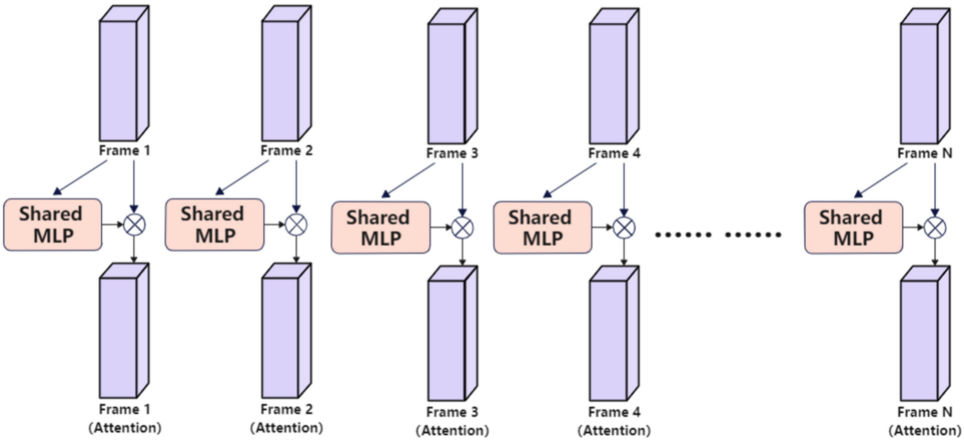


Figure 4. Schematic diagram of the fully connected network(FC) part of the baseline. In this part, the feature vectors output by the convolutional neural network are spliced together by frame to obtain the feature vector of the current video, and the final classification result is obtained through the MLP calculation.

4. Results

In this section, we will evaluate our model on datasets CK+ and eNTERFACE'05. In order to classify the video-based facial emotion data, we decompose the videos into several single-frame images and evenly sample 16 of them as inputs to the emotion recognition model to obtain the final prediction results. For the purpose of evaluating the impact of each attention module on the experimental results, we will add space, channel, and attention modules to the baseline, and observe the changes in the accuracy of the model predictions. At the same time, we will compare the performance of the FER proposed in this paper with other FER methods that introduce the attention mechanism to the eNTERFACE'05 dataset.

4.1. Datasets

We select datasets CK+ and eNTERFACE'05 to evaluate our models. The following is a brief introduction to these two models.

CK+ contains 593 video sequences from 123 subjects. In these videos, 327 clips from 118 subjects are labeled into seven categories, including anger, contempt, disgust, fear, happiness, sadness, and surprise. Under the guidance of the researchers, the subjects make various expression labels by combining various facial action units. The examples of this dataset are shown in Figure 5. Since CK+ does not divide the training set and the test set, we divide the dataset according to the ratio of 8:2 and selected 20% of the training set for verification, then conducted ten training, verification, and testing procedures. The average prediction accuracy on the test set is used as the final result.

Figure 5. Example of CK+. Subjects are first familiarized with various facial action units, and then combine facial action units to make expressions under the direction of the researcher.

Here, eNTERFACE'05 contains 1166 video sequences from 42 subjects of 14 different nationalities, of which 81% are male, 19% are female, 31% wear glasses, and 17% have a beard. The dataset contains six categories, namely angry, disgusted, fearful, happy, sad, and surprised. The researchers give subjects a text containing a scene and five reactions. After the subjects read the material and prepared their emotions, five reactions are read out as emotional data. An example of eNTERFACE'05 is shown in Figure 6. Like CK+, eNTERFACE'05 also does not divide training and test sets. We adopt the same evaluation method for evaluating model.

4.2. Experiment

Next, we evaluate our model on CK+ and eNTERFACE'05 and compare the performance of the baseline on the test set with different combinations of attention modules applied. Since neither CK+ nor eNTERFACE'05 is divided into a training set and test set, we obtain 10 groups of training set and test set divisions via random sampling and conduct independent experiments on each group to obtain the final result. We add a spatial attention module and a channel attention module to each basic block of ResNet18 and add a frame attention module to the output of ResNet18.

Figure 6. Example of eNTERFACE'05. Subjects perform emotions in preset scenes and read preset lines.

To overcome overfitting during training, we use data augmentation operations, including rotation, folding, and resizing of images to a resolution of 224 × 224. We use Adam as the optimization algorithm and set the weight decay to 0.000005 and the learning rate decay to 0.1 times when the loss function does not decline for 10 epochs. The ResNet18 backbone is pretrained for face recognition on the LFW dataset and fine-tuned for 50 epochs on the experimental dataset. We assign an initial learning rate of 0.0001 to the ResNet18 backbone and an initial learning rate of 0.001 to the rest.

4.2.1. Experiment on CK+

Our experimental results on the CK+ dataset are shown in Table 1. On the CK+ dataset, the accuracy of our baseline reaches 81.25%. We separately add a spatial attention module, a channel attention module, and a frame attention module to the baseline. The spatial attention shows the most obvious improvement in the prediction results, reaching 85.23%. On the basis of adding the spatial attention module, the network is added with the channel attention module. The recognition accuracy of the model remains basically unchanged, only increasing by 0.19% to 85.42%. Finally, the recognition accuracy shows further improvement, reaching 89.52%, when adding the frame attention module. Meanwhile, because the attention module uses a pooling operation and shallow convolution and full connection operation to extract key information, a large number of convolution operations are not performed, meaning the amount of calculation required by the model and the computing resources do not increase significantly.

4.2.2. Experiment on eNTERFACE'05

Our experimental results on the eNTERFACE'05 dataset are shown in Table 2. On the eNTERFACE'05 dataset, our baseline achieves 80.83% accuracy. As with the experiment on CK+, we separately add three attention modules to the baseline. The three modules show similar improvements in terms of network performance, and the channel attention module is the highest, reaching 85.38%. On the basis of increasing the channel attention, the network adds a spatial attention module and the recognition accuracy is increased by 0.87% to 86.25%. Finally, adding the frame attention module further improves the accuracy, reaching 88.33%. After experiments, the model using multiple attention modules shows a 5.41% improvement compared to the original baseline.

Table 1. Evaluation on the CK+ datasets of the recognition accuracy under different attention modules.

Model	Acc.
ResNet18+FC (Baseline)	81.25%
Baseline with SA	85.23%
Baseline with CA	82.34%
Baseline with FA	84.38%
Baseline with SA, CA	85.42%
Ours (Baseline with SA, CA, FA)	89.52%
CNN-RNN	87.52%
FAN (with Relation-attention)	83.28%
ResNet18+FC with CBAM [22]	85.38%
Vgg16+Bi-LSTM with Attention	87.26%
RAN+TS-SATCN	90.17%

Table 2. Evaluation on the eNTERFACE'05 datasets of the recognition accuracy under different attention modules and comparison of the performances of our method and state-of-the-art FER methods.

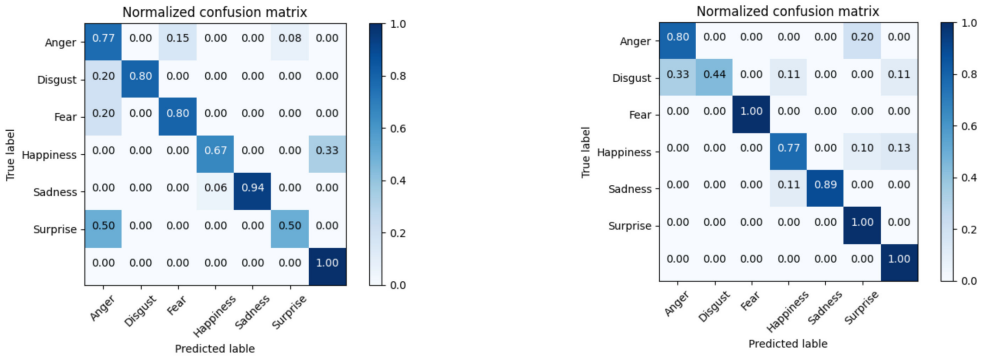
Model	Acc.
Vgg16 [23] + FC	72.18%
ResNet18+FC (Baseline)	80.83%
Baseline with SA	85.00%
Baseline with CA	85.38%
Baseline with FA	84.17%
Baseline with SA, CA	86.25%
Ours (Baseline with SA, CA, FA)	88.33%
CNN-RNN	86.18%
FAN (with Relation-attention)	82.08%
ResNet18+FC with CBAM	85.41%
Vgg16+Bi-LSTM with Attention	85.83%
RAN+TS-SATCN	89.25%

On the CK+ and eNTERFACE'05 dataset, we also compare the effectiveness of our method with some other FER methods, including CNN-RNN, FAN, ResNet18 with CBAM, Vgg16+Bi-LSTM with Attention, and RAN+TS-SATCN. In contrast, our method outperforms these methods. Although the effect of RAN+TS-SATCN is slightly better than our method, the network sizes of both RAN and TS-SATCN are much larger than ours. During the training process, RAN+TS-SATCN takes up 8–10 times more computational resources than ours. The cost of this accuracy is not what we want. Notably, our method outperforms the CNN-RNN structure without paying attention to the deterministic temporal order of video frames. This shows that in the FER task, the sequence of time series may not be more important information than the information contained in key frames.

After the experimental comparison, it can be seen that our method with the addition of spatial, channel, and frame attention modules is better than the method without multiple attention and other advanced FER methods. Compared to FAN and CBAM, our approach gives attention to more dimensions. Intuitively, spatial and frame attention mechanisms can help us to better locate key facial movements and key frames. Channel attention, by modeling the importance of each feature channel, focuses the network on effective information and inhibits the propagation of harmful information to the later layers of the network.

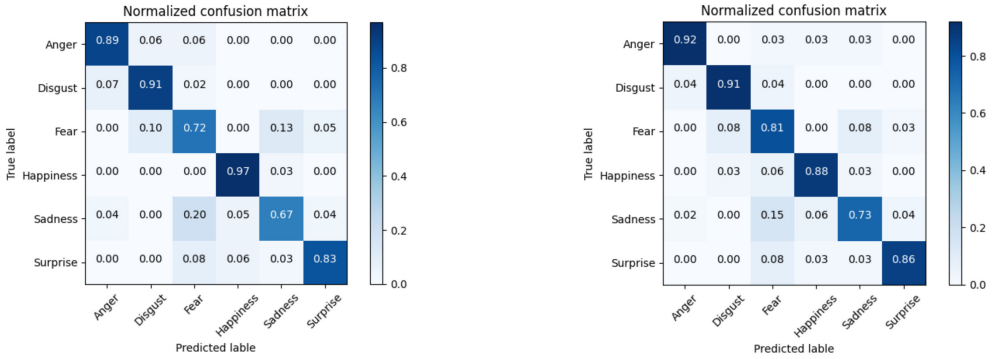
Next, we show the confusion matrices of the experimental results of the baseline and our method for CK+ and eNTERFACE'05 in Figures 7 and 8. It can be seen that our method identifies most emotions better than the baseline method. For eNTERFACE'05, although the baseline is more accurate than our method for 'happiness' alone, it also misclassifies more of the other emotions as 'happiness'. In the CK+ experiments there are more 'natural' emotion patterns (not labeled in the figure). Our method still outperforms the baseline

method for most of the emotions. For a few emotions, such as 'disgust' and 'sadness' in CK+, the baseline is slightly more accurate. However, for most emotions, the classification accuracy is higher when using our method. This suggests that our method could more effectively utilize the information for most emotions for classification tasks.



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

Figure 7. The confusion matrices of different methods on the CK+ dataset: (**a**) baseline. (**b**) our proposed method.



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

Figure 8. The confusion matrices of different methods on the eNTERFACE'05 dataset: (**a**) baseline; (**b**) our proposed method.

We show the advantages of our method in capturing facial actions in Figure 9. We overlay the weight matrix output from the spatial attention module trained to the best state onto the original image as a soft mask. The brightness of each region on the image represents how much attention the network pays to it. Completely irrelevant parts of the image such as hair and clothes are totally covered by the mask. Most areas of the subject's face are also darker, which means that these areas are also relatively underweighted. The areas around the mouth, eyes, eyebrows, and other areas that are prone to muscle movements are given higher weights and are highlighted in the figure.

In Table 3, we compare the number of parameters of our method with some advanced FER methods. ResNet18+FC with CBAM and ResNet18+FC with FAN are optimized for FER methods from different perspectives by introducing attention mechanisms in frame images or interframes, respectively. CNN-RNN is a more complex FER method than the above methods, but the experiments on eNTERFACE'05 and the comparison of the number of parameters show that our method performs better while being lighter.

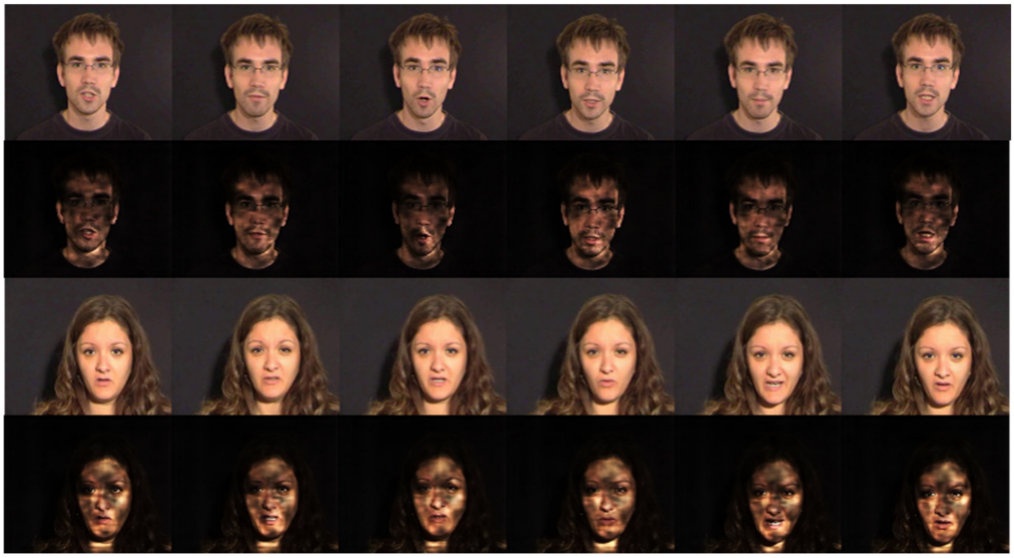


Figure 9. Comparison of the original data in the dataset before and after using applied spatial attention weights. The images are overlaid with a soft mask based on the values of the spatial attention matrix to demonstrate the role of spatial attention in capturing facial action regions.

Table 3. Comparison of the number of parameters for different FER methods, including ResNet18+FC (baseline), ResNet18+FC with CBAM, FAN, CNN-RNN, and our proposed method.

ResNet18+FC	ResNet18+FC with CBAM	FAN	Ours	CNN-RNN
11.78 M	11.87 M	11.79 M	13.39 M	30.99 M

5. Conclusions

In this paper, we propose a facial emotion recognition algorithm to enhance the feature extraction capability of CNN by introducing multiple attention modules. We propose three separate attention modules, namely spatial, channel and frame, based on video characteristics. In the spatial and channel dimensions, we use average pooling and maximum pooling to extract key information and superimpose it on the feature map after simple computation. In the frame dimension, we use MLP to calculate the weight of the current frame based on the feature vector output from CNN and overlay it on the feature vector. We conduct controlled experiments on the CK+ and eNTERFACE'05 datasets. The results show that the attention module improves the recognition accuracy to varying degrees. In addition, we also experiment with the extraction method of key information, and the results show that using the "network in network" approach would be slightly better than the pooling operation.

We also note that there are some limitations in our work, mainly in terms of data and model training. We find that although the proposed method performs well on the experiment dataset, the recognition accuracy is not satisfactory for our locally collected small-scale data. This is caused by inconsistent data distribution. Applying this work to other application scenarios requires adequate datasets and a reasonable model training strategy. In the next step, we plan to design and refine the facial data collection system and continue our research to solve medical problems. Simultaneously, we are moving towards designing models that are lightweight and target small datasets.

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

Funding: This research was supported by the Key Research Program of the Chinese Academy of Sciences, Grant NO.ZDRW-ZS-2021-1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://www.enterface.net/results/, accessed on 8 April 2022.

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

References

1. Saravanan, S.; Ramkumar, K.; Adalarasu, K.; Sivanandam, V.; Kumar, S.R.; Stalin, S.; Amirtharajan, R. A Systematic Review of Artificial Intelligence (AI) Based Approaches for the Diagnosis of Parkinson's Disease. *Arch. Comput. Methods Eng.* **2022**, *1*, 1–15. [CrossRef]
2. Jiang, Z.; Seyedi, S.; Haque, R.U.; Pongos, A.L.; Vickers, K.L.; Manzanares, C.M.; Lah, J.J.; Levey, A.I.; Clifford, G.D. Automated analysis of facial emotions in subjects with cognitive impairment. *PLoS ONE* **2022**, *17*, e0262527. [CrossRef] [PubMed]
3. Cecchetto, C.; Aiello, M.; D'Amico, D.; Cutuli, D.; Cargnelutti, D.; Eleopra, R.; Rumiati, R.I. Facial and bodily emotion recognition in multiple sclerosis: The role of alexithymia and other characteristics of the disease. *J. Int. Neuropsychol. Soc.* **2014**, *20*, 1004–1014. [CrossRef] [PubMed]
4. Shan, L.; Deng, W. Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing. 2020. Available online: https://ieeexplore.ieee.org/abstract/document/9039580 (accessed on 1 March 2022).
5. Ekman, R. *What the Face RevealsBasic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: New York, NY, USA, 1997.
6. Littlewort, G.; Bartlett, M.S.; Fasel, I.; Susskind, J.; Movellan, J. Dynamics of facial expression extracted automatically from video. In Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004.
7. 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]
8. Kahou, S.E.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent Neural Networks for Emotion Recognition in Video. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, New York, NY, USA, 9 November 2015; pp. 467–474.
9. Byeon, Y.-H.; Kwak, K.-C. Facial Expression Recognition Using 3D Convolutional Neural Network. *Int. J. Adv. Comput. Sci. Appl.* **2014**, *5*, 12. [CrossRef]
10. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; Nakano, Y., Ed.; The Association for Computing Machinery Inc.: New York, NY, USA, 2016; pp. 445–450, ISBN 9781450345569.
11. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Audio-Visual Emotion Recognition in Video Clips. *IEEE Trans. Affect. Comput.* **2017**, *10*, 60–75. [CrossRef]
12. Ma, F.; Li, Y.; Ni, S.; Huang, S.; Zhang, L. Data Augmentation for Audio–Visual Emotion Recognition with an Efficient Multimodal Conditional GAN. *Appl. Sci.* **2022**, *12*, 527. [CrossRef]
13. Kanade, T.; Tian, Y.; Cohn, J.F. Comprehensive database for facial expression analysis. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, Grenoble, France, 28–30 March 2002.
14. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE'05 Audio-Visual Emotion Database, International Conference on Data Engineering Workshops. *IEEE Comput. Soc.* **2006**, *8*, 383–388. [CrossRef]
15. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame Attention Networks for Facial Expression Recognition in Videos. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019. [CrossRef]
16. Sepas-Moghaddam, A.; Etemad, A.; Pereira, F.; Correia, L.P. Facial emotion recognition using light field images with deep attention-based bidirectional LSTM. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
17. Aminbeidokhti, M.; Pedersoli, M.; Cardinal, P.; Granger, E. Emotion recognition with spatial attention and temporal softmax pooling. In *International Conference on Image Analysis and Recognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 323–331.
18. Hu, M.; Chu, Q.; Wang, X.; He, L.; Ren, F. A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video. *IEEE Signal Process. Lett.* **2021**, *28*, 698–702. [CrossRef]
19. Wang, Y.; Wu, J.; Hoashi, K. Multi-attention fusion network for video-based emotion recognition. In Proceedings of the 2019 International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA, 14 October 2019; pp. 595–601. [CrossRef]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
21. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.

22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. *arXiv* **2018**, arXiv:1807.06521.
23. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.