information

Article

Spatiotemporal Convolutional Neural Network with Convolutional Block Attention Module for Micro-Expression Recognition

Boyu Chen [1,2], Zhihao Zhang [1,2,3], Nian Liu [1,2], Yang Tan [1,2], Xinyu Liu [1,2] and Tong Chen [1,2,3,*]

1 School of Electronic and Information Engineering, Southwest University, Chongqing 400715, China; boyuchen@email.swu.edu.cn (B.C.); zhangzh@psych.ac.cn (Z.Z.); nian206@email.swu.edu.cn (N.L.); tanyang1995@email.swu.edu.cn (Y.T.); liuxinyu1223@email.swu.edu.cn (X.L.)
2 Chongqing Key Laboratory of Non-Linear Circuit and Intelligent Information Processing, Southwest University, Chongqing 400715, China
3 Institute of Psychology, CAS, Beijing 100101, China
* Correspondence: c_tong@swu.edu.cn

Received: 29 May 2020; Accepted: 24 July 2020; Published: 29 July 2020

Abstract: A micro-expression is defined as an uncontrollable muscular movement shown on the face of humans when one is trying to conceal or repress his true emotions. Many researchers have applied the deep learning framework to micro-expression recognition in recent years. However, few have introduced the human visual attention mechanism to micro-expression recognition. In this study, we propose a three-dimensional (3D) spatiotemporal convolutional neural network with the convolutional block attention module (CBAM) for micro-expression recognition. First image sequences were input to a medium-sized convolutional neural network (CNN) to extract visual features. Afterwards, it learned to allocate the feature weights in an adaptive manner with the help of a convolutional block attention module. The method was testified in spontaneous micro-expression databases (Chinese Academy of Sciences Micro-expression II (CASME II), Spontaneous Micro-expression Database (SMIC)). The experimental results show that the 3D CNN with convolutional block attention module outperformed other algorithms in micro-expression recognition.

Keywords: micro-expression recognition; 3D convolutional neural network (3D CNN); convolutional block attention module (CBAM); adaptive feature weights; spatiotemporal features

1. Introduction

Emotions are the inner feelings of human beings and expressions are the windows of human emotions. A micro-expression on the face is a unique expression that happens spontaneously. When humans try to conceal their true emotions, the restrained feelings are shown by fast muscular movement out of a spontaneous physical reaction [1]. Therefore, a micro-expression is one of the foundations for the judgment of human psychological status.

In contrast to normal facial expressions that usually sustained 1/2 s to 4 s [2], the micro-expressions of humans last so short that they tend to be neglected. The duration of a micro-expression is only 1/25 s to 1/5 s with low intensity [3]. Porter suggested that micro-expressions are generated by parts of human faces as the result of muscular movement [4]. The micro-expression recognition is more difficult than the macro one because of its features.

Thanks to the development of artificial intelligence, the human–computer interaction better facilitates the study of micro-expression recognition. Specifically, Hinton proposed the concept of deep learning in 2006 [5], which is an important branch of machine learning. Deep learning excelled at feature

extraction and categorization in image recognition. More and more researchers made breakthroughs by combining machine learning or deep learning with micro-expression recognition [6–8].

Different from macro-expressions, micro-expressions consist of facial muscle movement with low intensity and short duration. What is more, this facial muscle movement tends to happen in some small but specific facial regions. For example, the macro-expression of happiness will involve the movement of action unit (AU)6 and AU12 with high intensity, which results in the raise of cheeks, wrinkling of outer corner of eyes, smaller of eye region, open of mouth or raise of mouth corner. However, the micro-expression of happiness will involve low intensity of AU12 or AU6, which results in the slight raise of the mouth corner or smaller of eye area. Therefore, for the recognition of micro-expression, we may focus on the mouth corner and eye regions without paying attention to other regions of the face. The attention model is a suitable method for performing this attention mechanism. It is thus used in this research.

The attention mechanism was added to the designed deep convolutional neural network (CNN) for micro-expression recognition. This method can not only extract the overall features of human faces, but also concentrate on some key features. Since micro-expressions only occur in parts of the human face, the attention mechanism helps to focus on specific facial regions, learning and acquiring the important features.

In our work, we designed a three-dimensional convolutional neural network (3D CNN) to learn spatiotemporal information, and a convolutional block attention module (CBAM) was appended after the 3D CNN. The proposed method was able to learn the information at the target domain effectively and to emphasize the features at important regions and this improved the ability of feature extraction.

The contributions of this study contain three aspects:

(1) this study designed a six-layer 3D CNN and took the whole face into the network for micro-expression recognition;
(2) the researchers reached the optimized network structure by monitoring the recognition accuracy while decreasing the convolutional layers successively on the basis of the existed network;
(3) this study combined the convolutional block attention module (CBAM) with its 3D CNN to simulate visual attention mechanism and enhance the information flows in channels and spaces.

2. Related Works

Polikovsky and other researchers [9] divided human faces into specific regions according to facial action coding system (FACS), based on which they proposed the 3D-gradient direction histogram for movement description. To extend the general texture features to the dynamic ones, Zhao et al. [10] proposed a feature descriptor named as local binary patterns on three orthogonal planes (LBP-TOP) for micro-expression recognition. This method achieved better accuracy. Pfister et al. [11] normalized videos of different lengths on the basis of a temporal interpolation model. Meanwhile, they extracted the image features with the help of the spatiotemporal local texture descriptors (SLTDs) in combination with multi-kernel learning (MKL) to recognize micro-expressions. Furthermore, Wang et al. [12] proposed the method of tensor independent color space (TICS) from the perspective of color space. This model extracted the dynamic texture features from the color components with better performance. Moreover, Liu et al. [13] proposed the method of main directional mean optical flow (MDMO), which used optical flow estimation technology to calculate the subtle motion of 36 regions of interest (ROIs). In addition, they aligned all the frames in the video clips of micro-expressions in the approach driven by optical flows.

These traditional methods for feature extraction have contributed significantly to the micro-expression researches. However, the approaches mentioned above do not achieve high accuracy in micro-expression recognition. As compensation for inefficiency, deep learning is able to advance the capability of feature presentations.

Deep learning has been gradually applied to computer vision [14], natural language processing [15], and other fields. In this context, a growing number of researchers have applied neural network to various tasks in the study of micro-expressions, such as detection, recognition, etc.

One of the most popular methods in the study of deep learning is CNN. Since the popularity of the LeNet [16], designed by LeCun in 1998, various network structures have been designed (e.g., AlexNet [17], GoogleNet [14], VGG-Net [18], etc.). They have been widely applied in fields of facial recognition and voice recognition. However, these models are confined to two-dimensional data processing. In recent years, researchers have utilized 3D CNN in consideration of the temporal dimension.

Peng et al. [19] employed a 3D CNN with the dual temporal scale in micro-expression recognition. After computing the optical flow of two video clips which had different numbers of frames, the dual temporal scale network with the support vector machine (SVM) generated satisfactory results. In addition, Li et al. [20] presented a micro-expression recognition method based on the image sequence (i.e., introducing both the gray sequence and the optical flow sequence to a 3D CNN). This 3D CNN could catch subtle motion flows, promoting recognition accuracy. On the other hand, Reddy et al. [21] testified the feature fusion of eyes and mouth as well as features on the whole face, concluding that learning the facial features on the whole face accomplished the micro-expression recognition task better with the help of two kinds of spatiotemporal CNNs. In consideration of overfitting that might occur when using small sample data in deep network, Peng et al. [22] applied fine-tuning on the micro-expression database after pre-training ResNet 10 on the macro-expression database from the perspective of transfer learning. Xia et al. [7] proposed spatiotemporal recurrent convolutional networks to capture the spatiotemporal feature from micro-expression video clips. They also adopted temporal data augmentation strategies to enlarge training data and proposed a balanced loss mechanism, which showed its effectiveness on spontaneous micro-expression databases. Verma et al. [6] utilized a dynamic imaging technique to convert the sequence into a frame of image sequences, and proposed a Lateral Accretive Hybrid Network (LEARNet) to learn the subtle features of the face area, which involves the cross decoupled relationship between convolution layers. Based on the standard micro-expression databases, the results of the proposed algorithm were improved to a certain extent compared with ResNet.

3. Materials and Methods

3.1. 3D CNN

CNN processes large amounts of data by utilizing convolutional layers and pooling layers continuously, which has shown better performance than traditional algorithms in feature extraction.

In many micro-expression recognition studies of CNN, researchers applied 2D CNN to extract features from apex frames. However, as a micro-expression is a movement of facial muscles, the successive video frames contain temporal information. The more dynamic information about micro-expressions is captured, the more key data acquired. To obtain dynamic information, Ji et al. [23] updated 2D CNN to 3D CNN to extract features from the temporal domain and the spatial domain of the image sequence at the same time for analysis of spatiotemporal information.

3.2. Convolutional Block Attention Module (CBAM)

Woo et al. [24] proposed the idea of a convolutional block attention module. Figure 1 demonstrates the structure of CBAM.

Convolutional block attention module (CBAM) enables the attention module to be applied to both the channel dimension and the spatial dimension. We will elaborate how it works from aspects of channel attention module and spatial attention module.

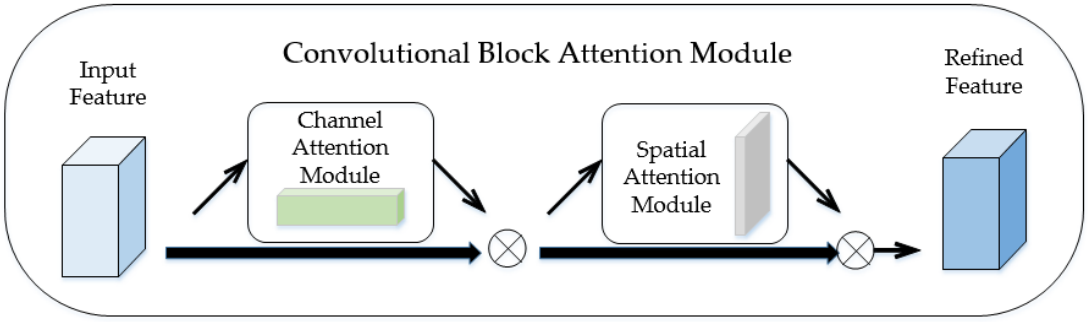

Figure 1. Convolutional block attention module (CBAM) structure.

Firstly, the preceding 3D CNN transferred the input data X into the feature map $F \in R^{C\times H\times W}$ (i.e., C represents the number of channels, H represents the height, and W the width, of the feature map) before it entered the CBAM. F was then processed by the channel attention module and spatial attention module as follows:

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F') \otimes F', \end{aligned} \tag{1}$$

Where $\otimes$ represents the element-wise multiplication, F' represents the result of the feature map multiplying the channel attention map, and F'' represents the result of the spatial attention map multiplying F' or the final output.

(i) Channel attention module

As each channel in the feature map represents one specific detector, the attention sector helped to extract the channels that contained the useful information.

To complete the feature extraction and reduce data lost, the channel attention module squeezed the feature map in the spatial dimension by using both global average pooling layer and global max pooling layer. Figure 2 shows the channel attention module. The global average pooling layer obtained the overall information, whereas the global max pooling layer captured information regarding differences of the feature. The combination of the two layers worked better than any single one.

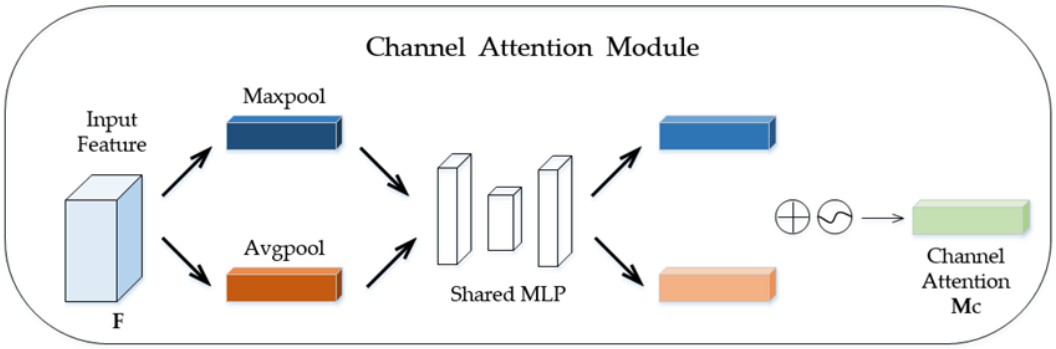

Figure 2. Channel attention module.

Afterwards, the squeezed feature maps F^c_{avg} and $F^c_{\max}$ were sent to the shared network which was constituted by the multi-layer perceptron (MLP) with one hidden layer. MLP was set at certain compression ratio to reduce the parameter and computation. The sigmoid function then computed the channel attention map of $M_c(F) \in R^{C\times 1\times 1}$, the process of which was as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(Avgpool(F)) + MLP(Maxpool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{2}$$

F^c_{avg} and $F^c_{\max}$ represent the squeezed feature maps of the two pooling layers, with σ demonstrating the sigmoid function. $W_0 \in R^{C/r\times C}$ and $W_1 \in R^{C\times C/r}$ are the weights of the MLP respectively.

(ii) Spatial attention module

As illustrated in Figure 3, the spatial attention module laid more emphasis on the parts of the feature map, which attained better reaction, in comparison with the channel attention module. For the feature maps generated in the channel attention module, the global average pooling layer and the global max pooling layer squeezed into two 2D feature maps: F^s_{avg} and $F^s_{\max}$ along the channel dimension, in order to highlight the regions containing key information.

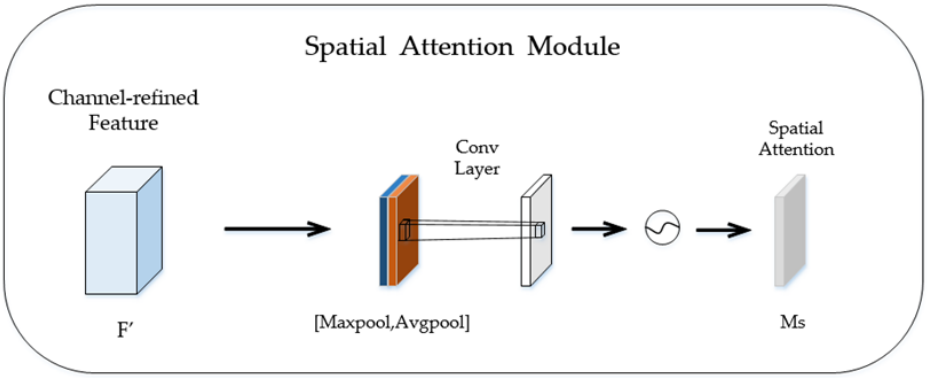

Figure 3. Spatial attention module.

The two 2D feature maps were concatenated to generate the effective feature map for convolution. After the convolution operation, the sigmoid function computed the spatial attention map: $M_s(F) \in R^{1\times H\times W}$ as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times 7}([F^s_{avg}; F^s_{\max}])), \end{aligned} \tag{3}$$

where F^s_{avg} and $F^s_{\max}$ are the feature maps squeezed in the channel dimension, and σ represents the sigmoid function.

CBAM introduced the attention mechanism to the CNN, making the channel module and the spatial module cooperate. The two modules learned the key information in both the channel dimension and the spatial dimension and redistributed the weight of the features in an adaptive way.

3.3. The Proposed System

In this study, we propose a 3D spatiotemporal CNN with convolutional block attention module, which was named as Convolutional Block Attention Module Network (CBAMNet).

The Basic Network extracted the overall spatiotemporal features of all the input data. Next, CBAM processed the feature maps and distributed the weights of the channel dimension and spatial dimension in an adaptive way.

The Basic Network and CBAM were both indispensable in that the Basic Network focused on the global information in the spatiotemporal dimension whereas CBAM highlighted the features. The two parts complemented one another, improving the function of the network.

Overfitting problems may happen when small data were used in a deep network. According to previous work [19,25], it confirms that the medium-size network trained on a small database may get higher recognition accuracy than a large or complicated network trained on the same dataset. We therefore designed a medium-size CNN. Figure 4 shows the overall structure of the CBAMNet, which includes six convolutional layers and six pooling layers.

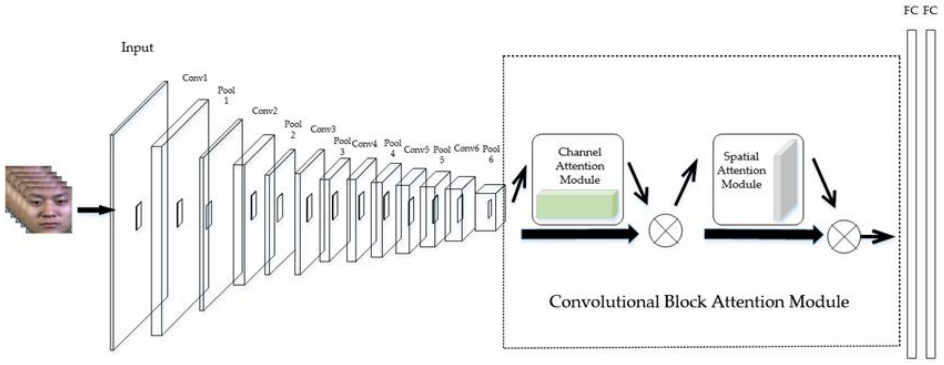

Figure 4. The overall structure of the proposed convolutional block attention module network (CBAMNet). The convolutional block attention module is shown within the dotted box.

Basic Network

The Basic Network is assigned as the front part of CBAMNet as shown in Figure 5. It is a 3D CNN, constituted by six 3D convolutional layers and six 3D pooling layers.

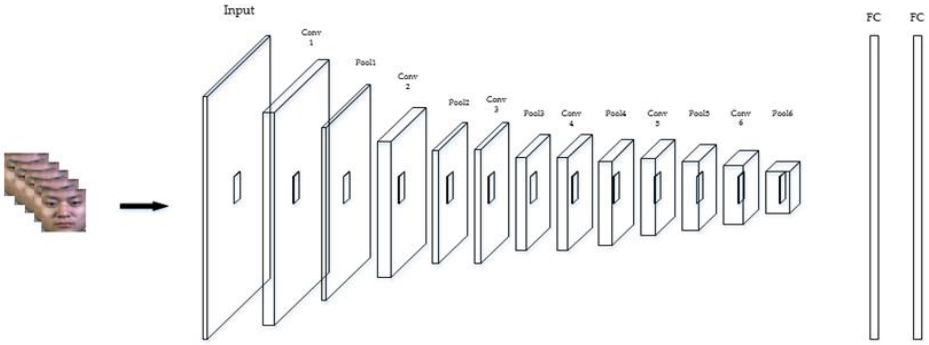

Figure 5. Basic network.

The number of filters in the convolutional layer in the Basic Network is set as 32, 32, 32, 64, 64, and 64 from the first to the last layer, respectively. The kernel of each convolutional layer is set as $3 \times 3 \times 3$ and that of each pooling layer is set as $2 \times 2 \times 2$. A batch normalization layer normalized the feature maps, which were generated from the convolutional layer and sent the maps to the Rectified Linear Units (ReLU) activation function to accelerate the training process. The dropout layer could deactivate some neurons to avoid overfitting. Finally, the softmax layer classified the image sequences into different types of micro-expressions. Table 1 concludes the detailed information of the Basic Network.

Table 1. The detailed configuration of Basic Network.

Layers	Kernel Parameter Settings	Number of Kernels	Output
Data			$64 \times 64 \times 16$
Conv1	$3 \times 3 \times 3$	32	$64 \times 64 \times 16$
Pool1	$2 \times 2 \times 2$	-	$64 \times 64 \times 16$
Conv2	$3 \times 3 \times 3$	32	$64 \times 64 \times 16$
Pool2	$2 \times 2 \times 2$	-	$32 \times 32 \times 8$
Conv3	$3 \times 3 \times 3$	32	$32 \times 32 \times 8$
Pool3	$2 \times 2 \times 2$	-	$16 \times 16 \times 4$
Conv4	$3 \times 3 \times 3$	64	$16 \times 16 \times 4$
Pool4	$2 \times 2 \times 2$	-	$16 \times 16 \times 4$
Conv5	$3 \times 3 \times 3$	64	$16 \times 16 \times 4$
Pool5	$2 \times 2 \times 2$	-	$8 \times 8 \times 2$
Conv6	$3 \times 3 \times 3$	64	$8 \times 8 \times 2$
Pool6	$2 \times 2 \times 2$	-	$4 \times 4 \times 1$

3.4. Experiments

The network was run in the TensorFlow environment. A desktop computer with graphics processing unit (GPU) of NVIDIA Geforce GTX 1080 processed all the data.

3.4.1. Database

CASME II [26] is a spontaneous micro-expression database. This database was upgraded from CASME [27] by Yan et al. in 2014. It contains 247 micro-expression samples from 26 participants. CASME II includes emotions of disgust, happiness, repression, surprise, sadness, fear, and others. The frame rate of the camera used for collecting the data is 200/fps and the image resolution is 640 × 480 to acquire information of the facial muscular movement clearly and effectively. According to previous work [28,29], we divided all the video sequences into three categories: negative (N), positive (P), and surprise (S). Negative video sequences include disgust and repression video sequences, and positive only contains happiness.

Spontaneous Micro-expression Database (SMIC) [30] was created by Zhao's team at Oulu University. The high speed (HS) dataset consists of 164 micro-expression sequences from 16 subjects. The camera frame rate is 100/fps and the image resolution is 640 × 480. The database divides micro-expression sequences into three categories: positive, negative, and surprise.

3.4.2. Data Pre-Processing

Before inputting to the network, the data were pre-processed by the following steps, which are illustrated in Figure 6.

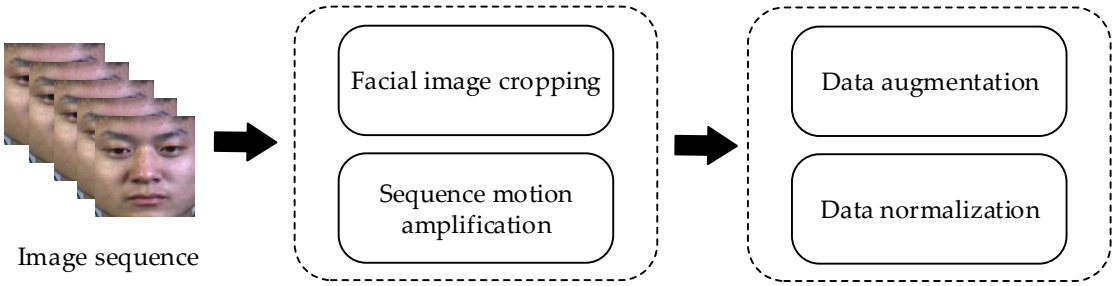

Figure 6. Steps of data pre-processing.

(1) Facial image cropping

The key facial regions were cropped in advance to avoid the interference of background factors other than human faces. Firstly, we determined the key regions by discriminative response map fitting (DRMF) [31]. DRMF detected the image sequence frame by frame and testified 66 landmark points on the face for each frame. Figure 7 shows the 66 landmark points on the face. Next, all the human face regions were kept according to the landmark points. The cropped area is demonstrated in Figure 8.

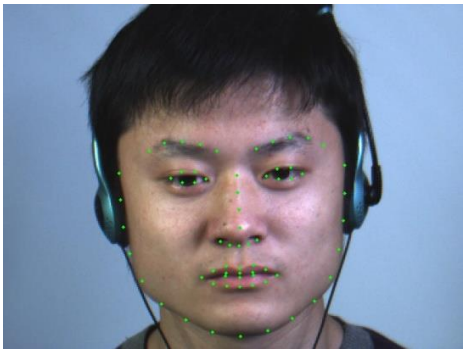

Figure 7. Sixty-six feature points.

Figure 8. The cropped area.

(2) Sequence motion amplification

To learn about the temporal motion information that was difficult to detect in the image sequence, we applied the Eulerian video magnification method proposed by Wu et al. [32] to amplify the hidden movement information in the adjacent frames.

The cropped regions of human faces were processed by down-sampling to generate the Laplacian pyramid. Then the IIR filter worked on the images of all dimensions by filtering of the temporal dimension. The 2D image signals were approached by Taylor series with fixed magnification factor (αk) on each band. The final step was to reconstruct the micro-expression sequence from the Laplacian pyramid images.

(3) Data augmentation

We randomly sampled the images' sequences and regained the length of each sub-sequence to 16 frames by using the linear interpolation method [33]. In this way, the number of subsamples of every category almost reached 5000.

(4) Data normalization

Normalization included two procedures: both the length of sequences and the size of images are normalized. The length of the micro-expression sequences in the database differed from each other. We thus normalized the image sequences into 16 frames. Furthermore, in order to reduce the amount of calculation, the size of the image was normalized to 64×64.

4. Results and Discussion

All the data after enhancement were divided into 10 groups according to the numbering of the images. The cross-validation was performed by taking nine groups of data as the training set, and the rest as the testing set. Each group of the data was the testing set. Since the images are numbered according to the subjects (i.e., the images of the same subject will have continuous numbering), the training set and the testing set will include different subjects. The cross-validation in this research is thus a cross-dataset validation.

4.1. Comparison with the Basic Network

The image sequence was sized into $16 \times 64 \times 64 \times 3$. The Basic Network and the CBAMNet proposed were both tested and verified on CASME II. Table 2 shows the average results of ten-fold cross-validation of the single network (Basic Network) and the CBAMNet on CASME II.

Table 2. Comparison of Basic Network.

Architecture	Accuracy
Basic Network	67.78%
CBAMNet	69.92%

The Basic Network used CNN as the only way to extract the overall features of human faces. The accuracy of this method was 67.78%. In contrast, the recognition rate of CBAMNet reached 69.92%. The results indicated that under the same 3D spatiotemporal CNN, the network which added the convolutional block attention module improved the recognition performance.

4.2. Comparisons with Different Basic Network Structures

The Basic Network is a neural network comprised of six convolutional layers. We decreased the number of the convolutional layers one by one to study the influence of different layers on CNN recognition. Other parameters on the network remained unchanged. Table 3 shows the network structures of the sub-networks.

Table 3. The network structures of sub-networks.

Sub-Network 1		**Sub-Network 2**		**Sub-Network 3**		**Sub-Network 4**	
2 Layers		**3 Layers**		**4 Layers**		**5 Layers**	
Conv1	$3 \times 3 \times 3$	Conv1	$3 \times 3 \times 3$	Conv1	$3 \times 3 \times 3$	Conv1	$3 \times 3 \times 3$
Pool1	$2 \times 2 \times 2$	Pool1	$2 \times 2 \times 2$	Pool1	$2 \times 2 \times 2$	Pool1	$2 \times 2 \times 2$
Conv2	$3 \times 3 \times 3$	Conv2	$3 \times 3 \times 3$	Conv2	$3 \times 3 \times 3$	Conv2	$3 \times 3 \times 3$
Pool2	$2 \times 2 \times 2$	Pool2	$2 \times 2 \times 2$	Pool2	$2 \times 2 \times 2$	Pool2	$2 \times 2 \times 2$
		Conv3	$3 \times 3 \times 3$	Conv3	$3 \times 3 \times 3$	Conv3	$3 \times 3 \times 3$
		Pool3	$2 \times 2 \times 2$	Pool3	$2 \times 2 \times 2$	Pool3	$2 \times 2 \times 2$
				Conv4	$3 \times 3 \times 3$	Conv4	$3 \times 3 \times 3$
				Pool4	$2 \times 2 \times 2$	Pool4	$2 \times 2 \times 2$
						Conv5	$3 \times 3 \times 3$
						Pool5	$2 \times 2 \times 2$

The sub-networks were tested on CASME II by experiments, and the average results of ten-fold cross-validation are shown in the following figure.

Figure 9 illustrates the recognition accuracy of the Basic Network on CASME II after deleting one, two, three, and four convolutional layers, respectively. It is seen that when the last layer is deleted (five convolutional layers), the performance is worse than the Basic Network. The recognition rate with four convolutional layers is higher than that with five convolutional layers. The rate declines as the convolutional layers are deleted gradually. However, the overall trend shows that the distinction between sub-networks is subtle. This indicates that the network functioned stably. The reason might be that the batch normalization layer accelerated the convergence speed and that the dropout layer improved the generalization function. Judging from the recognition performance, the network with six convolutional layers is suitable for the recognition. As a result, we chose the six-layer network as the Basic Network.

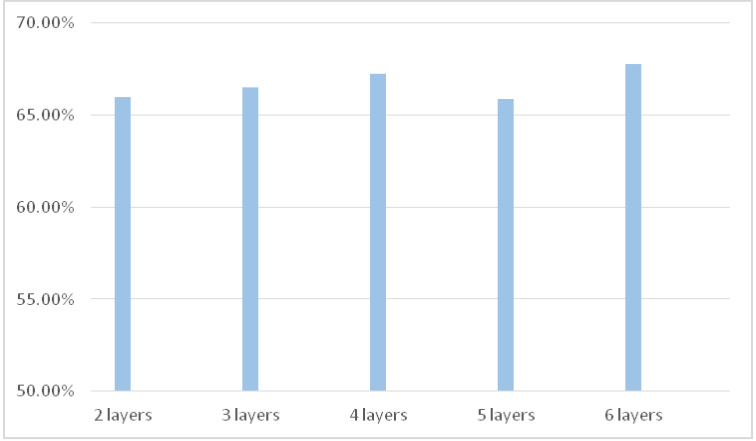

Figure 9. Performance comparison of sub-networks with different convolutional layers.

4.3. Comparison with Other Methods on CASME II

We also compared the CBAMNet with other methods on CASME II. First of all, the LBP-TOP was studied. LBP-TOP was upgraded from static local binary pattern to learn the dynamic texture features.

Secondly, the MicroExpSTCNN [21] was tested. It was a deep-learning model with 3D CNN that extracted spatiotemporal features and classified the micro-expression sequences.

In addition, the Residual Network (RESNET) [34] was measured. It is a classification network proposed by He et al. and it has been widely used in image recognition. To some degree, the RESNET has solved some training difficulties in CNN. Since the target object was image sequence, we extended RESNET to RESNET-3D in this research. We used the version of ResNet18. First the convolution kernel and pooling kernel were converted into 3D form. The convolution kernel changed from 3×3 to $3 \times 3 \times 3$, and the pooling kernel changed from 2×2 to $2 \times 2 \times 2$. In addition, some other parameters were changed accordingly. Table 4 shows the accuracy of different micro-expression recognition methods on CASME II.

Table 4. Comparison of average recognition rates of different methods. LBP-TOP: local binary patterns on three orthogonal planes; MicroExpSTCNN: xxx; RESNET-3D: Residual Network-3D; CBAMNet: convolutional block attention module network.

Method	Accuracy
LBP-TOP [10]	52.98%
MicroExpSTCNN [21]	66.07%
RESNET-3D [34]	65.01%
Basic Network	67.78%
CBAMNet	69.92%

The recognition performance achieved by MicroExpSTCNN (66.07%), RESNET-3D (65.01%), and Basic Network (67.78%) do not distinct from each other significantly, which could be that they both used neural networks. However, the Basic Network still slightly outperformed the other networks. This indicates that our design in the network functions (i.e., the optimal arrangement of the layers), improved the recognition rates and some batch normalization layers, and the drop out technology helped to prevent over fitting. The Basic Network outperforms the RESNET-3D, though it has simple network structure.

CBAMNet achieved the highest recognition rate (69.92%). This proves that the attention mechanism after the Basic Network does help to improve the recognition. By adaptively redistributing the weights of the channel features and the spatial features, the CBAM amplifies the specific information of micro-expression, thus pushing the recognition to a new level.

4.4. Comparison with Other Methods on SMIC

In this section, we also tested all the algorithms on SMIC. The results are shown in Table 5. For the SMIC database, it exhibits lower recognition performance compared to the CASME II. This performance could be due to a lower camera frame rate (100/fps) and background noises, such as illumination, shadows and so on. The recognition performance achieved by MicroExpSTCNN (50.92%), RESNET-3D (49.53%), and Basic Network (52.03%) did not show significant difference. However, the proposed CBAMNet achieved a higher recognition rate than the other models.

Table 5. Comparison of average recognition rates of different methods.

Method	Accuracy
LBP-TOP [10]	41.65%
MicroExpSTCNN [21]	50.92%
RESNET-3D [34]	49.53%
Basic Network	52.03%
CBAMNet	54.84%

4.5. Confusion Matrixes on CASME II

We also computed the confusion matrix of each algorithm on CASME II. Table 6 presents the confusion matrix of the CBAMNet. The labels at the bottom represent the predicting categories and those on the left represent the actual categories. The letters 'N', 'P' and 'S' represent negative, positive, and surprise, respectively. Table 7 demonstrates the performance of different methods on CASME II. It is seen that CBAMNet (N 69.44%, P 70.69%, S 69.65%) can achieve better performance than the Basic Network (N 66.56%, P 69.47%, S 67.33%) in the recognition in every category.

Table 6. Confusion matrix of CBAMNet. N: negative; P: positive; S: surprise.

N	69.44%	13.17%	17.39%
P	23.98%	70.69%	5.33%
S	13.33%	17.02%	69.65%
	N	P	S

Table 7. Comparison of different methods on CASME II.

Method	N	P	S
LBP-TOP [10]	56.65%	47.76%	54.55%
MicroExpSTCNN [21]	63.45%	68.98%	65.78%
RESNET-3D [34]	70.38%	62.17%	62.50%
Basic Network	66.56%	69.47%	67.33%
CBAMNet	69.44%	70.69%	69.65%

The accuracy of MicroExpSTCNN, Basic Network, and CBAMNet is higher than LBP-TOP, suggesting the spatiotemporal information might assure the stability of the network. Additionally, in all the methods based on CNN, the accuracy of the CBAMNet topped among the other methods in almost all the expressions (except that RESNET-3D outperformed others in the negative expressions). This suggests that the attention mechanism introduced to CBAMNet highlights some useful features that helped recognition.

The recognition performance of proposed CBAMNet or Basic Network is more stable than those of other methods by looking at the variations of the recognition rates of different categories. CBAMNet achieved recognition rates of 69.44%, 70.69%, and 69.65%, and the standard deviation is 0.55%. The standard deviation values of Basic Network, LBP-TOP, MicroExpSTCNN, and RESNET-3D are 1.23%, 3.79%, 2.27%, and 3.79%, respectively. This suggests that the proposed method may achieve more stable recognition results in unbalanced sample states.

4.6. Confusion Matrixes on SMIC

We computed the confusion matrixes of all the algorithms above on SMIC. Table 8 presents the confusion matrix of the CBAMNet on SMIC.

Table 8. Confusion matrix of CBAMNet.

N	54.90%	28.00%	17.10%
P	34.50%	55.50%	10.00%
S	36.71%	9.17%	54.12%
	N	P	S

To further analyze the recognition performance, Table 9 demonstrates the performance of different methods on SMIC. The letters 'N', 'P', and 'S' represent negative, positive, and surprise, respectively. It is seen that the recognition rates achieved by CBAMNet (N 54.90%, P 55.50%, S 54.12%) are more stable than those of other methods. The standard deviation of recognition rates of CBAMNet is 0.56%, while the standard deviation values of Basic Network, LBP-TOP, MicroExpSTCNN, and RESNET-3D are 2.75%, 1.62%, 0.76%, and 2.94%, respectively.

Table 9. Comparison of different methods on SMIC.

Method	N	P	S
LBP-TOP [10]	41.43%	43.74%	39.78%
MicroExpSTCNN [21]	51.33%	51.58%	49.86%
RESNET-3D [34]	53.67%	47.75%	47.17%
Basic Network	55.92%	49.92%	50.27%
CBAMNet	54.90%	55.50%	54.12%

4.7. Potential Application and Improvement

The method proposed by Zhi et al. [35] achieved remarkable results on micro-expression databases. The overall accuracy is 97.6% and 97.4% on CASME II and SMIC, respectively. They proposed a recognition algorithm for micro-expression sequences, which combines 3D CNNs with transfer learning. Firstly, the normal facial expression database is utilized as the source data, and 3D CNN model is pre-trained on it. Then the parameters of the overall network are stored and transferred to the micro-expression databases.

The reason for the remarkable results of the algorithm may be due to the reasonable design of 3D CNNs and the strong ability of transfer learning. The high level features of facial expression are learned in larger datasets (normal facial expression database) and are then transferred to the micro-expression recognition task. Though the accuracy achieved by our method is lower, our work proposes an algorithm of integrating attention mechanism to emphasize useful information. The emphasis of our work is the middle-size but suitable 3D CNN structure design and the attention mechanism. Combing the attention mechanism with the other method, such as transfer learning, might further improve the recognition rate. In the future we may pre-train a deeper network on larger normal facial expression databases, and then fine-tune the network together with attention block on the micro-expression database.

5. Conclusions

We introduced an attention mechanism into the micro-expression recognition in this research to improve recognition performance. The deep model of 3D spatiotemporal CNN with convolutional block attention module (CBAM) was proposed. The CBAMNet was constituted by two parts: the Basic Network at the front which learned the overall movement of the micro-expressions and extracted the spatiotemporal features from the image sequence; and CBAM following it which reinforced the features and accelerated the information flow in the network. The network distributed the weight of the features in an adaptive manner with the help of serial fusion. We further fused the Basic Network and CBAM to extract expression features in motion. The experiment in the CASME II indicated that CBAMNet outperformed other networks in micro-expression recognition.

Author Contributions: Conceptualization, B.C., Z.Z. and T.C.; Methodology, B.C., and T.C.; Data curation, Z.Z. and X.L.; Formal analysis, N.L. and Y.T.; Validation, B.C.; all authors wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Ekman, P. Darwin, deception, and facial expression. *Ann. N. Y. Acad. Sci.* **2003**, *1000*, 205–221. [CrossRef] [PubMed]
2. Ekman, P. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*; Henry Holt and Company: New York, NY, USA, 2003.
3. Yan, W.J.; Wu, Q.; Liang, J.; Chen, Y.H.; Fu, X. How fast are the leaked facial expressions: The duration of micro-expressions. *J. Nonverbal. Behav.* **2013**, *37*, 217–230. [CrossRef]
4. Porter, S.; Brinke, L.T. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychol. Sci.* **2008**, *19*, 508–514. [CrossRef]
5. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural. Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
6. Verma, M.; Vipparthi, S.K.; Singh, G.; Murala, S. Learnet: Dynamic imaging network for micro expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 1618–1627. [CrossRef]
7. Xia, Z.Q.; Hong, X.P.; Gao, X.Y.; Feng, X.Y.; Zhao, G.Y. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions (vol 22, pg 626, 2020). *IEEE Trans. Multimed.* **2020**, *22*, 1111. [CrossRef]
8. Wang, S.J.; Li, B.J.; Liu, Y.J.; Yan, W.J.; Ou, X.Y.; Huang, X.H.; Xu, F.; Fu, X.L. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* **2018**, *312*, 251–262. [CrossRef]
9. Polikovsky, S.; Kameda, Y.; Ohta, Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), London, UK, 3 December 2009.
10. Zhao, G.Y.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal.* **2007**, *29*, 915–928. [CrossRef]
11. Pfister, T.; Li, X.; Zhao, G.; Pietikainen, M. Recognising Spontaneous Facial Micro-Expressions. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1449–1456.
12. Wang, S.; Yan, W.; Li, X.; Zhao, G.; Fu, X. Micro-expression recognition using dynamic textures on tensor independent color space. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4678–4683.
13. Liu, Y.J.; Zhang, J.K.; Yan, W.J.; Wang, S.J.; Zhao, G.Y.; Fu, X.L. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* **2016**, *7*, 299–310. [CrossRef]
14. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
15. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Montreal, QC, Canada, 2014.
16. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
18. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
19. Peng, M.; Wang, C.Y.; Chen, T.; Liu, G.Y.; Fu, X.L. Dual temporal scale convolutional neural network for micro-expression recognition. *Front. Psychol.* **2017**, *8*, 1745. [CrossRef] [PubMed]
20. Li, J.; Wang, Y.D.; See, J.; Liu, W.B. Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Anal. Appl.* **2019**, *22*, 1331–1339. [CrossRef]

21. Reddy, S.P.T.; Karri, S.T.; Dubey, S.R.; Mukherjee, S. Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks. In Proceedings of the 2019 International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019; pp. 1–8.
22. Peng, M.; Wu, Z.; Zhang, Z.; Chen, T. From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, Xi'an, China, 15–19 May 2018; pp. 657–661.
23. Ji, S.W.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal.* **2013**, *35*, 221–231. [CrossRef]
24. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
25. Peng, M.; Wang, C.; Chen, T.; Liu, G. Nirfacenet: A convolutional neural network for near-infrared face identification. *Inf. Int. Interdiscip. J.* **2016**, *7*, 61. [CrossRef]
26. Yan, W.J.; Li, X.B.; Wang, S.J.; Zhao, G.Y.; Liu, Y.J.; Chen, Y.H.; Fu, X.L. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **2014**, *9*, e86041. [CrossRef]
27. Yan, W.; Wu, Q.; Liu, Y.; Wang, S.; Fu, X. Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–7.
28. Gan, Y.S.; Liong, S.; Yau, W.; Huang, Y.; Tan, L. Off-apexnet on micro-expression recognition system. *Signal Process. Image Commun.* **2019**, *74*, 129–139. [CrossRef]
29. Liong, S.; Gan, Y.S.; See, J.; Khor, H.; Huang, Y. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, Lille, France, 14–18 May 2019; pp. 1–5.
30. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikainen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–6.
31. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Robust Discriminative Response Map Fitting with Constrained Local Models. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3444–3451.
32. Wu, H.Y.; Rubinstein, M.; Shih, E.; Guttag, J.; Durand, F.; Freeman, W. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* **2012**, *31*, 1–48. [CrossRef]
33. Smolic, A.; Muller, K.; Dix, K.; Merkle, P.; Kauff, P.; Wiegand, T. Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems. In Proceedings of the International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 2448–2451.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. ZHI, R.; XU, H.; WAN, M.; LI, T. Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition. *IEICE Trans. Inf. Syst.* **2019**, *102*, 1054–1064. [CrossRef]