Motion magnification multi-feature relation network for facial microexpression recognition

Jing Zhang¹ · Boyun Yan¹ · Xiaohui Du¹ · Quanhao Guo¹ · Ruqian Hao¹ · Juanxiu Liu¹ · Lin Liu¹ · Guangming Ni¹ · Xiechuan Weng² · Yong Liu¹

Received: 1 September 2021 / Accepted: 30 January 2022 / Published online: 20 February 2022
© The Author(s) 2022

Abstract

Microexpressions cannot be observed easily due to their short duration and small-expression range. These properties pose considerable challenges for the recognition of microexpressions. Thus, video motion magnification techniques help us to see small motions previously invisible to the naked eye. This study aimed to enhance the microexpression features with the help of motion amplification technology. Also, a motion magnification multi-feature relation network (MMFRN) combining video motion amplification and two feature relation modules was proposed. The spatial feature is enlarged while completing the spatial feature extraction, which is used for classification. In addition, we transferred Resnet50 network to extract the global features and improve feature comprehensiveness. The magnification of the features is controlled through hyperparameter amplification factor α . The effects of different magnification factors on the results are compared, and the best is selected. The experiments have verified that the enlarged network can resolve the misclassification problem caused by the one-to-one correspondence between microexpressions and facial action coding units. On CASME II datasets, MMFRN outperforms the traditional recognition methods and other neural networks.

Keywords Micro-expression · Motion amplification · Relation module · Resnet · Transfer learning

Introduction

Microexpressions are small movements of human facial muscles that usually occur when someone deliberately conceals or suppresses their true feelings. Compared to macroexpressions, microexpressions are of short duration, have a small-expression amplitude, and are difficult to observe. A microexpression usually occurs for 1/25 to 1/5 s [1]; therefore, it is difficult to capture, even for well-trained professionals.

The study by Ekman and Friesen [2] has shown that microexpressions are expressions that occur quickly and are unconscious for people. When people are trying to suppress their current emotions, their true emotions could be expressed through microexpressions. Thus, an accurate judgment of

microexpressions is significant in the fields of public security, polygraph detection, and mental health.

In 2009, Shreve et al. [3] first presented a microexpression recognition technique based on optical flow, which detected the optical flow in key regions of face such as chin, cheek, and forehead in image sequence, and used central difference method to recognize microexpressions. In 2014, Gautam Krishna et al. [4] proposed a method using Euler video amplification technology to amplify microexpression features, and then extract Histogram of Oriented Gradients (HOG) features of enlarged images for microexpression recognition. In 2016, Kim et al. [5] put forward a model combining Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM) to classify microexpressions. The spatial and temporal characteristics of microexpressions were used, which improved the classification performance compared with the traditional methods. This work has inspired many researchers, and many subsequent methods [6] have been improved on the basis of CNN combined with LSTM network structure. In 2019, Li et al. [7] applied the 3D flow-based CNN model to extract deeply learned features that are able to characterize fine motion flow arising from minute

✉ Xiechuan Weng
wengxc2000@163.com

¹ MOEMIL Laboratory, School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

² Beijing Institute of Basic Medical Sciences, Beijing, China

facial movements, and finally achieved a good performance of 59.11% accuracy. In 2021, Wu et al. [8] used the temporal sampling deformation (TSD) to normalize the temporal lengths and conserve time-domain information for microexpression sequences and optimized a three-stream combining 2D and 3D convolutional neural network (TSNN) to classify the expressions as well, which achieved 75.49% accuracy and 61.42% F1-score.

Video motion magnification techniques can help people to observe the imperceptible movements by magnifying the small changes in the videos, such as those of vibrating airplane wings or buildings swaying under the influence of the wind [9]. Currently, there are two mainstream theories of motion amplification technology: Lagrange theory based on the prediction of motion behavior and Euler theory based on the frequency and time domains. In recent years, motion amplification algorithms have been developed based on Euler's theory. In 2018, Oh et al. [9] proposed a deep learning-based motion amplification network to replace the manual filter that was designed before convolutional neural network prevailing. This study achieved high-quality magnification in tests. We expect to use the motion amplification algorithm to enlarge the microexpression features and facilitate the recognition and classification of microexpressions. On the other hand, based on the neural network video magnification algorithm, we performed feature amplification while performing feature extraction instead of using video magnification as a preprocessing step.

Relation network appeared in the field of object detection first. Influenced by the attention model in natural language processing, object detection utilizes the correlation between objects in the image or the image context to optimize the detection effect. In 2018, Hu et al. [10] presented an object relation module to describe the relative geometry and appearance feature between the objects, which were appended as attention to the original features for regression and classification. This module realized the end-to-end training and also improved the detection effect. Thus, applying the relation module to extract the relative geometry and appearance feature between the microexpression frames improves the classification accuracy.

The present study proposes a motion magnification multi-feature relation network (MMFRN) based on a combination of motion amplification networks and two feature relation modules. MMFRN combines motion amplification and multi-feature relation module to amplify small facial movements and extract inter-frame global and local relation for microexpression recognition. Finally, the evaluations and experiments on CASME II datasets revealed that MMFRN is superior to the traditional recognition methods and other neural networks.

The contribution of our work can be summarized in three aspects. First, we innovatively introduced motion magnification networks to amplify the microexpression features, which is conducive to subsequent recognition and classification. In magnification module, the enlarged difference shape features are no longer superimposed on the original image, but are directly used for recognition, which greatly reduced the artifacts generated by the environmental background in the amplification process and improved the accuracy of subsequent recognition. The magnification of the features is controlled through hyperparameter amplification factor α . The effects of different magnification factors on the results are compared, and the best is selected. The results showed that the magnification network improves the misclassification problem caused by the one-to-one correspondence between microexpressions and facial coding units when the magnification is appropriate.

Second, we designed a motion magnification multi-feature relation network (MMFRN) combining video motion amplification, a global feature extraction module, and two feature relation modules. The network applies amplification module to amplify the features of microexpressions, and then classifies them according to the relation between the global and local features of microexpressions. Compared with other methods, the proposed network achieves the state-of-the-art results.

Third, according to the requirements of Facial Microexpression Grand Challenge 2019 (MEGC2019) [11], we used CASME II dataset to identify microexpressions in surprise, positive and negative categories. The experimental results indicate that the proposed network has a high recognition effect on the newly divided dataset three categories' experiment. Besides, we explored the influence of the apex frame in input, and introduced the SSIM parameters to illustrate the similarity of the two frames in terms of brightness, contrast, and structure. We calculated the SSIM parameters of each frame and the first frame, and then selected the different frames with the lowest SSIM parameter as input. The experimental results demonstrated that SSIM method to filter the input can replace the apex frame as the input.

Materials and methods

The used dataset and the proposed network structure are described. As mentioned above, microexpressions involve small human facial movements that are hidden deliberately and cannot be recognized or observed easily. In this instance, we apply motion video magnification technology to magnify small motions for easy observation and feature relation extraction in subsequent algorithms.

Dataset

The Fu Xiaolan team proposed three datasets, CASME, CAS (ME) [12], and CASME II [13], in 2011, 2013, and 2014, respectively. The CASME II dataset used a 200 fps high-speed camera to collect 255 microexpressions from 26 volunteers under good lighting conditions. The dataset was marked with a microexpression start frame, a microexpression end frame, and a frame with the strongest microexpression manifestation for each microexpression. The facial motion coding system associates facial muscle movements with facial expressions. An expression can be created by multiple facial muscle actions. In the CASME II dataset, researchers labeled the facial muscle movements for each sample. Although samples can be represented by a combination of facial action unit codes, the dataset still has a small number of samples that cannot be accurately determined by action units (AU), and this situation is classified under “Others,” as shown in Table 1. No one-to-one correspondence is observed between the expressions and facial coding. A facial AU appears in multiple expressions, such as AU4 and AU12; an expression is also represented by a combination of multiple action units. This gives rise to several challenges for the classification of microexpressions. Therefore, facial AUs should be classified instead of expressions. However, it is difficult for nonpsychological researchers to judge the meaning of an expression based on the combination of facial AUs, and hence, we chose to classify the expressions.

The dataset contains seven microexpressions: happiness, disgust, fear, sadness, surprise, repression, and others. Since the amount of fear and sadness is small in CASME II, these expressions were manually removed in several studies [14]. However, in this study, we continued this approach. After removing fear and sadness, 246 instances of the remaining five expressions were recorded (happiness, disgust, surprise, repression, and others). Finally, we selected the data of a total of 246 microexpressions from the five categories in CASME II as the evaluation dataset.

However, the amount of data between the five categories is not balanced. The specific data are shown in Table 1. Others comprise 40.2% of the data, and surprise comprises only 10.1%. This data imbalance increases the bias of network learning.

MMFRN structure

The MMFRN we proposed is composed of three parts: a spatial feature amplification network, a global feature extraction module, and a multi-feature relation module. As shown in Fig. 1, the magnification module is a spatial feature amplification network, Resnet50 is a global feature extraction network, and the relation module is a multi-feature relation

network. These three networks are connected through concatenation functions. After the relation network, the Flatten layer is added, and finally, softmax is applied to recognize the microexpressions.

Micro-expression motion magnification

The definition of motion magnification was followed of the study by Wu et al. [15] and Wadhwa et al. [16]. Each frame of a video containing motion can be represented by Eq. (1)

$$I(x, y, t) = f(M(x, y) + \delta(x, y, t)), \quad (1)$$

where (x, y) represents the pixel coordinates in the image. $M(x, y)$ represents the pixel space in the image that is not related to time, and $\delta(x, y, t)$ represents the pixel space in the image sequence that changes with time. $f()$ represents the Fourier function.

After motion magnification, the enlarged image can be expressed by Eq. (2)

$$\tilde{I}(x, y, t) = f(M(x, y) + \alpha \cdot \delta(x, y, t)), \quad (2)$$

where α is the amplification factor.

The movement offset of two similar moments can be calculated by Eq. (3)

$$\Delta h(x, y, t) = h(x, y, t) - h(x, y, t'), \quad (3)$$

where $h(x, y, t)$ represents the displacement function of the target at (x, y) with time t .

Equation (3) shows that $\Delta h(x, y, t)$ represents the deviation of the displacement between the two time points. The magnification of the motion amplitude is also the magnification of the displacement deviation between the two time points, such that the motion amplification can be expressed by Eq. (4)

$$\tilde{h}(x, y, t) = h(x, y, t) + \beta \cdot \Delta h(x, y, t), \quad (4)$$

where β is the amplification factor.

The facial expression images were collected continuously to form an image sequence, and each frame reflects the facial expression change. Therefore, any two frames in the image sequence could be used to represent the changes in facial expression at two moments. Then, Eq. (3) is applied to estimate the difference in the two images at different moments, as shown in Eq. (5)

$$\Delta h(x, y, t) \approx \delta(x, y, t). \quad (5)$$

Table 1 CASME II face coding expression and distribution

	Happiness	Disgust	Surprise	Repression	Others	Sum
Action units	1, 6, 7, 12, 14, 15	2, 4, 5, 6, 7, 9, 10	1, 2, 12, 14, 25, 26, 38	12, 14, 15, 17	4, 6, 7, 12, 14, 15, 17, 38	
Class codes	0	1	2	3	4	
Number/percent	32/13.0%	63/25.6%	25/10.1%	27/10.9%	99/40.2%	246

Fig. 1 MMFRN structure

According to Eqs. (4) and (5), the small motion in the captured image sequence can be enlarged by Eq. (6) as follows:

$$\tilde{h}(x, y, t) = f(M(x, y) + \alpha \cdot \delta(x, y, t)). \quad (6)$$

The facial movements represented by the facial expressions corresponding to the microexpressions show that the facial features play a critical role in expressions. For example, a clear change in the common facial action AU4 is the reduction of eyebrows, and a clear change in AU12 is the rise of the corners of the mouth. In the image, the shape of the facial features is more prominent than the texture features. Thus, we speculated that the shape features are more important than the texture features in microexpression recognition. This phenomenon is also displayed by the output image of the network encoder module during training (Fig. 2).

Although their duration is short, the microexpressions captured by a high-speed camera can still last for dozens of frames. These images can be considered as having time continuity. A microexpression does not change suddenly within its duration, and therefore, after extracting and enlarging the

spatial features of a single frame image, the relation network can be used to extract the relative geometry and appearance relations of the enlarged features.

Mainstream microexpression datasets use high-speed cameras (200 fps) to collect samples; this ensures the collection of changes in the microexpressions. However, due to the tiny changes in facial movements of the microexpressions and the high-speed camera having a short frame acquisition time (5 ms), the differences in the image between adjacent frames, i.e., the motion offset, Δh , is small, as shown in Fig. 3a, which is not conducive to subsequent feature extraction and amplification.

To increase the motion offset Δh , we used the microexpression start frame (onset frame), microexpression end frame (offset frame), microexpression maximum frame (apex frame), and the middle frame (mid-frame) between the microexpression start frame and the microexpression maximum frame with the microexpression labeled in the dataset. A total of four frames were arranged in chronological order, wherein the microexpressions occur as the input of MMFRN.

Fig. 2 Encoder output image

(a) Image difference between adjacent frames

(b) Network input with four frames of image difference

Fig. 3 Comparison of microexpression image difference. **a** Image difference between adjacent frames. **b** Network input with four frames of image difference

The inter-frame motion offset Δh input from the network is shown in Fig. 3b.

We magnified the image lossless by five times and used grid lines to show the tiny differences in the eyebrows of the four frames, as shown in Fig. 3b. The mesh is an enlargement of the eyebrows. It can be observed from Fig. 3b that the eyebrows in the apex frame exceeded the bottom grid line, while the other frames did not. In addition, the proportion

of black pixels in the fourth row and column of the apex frame is the largest, which is significantly higher than that of other frames. For the sunken area above the eyebrow, the sunken location of the largest frame exists more obvious in the second row and column of the grid, while others do not occupy a large proportion in this grid.

The spatial feature amplification network is designed based on deep learning which takes the feature extraction

Fig. 4 Spatial feature amplification network structure. **a** Magnification overall framework. **b** Encoder module. **c** Res.Blk. **d** Manipulator module. **e** Decoder module

and feature amplification functions into account. It is also composed of three modules, as shown in Fig. 4a. However, due to its different tasks, this network has made some deletions and modifications in specific modules compared to the deep learning amplification network. In the original network, the encoder and decoder implemented end-to-end network inputs and outputs of arbitrary dimensions for the fully convolutional network. A residual network was used to ensure the quality of the output image. However, in the classification task, it is not necessary to ensure that the output and input dimensions are consistent; therefore, the decoder module was redesigned.

The encoder demonstrated that it could decompose the texture and shape features of the image in the network. In

the foregoing analysis, shape features are more important for facial expression analysis than texture features, and so only the shape part of the encoder module is used. Simultaneously, to reduce the feature dimension and increase the number of features, the convolutional layer before the two residual layers is modified to a step size of 2, and the convolutional kernel is set to 64. The specific network structure is shown in Fig. 4b. Conv $\langle c \rangle_k \times k_s \times s$ denotes a convolutional layer of c channels, $k \times k$ kernel size, and stride s . The internal structure of Res.Blk. is shown in Fig. 4c. The channel, kernel, and stride of the convolutional layer in Res.Blk. are consistent with the parameters of the previous layer.

The manipulator module is the core module for feature amplification. It also makes specific modifications based on the deep learning motion amplification network. The specific network structure is shown in Fig. 4d. For the detection and recognition of microexpressions, the changes in the expressions are under intensive focus; therefore, in the manipulator module, the motion is no longer superimposed on the original image after magnification. In addition, because the output dimension of the encoder module changes, the convolutional kernel dimensions of all convolutional layers in the manipulator are adjusted to 64.

The decoder module has been redesigned to match the input of the relation network. Then, two convolutional layers with a size of 3×3 are used to reduce overfitting and increase the nonlinearity of the network. To further reduce the feature size and increase the number of features, the unified step size of the convolutional layer is set to 2, and the convolutional kernel dimensions of the two convolutional layers are 128 and 256, respectively. The specific network structure is shown in Fig. 4e. To meet the input of the relation network, the concatenation function is connected after flattening.

Multi-feature relation module

The magnification module removes the texture features of microexpressions and only extracts the shape features for amplification. To reduce the dimension of the features and increase the number of features, the strides of many layers were set to 2, which led to the loss of many global features. To improve the comprehensive features, we transferred Resnet50 to extract the microexpressions global feature to assist in the recognition. After the magnification and global feature extraction, the global feature and the magnified changing shape feature were fed into the relation module, respectively, to extract the relations of global and local features between the microexpression frames, finally combining the two relations to classification.

The feature relation module refers to the object relation module proposed by Hu et al. [10]. First, the weight of relative geometry and appearance feature between the m th microexpression frame and the current n th frame was calculated, as shown in Eqs. (7) and (8). The computation of ω_G^{mn} requires the calculation of the relative geometrical features of two frames and then embedding the four-dimensional relative geometry features into higher dimensions to obtain the features $\epsilon_G(f_G^m, f_G^n)$. Then, the embedded feature was dotted with W_G to obtain the weight of scalar ω_G^{mn} . If the weight is < 0 , then assign 0. In the equation, f_A and f_G represent the appearance feature and relative geometry, and ω_A^{mn} and ω_G^{mn} are weights, respectively. W_K , W_Q , W_G , W_V are transformation matrix, $f_R(n)$ is the output of local relation module, N_r is the number of the local relation module, and d_k is the feature dimension

$$\omega_A^{mn} = \frac{\text{dot}(W_K f_A^m, W_Q f_A^n)}{\sqrt{d_k}} \quad (7)$$

$$\omega_G^{mn} = \max\{0, W_G \cdot \epsilon_G(f_G^m, f_G^n)\}. \quad (8)$$

The total weight combines the weights of these two features, which are then normalized as Eq. (9)

$$\omega^{mn} = \frac{\omega_G^{mn} \cdot \exp(\omega_A^{mn})}{\sum_k \omega_G^{kn} \cdot \exp(\omega_A^{kn})} \quad (9)$$

According to the total weight of the m th relative to the n th frame, the output of every local relation module was calculated using Eq. (10)

$$f_R(n) = \sum_m \omega^{mn} \cdot (W_V \cdot f_A^m). \quad (10)$$

Then, all the local relation module features were concated and superimposed on the n th frame's original feature to output the new features, which have the same number of channels as before (Eq. 11)

$$f_A^n = f_A^n + \text{Concat}[f_R^1(n), \dots, f_R^{N_r}(n)], \text{ for all } n \quad (11)$$

The structure of the feature relation module is shown in Fig. 5.

The flattened layer is added to the relation model to transform the relation features into one-dimensional vectors. These vectors of global and local feature relations are concated, and the output of the concatenate function is normalized by the softmax function to obtain the final prediction vector.

General network configuration

The MMFRN network is implemented in tensorflow-gpu = 2.2.0, python = 3.6, cuda = 10.0, and cudnn = 7.4. Next, we use Adam [17] as the optimizer to minimize the softmax cross-entropy loss function with batch size 4. A learning rate of 0.00001 was set at decay of 0.000001. Moreover, the learning rate was tuned to be smaller than the typical rates because of the subtleness of a microexpression, which poses difficulty for learning [14]. Specific settings are shown in Table 2.

Owing to the small dataset, the training epoch is set to 50, and as mentioned above, CASME II dataset contains 246 microexpressions, and the batch size is set to 4. Therefore, 62 steps are required to complete each epoch. We set the epoch to 50 and trained about 3000 steps. However, in the actual training process, the loss function decreases rapidly. To avoid meaningless training steps, the network is set to end training in advance when the average loss function of total training steps is < 0.05 . The loss-of-function is illustrated in Fig. 6.

Fig. 5 Feature relation module structure [10] **Table 2** Experimental environment and settings

Environments	Loss function	Optimizer	Learning rate	Decay	Epoch	Batch size
Tensorflow-gpu = 2.2.0, python = 3.6, cuda = 10.0, and cudnn = 7.4	The cross-entropy loss function	Adam	0.00001	0.000001	50	4

Fig. 6 Loss function convergence

Results

Experiments

CASME II-provided cropped video frames were used as the network input. To match the network's size requirements for input images, all images were resized to 256×256 . For machine learning problems and classification tasks on highly skewed databases (such as CASME II), the accuracy rate was an inadequate measure of the effectiveness of a classifier despite its popularity in the literature [18].

The accuracy cannot consider the classification performance of the method for each category. A small number of categories have a much smaller impact on accuracy than

a larger number of categories. In other words, accuracy is affected by the weight of the number of categories. In addition to using accuracy as the evaluation standard for the model, the Macro-F1-score and unweighted-average-recall (UAR) were also used as the evaluation standards for the model [19]. The evaluation bias due to imbalanced datasets is avoided.

Precision (Precision) reflects the proportion of all types of data in each category that is correctly predicted. Recall (Recall) reflects the proportion of correctly predicted data in all the categories of data predicted for the specific category. UAR calculates the recall rate for each category separately and then averages them by category. The unweighted representative average was not related to the number of each category but only the number of categories. The F1-score is a parameter that reflects the performance of the model by two indicators of comprehensive precision and recall, as shown in formula (12), where P represents precision and R represents recall

$$F1 = 2 \times \frac{P * R}{P + R} \quad (12)$$

The Macro-F1 score calculates the F1 score value of each category separately and an unweighted average. As shown in Eq. (13), N is the number of categories

$$Macro - F1 = \frac{\sum_{i=1}^N F1_i}{N} \quad (13)$$

Table 3 Performance of the proposed methods vs. other methods of microexpression recognition

Methods	F1-Score	UAR	Accuracy/WAR
LBP-TOP [20]	0.2941	0.3094	0.4595
LBP-SIP [21]	0.4480	N/A	0.4656
Adaptive MM + LBP-TOP [19]	N/A	N/A	0.5191
FDM [22]	0.4053	N/A	0.4593
EVM + HIGO [24]	N/A	N/A	0.6721
CNN-LSTM [25]	N/A	N/A	0.6098
ELRCN-TE [6]	0.5000	0.4396	0.5244
3D-FCNN[7]	N/A	N/A	0.5911
Residual Network with Attention[26]	0.5390	N/A	0.6590
TSNN-IF [8]	0.6142	0.6000	0.7549
MMSTMN-1	0.4950	0.4880	0.5104
MMSTMN-80	0.6328	0.6240	0.6351

To avoid deviations caused by different subjects, the performance of the network was tested using leave-one-subject-out (LOSO) cross-validation. Table 3 shows a comparison of the results of different methods. In addition to using accuracy as the evaluation standard for the model, the Macro-F1 score and UAR are also used as the evaluation standard for the model [19]. The evaluation errors due to imbalanced datasets are avoided.

At the amplification factor $\alpha = 1$, when the network is not amplified, the MMFRN method has advantages over some methods and is superior to LBP-TOP [20], LBP-SIP [21], and FDM methods [22]. This shows that MMFRN can extract and classify the microexpression-related features. When the amplification factor α is set to 80, the result of MMFRN is improved significantly compared to when it is not amplified. Also, compared to other methods, MMFRN has achieved better results on the Macro-F1 score and UAR.

Table 4 compares the network results of multiple magnification factors by changing the magnification factor α , followed by the analysis of the effect of different magnifications on the feature magnification, which would affect the MMFRN method. Thus, an optimal magnification factor should be selected. The results of multiple magnification factors are shown in Table 4. The histogram is shown in Fig. 7. When the magnification factor $\alpha = 1$, the feature is not magnified, and the result is used as a baseline value. Table 4 illustrates that when the magnification factor $\alpha = 80$, the results of the three indicators of Macro-F1 score, UAR, and accuracy are optimal and better than the baseline value. Figure 7 manifests that not all other values with a magnification factor > 1 are better than the baseline results. For example, when $\alpha = 70$, all three indicators are lower than the

Table 4 Comparison of results between various magnification factors

α	Macro -F1-score	UAR	Accuracy
1	0.494987	0.488	0.510406
10	0.523321	0.500	0.539782
20	0.541637	0.525	0.553265
30	0.560432	0.550	0.588373
40	0.542631	0.529	0.559413
50	0.556983	0.536	0.563183
60	0.591470	0.580	0.615691
70	0.480804	0.472	0.508130
80	0.632770	0.624	0.635122
90	0.505728	0.493	0.522448
100	0.544846	0.54	0.566911

baseline values. A comparison of the results between multiple magnifications indicates that magnification is not positively correlated with the results. Thus, selecting an appropriate magnification factor improves the final classification result.

To further analyze the impact of the magnification factor on the results of each classification, the baseline ($\alpha = 1$), the best overall performance ($\alpha = 80$), the worst overall performance ($\alpha = 70$), the attenuating performance ($\alpha = 100$), and randomly selected magnification ($\alpha = 30$ and 60) draw six confusion matrixes (Fig. 8). Figure 9 shows the plots of the UAR of these six magnification factors for each class.

As shown in Fig. 8, the classification effects of category 2 (surprise) and category 4 (others) are significantly better than that of the other three categories for each magnification factor, and the classification effect of category 3 (repression) is significantly lower than the remaining four categories, except for the baseline. The MMFRN network is good at classifying categories 2 and 4, but the classification effect of category 3 is poor.

The confusion matrix in Fig. 8 shows that in each amplification factor, most of the other classes are divided into category 4 (others), especially category 1 (disgust) and category 3 (repression). This finding could be ascribed to the fact that the others category contains the largest proportion of the dataset (40.2%).

Category 4 (others) in the CASME II dataset consists mostly of the AU4 action in the face coding system; moreover, AU4 appears in multiple expressions. In category 1 (disgust), only 3/63 categories do not contain action AU4. As shown in Fig. 8, in the baseline result, 56% of category 1 data were classified as category 4, which exceeded its own recall rate of 32%; after amplification, the above situation was improved. When the magnification factor $\alpha = 60$, the recall rate of category 1 is increased to 60% and mistakenly classified as category 4 and decreased to 33%.

Fig. 7 Comparison of results of different magnification factors

In categories 0, 2, and 3, the best performance is achieved when the magnification factor $\alpha = 80$; in category 1, the best performance is achieved when the magnification factor $\alpha = 60$; in category 4, the performance is best when the magnification factor $\alpha = 70$. Thus, the comparison of the six amplification factors did not reveal any single amplification factor with optimal performance in each category. The overall result is optimal when the magnification factor is 80.

Compared to the results of category 3, the results of the magnification factor $\alpha = 70$ are lower than the baseline results, indicating that the classification results after magnifying the feature are not better than the unmagnified results. Thus, it is necessary to choose an appropriate amplification factor. After using the appropriate magnification factor ($\alpha = 80$), compared to the baseline results, five categories showed a better performance; 20% improvement was achieved in categories 0 and 1. Compared to the other methods, the final results obtained by means of the appropriate amplification factors have obvious advantages.

Supplementary experiment

Three categories' experiment

Facial Micro-expression Grand Challenge 2019 (MEGC2019) simplified the microexpressions into general emotion categories, namely negative, positive, and surprise, to adapt to different stimuli and environmental setups of various datasets, which reduce ambiguity among the emotion categories [23]. According to MEGC2019 competition requirements, happiness in the CASME II dataset is divided into positive category, while disgust and repression are classified into negative category, and surprise is divided into surprise category. Then, we tested the performance of

MMFRN on the newly divided CASME II dataset at the optimal magnification factor. Table 5 reports our results compared to other methods, revealing that MMFRN is much better than the other four methods on the Macro-F1-score, UAR, and accuracy. To reflect the classification more intuitively, we draw the confusion matrix of MMFRN, as shown in Fig. 10, where categories 0, 1, and 2 represent surprise, negative, and positive, respectively.

As can be seen from the confusion matrix in Fig. 10, on the newly divided CASME II dataset, MMFRN has high recognition accuracy for all three categories. In the surprise category, accuracy is maintained as high. In both negative and positive categories, recognition accuracy improves, especially in the negative category, and the accuracy is increased by about 30%. This phenomenon could be attributed to the fact that on the newly divided dataset, the negative emotion category accounted for the largest proportion, and the bias effect of other categories in the original CASME II dataset was removed. In conclusion, the experimental results showed that the MMFRN network has excellent performance on the original CASME II dataset five categories experiment and a high recognition effect on the newly divided dataset three categories' experiment.

SSIM experiment

To increase the microexpression changes between input images, the apex frame was used as an input image of MMFRN; also, the influence of the apex frame as input was explored. The SSIM parameters illustrated the similarity of the two images in terms of brightness, contrast, and structure. As shown in Eq. (14), (x, y) represents the two input images, μ represents the mean, σ represents the variance, and

Fig. 8 Confusion matrix with different magnifications. **a** $\alpha = 1$. **b** $\alpha = 30$. **c** $\alpha = 60$. **d** $\alpha = 70$. **e** $\alpha = 80$. **f** $\alpha = 100$

c avoids division by zero

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (14)$$

Herein, we used the microexpression data to calculate the SSIM parameters of each frame and the first frame of the image. The experimental results demonstrated that the difference frame with the lowest SSIM parameter is not consistent

with the apex frame. Hence, the difference frame was used instead of the apex frame as input for the MMFRN. Then, we calculated the distance between the difference frame and the start and end frames. The middle frame between the greater distance frame and the difference frame was selected as the mid-frame. Therefore, the input of MMFRN becomes the start frame, new mid-frame, difference frame, and end frame.

Table 6 shows the comparison between the inputs using the apex and the difference frames. The results of the two

Fig. 9 Comparison chart of each type of recall with different magnifications

Fig. 10 Confusion matrix of three categories experiments

Table 5 Comparison of results on three categories

Methods	F1-Score	UAR	Accuracy/WAR
CapsuleNet2019 [27]	0.7068	0.7018	N/A
TSCNN2019 [28]	0.6913	0.6879	N/A
CBAMNet2020 [29]	N/A	N/A	0.6993
AU-ICGAN2020 [30]	0.3550	N/A	0.7120
MMFRN-80	0.7520	0.7530	0.7881

input methods were similar. Using the difference frame as input when magnifying by 100 yielded a better UAR result

than the apex frame input. Therefore, the SSIM method to filter the input can replace the apex frame as the input.

Discussion

As can be seen from Table 4, the classification result is not positively correlated with the magnification factor. With the increase in magnification factor, recognition accuracy increases, followed by a decrease, except when the amplification factor is 70. This phenomenon could be attributed to the increased magnification factor in an appropriate range; subsequently, the microexpression features are well amplified with an improved classification effect. However, when the amplification factor increases, the microexpression features amplify the noise significantly, which interferes with the classification.

As mentioned above, each type of microexpression has multiple facial action units. For example, both happiness and surprise categories contain AU12, which represents the variation of the corners of the mouth. In the surprise category, the corners of the mouth are raised slightly, while in the happiness category, these are raised markedly. However, when the magnification factor is large, such as 100, the network will magnify the extent to which the corners of the mouth are raised, such that surprise may be misclassified as happiness, and as a result, the network performance declines. The confusion matrix illustrates this phenomenon (Fig. 8). For example, when the magnification factor α is 1, 30, 60, 70, or 80, category 2 data (surprise) are hardly classified as category 0 (happiness), but when the magnification factor $\alpha = 100$, 18% of category 2 data are misclassified as category 0, leading to a decline in classification.

The experiments demonstrated that the motion magnification network improves the misclassification problem caused

Table 6 Comparison of input results

Methods	F1-score	UAR	Accuracy/WAR
MMFRN-1	0.4950	0.488	0.5104
MMFRN-80	0.6328	0.624	0.6351
MMFRN-100	0.5448	0.540	0.5669
MMFRN-SSIM-1	0.4834	0.493	0.5053
MMFRN-SSIM-80	0.6164	0.627	0.6185
MMFRN-SSIM-100	0.5729	0.639	0.5853

by the one-to-one correspondence between microexpressions and facial coding units when the magnification is appropriate. The experimental comparison between multiple amplification factors provides an overall optimal amplification factor α . Simultaneously, the optimal amplification factors corresponding to the expressions of each category were not consistent. If an optimal amplification coefficient for each category is detected, the performance can be improved further.

Thus, a large number of experiments need to be conducted to find the optimal amplification factor for each type of microexpression in the future. Hence, we design a classification amplification network, which uses the corresponding amplification factor to amplify each type of microexpression, such that the classification performance of the network is improved significantly.

Acknowledgements We would like to thank Professor Yutang Ye and all the staff at MOEMIL Laboratory for their help in this study.

Author contributions JZ, BY, and XD conceived the original method. JZ, BY, XD, and QG implemented the algorithm and performed the simulation. JZ and BY wrote the manuscript. All authors did the data analysis, contributed to the manuscript, and approved the submitted version.

Funding This research was supported partly by the National Natural Science Foundation of China (Nos. 61405028 and 82073833) and the Fundamental Research Funds for the Central Universities (University of Electronic Science and Technology of China) (Nos. ZYGX2015J046 and ZYGX2016J069, ZYGX2019J053).

Data availability The CASME II database is online public available now for testing (see <http://fu.psych.ac.cn/CASME/casme2-en.php> for details).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the

permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ekman P, Friesen WV (1980) Constants across cultures in the face and emotion.pdf. 124,129
- Ekman P, Friesen WV (1969) Nonverbal leakage and clues to deception. *Psychiatry* 32:88–106. <https://doi.org/10.1080/00332747.1969.11023575>
- Shreve M, Godavarthy S, Manohar V et al (2009) Towards macro- and micro-expression spotting in video using strain patterns. 2009 Work Appl Comput Vision. WACV 2009:26–31. <https://doi.org/10.1109/WACV.2009.5403044>
- Krishna G, V SKN (2014) Micro-expression extraction for lie detection using Eulerian video (motion and color) magnification
- Kim DH, Baddar WJ, Ro YM (2016) Micro-expression recognition with expression-state constrained spatio-temporal feature representations. MM 2016 - Proc 2016 ACM Multimed Conf 382–386. <https://doi.org/10.1145/2964284.2967247>
- Khor HQ, See J, Phan RCW, Lin W (2018) Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: Proceedings of 13th IEEE Int Conf Autom Face Gesture Recognition, FG 2018 667–674. <https://doi.org/10.1109/FG.2018.00105>
- Li J, Wang Y, See J, Liu W (2019) Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Anal Appl* 22:1331–1339. <https://doi.org/10.1007/s10044-018-0757-5>
- Wu C, Guo F (2021) TSNN: three-stream combining 2D and 3D convolutional neural network for micro-expression recognition. *IEEJ Trans Electr Electron Eng* 16:98–107. <https://doi.org/10.1002/tee.23272>
- Oh TH, Jaroensri R, Kim C et al (2018) Learning-based video motion magnification. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 11208 LNCS:663–679. https://doi.org/10.1007/978-3-030-01225-0_39
- Hu H, Gu J, Zhang Z et al (2018) Relation networks for object detection. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. <https://doi.org/10.1109/CVPR.2018.00378>
- Xia Z, Peng W, Khor HQ et al (2020) Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans Image Process* 29:8590–8605. <https://doi.org/10.1109/TIP.2020.3018222>
- Yan WJ, Wu Q, Liu YJ, et al (2013) CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In: 10th IEEE Int Conf Work Autom Face Gesture Recognition, FG 2013. <https://doi.org/10.1109/FG.2013.6553799>

13. Yan WJ, Li X, Wang SJ et al (2014) CASME II: an improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* 9:1–8. <https://doi.org/10.1371/journal.pone.0086041>
14. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24:971–987. <https://doi.org/10.1109/TPAMI.2002.1017623>
15. Wu HY, Rubinstein M, Shih E et al (2012) Eulerian video magnification for revealing subtle changes in the world. *ACM Trans Graph*. <https://doi.org/10.1145/2185520.2185561>
16. Wadhwa N, Rubinstein M, Durand F, Freeman WT (2013) Phase-based video motion processing. *ACM Trans Graph*. <https://doi.org/10.1145/2461912.2461966>
17. Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. In: 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc 1–15
18. Liong ST, See J, Phan RCW et al (2016) Spontaneous subtle expression detection and recognition based on facial strain. *Signal Process Image Commun* 47:170–182. <https://doi.org/10.1016/j.image.2016.06.004>
19. Park SY, Lee SH, Ro YM (2015) Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In: *MM 2015 - Proc 2015 ACM Multimed Conf* 911–914. <https://doi.org/10.1145/2733373.2806362>
20. Zhao G, Pietik M (2007) Patterns with an application to facial expressions. *Most* 29:1–14
21. Cremers D, Reid I, Saito H, Yang MH (2014) Computer Vision-ACCV 2014: 12th Asian Conference on Computer Vision Singapore, Singapore, November 1–5, 2014 Revised Selected Papers, Part I. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 9003:525–537. <https://doi.org/10.1007/978-3-319-16865-4>
22. Xu F, Zhang J, Wang JZ (2017) Microexpression identification and categorization using a facial dynamics map. *IEEE Trans Affect Comput* 8:254–267. <https://doi.org/10.1109/TAFFC.2016.2518162>
23. See J, Yap MH, Li J, et al (2019) MEGC 2019 - The second facial micro-expressions grand challenge. In: *Proceedings of 14th IEEE Int Conf Autom Face Gesture Recognition, FG 2019* 1–5. <https://doi.org/10.1109/FG.2019.8756611>
24. Li X, Hong X, Moilanen A et al (2018) Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Trans Affect Comput* 9:563–577. <https://doi.org/10.1109/TAFFC.2017.2667642>
25. Wang SJ, Li BJ, Liu YJ et al (2018) Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* 312:251–262. <https://doi.org/10.1016/j.neucom.2018.05.107>
26. Wang C, Peng M, Bi T et al Micro-Attention for Micro-Expression Recognition. 1–17
27. Van Quang N, Chun J, Tokuyama T (2019) CapsuleNet for micro-expression recognition. *Proc - 14th IEEE Int Conf Autom Face Gesture Recognit FG 2019*. <https://doi.org/10.1109/FG.2019.8756544>
28. Song B, Li K, Zong Y et al (2019) Recognizing spontaneous micro-expression using a three-stream convolutional neural network. *IEEE Access* 7:184537–184551. <https://doi.org/10.1109/ACCESS.2019.2960629>
29. Chen B, Zhang Z, Liu N et al (2020) Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition. *Informatics*. <https://doi.org/10.3390/INFO11080380>
30. Xie H-X, Lo L, Shuai H-H, Cheng W-H (2020) AU-assisted graph attention convolutional network for micro-expression recognition. 2871–2880. <https://doi.org/10.1145/3394171.3414012>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.