Micro-expression recognition by two-stream difference network

Hang Pan 1 | Lun Xie 1 | Juan Li 2 | Zeping Lv 3 | Zhiliang Wang 1

Abstract
Facial micro-expression is a superposition of micro-expression features and identity information of a subject. For the problem of identity information interference in micro-expression recognition, this study proposes a new method for facial micro-expression recognition by de-identity information, called two-stream difference network (TSDN). First, a two-stream encoder-decoder network is trained by a convolutional neural network, where the input of the micro-expression stream is a micro-expression image, and the identity stream is a facial identity image. The micro-expression image is the apex image, and the identity image is the onset image in the micro-expression sequence. The identity information and micro-expression features are recorded in the intermediate layer of the micro-expression stream, while the intermediate layer of the identity stream contains only the identity information of a subject. Then, the identity information is removed by the difference network, but micro-expression features are stored in the intermediate layer of the micro-expression stream. Given the sequence of the micro-expressions, the TSDN model of de-identity information learns the difference that stores in the expression stream. Two public spontaneous facial micro-expression data sets (SMIC and CASME II) are employed in our experiments. The experiment results show that our model can achieve a superior performance in micro-expression recognition.

1 | INTRODUCTION

The facial expression reflects a person’s emotional state through the movement of facial muscles. Generally, facial expressions are classified into macro-expressions and micro-expressions according to the range of facial muscle movement areas, intensity level, and movement time. The macro-expression is that the facial muscles have more movement, high intensity, and long duration. Therefore, people can easily perceive the facial macro-expression during communication and use it for intelligent interaction [1], smart medical [2], social security [3], and online education [4]. Compared with the macro-expression, micro-expression occurs only in small areas of facial muscle movement, with low intensity and short duration [5]. These characteristics make the micro-expression difficult to be imitated, but have potential application value for emotion recognition tasks, such as exchange negotiations [6], mental illness [7], and criminal investigations [8]. In recent years, facial images have been acquired by various imaging devices for facial expression recognition in various research backgrounds and application scenarios [9–13]. Facial micro-expression recognition has achieved surprising performance [14–20], but it still requires further research.

The facial expression occurs as the facial muscle movement. Psychologists believe that facial muscle movement is not directly related to individual attributes, that is gender, age, and ethnicity when micro-expressions occur [5]. However, when micro-expression recognition is performed on the facial image acquired by various imaging conditions, the micro-expression image is expressed as a combination of the personal identity attribute and facial muscle movements. Therefore, individual attributes will interfere with micro-expression recognition.

Some research works have achieved surprising results in improving macro-expression recognition by considering the identity information of subjects such as age, gender, and ethnic background [21–24]. However, these methods do not apply to micro-expression recognition. Different from facial expressions, the micro-expression occurs only in small areas of the facial muscle movement, and these movements are very subtle.
that making micro-expression recognition has more attention to local facial features. These characteristics make the identity information in the micro-expression image very significant. Therefore, how to reduce the interference of identity information is one of the main research problems in micro-expression recognition, which is called the 'identity information interference problem'.

In response to this problem, this study proposes a new de-identity approach called two-stream difference network (TSDN). This method recognises micro-expression by removing identity information to extract micro-expression features. We believe that facial micro-expression is the superposition of identity information and micro-expression features. As illustrated in Figure 1, given a video sequence of the micro-expression, the onset and apex frame are extracted. The apex image as facial micro-expression image, and the onset frame as facial identity image. The TSDN model extracted micro-expression features by removing the subject's identity information. We called this de-identity.

The TSDN model consists of two parts, a two-stream encoder-decoder network, and a difference network. The two-stream encoder-decoder network is divided into a micro-expression stream and identity stream. In TSDN, the two-stream encoder-decoder network is divided into micro-expression stream and identity stream. The micro-expression stream is used to encode the apex image, and the identity stream is used to the onset frame. The identity information and micro-expression features are recorded in the intermediate layers of the micro-expression stream, while the intermediate layers of the identity stream only contain the identity information. To obtain the micro-expression features, it is needed to calculate the difference between the micro-expression stream and the identity stream of encoding and the intermediate layers. We aim to use this difference to recognise micro-expression through the difference network.

In general, this study attempts to propose a TSDN model for solving the 'identity information interference problem', and analyses the influence of this model on micro-expression recognition. The main contributions of this study are summarised as follows:

1. The influence of identity information on micro-expression recognition is analysed, and a learning model of de-identity information is proposed.
2. The de-identity information model based on the TSDN is proposed to reduce the interference of identity information on micro-expression recognition, thus improving the accuracy of recognition.
3. The superior performance of the de-identity information model—TSDN is verified on two public micro-expression data sets.

The remaining part of the study is organised as follows: In Section 2, a brief review of related research on micro-expressions recognition. Section 3 provides a complete introduction to the algorithm proposed. Section 4 shows the data sets, details, and results of the experiment. Finally, Section 5 presents the conclusions of this research method.

## 2 | RELATED WORK

The history of micro-expressions is relatively short compared to macro-expressions. In 1966, Haggard et al. proposed the concept of micro-expressions and found a short, hard-to-detect facial expression. Pfister et al. [25] first proposed the spontaneous facial micro-expressions recognition framework and that used temporal interpolation model (TIM) [26] to counter short micro-expression video lengths, spatiotemporal local texture descriptors (local binary pattern-three orthogonal planes, [LBP-TOP] [27]) to handle dynamic features and support the vector machine to perform classification, and the first spontaneous micro-expression database (SMIC) was published that achieves very promising results on this database. Since then, some micro-expression recognition is based on this framework. Huang et al. [28] proposed the spatio-temporal completed local quantized pattern (STCLQP) for micro-expression recognition by extended Completed Local Quantized Pattern (CLBP) to three-dimensional space. Wang et al. [29] proposed local binary pattern with six intersections point (LBP-SIP) to reduce the information redundancy of the LBP-TOP feature, then reducing the time complexity and space complexity. Previous research work has shown that the difference between facial micro-expression images and facial identity images at the pixel-level facilitates micro-expression recognition. To solve the sparse problem of pixel-level differences between identity image and micro-expression image in image sequences, Wang et al. [30] employed robust principal component analysis (RPCA) to feature pixel-level differences. Le Ngo et al. [31] achieved a good recognition rate by applying Sparsity-Promoting Dynamic Mode Decomposition (DMDSP) to remove neutral expressions from micro-expression videos. Liu et al. [32] proposed a sparse main direction mean optical-flow (MDMO) feature that learns an effective dictionary [33] from the micro-expression dataset. For the problem of handcrafted features that cannot well capture the subtle facial deformations of micro-expressions, Xia et al. [34] utilised recurrent neural network to learn...
subtle changes from image sequences to recognise micro-expressions, and for the problem of subtle changes in micro-expressions captured by different environments in cross-database tasks of micro-expression recognition, they input the optical flow features of the micro-expression sequence into the deep recurrent convolution network for micro-expression recognition [35].

Although micro-expression recognition based on video sequences can achieve very promising results, this introduces the problem of information redundancy. Therefore, Lioin et al. [36] proposed a new feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF), to encode essential expressiveness of the apex frame. Then, they introduced deep learning to propose an Optical Flow Features from Apex frame (OFF-Apex) [37] method for micro-expression recognition. The method uses the optical flow feature map of the micro-expression apex frame as the input of the convolutional neural network (CNN) to enhance the feature of the optical flow. It is noted that, unlike the above methods, these two methods utilise only two frames of the video sequence, namely, the apex frame and the onset frame instead of the complete video sequence for micro-expression recognition.

So far, the pixel-level difference has been used to extract micro-expression features. However, most of these are based on video sequences and information redundancy is higher. Meanwhile, the method based on a single-frame image does not consider the influence of identity information, that is gender, age, and ethnicity. Since the micro-expression image is a superposition of micro-expression features and identity information. Therefore, this study suggests extracting micro-expression features from the difference of the feature-level. We obtained the difference, that is the micro-expression features with more robust and effective from the intermediate layers in the two-stream encoder-decoder network of onset and apex frame, by the data-driven method.

### 3 | TWO-STREAM DIFFERENCE NETWORK—TSDN

In the early research study, the impressive recognition performance is presented through apex frames for micro-expression recognition. However, it is worth noting that it does not consider the impact of subject identifiable information, that is gender, age, and ethnicity on micro-expression recognition [36,37]. Although there are literatures that consider the impact of identity information on micro-expression recognition [29,30]. They calculated pixel-level differences directly for a complete video sequence. But these differences are often unreliable, and calculations of complete video sequences can cause redundancy. Therefore, this study proposes an architecture to recognise micro-expressions by removed identity information at the feature-level differences of onset and apex image.

The architecture of TSDN consists of two parts: the two-stream encoder-decoder network and the difference network. The two-stream encoder-decoder network is used to learn the intermediate layers features of micro-expressions and identities. The difference network utilises the difference information of the intermediate layers features of a two-stream encoder-decoder network for micro-expression recognition. We combine these two processes into the TSDN model. The architecture of TSDN illustrated in Figure 2. For each video sample, we locate the rectangular box that exactly bounds facial regions in the first frame, and then all the frames of the video are cropped and resized according to the box located in the first frame.

The identity stream in the two-stream coding network is used to encode the identity information, and the micro-expression stream also contains the coding of identity information and micro-expression features. The apex frame $I_{\text{apex}}$ is input to the micro-expression stream, and the onset frame $I_{\text{onset}}$ of the same subject is input to the identity stream for training. Then, the parameters of the encoder-decoder network are fixed, and the difference network training is performed. The features distance of Euclidean distance and absolute value distance output from the intermediate layers of the two-stream network is calculated as an input to the difference network for micro-expression classification.

#### 3.1 | Two-stream encoder-decoder network

The two-stream encoder-decoder network encodes the given apex frame $I_{\text{apex}}$ and an onset frame $I_{\text{onset}}$. Since the micro-expression data sets are small, to avoid over-fitting problems, we use a small network model to train the encoder-decoder. The encoder model only includes three convolutional layers (conv1, conv2, conv3). The decoder model also includes three convolutional layers (conv4, conv5, conv6). The kernel sizes of conv1 and conv6 are 5 * 5, and the kernel sizes of conv2, conv3, conv4, and conv5 are 3 * 3, and their stride is 1. The downsampling layer with a kernel size of 2 * 2 is added after each convolutional layer of the encoder, and the upsampling layer with the same kernel size is added after each convolutional layer of the decoder.

Image pairs $(I_{\text{apex}}, I_{\text{onset}})$ are provided to train two encode-decoder networks. First, the onset frame $I_{\text{onset}}$ input into the encoder-decoder model of the subject identity, and the high-dimensional feature distribution of the identity information is obtained by the encoder. Then the original data is restored by the decoder so that the distribution of the coding features is closer to the real data distribution. Similarly, we input the apex frame $I_{\text{apex}}$ into the micro-expression coding model to obtain the mixed feature distribution of the micro-expression and identity information.

In the image encoder-decoder process, we normalised the onset frame $I_{\text{onset}}$ to the range $[0, 1]$, and then input it to identity stream for training. The output of the decoder model is normalised by the SoftMax function to obtain $I_{\text{onset}}$. The cross-entropy [38] of $I_{\text{onset}}$ and $I_{\text{onset}}$ was calculated by summing pixel-wise as the objective function of identity information encoding. This process is represented by Equation (1).
Two-stream encoder-decoder Network

\[ L_{\text{onset}} = \sum_{i=1}^{N} \left( I_{\text{onset}} \log I'_{\text{onset}} + (1 - I_{\text{onset}}) \log (1 - I'_{\text{onset}}) \right) \]  

where \( I'_{\text{onset}} \) is identity facial image generated by the identity stream decoder, \( I'_{\text{apex}} \) is a micro-expression facial image generated by the expression stream decoder, and \( N \) is the number of training samples. The objective for the mixed features of micro-expression and identity information encoding is expressed as:

\[ L_{\text{apex}} = \sum_{i=1}^{N} \left( I_{\text{apex}} \log I'_{\text{apex}} + (1 - I_{\text{apex}}) \log (1 - I'_{\text{apex}}) \right) \]  

We denote the apex frame as \( I_{\text{apex}}^{\text{id}=S_{\text{exp}=\text{ME}}} \), onset frame as \( I_{\text{onset}}^{\text{id}=S_{\text{exp}=\text{neutral}}} \). The apex and onset frame are input into the two-stream encoder-decoder model:

\[ I_{\text{id}=S_{\text{exp}=\text{ME}}}^{\text{onset}} = \text{EncoderDecoder}_{\text{id}} \left( I_{\text{id}=S_{\text{exp}=\text{neutral}}}^{\text{onset}} \right) \]  

where \( \text{EncoderDecoder}_{\text{id}} \) is an encoder-decoder of the micro-expression stream, and \( \text{EncoderDecoder}_{\text{id}} \) is an encoder-decoder model of the identity stream, \( S \) belongs to any subject, \( \text{ME} \) belongs to any micro-expression. The onset image has no micro-expression occurs, and we recorded it as neutral. From Equation (3), we encode the apex frame that the micro-expression features and identity information of the subject is recorded in the intermediate layer of the micro-expression stream. But from the Equation (4), we only recorded the identity information in the intermediate layer of the identity stream through the encoded onset frame.

3.2 | Micro-expression recognition based on difference network

The micro-expression feature can be obtained by the difference between the pixel-level or the feature-level of the apex image and the onset image. Because of the rotation, translation, and illumination conditions of the image, pixel-level
differences between images are unreliable. Even with the same classification of micro-expression, pixel-level differences can be large. Therefore, this study uses feature-level differences to extract micro-expression features.

Therefore, the identity information of the onset frame encoded is stored in the identity stream, and the mixed features of the micro-expression and identity of the apex frame encoded are stored in the expression stream. We extracted the difference features by calculating the Euclidean distance and the absolute value distance of the intermediate layers of the micro-expression stream and identity stream. Then, the difference features are input to the Difference Network (DN) for micro-expression recognition. The difference features calculation method is as shown in Equation (5).

\[ FC_{distance} = \theta \sqrt{(FC_{exp+id} - FC_{id})^2} + (1 - \theta)(FC_{exp+id} - FC_{id}) \]  

where \( FC_{exp+id} \) represents the intermediate layer features of the micro-expression stream, \( FC_{id} \) represents the intermediate layer features of the identity stream. \( FC_{distance} \) represents difference features determined by the Euclidean distance and absolute value distance of the intermediate layers of the two-stream encoder-decoder network. The range of parameter \( \theta \) is \([0, 1]\).

Figure 2 shows that we input the difference features to the local difference network (LDN) model. All LDN models are combined into a difference network. The fully connected layer of the LDN model is a 1 * 4096 vector, and the output layer is the micro-expression class (happiness, disgust, surprise, depression, other). The loss function of the difference model is noted as \( L_{diff} \).

\[
L_{diff} = \sum_{i=1}^{N} \{ Y_{me} \log y_{me} + (1 - Y_{me}) \log (1 - y_{me}) \} 
\]

where \( Y_{me} \) is the ground truth of the micro-expression category, \( y_{me} \) is the output of the LDN model. The range of \( y_{me} \) is \([0, 1]\), and \( Y_{me} \) value is 0 or 1. For each local difference network (LDN) model, the loss function is defined as \( Loss_{diff} \), \( i \in \{1, 2, 3, 4, 5, 6\} \), where index \( i \) corresponds to the intermediate layer conv1, and so on. The coding features loss is defined as \( Loss_{code} \). All LDN models and the coding features difference network (LDN-code) model combine to form a complete differential network for micro-expression recognition. Consequently, each local loss and the feature coding loss are accumulated to generate a total loss function, which is specifically defined as:

\[
Loss_{total} = \delta_1 Loss_{diff-1} + \delta_2 Loss_{diff-2} + \delta_3 Loss_{diff-3} + \delta_4 Loss_{diff-4} + \delta_5 Loss_{diff-5} + \delta_6 Loss_{diff-6} + \delta_{code} Loss_{diff-code} 
\]

### 4 EXPERIMENTAL RESULTS

The proposed TSDN approach is evaluated on two public facial micro-expression databases. The experimental results show that our method can achieve state-of-the-art performance in micro-expression recognition.

#### 4.1 Experimental details

Due to the difficulty in collected micro-expression data, the micro-expression datasets are small, and the training used directly causes the model to be over-fitting. Therefore, we applied a data augmentation method to extend the training samples. First, 10 images of size 224 × 224 are randomly cropped out from the centre of each training sample. Each cropped image is rotated by \( \{-9°, -6°, -3°, 0°, 3°, 6°, 9°\} \) respectively. Then, each image is adjusted for contrast by \( \{95%, 100%, 105%\} \). A relatively large training dataset is generated by the data augmentation method, which reduces the over-fitting problem of the model. Additionally, we use leave-one-subject-out (LOSO) cross-validation to evaluate the de-identity model of TSDN. In each cross-validation, one subject's video clips are used as the test data and the others are used as the training data. By tuning the initialisation learning rate, batch size, and dropout on the training set, the model can get the best results on the validation set. The two-stream encoder-decoder model is first trained by the onset and apex frame of the micro-expression in all training videos through data augmentation. Then, the intermediate layer of the encoder-decoder model is user input, and the differential model is trained for micro-expression classification. To obtain the optimal model, we divide the samples used for training into the training set and validation set.

For the training process of two-stream encoder-decoder networks, we use Adam optimiser with a batch size of 64 an initialisation learning rate of 0.001. In addition to added the dropout of 0.5 in the fully connected layer, the training settings of each LDN are consistent with the above 300 epochs to train the stream encoder-decoder network and difference network. For the classification loss function of each LDN, \( \delta_1 = 0.7 \), \( \delta_2 = 0.5 \), \( \delta_3 = 0.3 \), \( \delta_4 = 0.3 \), \( \delta_5 = 0.5 \), \( \delta_6 = 0.7 \), \( \delta_{code} = 1.0 \) are respectively set. The proposed model is trained and predicted using PyTorch on the GeForce GTX 1080 platform.

#### 4.2 Expression recognition results

The SMIC is a public data set commonly used for micro-expression recognition. The apex frame position is not provided for the SMIC. We first use the TSM method proposed in [41] for face detection and landmark localisation. Then, according to the face detection result, the face region is cropped and normalised to a size of 234 × 234, and the face is aligned according to the landmark of the centre of eyes and the nose. Finally, the apex frame automatic recognition system (AFRS) [42] for determining the location of the apex frame through the region of
interest of ten landmark points (‘left eye + eyebrow’, ‘right eye + eyebrow’, ‘nose’ and ‘mouth’). In the process of data collection, the micro-expression video sequence is collected by watching a film to elicit the emotion of the subject. Twenty subjects were elicited with emotion and 328 video sequences were collected. These are recorded with a resolution of 640 x 480, the frame rate of 100 fps camera. From all of the 328 video sequences collected, 164 samples are labelled for micro-expression recognition. The sample is labelled as one of three micro-expressions, that is negative, positive, and surprise, from onset to offset micro-expression. The SMIC does not label the apex frame, so we apply the AFRS [43] to obtain the location of the apex frame, as a general procedure, we use the before and after four frames of the apex location with the calculated micro-expression label. Meanwhile, we use the first two frames on the onset frame of each micro-expression sequence and the last two frames of the offset frame as the facial identity image, which results in 820 image pairs as training samples for the two-stream encoder-decoder model.

In the micro-expression classification based on the difference model, we use the LOSO cross-validation method to obtain the final recognition result. As shown in Table 1, the TSDN method is optimal at 100 epochs of two-stream encoder-decoder networks and a difference network. In evaluating micro-expression recognition, to deal with the imbalanced class distribution, we also use the accuracy and F1_score for performance evaluation. Specifically, F1_score is expressed as:

\[
F1_{\text{score}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]  

(8)

\[
\text{Recall} = \frac{TP}{TP + FN}
\]  

(9)

And

\[
\text{Precision} = \frac{TP}{TP + FP}
\]  

(10)

Accuracy is expressed as:

\[
\text{Acc} = \frac{TP + TN}{TP + FP + TN + FN}
\]  

(11)

where TP is true positive, TN is true negative, FP is the false positive and FN is false negative. As shown in Table 2, our proposed TSDN model achieves an F1_score of 69.52% and a recognition rate of 68.90%, which is better than the state-of-the-art method that has been compared. Note that all of these methods take advantage of the temporal features extracted from the image sequence, except for Bi-WOOF, OFF-ApexNet, and our TSDN model. Compared with all other methods, our proposed model shows superior performance on facial micro-expression recognition on SMIC datasets. Our model achieves an improvement of about 2% in the recognition rate and 2.5% in F1_score followed by the OFF-ApexNet. The confusion matrix of the TSDN model is shown in Figure 3, where the positive classification shows the lowest recognition rate with 59%, and the surprising is highest with 84%. This may be related to the intensity of the micro-expression. By observing the data, we found that the probability of positive is lower than surprise and negative, which is consistent with the experimental results.

The Chinese Academy of Sciences Micro-expression database (CASME II) [44] contains 247 video sequences collected from 26 subjects. These video sequences are recorded with a 200fps camera. This data set contains five classes of micro-expressions: happiness (32 samples), surprise (25 samples), disgust (64 samples), repression (27 samples), and others (99 samples). To get more data, we use the before and after four frames on the apex location with the provided micro-expression label, and the first two frames on the offset frame of each micro-expression sequence and the last two frames of the offset frame with the provided identity label. The LOSO cross-validation is performed similar to the experimental setting in SMIC. As we can see in Table 3, the TSDN method is optimal.

| Methods          | Acc (%) | F1_score |
|------------------|---------|----------|
| LBP-TOP [25]     | 45.73   | 0.4600   |
| OSW [39]         | 53.05   | 0.5431   |
| LBP-SIP [29]     | 54.88   | 0.5502   |
| STLBP-IP [40]    | 57.93   | 0.5829   |
| RPCA [30]        | 58.00   | 0.6000   |
| STCLQP [28]      | 64.02   | 0.6381   |
| Bi-WOOF [36]     | 61.59   | 0.6110   |
| OFF-ApexNet [37] | 67.68   | 0.6709   |
| TSDN (pre-trained)| 52.66   | 0.4860   |
| TSDN (ours)      | 68.90   | 0.6952   |

Table 1 Comparing the impact of all samples training times on micro-expression recognition accuracy on SMIC using TSDN. The horizontal represents the two-stream encoder-decoder network, and the vertical represents the difference network.
at 200 epochs of the two-stream encoder-decoder network and 100 epochs difference network.

The F1_score and accuracy of recognising five micro-expressions on CASME II are shown in Table 4. We have added the state-of-the-art method of CNN + LSTM [43] and LTOGP [45] for comparison. The experimental results demonstrate that our TSDN model can achieve the highest accuracy and F1_score compared with CNN-based methods (CNN + LSTM, OFF-ApexNet) and handcrafted feature-based methods (i.e. LBP-TOP, RPCA, STCLQP, Bi-WOOF, and LTOGP). In addition to Bi-WOOF, OFF-ApexNet and our TSDN model recognise micro-expression recognition for a static image, while other methods are performed for video sequences. Figure 4 is the confusion matrix of the TSDN model, consistent with the SMIC dataset, whereas the recognition rate of our proposed TSDN model, negative emotions (repression and disgust) and surprizes emotions are higher than positive emotions (happiness).

The results of data augmentation for database SMIC and CASME II are shown in Table 5. It can be proved that this data augmentation method can achieve superior results.

### Table 3
Comparing the impact of all samples’ training times on micro-expression recognition accuracy on CASME II using TSDN. The horizontal represents the two-stream encoder-decoder network, and the vertical represents the difference network.

| Epoch | 30   | 50   | 100  | 200  | 300  |
|-------|------|------|------|------|------|
| 30    | 68.32| 68.75| 69.43| 69.56| 69.32|
| 50    | 68.54| 69.70| 70.42| 70.36| 70.94|
| 100   | 69.09| 70.62| 71.24| **71.49**| 71.06|
| 200   | 69.86| 70.94| 71.19| 71.04| 70.63|
| 300   | 69.43| 70.87| 70.96| 71.12| 70.79|

### Table 4
Micro-expression recognition performance comparison on the CASME II databases.

| Methods     | Acc (%) | F1_score |
|-------------|---------|----------|
| LBP-TOP [44] | 39.68   | 0.3589   |
| OSW [39]    | 41.70   | 0.3820   |
| LBP-SIP [29] | 43.32   | 0.3976   |
| STLBP-IP [40]| 59.51   | 0.5679   |
| RPCA [30]   | 49.00   | 0.5100   |
| STCLQP [28] | 58.39   | 0.5836   |
| CNN + LSTM [43]| 60.98 | -        |
| LTOGP [45]  | 66.00   | -        |
| Bi-WOOF [36]| 57.89   | 0.6125   |
| OFF-Apex Net [37]| 68.94 | 0.6967   |
| TSDN (pre-trained) | 49.39 | 0.4762   |
| TSDN (ours) | 71.49   | 0.7024   |

### Table 5
Micro-expression recognition performance comparison on the CASME II databases.

| Methods     | SMIC Accuracy (%) | F1 Score | CASME II Accuracy (%) | F1_score |
|-------------|-------------------|----------|------------------------|----------|
| TSDN        | 63.27             | 0.6483   | 62.84                  | 0.6431   |
| TSDN (DA)   | 68.90             | 0.6952   | 71.49                  | 0.7024   |

The intermediate layer of the two-stream encoder-decoder model has a different contribution to the micro-expression recognition rate. That is to say, the expression ability of different levels of features in CNNs is different.

The effect of each LDN, DN, and CNN on the micro-expression recognition results is shown in Figure 5. The recognition result of the apex frames directly using the CNN model is lower than the influence of coding feature difference network. It shows that removing identity information can improve the accuracy of micro-expression recognition. The influence of the local difference network of the two-stream encoder-decoder intermediate layer on the recognition result is the LDN-1 and LDN-6 models higher than the LDN-2, LDN-3, LDN-4, and LDN-5. The LDN-2 and LDN-5 are in...
the middle position. Thus, it justifies the $\delta_1$ and $\delta_2$ having bigger weights than $\delta_3$, $\delta_4$ and $\delta_5$ for the total loss in Equation (7). At the same time, we analysed the influence of the parameter $\theta$ on micro-expression recognition. The experimental results are shown in Figure 6. We find that when $\theta = 0.7$, the recognition result of the SMIC database is the best, and when $\theta = 0.8$, the recognition result of the CASME II database is the best.

5 | CONCLUSION

In this study, we propose a de-identity information method for the spontaneous facial micro-expression recognition which is based on the TSDN. Firstly, a two-stream encoder-decoder model is trained by the CNN. The expression stream in the two-stream encoder-decoder extracts the expression feature and identity information from a micro-expression facial image and the identity stream encoder-decoder extracts the identity information from an identity facial image. Then, a local difference network is trained on the intermediate layers of the two-stream encoder-decoder model. In this process, the micro-expression feature of the expression stream that has been recorded in the two-stream encoder-decoder model can be extracted through the local difference model which removes the identity information in the expression stream. We combine these two processes into a two-stream difference network.

The TSDN model was evaluated on SMIC and CASME II of two published micro-expressions datasets. This method does not utilise the temporal information of the video sequence. However, it is better than the state-of-the-art sequence-based methods. On the contrary, this method only uses the onset frame and apex frame in the micro-expression sequence and is also higher than the current best method based on the apex frame.

From the recognition result, it can solve the ‘identity information interference problem’. Meanwhile, it is clear that the hypothesis of ‘the micro-expression facial image is a combination of micro-expression features and identity information, and the micro-expression features can be obtained by removing identity information approach’ is correct. In the future work, we will further consider if the intermediate layer feature of the encoder-decoder model is replaced by mixed features of depth learning features and manual features (i.e. LBP, HOG, Optical Flow etc). More detailed facial micro-expression features are input to the differential model to obtain finer micro-expression features, thereby improve the classification performance. Besides, our future work will combine the micro-expressive features and the temporal information for spotted micro-expressions.

ACKNOWLEDGEMENTS

This study was supported by the National Key R&D Programme of China (No. 2018YFC0100700), the National Natural Science Foundation of China (No. 61672093) and Advanced Innovation Centre for Intelligent Robots and Systems Open Research Project (No. 2018IRS01).

REFERENCES

1. Yuan, S., Mao, X.: Exponential elastic preserving projections for facial expression recognition. Neurocomputing. 275(1), 711–724 (2018)
2. Piparsaniy, Y., et al.: Robust facial expression recognition using Gabor feature and Bayesian discriminating classifier. In: IEEE International Conference on Communications and Signal Processing (ICCSPP), pp. 538–541. IEEE (2014)
3. Owusu, E., et al.: A neural-AdaBoost based facial expression recognition system. Expert. Syst. Appl. 41(7), 3383–3390 (2014)
4. Bian, C., et al.: Spontaneous facial expression database for academic emotion inference in online learning. IET Comput. Vis. 13(3), 329–337 (2018).
5. Ekman, P., Friesen, WV.: Constants across cultures in the face and emotion. J Pers. Soc. Psychol. 17(2), 124–129 (1971)
6. Russell, T.A., et al.: A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. Br. J. Clin. Psychol. 45(4), 579–583 (2006)
7. Salter, E., et al.: Sex differences in negotiating with powerful males. Hum. Nat. 16(3), 306–321 (2005)
8. O’Sullivan, M., et al.: Police lie detection accuracy: the effect of lie scenario. Law Hum. Behav. 33(6), 530 (2009)
9. Jung, H., et al.: Joint Fine-tuning in deep neural networks for facial expression recognition. In: IEEE International Conference on Computer Vision (ICCV), pp. 2983–2991. IEEE (2015)
10. Zeng, N., et al.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing. 273, 643–649 (2017)
11. Minae, S., et al.: Biometric Recognition Using Deep Learning: A Survey. arXiv (2019)
12. Shabat, A.M., Tapamo, J.: Angled local directional pattern for texture Analysis with an application to facial expression recognition. IET Comput. Vis. 12(5), 603–608 (2018)
13. Minae, S., Abdolrahshidi, A.: Deep-emotion: Facial Expression Recognition Using Attentional Convolutional Network. arXiv (2019)
14. Yuan, Z., et al.: Cross-database micro-expression recognition: a benchmark. In: ACM International Conference on Multimedia Retrieval, pp. 354–363 (2019)
15. Mezghani, W., et al.: The implication of spatial temporal changes on facial micro-expression analysis. Multimed Tool Appl. 78(15), 21613–21628 (2019)
16. Li, Q., et al.: Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. Multimed. Tool Appl. 78(20), 29307–29322 (2019)
17. Zong, Y., et al.: Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. IEEE Trans. Multimed. 20(11), 3160–3172 (2018)
18. Liu, Y., et al.: A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans. Affective Comput. 7(4), 299–310 (2015)
19. Han, J., et al.: Personalised broad learning system for facial expression, 1–18. Multimed. Tools Appl. (2019)
20. Xia, Z., et al.: Spatiotemporal recurrent convolutional networks for recognising spontaneous micro-expressions. IEEE Trans. Multimed. 22(3), 626–640 (2020)
21. Zhang, C., et al.: Identity-aware convolutional neural networks for facial expression recognition. J. Syst. Eng. Electron. 28(4), 784–792 (2017)
22. Kim, Y., et al.: Deep Generative-contrastive Networks for Facial Expression Recognition. arXiv (2017)
23. Yang, H., et al.: Facial expression recognition by de-expression residue learning. In: IEEE International Computer Vision and Pattern Recognition, pp. 2168–2177. IEEE (2018)
24. Xie, S., et al.: Facial expression recognition using intra-class variation reduced features and manifold regularisation dictionary pair learning. IET Comput. Vis. 12(4), 458–465 (2017)
25. Prister, T., et al.: Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In: IEEE International Conference on Computer Vision Workshops, 868–875. IEEE (2011)
26. Zhou, Z., et al.: Towards a practical lipreading system. In: IEEE International Computer Vision and Pattern Recognition, pp. 137–144. IEEE (2011)
27. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29(6), 915–928 (2007)
28. Huang, X., et al.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing, 175, 564–578 (2016)
29. Wang, Y., et al.: Lhp with six intersection points: reducing redundant information in Lhp-Top for micro-expression recognition. In: Asian Conference on Computer Vision, pp. 525–537. Springer, Cham (2014)
30. Wang, S., et al.: Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In: European Conference on Computer Vision, pp. 325–338. Springer, Cham (2014)
31. Le Ngo, A.C., et al.: Sparsity in dynamics of spontaneous subtle emotions: analysis and application. IEEE Trans. Affective Comput. 8(3), 396–411 (2016)
32. Liu, Y., et al.: Sparse MDMO: learning a discriminative feature for spontaneous micro-expression recognition. IEEE Trans. Affective Comput., 1–18 (2018)
33. Li, X., et al.: Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. IEEE Trans. Geosci. Rem. Sens. 52(11), 7086–7098 (2014)
34. Xia, Z., et al.: Spontaneous facial micro-expression recognition via deep convolutional network. In: IEEE Eighth International Conference on Image Processing Theory, Tools and Applications, pp. 1–6. IEEE (2018)
35. Xia, Z., et al.: Cross-database micro-expression recognition with deep convolutional networks. In: IEEE International Conference on Biometric Engineering and Applications, pp. 56–60. IEEE (2019)
36. Liong, S.T., et al.: Less is more: micro-expression recognition from video using apex frame. Signal Process. Image Commun. 62, 82–92 (2018)
37. Gan, Y.S., et al.: Off-apexnet on micro-expression recognition system. Signal Process. Image Commun. 74, 129–139 (2019)
38. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inf. Theor. 26(1), 26–37 (1980)
39. Liong, S.T., et al.: Optical strain based recognition of subtle emotions. In: IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp. 180–184. IEEE (2014)
40. Huang, X., et al.: Facial micro-expression recognition using spatiotemporal binary pattern with integral projection. In: International Conference on Computer Vision Workshops, pp. 1–9. IEEE (2015)
41. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE International Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)
42. Liong, S.T., et al.: Automatic apex frame spotting in micro-expression database. In: Asian Conference on Pattern Recognition, pp. 665–669. Springer, Cham (2015)
43. Kim, D.H., Baddar, W.J., Ro, YM.: Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In: ACM International Conference on Multimedia, pp. 382–386 (2016)
44. Yan, W., et al.: CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLOS One 9(1), 1–8 (2014)
45. Niu, M., et al.: Discriminative video representation with temporal order information in Lbp for micro-expression recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2112–2116. IEEE (2019)

How to cite this article: Pan H, Xie L, Li J, Ly Z, Wang Z. Micro-expression recognition by two-stream difference network. IET Comput. Vis. 2021;1–9. https://doi.org/10.1049/cvi2.12030